Revenue growth prediction for small and medium-sized enterprises: a data mining approach for the insurance industry

Author(s):
Müller, Daniel

Publication Date:
2018

Permanent Link:

https://doi.org/10.3929/ethz-b-000284491 →

DISS. ETH NO. 25244

REVENUE GROWTH PREDICTION FOR SMALL AND MEDIUM-SIZED ENTERPRISES: A DATA MINING APPROACH FOR THE INSURANCE INDUSTRY

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

DANIEL MÜLLER

M.A. HSG in BANKING AND FINANCE

UNIVERSITY OF SANKT GALLEN

born on 21.11.1985

citizen of Germany

accepted on the recommendation of

Prof. Dr. Elgar Fleisch

Prof. Dr. Florian von Wangenheim

2018

Acknowledgments

This dissertation is the result of my tenure at the Department of Management, Technology and Economics (D-MTEC) at ETH Zurich from December 2014 until June 2018, where I was a PhD student at the Mobiliar Lab for Analytics. The collaboration with La Mobilière insurance group through the Mobiliar Lab for Analytics provided me with the opportunity to conduct research closely related to actual business questions in the insurance sector. Hence, I would like to thank La Mobilière for providing funding and company resources for this research project.

Further, I would like to thank my supervisors, Prof. Dr. Elgar Fleisch, Prof. Dr. Florian von Wangenheim and Dr. Irena Pletikosa. I am thankful for the guidance and advice given to my research and the support I received during the conducted studies. I would also like to say a big thank-you to Dr. Gundula Heinatz, and to her successor, Dr. Erika Meins, the head of the Mobiliar Lab for Analytics at La Mobilière. Their efforts enabled me to discuss and present my interim findings regularly to insurance agents, the managers responsible for commercial clients and senior managers of the insurance group. They also enabled me to access various business resources, including the databases that store current and historic information about the insurance group's customer base and business transactions with them. To Dr. Andrea Ferrario, I also want to express my gratitude. Thank you for your guidance and the many insightful conversations we had about data and statistics.

Moreover, I want to thank Barbara Agoba, Matyas Filip, Jan Goldmann, Roland Halter, Andreas Hölzli, Dr. Jose Iria, Jerome Koller, Patrik Linder, Lara Müller, Markus Odermatt, Malte Sandow, Lukas Stöcklin, Sebastian Wyder and many more from La Mobilière for their support of this research.

The past years would have been just half as impressive without the company of my colleagues. You have made my PhD tenure a lasting experience; thanks to Liliane Ableitner, Filipe Barata, Dr. Thomas von Bomhard, Raquel Brüggner, Mathieu Chanson, Shih (Iris) Chen-Hsuan, Dr. Andre Dahlinger, Dr. Remo Frey, Klaus Fuchs, Bernhard Gahr, Johannes Hübner, Cristina Kadar, Jan-Niklas Kramer, Florian Künzler, Arne Meeuw, Dominik Rüegger, Benjamin Ryder, Dr. Anne Scherer, Sandro Schopfer, Yiea-Funk Te, Dr. Verena Tiefenbeck, Peter Tinschert, Dr. Dirk Volland, Denis Vukovac, Anselma Wörner, Dr. Dominik Wörner and many more.

Finally, I want to express my deepest gratitude to my family and my future wife, Anna. Thank you for understanding, support and positive energy during this time.

Zurich, May 2018

Daniel Müller

Disclaimer

This thesis contains text parts of previously written publications by the author of this thesis.

A major part of Chapters I and III contain passages from the following publication: Müller, D., Te, Y.-F., Jain, P. (2017). *Insurance premium optimization using motor insurance policies – a business growth classification approach.* Proceedings of the 2017 IEEE International Conference on Big Data, Boston.

A part of Chapter I contains passages from the following publications: Müller, D., Mau, S., Pletikosa Cvijikj, I (2015). *A framework for consensual and online privacy preserving record linkage in real-time.* Big Data (Big Data), 2015 IEEE International Conference on Big Data (2015); Mau, S., Müller, D. Pletikosa Cvijikj, I, Wagner, J. (2016). *Anticipating insurance customers' next likely purchase events.* Second International Conference for Marketing in the Insurance Industry (ICMI 2016).

A major part of Chapter IV contains passages from the following publication: Müller, D., Te, Y.-F., Jain, P. (2018). *Augmenting data quality through high-precision gender categorization.* ACM, Journal of Data and Information Quality (JDIQ). *Manuscript submitted for publication.*

A part of Chapter IV also contains passages from the following publication: Müller, D., Te, Y.-F., Jain, P. (2017). *Predicting business performance through patent applications.* Proceedings of the 2017 IEEE International Conference on Big Data, Boston, 2017.

A major part of Chapter V contains passages from the following publication: Müller, D., Te, Y.-F., Bettler, D., Heinzer, M. (2018). *Predicting the business performance of hotels and restaurants using publicly available data.* International Journal of Hospitality Management. *Manuscript submitted for publication.*

Abstract

Commercial insurance companies are exposed to margin pressure due to customers re-negotiating premiums, new competitors entering the market, and increased market transparency. Traditional insurance companies with historic information about their commercial customers can use internal and external data to improve competitiveness by predicting the revenue growth of their clients. Insights from these data improve competitiveness because insurance companies can use it to ambitiously price the policies of the customers they predict to be most valuable in the future and gear relationship efforts toward retaining those valuable customers. Furthermore, identifying customers with past revenue growth not reflected in the insurance policy allows insurance companies to select those customers who are most likely to have understated their current revenues. These underinsured customers do not pay sufficient premiums and are therefore not fully covered in the event of a claim. It is in the interest of commercial insurance companies to avoid such circumstances and to identify these customers as well.

This thesis aims to determine how and with which data sources growing customers can be identified. To this end, the research question is as follows: How can insurance companies predict the revenue growth of small and medium-sized enterprises using internal and external data sources? The research question will be answered first by developing a process to select those customers who are most suitable for a prediction and then with three examples of predictions performed using internal and external data sources.

The findings of this research suggest that, with the help of general liability insurance policies, the customer cohorts most suitable for a prediction can be identified. These cohorts are commercial agencies, motor vehicle garages, consultancies, and restaurants. The thesis then illustrates (1) how data from vehicle insurance policies can be used to identify growing businesses, (2) how growth prediction can be improved with the help of augmented patent-filing data, and (3) how information collected from online review websites contributes to identifying growing restaurants and hotels.

On this basis, it is recommended that traditional commercial insurance companies replicate this approach and identify the cohorts most suitable for growth predictions. Then, using internal and external data sources, cohort- and data-source-specific growth-prediction models that work together over the whole portfolio can be built. Vehicle insurance policies especially are a valuable source of information that can be applied to a large cohort of customers and accurately identify growing customers. In addition, the high predictive capacity of patent filings with respect to revenue growth should encourage insurance companies to display new patent filings in their internal customer

relationship management systems and trigger a notification to the agents when a customer has filed a patent. The notification indicates increased customer importance in the future.

Further research could be undertaken to identify other data sources with a high predictive capacity. In addition, a field test could investigate how insurance agents react to information indicating customer revenue growth and the impact on profitability and customer loyalty.

Zusammenfassung

Anbieter von Betriebsversicherungen sind erhöhtem Margendruck ausgesetzt. Dies ist der Fall da Geschäftskunden die zu zahlenden Prämien nachverhandeln, neue Wettbewerber in den Markt strömen und dieser zunehmend transparenter wird. Zur Verbesserung der Wettbewerbsfähigkeit können Versicherer mit bestehenden Geschäftsbeziehungen und historischen Informationen über Geschäftskunden interne und externe Daten verwenden, um erwartete Umsatzveränderungen von Kunden vorherzusagen. Dadurch gewonnene Erkenntnisse ermöglichen es, die Policen von in Zukunft sehr wertvollen Kunden, wettbewerbsfähig zu berechnen und etwaige Kundenbindungsmassnahmen anzupassen. Darüber hinaus ermöglicht ein solcher Prozess, Kunden mit vergangenem Umsatzwachstum, das jedoch in den Geschäftspolicen nicht abgebildet ist, zu identifizieren. Geschäftskunden mit versicherten Umsätzen, die kleiner sind als die effektiven Umsätze, sind unterversichert. Sie zahlen daher weniger Prämie und sind im Schadensfall nicht hinreichend gedeckt. Folglich liegt es im Interesse von Versicherern, unterversicherte Kunden zu erkennen.

Ziel dieser Arbeit ist es, zu ermitteln, wie und mit Hilfe welcher Datenquellen wachsende Kunden identifiziert werden können. Die zugrundeliegende Forschungsfrage lautet: Wie können Versicherer das Umsatzwachstum von kleinen und mittelständischen Unternehmen mit Hilfe interner und externer Datenquellen vorhersagen? Die Forschungsfrage wird beantwortet, indem zunächst erklärt wird, wie Kunden, die für die Wachstumsvorhersage besonders gut geeignet sind, ausgewählt werden können. In einem zweiten Schritt werden dann mit Hilfe von internen und externen Datenquellen drei unterschiedliche Modelle für die Umsatzwachstumsprognose entwickelt.

Die Ergebnisse dieser Forschungsarbeit legen nahe, dass mit Hilfe von Informationen, die aus Betriebsversicherungen entnommen wurden, Kundensegmente identifiziert werden können, die für eine Vorhersage in Frage kommen. Die identifizierten Kundensegmente umfassen Handelsvertretungen, Fahrzeug-Werkstätten, Beratungen und Restaurants. Weiterhin zeigen die Forschungsergebnisse, dass (1) Daten aus Fahrzeug-Versicherungspolicen sich eignen, um wachsende Unternehmen zu identifizieren, (2) Wachstumsprognosen durch Patentdaten verbessert werden können, und (3) mittels Daten von Online-Bewertungs-Webseiten wachsende Hotels und Restaurants identifiziert werden können.

Basierend auf diesen Erkenntnissen wird empfohlen, dass bestehende Betriebsversicherer den erläuterten Ansatz auf das eigene Kundenportfolio anwenden und so die geeignetsten Kundensegmente ermitteln. Mit Hilfe interner und externer Datenquellen können Versicherer dann segment- und datenquellenspezifische Wachstumsvorhersagemodelle erstellen. Die unterschiedlichen Modelle können daraufhin zusammengeführt und für die Identifikation von wachsenden

Geschäftskunden genutzt werden. Insbesondere Fahrzeugversicherungen sind eine wertvolle Informationsquelle, die aufgrund ihrer Breite für viele Geschäftskunden relevant ist sowie eine hohe Vorhersagegenauigkeit verspricht. Zudem sollten Patentneuanmeldungen in die Agenturinformationssysteme einfliessen und eine Benachrichtigung der Versicherungsberater auslösen. So können Agenturen über die Steigerung der Bedeutung eines Kunden informiert werden.

Zukünftige Studien könnten weitere Datenquellen sichten und deren Vorhersagekapazität in Hinsicht auf Umsatzwachstum testen. Zudem ist es von Relevanz, die Auswirkungen der Bereitstellung von erwarteten Umsatzentwicklungen von Geschäftskunden im Hinblick Veränderungen der Profitabilität und Kundenloyalität zu messen.

Table of Contents

ACKNOWLEDGMENTS ..... **IV**

DISCLAIMER ..... **VI**

ABSTRACT ..... **VIII**

ZUSAMMENFASSUNG ..... **XI**

TABLE OF CONTENTS ..... **XIV**

LIST OF FIGURES ..... **XVIII**

I) INTRODUCTION **1**

I.1 MOTIVATION 1

I.1.1 Managerial Motivation *1*

I.1.2 Research Motivation *4*

I.2 RESEARCH QUESTIONS 6

I.2.1 Research Question 1 ..... *7*

I.2.2 Research Question 2 ..... *8*

I.2.3 Research Question 3 ..... *8*

I.2.4 Research Question 4 ..... *8*

I.3 RESEARCH DESIGN 9

I.3.1 Statistical Modeling of Data Problems *9*

I.3.2 Relevance of Business Analytics for an Insurer *11*

I.3.3 Record Linkage *12*

I.3.4 Data quality assessment *14*

I.4 CONTEXT AND STRUCTURE OF WORK 17

I.4.1 Context of the Research Project *17*

I.4.2 Document Structure *18*

II) GROWTH PREDICTION FOR WHICH KIND OF SMES? **21**

II.1 INTRODUCTION 21

II.2 LITERATURE REVIEW AND HYPOTHESIS 23

II.2.1 Related work of SME Growth Prediction *23*

II.2.2 Corporate Evaluation of SME Growth Through Business Analytics *23*

II.3 RESEARCH DESIGN 25

II.3.1 Data Set and Data Linkage *25*

II.3.2 Used Variables *26*

II.3.3 Methodology *26*

II.4 RESULTS 28

II.4.1 Economic relevancy *28*

II.4.2 Statistical Relevancy *29*

II.5 SUMMARY 34

II.6 IMPLICATIONS 34

II.7 Limitations and Future Work 34

III) GROWTH PREDICTION WITH INSURANCE DATA 37

III.1 Introduction 37

III.2 Literature Review and Hypothesis 38

III.2.1 Social Identity and Brand Personality Theory 38
III.2.2 Insurance Data Mining 39

III.3 Research Design 40

III.3.1 Data Set and Record Linkage 40
III.3.2 Covariates 41
III.3.3 Methodology 42

III.4 Results 46

III.4.1 Descriptive Statistics 46
III.4.2 Prediction Results 51

III.5 Summary 52

III.5.1 Discussion and Conclusion 52
III.5.2 Implications for Research and Practice 55
III.5.3 Limitations and Future Work 55

IV) IMPROVING GROWTH PREDICTION WITH PATENT DATA 58

IV.1 Introduction 58

IV.2 Literature Review and Hypothesis 59

IV.3 Research Design 61

IV.3.1 Data Set and Record Linkage 61
IV.3.2 Covariates for Prediction 62
IV.3.3 Methodology 63

IV.4 Results 74

IV.4.1 Descriptive Statistics 74
IV.4.2 Imputation Results 79
IV.4.3 Prediction Results 87

IV.5 Summary 89

IV.5.1 Discussion and Conclusion 89
IV.5.2 Implications for Research and Practice 89
IV.5.3 Limitations and Future Work 90

V) GROWTH PREDICTION WITH TOURISM DATA 92

V.1 Introduction 92

V.2 Literature Review and Hypothesis 94

V.2.1 Electronic Word-of-Mouth 94
V.2.2 Studies of eWOM in Hospitality and Gastronomy 99

V.3 Research Design 102

V.3.1 Data Set and Context of Study 102
V.3.2 Covariates for Prediction 104
V.3.3 Methodology 112

V.4 RESULTS 112
V.4.1 Hotel Revenue Growth 113
V.4.2 Restaurant Revenue Growth 115
V.5 SUMMARY 117
V.5.1 Discussion and Conclusion 117
V.5.2 Implications 118
V.5.3 Limitations and Future Work 118
V.5.4 Conclusion 119
VI) DISCUSSION AND CONCLUSION 122
VI.1 SUMMARY OF KEY FINDINGS 122
VI.2 ANSWERS TO RESEARCH QUESTIONS 123
VI.3 IMPLICATION 130
VI.3.1 Implications for Practice 130
VI.3.2 Implications for Research 133
VI.4 LIMITATIONS AND FUTURE RESEARCH 135
REFERENCES 138
APPENDIX 150
CURRICULUM VITAE 181

List of Figures

Figure 1 - Record linkage framework 14
Figure 2 - SME growth modeling selection criteria 35
Figure 3 - CAGR by age of driver 46
Figure 4 - Revenue by age of driver 47
Figure 5 - Variable importance of LogitBoost growth prediction model 52
Figure 6 - Framework of growth prediction study using BFS, HReg and patent data 64
Figure 7 - SME growth by patent count 75
Figure 8 - Data description and first name distribution 77
Figure 9 - Percentage of females insured to total 77
Figure 10 - Female share of the name “Gabriele” 78
Figure 11 - Nationality differences for the name “Andrea” 79
Figure 12 - SME growth by patent application gender 86
Figure 13 - Random forest feature importance without patent data 88
Figure 14 - Random forest feature importance with patent data 88
Figure 15 - The conceptual model of word-of-mouth by Litvin et al. (2008) 95
Figure 16 - eWOM research streams: perspective 1/4 96
Figure 17 - eWOM research streams: perspective 2/4 97
Figure 18 - eWOM research streams: perspective 3/4 97
Figure 19 - eWOM research streams: perspective 4/4 99
Figure 20 - Data sources overview and gross value added Hotels and Restaurants 111
Figure 21 - Hotel growth prediction feature importance 114
Figure 22 - Restaurant growth prediction feature importance 116

List of Tables

Table 1 - Sectors of the Swiss economy (2013) 22
Table 2 - Legal forms of Swiss SMEs (2013) 22
Table 3 - Most frequent industry types (year 2015) 28
Table 4 - Highest revenue risk industries (year 2015) 29
Table 5 - Lowest CV industries (year 2015) 30
Table 6 - Highest CV industries (year 2015) 30
Table 7 - CAGR most common industries (all years) 31
Table 8 - Highest CAGR industries (all years) 31
Table 9 - Lowest CAGR industries (all years) 32
Table 10 - Lowest multiplication score 33
Table 11 - Covariates for growth prediction with policy data 41
Table 12 - CAGR by age of car and leasing status 47
Table 13 - CAGR and age of driving test 48
Table 14 - CAGR and vehicle type 49
Table 15 - CAGR and fuel type, leasing, age of driver and car 49
Table 16 - CAGR, car brand and age of car 50
Table 17 - Confusion matrix vehicle policies 51
Table 18 - Insurance data set covariates 62
Table 19 - Commercial register (HREG) covariates 62
Table 20 - Swiss federal statistical office (BFS) covariates 63
Table 21 - Patent database (LENS) covariates 63
Table 22 - Distribution of the name "Peter" 69
Table 23 - Distribution of the name "Gabriele" by decade 70
Table 24 - Distribution of the name "Andrea" by nationality 70
Table 25 - Non-Swiss dominant names 71
Table 26 - Patent and SME counts of industry groups 75
Table 27 - Nationality imputation with no information replacement of non-mapped names 81
Table 28 - Nationality imputation without non-mapped names 81
Table 29 - Overall nationality mapping evaluation 81
Table 30 - 2-label model performance 83
Table 31 - 3-label model performance 84
Table 32 - Counts of SMEs and patents by gender 85

Table 33 - Random forest prediction results with and without patent attributes 87
Table 34 - Annual changes in gastronomy hospitality revenues in Switzerland 102
Table 35 - Average revenue per room night available for hotels in Switzerland 103
Table 36 - Description of additional data 105
Table 37 - Description of swisshotel data 106
Table 38 - Description of Swiss Federal Office of Statistics data 107
Table 39 - Description of TripAdvisor data 108
Table 40 - Study results hotel growth prediction 113
Table 41 - Study results restaurant growth prediction 115
Table 42 - Summary of research questions and findings 129

Abbreviations

AG	Aktiengesellschaft
ANOVA	Analysis of variance
B2B	Business to business
B2C	Business to customer
BA	Business analytics
BFS	Swiss Federal Statistical Office
BI	Business intelligence
CAGR	Compound annual growth rate
CASCO	Casualty and Collision
CEO	Chief executive officer
CHF	Swiss franc
CRM	Customer relationship management
CV	Coefficient of variation
D-MTEC	Department of Management, Technology and Economics
ETH	Swiss Federal Institute of Technology
eWOM	electronic word-of-mouth
Fis	Financial institutions
FN	False negative
FP	False positive
GmbH	Gesellschaft mit begrenzter Haftung
GVA	Gross value added
H&R	Hotels and restaurants
ID	Identity document
IGE	Institut für geistiges Eigentum
IN	Imputed nationality
IP	Intellectual property
LENS	Joint initiative by Cambia and Queensland University of Technology to build a global patent database
LI	Liability insurance
LMT	Logistic model tree
MAE	Mean average error
ML	Machine learning
MTEC	Management, Technology and Economics
NIR	No information rate
P( )	Probability (of)
R&D	Research and development
RL	Record linkage
RMSE	Residual mean square error
RQ	Research questions

RSD	Relative standard deviation
SECO	State Secretariat for Economic Affairs
SME	Small and medium-sized enterprises
SUV	Sport utility vehicle
TFP	Total-factor productivity
TN	True negative
TP	True positive
URL	Uniform Resource Locator
WOM	Word-of-mouth
w/o	without
yoy	Year-on-year

I) Introduction

"If we have data, let's look at data. If all we have are opinions, let's go with mine."

– Jim Barksdale, president and CEO of Netscape from January 1995 until the company merged with AOL in March 1999

I.1 Motivation

I.1.1 Managerial Motivation

The small business insurance market is currently largely operating through offline agents and insurance brokers (Hobey, 2017). They generally act independently and serve their local customers, prioritizing their portfolios according to their own judgment. In many cases, the portfolios contain several hundred customers, forcing the agents to prioritize among them. Further, due to the competitive nature of the commercial insurance market as well as the maturing digitalization of insurance sales channels, margins are under pressure. The already intense competition is likely to ramp up further over the next few years (McKinsey and Company, 2016). The market is both fragmented and profitable, a scenario that is drawing attention from insurance companies, whose primary business lines are saturated and commoditized, as well as from attackers seeking fields that are open for innovation (McKinsey and Company). In addition, the increasing transparency and customers seeking offers from competitors when negotiating policy renewals has exposed traditional insurance to price pressure.

To remain competitive, and with the goal of keeping customers as loyal clients, insurers may have to consider the expected future premium payments of a client when pricing competitive offers under uncertainty and when budgeting agents' relationship-building efforts. Customer-retention activities are beneficial for business insurers since existing customers tend to have better risk profiles; assuming an insurance company does not have to adjust pricing to reflect the risk over time, the insurer can collect higher premiums from long-lasting business relationships (McKinsey and Company). It is therefore in the interest of the

insurance company to maintain a close relationship with all of its customers, provide suitable products and keep an up-to-date record of the company's current economic status and also actual risk. To achieve this goal currently, agents try to meet frequently with the most relevant customers in their portfolio, which is increasingly more difficult to do as portfolios become larger.

Within the business-client portfolios, the most commonly sold commercial risk insurance is general liability insurance, which provides businesses with coverage for damage caused to others as a result of their business operations (Henry, 2016). Other business insurance includes property, business owners' policy, commercial auto, group benefits, workers' compensation and professional liability (Cusick, 2016). The potential damages, which are most often covered by general liability are bodily injury, property damage, finished products liability, personal injury, slander or defamation (Henry, 2016). General liability insurance is primarily priced by two components. First, a merit rating, considering the class rates based on the industry risk average, which is then adjusted upward or downward based on individual loss experience. And second, the company's exposure to risk, which expresses risk as a quantity and is roughly proportional to the risk of a policyholder or a group of policyholders. The type of business is stable over time for most companies. However, risk exposure is measured as the insured revenue fluctuates and needs to be updated regularly. Especially for small companies, where insurance agents have less of an incentive to maintain a close relationship, insured revenues are often not updated upward in a timely manner. This prevents the insurance company from collecting appropriate premiums and leads eventually to a loss of profit for the insurance company.

Further, the insured company's maximum revenue is addressed as "maximum allowable turnover limit" in most insurance policies. If an insured company fails to report changes in its revenue, most insurers are no longer under the obligation to fully cover the company's risk. In the event of a claim, the customer and the insurance company then experience a challenging period in the business relationship. The insured expects the insurer to pay for the claim; the insurer, however, was not compensated appropriately for the risk supposedly covered. Hence, it is in the interest of both the customer and the insurance company to discover early if a customer's maximum allowable turnover limit will be or has been reached.

In order to determine whether a customer's maximum allowable turnover limit, by design equal to the company's revenue, needs to be updated, an insurance company has many options in receiving the necessary information. Avoiding the manual process of agents reaching out to all customers, the insurance company can use already collected information about a customer, which can be combined with other data sources to identify those customers who are expected to have increased revenues.

Integrating the outcome of this envisioned process into a CRM (customer relationship management) system represents a scalable solution to identify SMEs with growing revenues. Hence, a smart information technology infrastructure may provide the opportunity to allocate agents' resources efficiently and price policies more competitively by incorporating the expected economic future of clients, increasing the insurance premiums collected during the clients' business relationship and avoiding that customers are underinsured in the event of a claim.

Compared to other industries, insurance companies operate in a data-rich environment, which enables them as an industry to position themselves as pioneers when engaging in customer-insight activities. Data are of an increasingly central role in most organizations (Wedel and Kannan, 2016), but insurance companies especially have a promising position to use historic operational information to their advantage. They have by now several years of operational history about customers' transactions, communications and more digital data in large quantities and in structured and searchable form. Therefore, they are more empowered than ever to make use of advanced data analytics, which can generate accurate customer insight.

Moreover, insurance companies can expand their view toward external data sources, which may provide enhanced insights about their customers' current situations (Mau et al., 2016). This development involves using data collected on customers from inside and outside the firm environment (Wedel and Kannan, 2016). Applying automated probabilistic methods on the collected data can enable the process of determining which business customers are growing, hence contributing to achieving the elaborated goals, including improvements in agents' time allocation, competitive pricing, customer retention, relationship management, premium optimization and avoiding underinsurance. Therefore, such methods provide an economically useful solution for traditional insurance, with large volumes of historic customer data to increase premiums collected and maintain good customer relationships by

improving the functionality of their relationship-management systems with the help of growth-prediction models.

I.1.2 Research Motivation

Outside of the commercial insurance industry, the need for adoption of data mining techniques to support decision-making in a customer-driven industry has already been recognized as essential for targeting customers effectively. Due to the emergence of novel techniques and computational power, attention was turned to less intuitive methods (Williams et al., 2006) that offered improved prediction results and the ability to perform large-scale calculations on very large data sets with high dimensionality. Examples include supervised and unsupervised machine learning (ML) methodologies that have been used extensively to improve economic decision-making problems. Various studies have shown that ML techniques such as decision tree algorithms can be employed as an alternative method to resolving classification problems instead of the traditional statistical methods (Brachman et al., 1996). Traditional statistical methods use restrictive assumption (Gordini, 2014), whereas ML techniques are characterized by a high degree of flexibility when implementing different algorithms. Current research regarding the use of predictive models for business intelligence purposes covers many different types of machine-learning methods. For example, decision trees were applied to address a common task of customer classification. Yet, this task was shown to be highly domain-specific, thus indicating a need for sector-specific implementation and domain knowledge as key factors for success (Wu et al., 2005).

In the financial industry, there is plentiful documentation for the application of data-mining and ML methods in optimizing operational outcomes. Relevant studies can be found in the domains of marketing, bankruptcy prediction, fraud detection, and operational research. Smith et al. (2000) had the objective to predict the retention probabilities of motor insurance customers using a sample of an Australian insurance company. Their goal was to increase efficiency and achieve market growth through data mining by detecting specific kinds of customers through a predictive model. Only personal customer data (e.g., age of vehicle, gender, etc.) was used. The authors demonstrated three different data-mining techniques for classification: logistic regression, decision trees, and neural networks. The latter one outperformed the others and was applied to a holdout sample (test set). Still, the chosen neural net model suffered from a low recall (<25%) when classifying customers who

had terminated their policy. In practice, three out of four customers could not be identified by the model or targeted through marketing activities.

ML techniques are applied in the business growth literature mostly to classify between successful or failing firms and to a lesser extent to provide numerical evidence comparable to regression results (Kurt et al., 2008). A prominent study was conducted by Gordini (2014) who applied classification algorithms to identify bankrupt SMEs in the manufacturing sector. Only traditional financial ratios, derived from the income statements and the balance sheets of the SMEs, were included in the model. The author tested several models and achieved 78.8% accuracy. This and similar studies refer to other authors investigating the underlying causes of business success. These date back to the beginnings for corporate finance. Those studies published from the 1930s to the 1970s (Jardin and Séverin, 2010) investigated the information capacity of income statements and balance sheets, found reliable performance predictors, and established the use of multivariate statistical methods in the field of economic growth prediction of companies. In newer studies on economic forecasting or operations research, the variables most often used to predict the future state of a company are based on the findings by the very first authors that published between the 1930s and the 1970s (Jardin and Séverin, 2010). These authors investigate the information capacity of financial-reporting data sources, replicating or expanding the previous findings, utilizing the plethora of possible ratios based on the financial numbers.

The variable selection of the many authors that only looked at publicly available financial information, is somewhat restricted in that it captures only financial information, despite the many available data sources provided by today's information systems. Also, the limitation of research to large companies that reveal such financial information restricts the generalization of potential findings since most small companies are informationally opaque. However, at the same time, they are much more representative of the average firm within most countries (Capon et al., 1990).

While previously listed studies apply different methods on mainly financial data sources, many of them suffer from the above-mentioned variable selection biases (survivorship of historically important financial variables). Hence, they are restricted to the discovery of new indicators to explain variance of SME growth.

Some authors have ventured into other data sources and alternative methods to select variables (Marom and Lussier, 2014; Salavou et al., 2004). The investigation of other data

sources to predict business performance is of particular interest since 98% of all companies are small businesses and do not produce income statements or balance sheets and are deeply interwoven with the economic situation of the founder (Brandstätter, 2011). Initially addressed by Westhead et al. (2001), there is potential to discover new information indicating revenue growth using many other data sources. However, these premises have not been thoroughly investigated in academic studies and thus lack a theoretical confirmation (Simpson et al., 2013). To the best of my knowledge, there is no research on the relationship between insurance internal data sources about SMEs, augmented or not by additional publicly available information, and SME performance. Research in the direction of the effectiveness of other data sources exists but is still very limited (Hussain Naqvi, 2011). As a result of my systematic review of related literature, I conclude that none of the studies conducted an evaluation of the for-insurance available data sources to describe SMEs' financial success, which further encouraged me to address this gap. According to Reimer and Becker (2015), companies and researchers need to consider adequate data sources to achieve their customer analytics objectives. The study hinted that customers' action data possess greater predictive power compared to their personal data. Following their recommendation, the guiding theme of this thesis is to combine personal data (i.e., a customer's covariates from the operational database) with other data sources, considered to be indicative of a customers' business activities and subsequently its success.

To address the above-identified research gap, this thesis aims to evaluate the potential data sources commonly available for insurance companies as well as other accessible sources. I will further describe the research questions that lead toward addressing the mentioned research gap. Building upon these opportunities, this thesis aims to answer the following research question:

> *How can insurance companies predict the business growth of small and medium-sized-enterprises using internal and external data sources?*

I.2 Research Questions

Section I.1.1 has provided an overview of why insurance companies have an incentive to predict and update the insured revenues of their SME customers. Also, it was argued that insurance companies are at a promising starting point to forecast the revenue growth of their

SME customers. They have large volumes of historic customer data that can be augmented with external data and analyzed to derive insights. In order to find answers on how to achieve this goal, I will answer to the following research questions in the form of four case studies.

I.2.1 Research Question 1

RQ1: For which SMEs should insurance companies predict revenue growth?

When engaging in efforts to predict revenue changes of SMEs, it may not be advisable for insurance companies to equally address all industries or SMEs of all sizes. “Divide and conquer” is a widely adopted strategy to deal with large-scale optimization of this kind, where the key issue is to detect the correlational relationship between the decision variables so that correlated relationships are grouped into the same subpopulation and independent relationships grouped into different subpopulations (Zhou et al., 2014). The challenge here is how to identify the SME cohorts with which to build models. Instead of building a scaled predictive model for all SMEs, I suggest building a community of micromodels, which as an ensemble operate over all business clients of the insurance company. This is done by understanding and dealing with the nuances and idiosyncrasies of the different SME cohorts. It is within the much smaller subpopulations where the predictive models are trained and evaluated. The final is then the ensemble of individual models, which can be applied to each new observation (Zhou et al., 2014). Without prior insight in building subpopulations, insurance companies can apply economic and statistical considerations that are beneficial for the outcome and relevance of the predictive model. This has several implications. For example, SMEs that have very low revenues, particularly if there is only one person selling his or her own (unskilled) labor through the SME, may be economically irrelevant for an insurance company as it is unlikely to grow. Also, SME cohorts that are very low in overall counts, such as rare industries, may not be relevant for the insurance. When SMEs within an industry are very similar in terms of individual growth or declining revenues over several years, a prediction may not be advisable because there is no expectation of positive change. By answering RQ1, I will investigate for which kind of SMEs it is worthwhile to forecast revenues and how to group them in subpopulations.

I.2.2 Research Question 2

RQ2: *How can insurance companies predict the future revenue growth of SMEs with vehicle insurance data, and what are the benefits of such models?*

Before engaging in costly external data-collection endeavors, insurance companies favor the exploitation of already existing customer records such as vehicle insurance policies. Many attributes related to the customers and the insured objects are stored in these policies and insurance companies may find valuable insights in the archives of current and old business transactions. These can be used as input for revenue growth prediction models.

I.2.3 Research Question 3

RQ3: *How can insurance companies improve revenue prediction models with the help of patent filings and how can these models be applied?*

In economics, total-factor productivity (TFP) is the portion of economic output not explained by traditionally measured inputs of labor and capital used in production (Comin, 2008). By and large, what is not measured is the economic impact of innovation. SMEs with a high level of innovation and the ability to commercialize their innovations should therefore be growing at a higher rate compared to their not-so-innovative peers. By answering RQ3, I will propose patent filings as a proxy for innovation of SMEs, and investigate the prediction capacity of this information source.

I.2.4 Research Question 4

RQ4: *How can insurance companies predict the future revenue growth of hotels and restaurants?*

Having investigated for which kind of SME customers it makes economic and statistical sense to predict revenue growth, I answer RQ4 on the example of the hotels and restaurants, which has been identified as important (see RQ1). By collecting and evaluating publicly available data which reveal insights about hotels and restaurants, I investigate whether I can predict the future state of a business customer in that industry.

I.3 Research Design

I.3.1 Statistical Modeling of Data Problems

In various scientific disciplines, statistical modeling is a widely applied and powerful tool to address empirical research questions such as the ones introduced about insurance companies in section I.2 (Shmueli and Koppius, 2010). There are two cultures in the use of statistical modeling to reach insight from data. One assumes that the data are generated by a given stochastic data model and underpins the need for a framework to understand the properties and scope of methods used in applications (Davison, 2003). The other one uses algorithmic models and treats the data mechanism as unknown (Breiman, 2001). Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics, such as medicine, economics, psychology, education, and environmental science (Jordan and Mitchell, 2015). The advocates of algorithmic modeling assume that the importance of "causality" will be overtaken by "correlation." Trying to discover causality represents a great search for an in-depth understanding of the data; however, it remains challenging in many real domains. For the statistical modeling of data problems such as the ones introduced, I emphasize that correlation is far from sufficient, and the role of causality cannot be replaced by correlation, especially when the purpose of statistical modeling is to find evidence that will later be applied in a business settings, such as insurance pricing and customer-relationship management (Zhou et al., 2014). The value of algorithmic modeling manifests wherever variable data must be summarized or is used to test or confirm theories or to inform decisions. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets (Breiman, 2001). The larger the samples, the easier it is to find evidence for a significant difference between cohorts, given the existence of an actual difference. Data size was long considered a bottleneck for traditional data analytics operations that were not able to analyze very complex data sets. These tasks can now be performed by machine learning, which leads in some cases to an augmentation of the cognitive ability of humans (Ben-David and Shalev-Shwartz, 2014) and helps to uncover more fine-grained patterns, enabling practitioners and researchers to make more timely and accurate predictions than ever before (Zhou et al., 2014). The key feature of machine learning is that variability is represented using probability distributions, which form the foundation from which models are later created. Typically, they accommodate both systematic and unsystematic variance. The randomness inherent in the probability distribution accounts for

the apparently random scatter in the data, and the systematic pattern is supposed to be recognized by the structure in the model (Davison, 2003). Applying the model to data, it has not seen yet and is often referred to as a "prediction." Findings of statistical modeling often provide a measure of quality with respect to the pattern or effect size they find. A common metric to evaluate the quality of a model is to provide accuracy scores (Jordan and Mitchell, 2015). Learning from data to achieve high accuracy scores is a complex task, and its validation is often disillusioning for those hypothesizing on what one can learn from the data. As such, and it is sometimes discouraging for those searching for insight, predictions are rarely perfect. There are usually many unmeasured variables whose effects are referred to as "noise." McCullagh and Nelder (1989) summarize in their book on statistical models that at first sight, it might seem as though a good model is one that fits the data very well—a model whose predicted value is very close to the response value. But the researchers note that the extent of the agreement is biased by the number of parameters used in the model and is therefore not a satisfactory measure. What the authors refer to, given a model that has too many parameters, is overfitting the data, which results in a biased estimate of accuracy and unsatisfying results when applied to data that the models have not seen. The concern with overfitting led to a natural favor for simpler models with less parameters (Wedel and Kannan, 2016). There are two alternatives to get a more unbiased estimate of a model's predictive accuracy while allowing larger quantities of parameters. First, the benefit of machine learning lies in the fact that with more and more samples available for learning, the risk of overfitting becomes smaller (Zhou et al., 2014). Hence, larger samples can be collected. Second, cross-validation can be performed as advocated in an important early work by (Stone, 1974). Cross-validation is especially suitable then, when data sets are larger and a test set can be put aside (Breiman, 2001).

Generally, the complexity of the model will depend on the problem at hand and the answer required, so different models and analyses may be appropriate for a single set of data (Davison, 2003). Taking the above into consideration, when answering the posed research questions, I combine stochastic data models to describe the data and algorithmic models when performing predictions based on insurance and additional data sources. With the help of modeling, it is my goal to show that positive or negative growth can be predicted with the available data sources. Further, I provide ideas of how and why additional models can be built with a high level of confidence. However, I also address the many reasons, including the

sample size, a small effect, and lagging effects, why those in search of patterns may not be able to find a good predictor for every SME. Therefore, when planning to add external data sources, it is prudent to choose SME cohorts sufficiently large in count and to choose a reliable data source that is available and contains information revealing indicators of growth that can be extracted.

I.3.2 Relevance of Business Analytics for an Insurer

As competition intensifies in the commercial insurance market, a winning strategy, as stated in a 2016 business report by McKinsey and Company, will be less defined by industry- or business-sized specialization and more defined by identifying and targeting specific customer segments. A first step for such a customer segmentation is to build in-house predictive analytical skills. They allow insurance companies to understand which customers are likely to generate the largest profits, which customers are most amenable to up-selling and cross-selling (Accenture, 2011), and how these customers react to premium amendments. For example, in the United States, a study relating to customer segmentation and competitive pricing found that especially small service and retail companies have simple, basic coverage needs (less than $250 000 in annual revenues) and are very price sensitive. But there are also larger businesses across a broad range of industries that are highly price sensitive (McKinsey and Company). Adopting a segmentation approach to other, more specific types of cohorts while adding revenue growth and insurance relevance as considerations eventually enables one to contact priority growing SME customers and see where the premiums can be increased without churning the customers. The insurance agents can then, besides advising and offering additional insurance products, accurately update the insured SME revenues and engage in competitive policy pricing whenever required to retain the customer. As a result, the insurer knows with more certainty whether or not a customer is worth keeping, how much investment in the customer is justified, and how to fortify strategies for winning (Accenture, 2011). Advances of customer retention measured in the commercial insurance segment is of particular importance for two reasons. First, existing customers tend to have better risk, which further improves over time. Second, the pool of potential switchers in any given renewal cycle represents almost 50% of the market (in the United States). To reach them, insurance companies must harness new and more granular insights about the customer (McKinsey and Company). At a high level, these predictive analytics empower insurers to invest in the best growth opportunities by using data from across the enterprise to make

accurate predictions about customer economic future (Accenture, 2011). At the same time, it is not sufficiently clear which types of analytics work for what types of problems and data, what new methods are needed for analyzing new types of data, or how companies and their management should evolve to develop and implement skills and procedures to compete in this new environment (Wedel and Kannan, 2016), with the goal of reigniting profitable growth (Accenture, 2011).

I.3.3 Record Linkage

The modeling of data to derive insights is generally contingent of high-quality data input. Before designing a statistical model to derive managerial insights, the relevant data must be gathered. In many cases, insurer business analytics outcomes can benefit from adopting an enterprise-level data-sourcing approach. This means avoiding a silo approach to analytics that can impede an insurer's ability to look across all of its data sources to make better decisions for the enterprise as a whole (Accenture, 2011). Insurers must understand their data, ensuring high-quality levels for internal data, and use a creative approach to sourcing external data (Accenture, 2011).

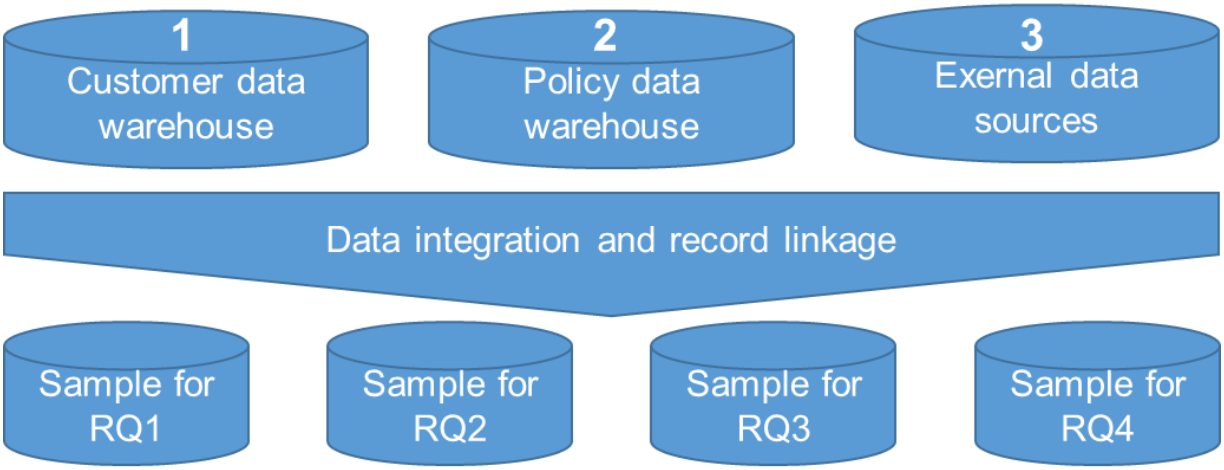
Therefore, in the quest for creating insightful complete data sets, the combination of different databases is a common task to be performed. The process of integrating different data sources is called record linkage (RL). It is also known as entity resolution or duplicate detection and is a deterministic or heuristic process in which tuples that represent the same real-world entity are identified (Altowim et al., 2014). Record linkage has applications in customer systems for marketing, fraud detection, customer-relationship management, data warehousing, and legal and government administration (Gu et al., 2003). Despite the maturity of this method, the high degree of heterogeneity in data structures, the absence of unique identifiers used for the matching process, and the cross-organizational collaboration on a data level still represents a challenging problem (Duggal and Soni, 2016). There has been extensive research on data integration architectures, RL methods, and schema matching techniques (see Duggal et al., 2016 and the references therein for a detailed overview). The integration of data sources has gained popularity as companies have realized that the data they own are becoming one of their most valuable assets (Herschel, 2008). Improving the quality of that data was shown to hold potential for reduction of operational inefficiencies (Kuzu et al., 2011). Moreover, to obtain regulatory compliance and improve their customer relationship strategies, many industries require accurate and complete customer information (Albrecht,

2016). However, acquiring such data is a task that is sometimes not feasible due to the limitations imposed by the design of existing systems and the level of details in the stored customer data, which is usually focused only on attributes relevant for the specific business context (Duggal and Soni, 2016). Some of these challenges might be addressed by merging data from external sources, thus gaining new insights and thereby enhancing the organization's record base. Instead of feeding solely records of traditional customer data from operational databases into statistical models, such customer information can be enriched with additional data from various sources. According to Verhoef et al. (2010), data integration at the individual customer level and across all operational databases is fundamental for firms to get a holistic view of all related customer activities. Still, not many companies have completed this task yet. The data sets that were provided by the insurance company and collected from other sources were not in a single database. Therefore, to investigate the research questions introduced before, several data sets had to be integrated in a preliminary step.

The data provided by the insurance company includes records of customers and their insurance policies. The first data set that is extracted from the operational database of a Swiss insurance company and includes various information (e.g., the name of a company/company owner, the owners' age/foundation year, the address, and how long the business is a customer with the insurance). The data sample spans a period from 2008 to 2016 and includes 111 284 SME policies. These records also include a unique identifier (ID) that links a customer to the purchased insurance products. The focus is especially on the insurance products general liability insurance and motor insurance. The general liability insurance product includes coverages of business liabilities, especially for damage caused to others as a result of their business operations. These include damages such as bodily injury, property damage, and damages on finished products. The vehicle insurance product contains the obligatory third-party liability coverage and optional coverage such as collision damage and accident. In order to address each research question adequately, the relevant customer and policy information is selected from the overall sample for each use case and described in detail in the corresponding chapter. With the focus of this thesis being on the information for the prediction and the application in the business insurance sector, I selected two additional data sources that will be discussed in detail. Further, a few auxiliary external public data sources are used to augment the data set. In combination, the joined data are expected to reveal information about SME business growth and allow me to answer RQ3 and RQ4. First,

however, the records of the external data sources have to be linked to records of existing customers from the operational database. The framework illustrated in Figure 1 shows the process of the record linkage and data extraction to answer the research questions.

Figure 1 - Record linkage framework



The first source (1) refers to the operational data of the insurance. The second source (2) includes data that relates to individual SME customers policies. Many SME clients have several policies. They are used to enrich the original customer data. In order to integrate these data sources with the operational data warehouse, the records are linked based on a unique common covariate, resulting in a 1-to-*n* relationship.

The third source (3) that relates to external sources is also integrated with the information from the customer data and policy data warehouse. However, due to the absence of a unique covariate (i.e., a customer ID), several approaches for the linkage of the customer data and additional data were chosen.

I.3.4 Data quality assessment

One main assumption of this study is that the business performance of executives who apply data driven decision making is influenced by the quality of the models they use. The used models are generally trained on available historic data. Lacking data quality, can cause the models to perform well on a given data set, but to underperform in the more generic business context. Therefore, data quality has to be evaluated, before answering the proposed research questions and making suggestions for business professionals.

The evaluation process of data quality is dependent on the business need and the data consumer. In this study, consumers are expected to be insurance executives who use the prediction models to identify growing and shrinking SME customers. The users assume the specified models are based on high quality data, accurately reflecting the objective information about SMEs. Users' measure the objectivity of data is based on the degree of objectivity versus the degree of judgment used in creating it (Fisher et al., 2012). Objectivity alone is not sufficient for data to be considered of high quality. Data quality assessment needs to include other aspects such as completeness, timeliness and accessibility (Samitsch, 2015). A conceptual framework to assess data quality, capturing the most relevant data quality aspects, has been developed by Wang and Strong (1996). The framework contains for main elements. These element are (1) intrinsic, (2) contextual, (3) representation, and (4) accessibility data quality.

Intrinsic data quality refers to the objectivity of data. In the context of the research proposed, the most central information are the revenue numbers contained inside the general liability policies. These revenue numbers serve as the studies' ground truth and are used to train and test models which are discussed throughout this dissertation. How similar SME's actual revenues are compared to reported revenues in the insurance policies, remains yet unanswered. This is because no other valid data to be considered objective is accessible for comparison. For most SMEs it is neither practical to report exact revenue numbers, nor do all SMEs follow the same revenue recognition principles which limits comparability. The businesses themselves, tax authorities and insurances accept some freedom of judgment, accounting for the periodicity and complexity of the business activities. Given the nature of the data creation process, I expect the insurance reported revenue numbers to reflect tax compliant revenue numbers, which are close to the actual revenue numbers, but rounded on thousand CHF. The expectation of proximity between reported and actual revenue numbers can be further justified with the help of the business logic. Business liability insurance is optional for most SMEs. Especially those SMEs do not have an incentive to provide false information. If SMEs do not want to report their actual revenue numbers to the insurance, they would either not buy the product at all or search for bespoke alternatives. Further, all SMEs insurance clients are informed that intentionally understating revenue numbers can lead to the revocation of an insurance's obligation to cover a future claim in full. Additionally, the insurance product is designed that if actual revenue numbers divert from reported

numbers, the insurance obligation is reduced overproportionally, due to first loss absorbing capacity of the difference between actual and reported revenue numbers. This also reduces the probability of a claim to be covered by the insurance to less than the probability the insurance was initially priced with. Hence, there is no economic incentive to pay for an insurance product and intentionally limiting the products' benefits. The existence of SMEs, that would provide false information can, despite the lack of rational reason to do so, however cannot be precluded. Due to the large sample size and the several years of observations, potential misrepresentations are mitigated. Therefore, intrinsic data quality with respect to business need and the data consumer is assumed to be high for all revenue numbers reported by the SMEs at the time of the insurance purchase.

Contextual data quality refers to the relevancy, completeness and timeliness of data. I assume the reported annual SME revenue numbers to be the most relevant source of information to identify growing and shrinking SMEs. The data set provided is complete, as all SMEs that bought the product had to report the revenue numbers for at least one accounting period. The timeliness of the data however remains questionable. Many SMEs report several years of identical revenues, followed by an increase or decline in revenues. This suggests that some of the annual revenues numbers stored in the operational database of the insurance company might not be up-to-date. This is because SME owners may not talk to their insurance agents every year and neglect to update their revenues in a timely fashion, despite the economical motivation to do so. In other words, the stored data sets do not always correspond with their real world counterparts with respect to timeliness. Overall, data quality in respect to the relevancy and completeness is high. Data quality in respect to timeliness of data is questionable. Lacking timeliness of data quality can be mitigated by observing longer periods of time. Further, specifying prediction models to identify growing and shrinking SMEs over longer periods of time bears less potential of over-specification of models, than building models calculating growth, assuming a lack of conformance for data timeliness. To further mitigate this concern, studies should not aim to predict an increase or decrease of revenue at a certain time, but rather predict whether SMEs are growing or not over a longer period of time.

Representational data quality refers to the representational consistency and conciseness of data. To ensure consistency in the collection process, only a single insurance product from the same insurance company has been considered. This ensures the

comparability of revenue numbers among SMEs and years observed. The central value studied is an integer, which is always reported to the nearest thousand CHF. Other currencies are not considered in the data set. Data quality in respect to representational consistency and conciseness is therefore high.

Accessibility data quality refers to how and if data is available, and how well data is secured against unauthorized access. Both, the insurance company when collecting the data and the researchers that worked with the raw data had to use a password protected login. Further, several redundant copies of the same raw data was used for the studies and no signs of manipulated or representation variations have been observed.

Overall the data quality according to Wang and Strong framework (1996) can be considered high. The lacking timeliness of revenue data however limits the scope of the prediction of SMEs' revenue growth. Suggestions how to mitigate issues related to the lacking timeliness have been proposed. In summary, there is no evidence of lacking believability and reputation of the data, which is used as the ground truth. Therefore, data quality is expected to be adequate for the model specification allowing to identify growing and shrinking SMEs with the goal of augmenting the organizational performance of commercial insurance companies.

I.4 Context and Structure of Work

I.4.1 Context of the Research Project

The current thesis and the connected research project were conducted at the Mobiliar Lab for Analytics, a joint research facility between the La Mobilière Insurance Group and the Chair of Information Management at the Department of Management, Technology and Economics (D-MTEC) at the Swiss Federal Institute of Technology in Zürich (ETH Zürich). The Mobiliar Lab of Analytics provides a basis for discussing research-relevant topics between experts from industry and academia. Moreover, through their cooperation, the research project had access to company data from various sources and to knowledge and insights from the practitioners. La Mobilière insurance was founded in 1826 and is the oldest operating private insurance company in Switzerland. Headquartered in Berne, La Mobilière operates exclusively in Switzerland and Liechtenstein and is one of the major nonlife insurers in the domestic market. The sales strategy is based on 160 offices in Switzerland (2015), which represents the most important distribution channel and touch point for business and private customers. The

company describes itself as the “most personal insurance” in Switzerland and actively promotes its local consultancy services (Schweizerische Mobiliar Holding AG, 2016). La Mobilière also sells the majority of its business insurance policies by directly approaching the owners of companies and has the goal of building long-lasting business relationships with them. Overall, the current work aims to provide tools to support the selection of which businesses the insurance should prioritize and how to find out whether the insurance company’s records reflect the actual economic truth of the insured clients, which results in advances in the customer relationship and operational performance. Based on the cooperation of La Mobilière, the topic of this thesis has been introduced and refined in joint discussions with insurance agents and the managers responsible for all commercial clients. In order to address this broad research objective as well as the questions of the mentioned stakeholders, I subdivided the topic into several case studies, which provide answers. Within the first case study, I answer how to determine which SME customers are relevant for the scope of this research. Then, with the help of the other three case studies, I investigate which data sources can be used to build or improve predictive statistical models. The derived insights increase the knowledge for researchers and practitioners of how insurance companies can benefit from collecting and analyzing data that relates to their insured business customers. If successful, the findings of this thesis can be used to increase insurance premiums paid, identify underinsured SMEs, provide strategies to improve customer retention, and prioritize among customers.

I.4.2 Document Structure

The overall structure of the document is aligned with the problem statement, the four research questions, and the corresponding case studies. The preface contains a summary of the overall thesis, the table of contents, the list of figures, the list of tables, and abbreviations. In Chapter I, the research idea of predicting the revenue growth of SME clients is motivated from a practitioner and a research perspective. Then the research questions are formulated. Further, the general methodology for the data collection and the statistical modeling is proposed. Chapter II describes what SMEs are and what their role is in Switzerland. Continuing on, Chapter II addresses research question one (RQ1) and investigates for which kind of SMEs a prediction seems economically and statistically feasible. Chapter III contains the study related

to research question two (RQ2) and shows how, with the help of insurance internal data, the growth of many SMEs can be predicted. Chapter IV investigates by answering research questions three (RQ3) how an indicator of the SMEs' innovation capacity can be used to predict revenue growth. With the last case-study, the focus shifts to a specific industry. By answering research questions four (RQ4), the content of Chapter V shows how a sector-related data source can be utilized to predict the revenue growth in the tourism industry. Finally, in Chapter VI the results are discussed and evaluated in the insurance context. Further, it summarizes the findings of the case studies, offers a critical reflection of the work, and discusses limitations and topics for future research.

II) Growth Prediction for Which Kind of SMEs?

II.1 Introduction

Insurance companies can use predictive analytics to segment its customers into groups sharing similar traits, which can, in a second step, be used to determine which customers offer the best growth opportunities to unlock their own full performance (Accenture, 2011). This chapter describes how to group commercial customers specifically for the purpose of growth prediction. First, I define the term *SME*, and then I describe their economic importance for Switzerland. Further, I will introduce related work discussing previous attempts to predict growth, with a focus on the use of data sources for model building. Continuing, I highlight which considerations have importance for insurance companies when segmenting commercial customers with the purpose of building growth-prediction models. Finally, the results are presented, which leads to a description of the most relevant cohorts that can be established if no prior knowledge exists that dictates segmentation criteria. The aspired findings of this study eventually guide insurance companies to reevaluate common segmentation strategies to reflect the problem characteristic that the segments are created for.

Common segmentation strategies, which are currently used holistically by insurance companies, group customers by business size (annual revenue, number of employees) or present customer importance (McKinsey and Company, 2016). Most companies are not very large companies. They have only a few employees relative and the premiums collected from each individual small company are only marginal to the insurance company. Still, despite their size, they have utmost economic importance to the insurance companies as a group. A further segmentation of this group of customers that are “not large customers” is less obvious. Insurance products are generally customized for large clients, however, become more standardized for the smaller equivalents. To address them as a group, they are often referred to as small- and medium-sized enterprises (SMEs). There is no universal understanding of what defines an SME. For Switzerland, the State Secretariat for Economic Affairs (SECO)

provides a single criterion: the number of employees. Every market-based company, independently from its legal form, is categorized as an SME if it employs less than 250 people. The European Union also uses this threshold value.

In Switzerland, 99.73% of all companies are SMEs according to this definition, and together (in 2013), they employed 67.9% of the working population. An overview of the employment and SME numbers, grouped by employee counts, can be found in Appendix 1. The Swiss economy is, like most advanced economies, a service-driven economy as indicated in Table 1.

Table 1 - Sectors of the Swiss economy (2013)

Sectors	Number of SMEs	Employees
Primary sector	54 565 (9.4%)	158 239 (5.3%)
Secondary sector	90 469 (15.6%)	768 987 (25.9%)
Third sector	435 357 (75.0%)	2 046 178 (68.8%)
	580 391 (100%)	2 973 403 (100%)

SMEs can actively choose their legal form when incorporating. The default when starting or taking over a business, without any preexisting legal form, is the legal form "sole proprietorship."[1] Most SMEs are sole proprietorships; however, joint-stock companies (in German-speaking Europe abbreviated as AGs), employ most people as illustrated in Table 2.

Table 2 - Legal forms of Swiss SMEs (2013)

SME classification	Number of SMEs	Employees
Sole proprietorship	330 978	638 737
Joint-stock company (AG)	115 056	1 546 082
Limited liability company (GmbH)	97 418	427 758
Corporation	10 032	40 828
Clubs and associations	11 240	120 204
General partnership	7 122	29 081
Cooperative	3 238	41 741
Foundation, endowment	1 677	65 835
Limited partnership	971	4 995
Other	2 659	61 493
	580 391	**2 976 754**

Source: Eidgenössisches Departement für Wirtschaft Bildung und Forschung WBF, 2013

[1] In German referred to as Einfache Gesellschaft.

Comparing the number of SMEs with the total employees allows for an estimate of the average employment of an SME by legal form. In respect to the growth, no data are publicly available; however, one could assume that on a granular level, those SMEs that typically only sell their labor time and do not employ other people, have often limited growth potential. As a rather simple approach, it can be used as a starting point for an insurance to select for which types of SMEs a growth prediction does not make economic sense. Since large insurance companies, however, have access to many observations and SME data points, they can also use their own records to find a more powerful metric on how to select a cohort of SMEs for which a growth prediction bears value.

II.2 Literature Review and Hypothesis

II.2.1 Related work of SME Growth Prediction

On a macroeconomic level, economists perform the growth prediction of countries', sectors' or even industries' gross value added (GVA). Predicting and explaining growth on a more granular microeconomic level is performed primarily by business economists investigating individual firm or cohort characteristics. The discussion around measuring and forecasting business performance of companies of any size is as old as the term strategy itself (Morgan and Strong, 2003). Capon et al. (1990) found in their meta-analysis on firm performance that the majority of research is based on market or accounting measures as an independent variable. Other studies also include the role of innovation (Roper, 1997), marketing (Morgan, 2012), quality management (Rust et al., 2016), and human capital (Prajogo et al., 2018), including many studies that examine the characteristics of founders and managers (i.e., by investigating the entrepreneurial orientation and performance relationship, in Covin and Miller, 2014; Covin and Wales, 2012); these mostly took place in Anglo-Saxon economies (Van Doorn et al., 2013).

II.2.2 Corporate Evaluation of SME Growth Through Business Analytics

Besides researchers, firms themselves are also engaged in forecasting the performance of other firms. This is for the purpose of budgeting their production strategies and creating and attributing marketing budgets and the strategic orientation a firm pursues. The analytical

investigation of firms with respect to the growth of their own or prospect customers is often referred to in the context of business analytics (BA). BA has become increasingly important as firms seek to improve firm performance, generally captured by financial (e.g., return on investment) or nonfinancial (e.g., customer satisfaction, sales growth) measures (Bharadwaj et al., 2013; Sidik, 2012). Research on BA, which lies at the junction of information science and customer behavior, is encouraged by practitioners and academics alike (Chen and Storey, 2012). In practice, this often means evaluating the use of internal and external data sources to provide behavioral and financial insights about customers in the governmental, private, and business sectors. These insights can be translated by business analysts to an organization's advantage. Insurance companies make heavy use of BA, striving to grow and improve the value-creation opportunities with their clients (Auge-Dickhut et al., 2016). The term *analytics* refers to the methods applied to discover hidden patterns in the collected data (Erevelles et al., 2016). While the majority of marketing research in the BA domain focuses on private households or individuals, there is also a growing body of BA research in the context of business customers (Bharadwaj et al., 2013). An application of BA in the business-to-business customer segmentation of customers based on expected growth. Market segmentation is a key decision area for organizations in all sectors (Weinstein, 2017). The concepts originates in economic pricing theory, which suggests that profits can be maximized when a firm discriminates between segments (Wind, 1978). Grouping customers with similar importance or behavior aids organizations in dealing with market heterogeneity (Venter et al., 2015), thereby focusing resources on relatively homogeneous customer segments (Venter et al, 2015) and thus ensuring an efficient allocation of resources.

In the past, insurance companies produced more data than they were able to analyze and use to their advantage (Pome and Bilderbeek, 2014). However, due to today's advanced computational methods, the data available from clients can be processed, leading to entirely new ways to analyze and classify customers (Erevelles et al., 2016). Although the potential of computational methods of large volumes of data may have been overhyped initially, and companies may have invested too much in data storage and not enough in analytics, it is becoming clear that the availability of large data sources can spawn data-driven decision-making for many operational tasks (Wedel and Kannan, 2016). These include, for example, customer segmentation with the goal of finding cohorts of SMEs that are most meaningful to insurance companies. As infrastructure availability and data availability are less of a concern,

insurance companies are at the forefront to build the analytics skills required to detect customer underinsurance, to collect additional premiums from existing customers, and to price customer policy more competitively to avoid churn. To be effective in building revenue growth prediction models, insurance companies have to consider statistical and economical aspects when determining suitable cohorts to which models can be applied to. By examining the below stated hypotheses H1.1 to H1.4, I will investigate how, with the help of business analytics, insurance companies can segment their commercial customer base specifically for the purpose of finding suitable cohorts to build growth-prediction models.

H1.1: *The relative industry frequency holds importance and can be evaluated in the context of SME growth prediction.*

H1.2: *The industry revenue risk contribution plays a role when building a growth prediction model.*

H1.3: *The heterogeneity of a cohort plays a role when predicting the revenue growth of SMEs.*

H1.4: *The historic cohort growth plays a role when predicting the revenue growth of SMEs.*

Combining the findings of H1.1 to H1.4 will answer to research question one.

II.3 Research Design

II.3.1 Data Set and Data Linkage

In order to investigate the stated hypothesis, the relevant data has to be extracted and combined to suit the requirements of the study. From the data warehouse of the insurance company, the name of the SME and its internal unique identifier was extracted on request. From the policy data warehouse, all attributes related to the SME's general liability insurance were extracted. The record linkage is performed on the unique identifier, which results in 111 284 matches for at least one year of revenue.

II.3.2 Used Variables

The merged data set from both data warehouses includes the following attributes:

- Unique identifier
- Industry type (distinct, one out of 1 129 possible choices)
- Revenues per year expressed in CHF
- Year in that the revenues have been produced

To answer research question 1, these four attributes are chosen and evaluated for all SMEs. I limit the selection of revenues and corresponding years to earliest year (t_{-i}) at which a record of revenue exists and the most recent year (t_0) a revenue record was stored in the database.

II.3.3 Methodology

From the extracted attributes, I calculate the SME's compound annual growth rate (CAGR) according to Equation 1. The CAGR calculation will serve as the most relevant growth indicator and is referenced throughout this thesis.

Equation 1 - Compound annual growth rate

$$CAGR = \left(\frac{SME\ revenue\ t_0}{SME\ revenue\ t_{-i}} \right)^{\frac{1}{i}} - 1$$

Further, the arithmetic mean of CAGRs and the arithmetic mean of the revenues of a cohort (i.e. an industry), as well as the standard deviation of the cohort, are calculated according to Equations 2, 3, 4 and 5. Summing the revenues of a particular industry of n observations in year t is thereby noted as t_n.

Equation 2 - Mean of CAGR

$$\mu(g) = \frac{\sum CAGR}{sample\ size}$$

Equation 3 - Mean revenue of cohort

$$\mu(r) = \frac{\sum SME\ revenue[t_n]}{sample\ size}$$

Equation 4 - Standard deviation of cohort revenue

$$\sigma(r) = \sqrt{\frac{\sum(SME\ revenue\ t_n - \mu(r))^2}{sample\ size - 1}}$$

Equation 5 - Standard deviation of cohort growth

$$\sigma(g) = \sqrt{\frac{\sum(CAGR - \mu(g))^2}{sample\ size - 1}}$$

Further, I calculate the coefficient of variation (CV) of the cohorts. The CV is the ratio of standard deviation (σ to mean μ). The main purpose of finding coefficient of variance is used to study quality assurance by measuring the dispersion of the population data of a probability of frequency distribution, or by determining the content or quality of the sample data of substances. The method of measuring the ratio of standard deviation to mean is also known as relative standard deviation, often abbreviated as RSD. It only uses positive numbers in the calculation and is expressed in percentage values. Therefore, the resultant value will be multiplied by 100. CV is important in the field of probability and statistics to measure the relative variability of the data sets on a ratio scale. In probability theory and statistics, it is also known as unitized risk or the variance coefficient. The CV is calculated as stated in Equation 6.

Equation 6 - Coefficient of variation of cohort

$$CV = \frac{\sigma(r)}{\mu(r)} * 100$$

Finally, the share of a cohort's accumulated revenue in respect to the overall insurance portfolio is calculated. This is for the purpose of weighing a cohort's relevance in respect to the overall volume of insured revenues, from which the collected insurance premiums depend upon (among others things). Equation 7 illustrates how the relative share of the insurance portfolio is calculated.

Equation 7 - Share of insurance portfolio

$$ps = \frac{\sum revenue_{cohort}}{\sum revenue_{total}]}$$

II.4 Results

II.4.1 Economic relevancy

Applying the calculations to the data set enables an industry comparison of the insured SMEs. For confidentiality reasons, the different samples illustrated below are purely indicative and are not explicitly representative for the insurance or Switzerland.[2] Ranking the different industries against each other and sorting by frequency count shows which industries occur frequently in a respective insurance portfolio.

Table 3 - Most frequent industry types (year 2015)

Industry	Frequency rank	Count of SMEs in sample (year 2015)	Share of all SMEs in portfolio	Share of all revenue in portfolio
Restaurant	1	1 610	6.08%	4.18%
Hair dresser	2	962	3.63%	0.47%
Motor vehicle garage (repair cars / motorcycles)	3	818	3.09%	7.42%
Consulting	4	677	2.56%	1.69%
Architecture firm	5	579	2.19%	0.81%
Naturopathic practice/acupuncture/kinesiology	6	525	1.98%	0.39%
General medicine practitioners (w.o. X-ray)	7	485	1.83%	1.86%
Software consulting and development	8	434	1.64%	2.02%
Building joinery/carpenter's shop	9	369	1.39%	1.29%
Civil engineering w/o special focus	10	336	1.27%	2.22%

n = 26 469

Restaurants, hair dressers, and motor vehicle garages are the most common businesses insured. Summing up their annual insured revenues also enables the insurance company to draw conclusions about the share in respect to the overall insured SME portfolio. The above-mentioned three industries together represent 12.79% of all insured businesses (accumulated); however, they do not proportionally contribute to insured revenues in the portfolio. Hair dressers, for example, represent 3.63% of all SMEs in the portfolio, however their combined revenues are only 0.47% of all insured revenues. Having an understanding which businesses occur quite frequently within the insurance company's SME portfolio, gives one guidance

[2] The data provider purposely left out some companies to obfuscate market share of individual customer segments.

whether a growth prediction of a cohort (such as an industry) is promising. This supports H1.1. Due to the unproportional revenue contribution of some industries, a second dimension has to be considered in order to find out which industries are worth investigating for the prediction of revenue growth. This can be done by sorting the industries according to their highest sum of insurance cohort revenues.

Table 4 - Highest revenue risk industries (year 2015)

Industry	Portfolio rank	Count of SMEs in sample (year 2015)	Share of all SMEs in portfolio	Share of all revenue in portfolio
Motor vehicle garage (repair cars / motorcycles)	1	818	3.09%	7.42%
Restaurant	2	1 610	6.08%	4.18%
Commercial agency (office only)	3	80	0.30%	3.63%
Civil engineering w/o special focus	4	336	1.27%	2.22%
Software consulting and development	5	434	1.64%	2.02%
Kiosk / Retail trade with magazine	6	247	0.93%	1.88%
General medicine practitioners (w.o. X-ray)	7	485	1.83%	1.86%
Hotel with restaurant/pool/spa	8	280	1.06%	1.81%
Consulting	9	677	2.56%	1.69%
Feed retail, Agricultural Cooperative	10	25	0.09%	1.52%

n = 26 469

Sorting the industry cohort by share of the insured industry risk alternates the previous ranking. Motor vehicle garages and restaurants remain the most important industries but in switched order. Commercial agencies are low in frequency (n = 80) and were not mentioned in the frequency ranking (Table 3), but they contribute a large share of the insured revenue-related business risk of SMEs. Having an understanding of which businesses cohorts contribute most to the overall insured risk in the SME business portfolio enables the insurance company to take into consideration the industry variations in average revenues of SMEs. This supports H1.2.

II.4.2 Statistical Relevancy

In order to find predictive attributes for revenue growth, the insurance company should also investigate whether a cohort includes SMEs with both, small and large revenues, or if the revenues are very similar. This relationship can be expressed with the CV. Industries with the lowest CV are homogeneous with respect to revenue similarity. The industries that have more than 30 observations in the sample, sorted by ascending CV, are listed in Table 5.

Table 5 - Lowest CV industries (year 2015)

Industry	CV rank (low)	Count of SMEs in sample (year 2015)	Share of all SMEs in portfolio	Coefficent of variation
Dental practice	1	329	1.24%	0.61
Pharmacy	2	63	0.24%	0.68
Practice of medical specialists (without X-ray)	3	114	0.43%	0.71
Food stall, take away	4	72	0.27%	0.73
Chimney sweep	5	38	0.14%	0.73
Art painting studio	6	55	0.21%	0.74
Veterinary practice / vet. affairs	7	59	0.22%	0.75
Manicure	8	61	0.23%	0.77
Pedicure	9	112	0.42%	0.78
Bakery with restaurant	10	32	0.12%	0.78

n = 26 469

I find dental practices, pharmacies, and practices of medical specialists (not general practitioners) to have very similar (in-group) revenues, expressed as a low coefficient of variation. For those industries, a prediction of revenue change does not seem appropriate as the insured revenue within a cohort are close to identical, and a good estimator of future revenues are the group measures of central tendency. The most heterogeneous industries (with more than 30 observations) in terms of revenue are listed in in Table 6.

Table 6 - Highest CV industries (year 2015)

Industry	CV rank	Count of SMEs in sample (year 2015)	Share of all SMEs in portfolio	Coefficent of variation
Business consulting	1	140	0.53%	5.56
Commercial agency (office only)	2	80	0.30%	5.25
Consulting	3	677	2.56%	3.94
Computer Retail	4	65	0.25%	3.71
Gift shop and souvenir retail	5	38	0.14%	3.66
Shipping and Internet retailers	6	55	0.21%	3.65
Cleaning company	7	240	0.91%	3.56
Flower and plant retail	8	105	0.40%	3.43
Travel agency with travel companion	9	83	0.31%	3.40
Painting and plastering	10	184	0.69%	3.34

n = 26 469

Industries such as business consulting, commercial agencies, and consulting SMEs are structurally very different. The sample has, for example, many commercial agencies that have very large revenues, but also many agencies that have much lower revenues. I assume that at each of the larger SMEs started out small, however grew over time and become bigger. The industries listed in Table 6, sorted by declining CV, have very heterogeneous within-industry revenues. Hence, they can be thought of to be more suitable for the investigation of changes in revenue. Having an understanding of which businesses cohorts have very similar revenues

and which do not enables the insurance company to take into consideration the industry variations in average revenue of SMEs. This supports H1.3.

Finally, another relevant consideration aspect in choosing a cohort for prediction based on industry is the structure and performance of the industry itself. Table 7 list the growth of the most common industries.

Table 7 - CAGR most common industries (all years)

Industry	Frequency rank	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR
Restaurant	1	7195	1.1%	17.7%
Hair dresser	2	3841	1.1%	11.9%
Motor vehicle garage (repair cars / motorcycles	3	3624	1.8%	18.0%
Consulting	4	2842	2.1%	22.5%
Architecture firm	5	2527	1.8%	12.5%
Naturopathic practice/acupuncture/kinesiology	6	1992	1.9%	10.5%
Software consulting and development	7	1842	1.9%	14.5%
General medicine practitioners (w.o. X-ray)	8	1766	1.6%	10.9%
Building joinery/carpenter's shop	9	1470	1.5%	11.4%
Accounting office	10	1467	1.9%	18.3%

n = 111 284

The average annual growth of very common businesses (top 10) such as restaurants, hairdressing studios, or motor garages is positive and lies between 1.1% and 2.1%. The highest average CAGR rates for businesses that occurred at least 50 times in the data sample are found for the industries listed in Table 8.

Table 8 - Highest CAGR industries (all years)

Industry	CAGR rank	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR
Brewery	1	69	15.2%	63.1%
Florist	2	124	13.8%	95.9%
Installation of ventilation systems	3	58	10.1%	54.9%
Butcher w/o shop	4	82	9.4%	68.1%
Kiosk / Retail trade with magazine	5	787	8.9%	42.0%
Heating, ventilation, AC engineering	6	89	8.7%	34.1%
Charitable organization	7	100	8.2%	41.6%
Nursery (private)	8	346	7.8%	25.5%
Milk intake / Milk collection point	9	76	7.7%	65.2%
Other technical advice and planning	10	85	7.2%	45.4%

n = 111 284

The illustrated ranking is sorted by the average CAGR of the highest-growing industries. As such, the ranking suffers from high-weight data points that especially influence the average CAGR of low-count SMEs. This issue can be addressed by filtering for very small businesses

or by further increasing the minimum cohort size. A very high standard deviation of the CAGR indicates the nonhomogeneous in-group CAGR, which could also be interpreted as a sign of industry cyclicality or consolidation. The industries (with more than 30 observations) that have been shrinking most on average are illustrated in Table 9.

Table 9 - Lowest CAGR industries (all years)

Industry	CAGR rank (neg)	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR
Electronic elements wholesale	1	33	-3.6%	8.9%
TV-Retail	2	150	-3.6%	11.9%
Interior design studio	3	64	-3.0%	12.3%
Textiles-Wholesale	4	78	-2.9%	11.1%
Colours retail	5	32	-2.6%	10.0%
Metal finishing	6	32	-2.2%	20.1%
Apparatus: Precision apparatus (w.o. electromed.)	7	100	-2.1%	12.9%
Production of other plastic goods	8	43	-2.0%	12.3%
Photographic laboratory (w/o sales)	9	39	-1.8%	5.0%
Carpet, floor and wall coverings retail	10	42	-1.7%	7.4%

n = 111 284

The ranking shows that electronics retailers, interior design studios, and textile wholesalers are among the shrinking industries. Having an understanding of which industries have been growing in the past and which have been shrinking enables insurance companies to take into consideration past industry growth patterns when deciding for which SMEs they want to predict growth. This supports H1.4. It is noteworthy that a low standard deviation of an industry's CAGR suggests that the mean CAGR values are somewhat representative for the overall sample. Compared to the fastest-growing SMEs (compare to Table 8), the industries described in Table 9 are less influenced by heavy weight data points that are far away from the mean CAGR. This can be explained by the way the data are collected. SMEs with high growth report their revenue changes. SMEs that go out of business, do not report the changes and just terminate their insurance policy. However, if they would report their revenue growth, it would lead to observations with -100% "growth." Hence, the standard deviations of shrinking industries are artificially low.

Interim results The data illustrated in Table 3-9 are available to most insurance companies that offer general liability insurance. Therefore, they can investigate the own portfolio composition, industry mean revenue growth, standard deviation of revenue growth, industry revenue heterogeneity as well as mean absolute revenue per business (see Appendix 3-Appendix 6). To investigate which SME industries seem worthwhile selecting for a prediction of business growth, I consider the ranking of the following four criteria:

- The industry occurs frequently (otherwise a case by case analysis would yield a better value versus cost consideration).
- The industry contributes a relevant share to the overall insured revenue portfolio risk (c.p. premiums related).
- Within the industry, the SMEs are structurally different with respect to revenue distribution (CV).
- The industry shows revenue growth (positive mean CAGR).

In order to determine which SME cohorts best suit those criteria, I weighted the ranks of sample frequency, portfolio share, cohort structure, and within-industry revenue distribution with equal importance. Multiplying the ranks and sorting them by the lowest count yields the following result.

Table 10 - Lowest multiplication score

Industry	Rank Multiplication	Rank Sample count	Rank Portfolio relevance	Rank CV Revenue	Rank CAGR
Consulting	1	4	9	3	333
Motor vehicle garage (repair cars / motorcycles)	2	3	1	59	378
Restaurant	3	1	2	96	480
Commercial agency (office only)	4	70	3	2	303
Business consulting	5	38	14	1	366
Software consulting and development	6	7	5	26	359
Civil engineering w/o special focus	7	12	4	33	307
Building joinery/carpenter's shop	8	9	12	43	421
Architecture firm	9	5	19	81	367
Painting and plastering	10	25	28	12	411

n = 111 284

I find that industries such as consulting, motor vehicle garages, and restaurants score the highest according to this self-defined metric. They occur relatively frequent in the sample portfolio, contribute a relevant amount to the insured revenue, have occurred revenue changes in the past, and show (mostly) positive growth. Insurance companies that combine business count, their relative contribution, the cohort similarity, and average historic growth can calculate a score to rank industries, combining the four attributes by multiplication. This yields a top-down order that provides guidance when selecting a promising cohort of SMEs for which to predict revenue growth.

II.5 Summary

Insurance companies aiming to predict the revenue growth of the insurance SME, in order to improve its own operational efficiency (I.1.1), can select cohorts for which predictions make economical and statistical sense. These cohorts have to have a relative frequency that justifies the efforts of collecting data and building models. The insured revenue risk of the cohort also has to be sufficiently large. Further, the within-group revenue similarity expressed by the cohorts' CV has to be large enough. Otherwise, prediction efforts that go beyond those measured of central tendency bear not much potential. Further, the cohort, assuming historic cohort growth, is indicative for future cohort growth and has to show a tendency to grow (or shrink). Combing all mentioned criteria for selecting suitable SMEs is possible by ranking the cohorts by criteria and multiplying the ranks. The lowest values, as a result of multiplication, can be interpreted as a list of suitable cohorts for the prediction. For the provided sample, the suggested top four cohorts are the industries of commercial agencies, restaurants, motor garages and consultancies.

II.6 Implications

Building an accurate universal growth model is a complex task as illustrated by the particularities of the SME portfolio examined. Therefore insurance companies can subdivide their portfolio of "all SMEs" into several smaller cohorts, for example defined by industry, as a basis to build a community of micromodels that, in combination, operate over the whole population (compare Zhou et al., 2014). Such a structured approach to evaluate whether a cohort of SMEs is suitable for prediction, enables insurance companies to guide their efforts to the more promising cohorts. This way, insurance resources can be pointed to where they provide the most value. *Industry* as a selection criteria for a cohort is generally applicable for any insurance company and is easy to interpret.

II.7 Limitations and Future Work

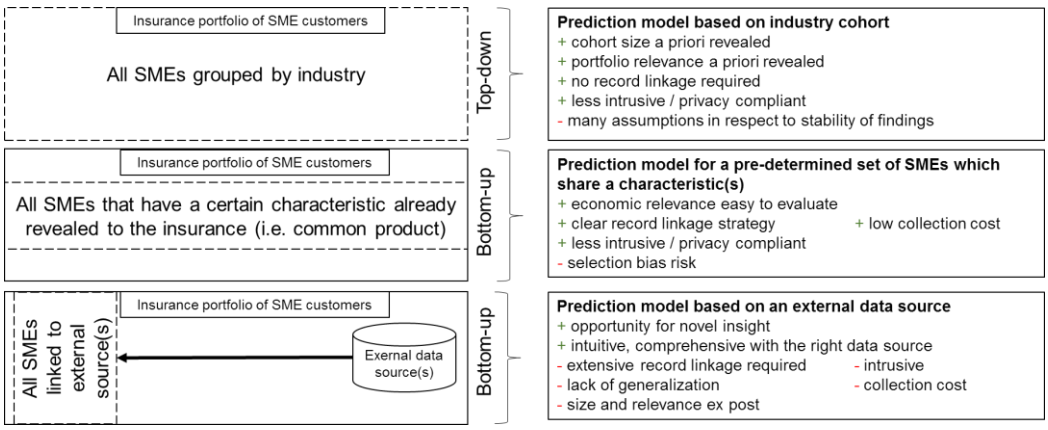
Cohorts solely defined by industry (top-down) demand many requirements to be fulfilled. The methodology presented in this chapter is based on the assumption that past business frequency, business size, industry growth, and cohort heterogeneity remain similar in the future. This may not be the case, as industries die slowly, others plateau, some are very

cyclical, and some are even disrupted abruptly. Therefore, when selecting cohorts to build growth-prediction models, continuous in-depth data analysis and reflection of the context should not be replaced by a rigid statistical methodology.

However, there are also alternative approaches to select a cohort. These can augment or replace *industry* as a sole (pre-) selection criteria. Given the nature of the problem, insurance companies may consider the existence of other data of a customer as a starting point and build cohorts according to the data availability (bottom-up). One such attribute could be the existence of a common product that the insurance company has sold to a customer or a common cross-industry attribute (i.e., legal form, location, etc.), which is already revealed to the insurance company and is analyzable at scale. This with the benefit of low effort required to collect the data as well as the absence of complex merging of the data sources. However, such selection criteria may also introduce bias to the models, which reduces their generalization.

Similarly, adding external data sources may result in superior suitability when forming cohorts. Insurance companies may acquire sources that are thought of as strong indicator of future growth and may use them instead or additionally when building predictive models (bottom-up). Examples are patent filings, registered brands, company news, management reports, the SME's website, the personality traits, or other attributes that can be linked to the person running the SME or the economic situation of the SME. Adding external data sources comes with the cost of collection and extensive matching efforts, as well as the risk of alienating customers, who may be offended by an insurance company's effort to gain insights into a customer from an insurance unrelated data source. A graphical summary of the three approaches and their characteristics is illustrated in Figure 2.

Figure 2 - SME growth modeling selection criteria



III) Growth Prediction with Insurance Data

III.1 Introduction

The size of a cohort (i.e., industries with many SMEs insured) is important when building growth-prediction models that have a practical implication for an insurance company. This is especially true when external data have to be gathered and matched. Because of the matching process, most often, not all observations can be linked to an external data source. This results in a loss of observations that can be used to build models. Less observations also limit the future economic relevance after a model is available. An alternative to the predetermined top-down cohort view (i.e., selecting large and relevant industries a-priori) is that insurance companies can also first select a potent data source and then evaluate whether a common attribute matches a larger sample of SMEs insured across several industries. If the matched observations are then sufficiently large and relevant, the data source can be exploited to build a scalable revenue growth-prediction model.

Leaving the perspective of industry-specific prediction, this chapter introduces an alternative that is relevant for insurance companies with the goal of predicting revenue growth of SMEs. Insurance companies make an effort to create long-lasting business relationships with their customers. Over the course of this business relationship, the SME client interacts with the insurance company, reports claims, buys new polices, and adapts or terminates existing ones. These business interactions generally create digital traces in the operational systems of the insurance, which can be mined. One example commonly found in an insurance data warehouse is the documentation of additional products purchased by SMEs. Insurance policies are data rich as they normally describe the characteristics of the insured good. In the case of SMEs, a common additional policy an SME has taken out with the same insurance company, besides the general liability insurance, is vehicle insurance.

Vehicle insurance policies contain information related to the insured good, which is in most cases a car, and information about the primary driver. Many of the SME owners only insure one car at a time and are stated as the only driver. In addition, other studies using

commercial motor vehicle insurance data reported that between 66% and 78% of all buyers of insurance policies are also the owner of the company (Deloitte, 2015). Further, more than half of all SMEs in Switzerland have only one or two employees (see Table 2). This implies that the primary driver is at least contributing to the success of SME. Therefore, I hypothesize that the choice of vehicle, its age, and other information that can be extracted from vehicle insurance policies may reveal important information to segment growing from shrinking SMEs.

III.2 Literature Review and Hypothesis

III.2.1 Social Identity and Brand Personality Theory

Company owners select their vehicles based on a range of reasons. Their choice is influenced by economic and emotional considerations. Purchasing choices of consumer brands have been analyzed and interpreted with the help of a construct referred to as social identity theory (Phillips, 2003). It describes certain groups' behaviors on the basis of perceived group status differences, the perceived legitimacy and stability of those status differences, and the perceived ability to move from one group to another.

The purchase of a specific brand (i.e., a luxury car brand) can be used as valid tool for status signaling in everyday circumstances (Cătălin and Andreea, 2014) and is considered more than just an instrument of hedonic experiences as Nelson and Meyvis (2008) stated. This is because the purchase of a brand has the power to harness and channel specific hedonistic desires in expressing a bigger sociological and psychological construct such as lifestyle (Cătălin and Andreea, 2014). Therefore, brands are a lifestyle "beacon" and a relevant mean of self-expression (Cătălin and Andreea, 2014). Each individual lifestyle reflects a person's values, life vision, aesthetic style, and life goal (Vyncke, 2002).

Further, consumers will often seek new ways in which they can express their personal identity (Cătălin and Andreea, 2014). Specific brands of things like cars tend to confer some distinguished features among certain classes of consumers (Cătălin and Andreea, 2014). One characteristic of a brand is the self-expressive function (Keller, 2008). Brands have the power to communicate valuable information and can be used and perceived in many different ways by those that buy them. These consumer (Cătălin and Andreea, 2014) tend to choose brands that are considered "appropriate for their self-image" (Cătălin and Andreea, 2014). The relationship between social identity theory and consumer choices with respect to brands and their function has also been addressed with the help of the concept of brand personality. It

stems from one stream of research that emphasized human qualities of brands inspiring personality-led branding and the creation of brand personality as a concept (Kuenzel and Halliday, 2010). Much has been written about how customers perceive brands and what human characteristics they associate with the brand. Attributing human features to inanimate objects is relevant as a number of studies have shown that a consumer's attitude is influenced by matching the perceived product user image with his or her own self-concept. A brand personality evaluation enables researchers to quantitatively relate brand personality to self-congruence of the consumer, which in turn can be put into perspective with the consumer's symbolic consumption benefits. Marketing applications take advantage of brand personality and communicate a brand's image mainly by utilizing user imagery (i.e., by portraying of a specific brand user who matches the brand's perceived personality, in Kuenzel and Halliday (2010)). Consumers try to reflect their own identity through choices (Cătălin and Andreea, 2014); hence, they buy brands with brand personalities that match with their own.

An SME owner's choice of a car brand can be interpreted with the help of social identify and brand personality theory, eventually allowing to derive insights about the economic status (and aspired future) of a company. This is because brands have become instruments of status signaling (Cătălin and Andreea, 2014) and life-style signaling.

III.2.2 Insurance Data Mining

As the insurance industry becomes more aware about the business value of emerging technologies like data mining, the number of publications from operational researchers and practitioners are increasing (Johnson et al., 2017). Reports of successful projects applying data-mining and machine-learning algorithms to problems such as fraud detection, underwriting, insolvency prediction, and customer segmentation (Smith et al., 2000) have been recognized within the insurance industry. Emerging techniques like data mining are proving to be of enormous benefit to the business world in terms of identifying hidden patterns in data as well as predicting future behaviors of customers (Gupta and Gupta, 2010). Insurance businesses are avant-garde application cases as they usually have large and effective data warehouses that record details of every financial transaction and claim (Smith et al., 2000). When these companies realize that valuable information is hidden in this data, data mining can help them achieve their objectives of market growth and profitability. The

purpose of this case study is then to demonstrate the potential of data mining as a means for predicting customer-revenue growth with the help of vehicle insurance data.

Building upon the general understanding that brand and product attribute choices of customers can be examined to derive insights about the consumers themselves (e.g., Cătălin and Andreea, 2014; Kuenzel and Halliday, 2010; Vyncke, 2002), the following hypotheses, hypothesis 1 (H2.1) and hypothesis 2 (H2.2), are proposed to address the expected observations in the data:

H2.1: *Consumer choices of motor vehicles and their characteristics allow the inference of the growth of SMEs.*

H2.2: *Motor vehicle characteristics found in insurance policies are valuable predictors of the business performance of SMEs.*

Combining the findings of H2.1 and H2.2 answers to research question two.

III.3 Research Design

To investigate the derived hypotheses, the full data sets of SME customers are extracted and merged with policy information from the general liability business insurances and motor vehicle insurances. The following sections provide details about the available variables in the sample and lay out the methodology used to derive the results. At first, the data and the record linkage process is presented. Subsequently, I provide a statistical description of the merged data set. Finally, the prediction is performed and evaluated with the help of a confusion matrix.

III.3.1 Data Set and Record Linkage

The utilized sample for this study contains data of customers from the operational and the policy information database of a Swiss insurer. The products provide coverage in different circumstances. The vehicle insurance policies include mandatory coverages for third-party liability and optional collision packages for station wagons and motorcycles. Swiss law stipulates that all vehicles must have liability insurance (LI) that covers damage inflicted on other drivers and their vehicles. Every vehicle insurance company offers this coverage along

with another two noncompulsory products: a limited comprehensive insurance (Mini-CASCO) that covers damage caused by act of God, fire, vandalism, or theft and a full comprehensive insurance (CASCO) that covers all risks, including damage at fault on the insured vehicle.

The general liability insurance is tailored to the needs of SMEs in a wide range of industries and are valid worldwide, with the exception of the United States and Canada. It contains machinery insurance for damage to stationary and circulating machines as a lump sum or as individual insurance. Also, the transport insurance is included. It insures goods damage on transports, at exhibitions, and even abroad (up to 100 kilometers from the Swiss border). Further, buildings can be insured, and employment protection is offered. Company liability insurance protects the insured against the financial consequences of personal injury, property damage, and financial loss for which the company would be held liable in case of a claim. The public liability also protects SMEs from the financial consequences of investment, operating, product, and environmental risks. Additionally, it insures special risks such as loss of use, dismantling and installation costs, damage due to connection and mixing, or damage to treatment and care can be added.

Due to the existing unique customer ID, the merge of the two data sets did not require special algorithms. I merged each SME with its general liability insurance policies (several years). To form the final data set, I merged the created data set with the vehicle insurance policies. Many SMEs insured many vehicles sequentially, and some SMEs insure several in parallel (a fleet).

III.3.2 Covariates

The observation across all insurance products include the following covariates.

Table 11 - Covariates for growth prediction with policy data

Attribute	Variable	Values
Overall covariates		
Company ID	ID	Unique numerical value
Industry	Industry	Insurance internal classification of the business type (323 categorical values)
Additional covariates for general liability insurance		
Revenue [2008-2016]	Revenue	SME revenue in respective year in CHF

Additional covariates for vehicle insurance		
Date of birth	Date_of_Birth	Date of birth of the driver
Age of driver	Age_Driver	Age of the driver (in years), calculated by subtracting Date_of_Birth from the last observed revenue year (2008-2016)
Gender of driver	Gender_Driver	Gender of the driver (male, female, unknown)
Driver's license since	Date_Driving_Test	Date at which the driver took the exam
Age at driving test	Age_Driving_Test	Age (in years) at which the driver took the exam, calculated by subtracting Date_of_Birth from the Date_Driving_Test
Type of vehicle	Car_Class	19 levels (i.e. middle class, SUV, scooter)
Vehicle first registration	Car_Registration	Year at which the car was registered for the first time.
Age of vehicle	Age_Car	Age of vehicle in years, calculated by subtracting Car_Registration from the last observed revenue year (2008-2016)
Fuel type	Fuel_Type	Fuel that a vehicle runs on (gas, diesel, electric, …)
Leasing status	Leasing	Vehicle status (leased, owned, unknown)
Annual kilometers	Annual_km	Estimate driving: <7'000 km, ≥7'000 or unknown)
Brand	Car_Brand	357 brands (i.e. Audi, VW, Subaru, …)
Model	Car_Model	3698 models (i.e. VW Polo 75, John Deere 5515)

III.3.3 Methodology

Growth This study continues to use the CAGR calculation as a measure of annual revenue growth (see Equation 1). The mean CAGR values for different attributes found in the insurance policies are then calculated.

Sample descriptives and classification process For the cohorts, I select the numerical and categorical values with at least 30 observations and report them in respect to mean CAGR and in a few cases in respect to the mean cohort revenues. The additional measure of mean revenue serves the purpose of evaluating the average size of a company in respect to the attribute. Further, from the CAGR values, I derive classes according to the following rules:

Classification Rules		**Label**	**Description**
CAGR < 0:	assign class -1	shrinking	SME revenues reported declining
CAGR = 0:	assign class 0	stable	SME revenues reported equal
CAGR > 0:	assign class 1	growing	SME revenues reported growing

These classes are assigned to each of the SMEs. Each SME is only assigned one class. All SMEs that only reported one revenue number at one point in time were not considered. For the descriptive study, I use all classes. For the prediction, I limit the data set to shrinking and growing SMEs. In reality, one would very rarely find a business that has identical revenues over several years. Within the insurance data set, many revenue numbers are rounded and updated inconsistently.

Prediction model selection For the prediction, I choose LogitBoost, a supervised learning algorithm. It is a popular machine learning ensemble algorithm for binary classification problems (Fraz et al., 2012). It is employed commonly in many areas (Cai et al., 2006). I selected the algorithm for the two-class classification problem because of its suitability, prediction power and algorithm's properties. This classification algorithm was initially formulated by Jerome Friedman, Trevor Hastie, and Robert Tibshirani (Friedman et al., 2000). It is a generalized additive model, combined with the cost function of logistic regression. It is the logistic variant of a logistic model tree (LMT), combining logistic regression and decision tree leaning. LMTs are derived from decision trees that have linear regression models at their leaves to provide piecewise linear regression models instead of constant models (Landwehr et al., 2005). At each note, the algorithm splits using the C4.5 criterion. Given by the greedy construction process, at each step, the combination of single best variable and optimal split point is selected. Then, generally, the tree is pruned (Marc Sumner, 2005).

The algorithm is considered an effective boosting technique and known to reduce bias and in particular variance (Friedman et al. 2000). The reduction of variance is achieved by reducing the chance of overfitting, known to occur when a model unknowingly extracts some of the residual variation (i.e. the noise) as if that variation represented underlying model structure. Friedman et al. (2000) found that using LogitBoost could reduce training errors linearly and hence yield better generalization (Cai et al., 2006), a property I aim to achieve by specifying a general classification model for growing and shrinking SMEs. Another benefit of the LogitBoost algorithm is that it is based on a loss function of the loglikelihood to diminish the sensitivity to extreme and outlier observations. It joins a number of weak learners to achieve powerful and robust learning (Cai et al., 2006). LogitBoost is further known to outperform standard logistic regression and is considered an influential boosting algorithm for classification.

The benefit of boosting in classification problems such as the identification of growing and shrinking SMEs, is that boosting increases the prediction performance. Boosting was proposed as a method that combines the classification results obtained using base classifiers, where the sample weights are sequentially adjusted based on the performance in previous iterations. In combining low quality classifiers with a voting scheme, boosting produces a classifier better than any of its components. Boosting has been successful in machine learning and industry practice (see examples in Li, 2012), which further encouraged me to apply this algorithm. Boosting is also among the most successful algorithm families used in Kaggle competitions for predictive modelling and analytics in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users.

Another benefit of using the LogitBoost algorithm is that it works well with categorical variables (Landwehr et al., 2005). The records I include in the prediction contain many vehicle brands, car types, and other higher dimension categorical variables; hence, LogitBoost is able to process the data collected and is expected to achieve good prediction results.

Data split The classification model is built using a training set of example instances that represent each class (shrinking and growing SMEs). The model is then able to predict class membership for new instances by examining the feature values of unknown observations. Therefore, I randomly have to assign instances to the training and testing set. The optimal proportion of instances for the training set is considered to be in the range of 40% to 80% for the wide range of conditions and domains studied (Dobbin and Simon, 2011). The number of cases needed for effective training depends on the "signal strength" or the extent of the separation of the classes with regard to relevant features. For this study, I split the newly created data set of 14 114 observations into an 80% training and 20% testing set. The training set contains 6 755 growing SME and 4 480 shrinking SME observations. The testing set contains 1 830 growing and 1 049 shrinking observations. As a precautionary measure, when assigning observations, I enforce that an SME with more than one vehicle insured; hence, more than one observation is either in the training or the testing set but not distributed over both.

Evaluation Classification problems can be evaluated by a metric that is calculated from true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs), all of

which are tabulated in the so-called confusion matrix. The relevance of each of these four quantities will depend on the purpose of the classifier and motivate the choice of metric (Lever et al., 2016). Further, I calculate the following evaluation measures:

Equation 8 - Accuracy

$$Accuracy = \frac{\sum TP + \sum TN}{\sum Total\ population}$$

Equation 9 - Precision

$$Precision = \frac{\sum TP}{\sum TP + \sum FP}$$

Equation 10 - Recall

$$Recall = \frac{\sum TP}{\sum TP + \sum TN}$$

Equation 11 - Positive predicted value

$$Positive\ predicted\ value = \frac{\sum TN}{\sum Prediction\ positive}$$

Equation 12 - Negative predicted value

$$Nagative\ predicted\ value = \frac{\sum TN}{\sum Prediction\ negative}$$

Equation 13 - Sensitivity

$$Sensitivity = \frac{\sum TP}{\sum TP + \sum FN}$$

Equation 14 - Specificity

$$Specificity = \frac{\sum TN}{\sum FP + \sum FN}$$

Equation 15 - Balanced accuracy

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2}$$

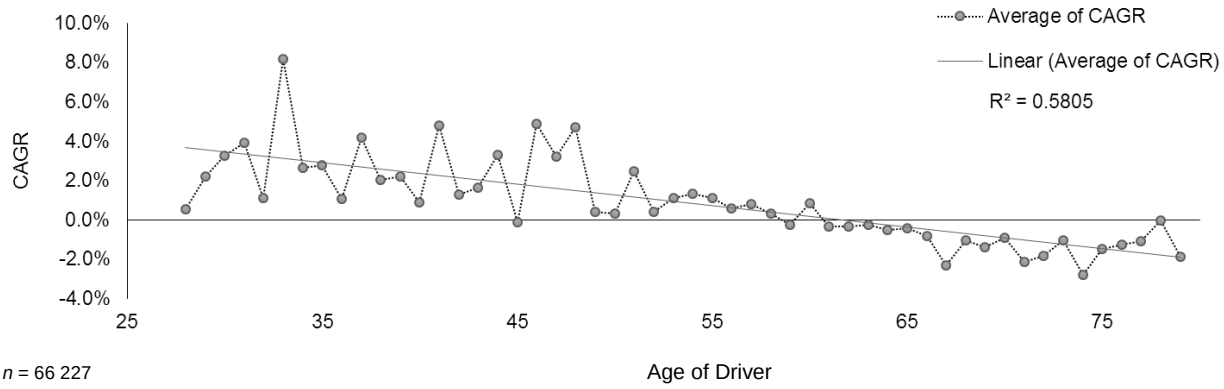
III.4 Results

In this section, the results of the analysis are presented. First, an overview of the descriptive statistics for each attribute is provided. Then, the results for the prediction are presented. Finally, the findings related to the prediction evaluation are reported.

III.4.1 Descriptive Statistics

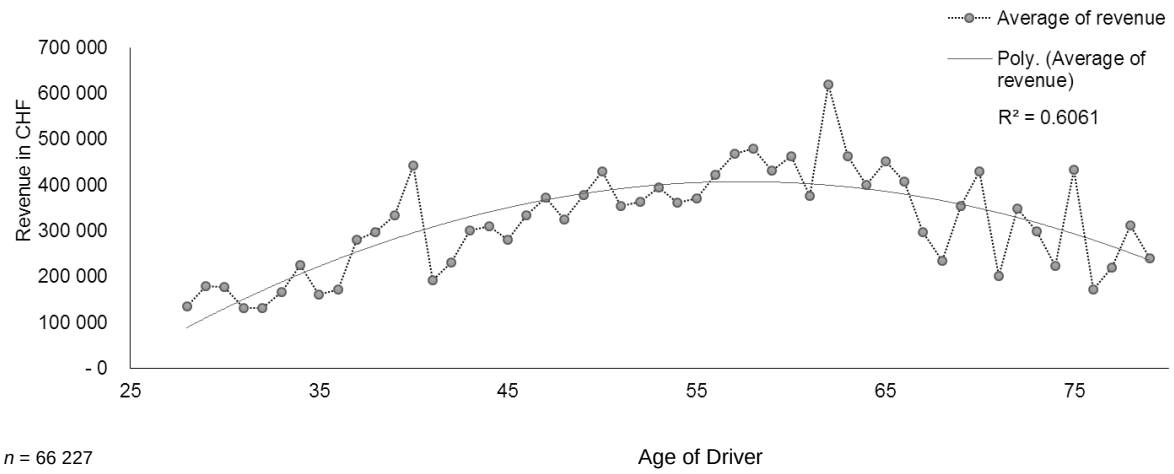
The first descriptives I provide show the relationship between the average age of an SME company vehicles and the CAGR of the SME. Figure 3 illustrates their relationship.

Figure 3 - CAGR by age of driver



A linear trend line shows a negative relationship between mean age and mean CAGR. The R^2 of the trend line is 0.58. Graphically evaluating the trend line and the actual data points shows that the relationship between CAGR and the age of drivers is somewhat ambiguous for younger drivers. For drivers between age 53 and 80, the trend line adequately reflects SME growth. SMEs with drivers that are older than the official retirement age (65 years for man), on average, shrink. To avoid false conclusion, the growth of SMEs needs to be put in economical perspective. Figure 4 therefore illustrates the relationship between age and average revenues of an SMEs in the data sample.

Figure 4 - Revenue by age of driver



n = 66 227

A second-order polynomial trend line describes the relationship between mean age and mean revenues, with an R^2 of 0.61. Revenues increase with age and start declining around age 63. Age and average revenue have a positive relationship for the ages between 25 to 63 and a negative relationship afterwards. The variance of revenues increase after the retirement age.

Next, I investigate the relationship between the age of the insured vehicles and the average CAGR of SMEs. I separate leased vehicles from owned vehicles and indicate how many vehicles are in each group. Table 12 shows the results for vehicles age 0 to 5 years.

Table 12 - CAGR by age of car and leasing status

Age of vehicle	CAGR leasing		CAGR owning	Observations leasing	Observations owning
new	↘	0.54%	↓	0.09%	207	438
1	↘	0.61%	↓	0.14%	381	821
2	→	1.01%	↓	0.19%	477	948
3	↓	0.38%	→	0.86%	661	1 339
4	↘	0.71%	↘	0.52%	896	2 111
5	↘	0.76%	↓	0.42%	1 002	2 471

n = 33 561

The color-coded arrows indicate the CAGR value relationship to the cohort's mean CAGR. Red arrows suggest below-average, yellow arrows suggest average growth and green arrows suggest high growth (see Appendix 26 for all years). Table 12 suggests a positive relationship between age of a vehicle and mean CAGR. For SMEs with leased vehicles of age 0 to 3, the average CAGR is lower than the CAGRs of SMEs with owned vehicles. From the observation

counts, one can derive that for vehicles aged 0 to 3 years, the distribution between leased and owned vehicles is approximately 1:2.

From the date of birth of the driver and the date of the driver taking the driving test, I was able to calculate the age at which the driving test was taken. Table 13 shows the results in relation to CAGR, CAGR standard deviation and average of the SME revenues.

Table 13 - CAGR and age of driving test

Age at driving test	Number of observations	Average of CAGR	Standard Deviation of CAGR	Average of Revenues in CHF
18	5276	1.41%	20.06%	370 368
19	5512	0.73%	10.35%	335 701
20	3570	0.50%	10.45%	331 257
21	1640	0.47%	8.47%	325 858
22	1195	-0.08%	5.37%	323 219
23	948	0.64%	11.35%	333 066
24	731	0.52%	5.87%	331 470
25	578	0.85%	9.06%	320 641
26	560	0.83%	10.85%	281 468

n = 22 689

Approximately the same amount of people in the sample took the driving test at the age of 18 and 19 years. With each additional year of age, the number of test takers declined. The average CAGR, the standard deviation of the CAGR and the average SME revenues decline as the age of test takers goes up. The highest revenues are observed in the group taking the test at 18 years. The lowest revenues were by those taking the test at 26 years.

Next, I investigate the type of vehicle and CAGR relationship. I further calculate average revenue and standard deviation of the revenue, grouped by vehicle type. The length of the green (red) bar indicates the cohorts' positive (negative) CAGR in respect to the CAGR of the whole sample. The results, which are sorted by vehicle type sample frequency, are found in Table 14.

Table 14 - CAGR and vehicle type

Vehicle Type	Number of observations	Average of CAGR	Standard deviation of CAGR	Average revenue	Standard deviation of revenue
Lower middle class	10 742	0.94%	16.46%	345 210	702 415
SUV	5 582	0.72%	9.75%	449 706	870 171
Middle class	4 586	0.69%	24.06%	325 869	614 471
Minivan	4 489	0.71%	10.35%	346 213	658 658
Sub-compact	3 828	1.57%	32.20%	310 894	756 930
Roadster (open)	1 780	0.54%	8.34%	465 608	934 831
Road machine	1 698	1.05%	10.81%	333 360	585 413
Scooter	1 685	0.89%	9.20%	315 340	556 566
Upper middle class	1 506	-0.32%	6.82%	420 424	711 052
Micro class	1 141	0.48%	8.22%	283 845	648 037
Chopper/Cruiser	796	0.22%	3.90%	356 640	885 697
Sportscar	684	-0.37%	5.60%	542 499	1 028 468
Construction machine	665	0.89%	8.66%	291 158	618 339
ATV/Quad	258	5.68%	25.69%	346 748	416 932
Luxury class	194	-0.91%	4.18%	893 121	1 365 834

n =66 227

Table 14 shows 15 different vehicle types. Most commonly, SMEs insure lower-middle-class and sport utility vehicles (SUVs). The highest CAGR rates occur with SMEs that insure ATVs/quads, subcompact vehicles, and road machines, all characterized by high standard deviations of the CAGR. The highest average revenues are observed for the luxury class, sportscars, and roadsters. The relationship for these types of vehicles and the SME CAGR is negative, with the exception of sportscars. Further, I examine the type of fuel, leasing status, and ages of the vehicles insured.

Table 15 - CAGR and fuel type, leasing, age of driver and car

Fuel type	CAGR of leased vehicles	CAGR of owned vehicles	Differance of vehicle age to mean age	Average Revenues
Regular gasoline	0.7%	1.0%	0.3	361 209
Diesel	0.8%	1.0%	-3.8	405 187
Unkown	1.2%	1.7%	2.6	340 035
Gasoline (leaded)	-0.5%	-0.6%	12.3	304 628
2-stroke mixture		-1.0%	7.4	270 885
Hybrid	1.3%	0.6%	-9.9	517 366
Electrical	0.0%	-0.6%	2.0	954 576
Ethanol E85 / gasoline	6.4%	2.3%	-2.7	175 380
Natural gas (CNG)	0.0%	0.5%	-5.7	593 377

n = 43 890

Sorted by observation frequency, Table 15 illustrates the CAGR of the cohorts by fuel type separately for leased and owned vehicles. The fuel types regular gas, diesel, and *unknown* together represent 98.5% of all insured vehicles. In all three categories, the SMEs with owned vehicles had a higher CAGR than the leased vehicles. Leaded gasoline vehicles, commonly

found in agriculture, had a negative CAGR. Vehicles with 2-stroke mixture engines, typical for smaller motorcycles, and electrical vehicles[3] also had a negative CAGR. Hybrid and leaded gasoline vehicles, on average, have the oldest drivers, while leaded gasoline vehicles and 2-stroke motorcycles are the oldest average vehicles. Hybrid cars have the youngest drivers in all cohorts. Further, SME cohorts with diesel vehicles have higher average revenues. The highest average revenues are found with electrical vehicles and natural gas / gasoline vehicles.

Finally, the vehicle brand choices of SMEs owners are put into perspective with the average CAGR. Besides the average CAGR for all observations, I separately show the average CAGR given that an SME is growing or shrinking. Further, the number of observations and average revenues are displayed for the most common car brands present in the data sample.

Table 16 - CAGR, car brand and age of car

Car brand	Rank	Average of CAGR	Average CAGR of growing SMEs	Average CAGR of shrinking SMEs	Observations of growing SMEs	Observations of shrinking SMEs	Average SME revenues
Volkswagen	1	0.75%	12%	-10%	257	173	393 160
Toyota	2	0.78%	15%	-9%	131	107	354 312
Renault	3	0.34%	10%	-10%	129	97	352 242
Opel	4	0.77%	15%	-12%	117	91	380 438
Mercedes-Benz	5	-0.27%	7%	-10%	101	94	526 528
Ford	6	0.32%	8%	-9%	113	68	394 924
Subaru	7	2.04%	18%	-9%	116	55	281 202
Peugeot	8	0.78%	13%	-10%	91	62	351 385
Audi	9	0.32%	10%	-10%	87	62	576 486
Fiat	10	1.13%	18%	-7%	75	73	448 160

n = 9 832

Table 16 states the 10 most commonly insured vehicle brands and the average CAGR. Within the most frequently observed car brands, Subarus, Fiats, Toyotas, and Peugeots had the highest average CAGR. The sample contained 357 brands that resulted in low counts in each brand cohort. Therefore, heavy data points influence the average much stronger compared to unranked categorical features such as the fuel type or vehicle class. To mitigate the overinterpretation of the stated results, Table 16 also indicates the average CAGR of all SMEs

[3] The most frequent observations in the cohort “electrical” are taxi and other transportation vehicles, commonly observed in emission-free mountain resorts in Switzerland. These vehicles are predominantly registered by hotels. In the sample I collected, the car brand Tesla is not present.

that have been growing/shrinking and had a vehicle of the mentioned brand insured. The highest average revenues are found for SMEs with Audis and Mercedes. Compared to the average CAGR of the total sample (0.95%), both brands show below average CAGR values.

III.4.2 Prediction Results

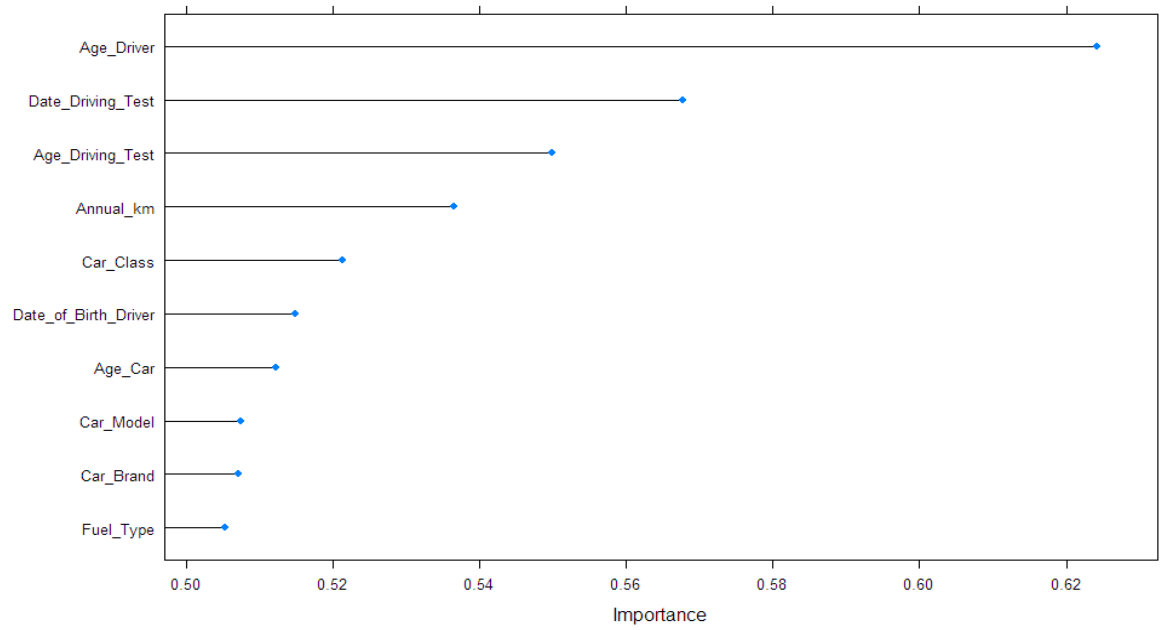
To evaluate the described SME data set in respect to its prediction capacity, the LogitBoost algorithm for classification is applied to the data sets and the prediction performance is compared. This simple and elegant algorithm became popular as one of the first practical implementation of boosting (Ben-David and Shalev-Shwartz, 2014; Kurt et al., 2008). The algorithm is closely related to classification trees and was determined as a suitable technique for predicting the growth of SMEs. Table 17 shows the classification outcome of the prediction of growing and shrinking SMEs.

Table 17 - Confusion matrix vehicle policies

Actual/Predicted	**Growing**	**Shrinking**	**Total**	**Recall**	
Growing SMEs	**1 619**	625	2 244	72.15%	Pos. Pred. Value
Shrinking SMEs	211	**424**	635	66.77%	Neg. Pred. Value
Total	1 830	1 049			
Precision	88.47%	40.42%			
	Sensitivity	Specificity			
Accuracy	70.96%				
No Information Rate	63.56%				
P-Value [Acc > NIR]	< 2.2e-16 ***		*** P ≤ 0.001 ** P ≤ 0.01 * P ≤ 0.05		
Balanced Accuracy	64.44%				

The model is able to predict with a 70.96% accuracy whether an SME is growing or shrinking. The “positive class” is defined as “growing,” and the 95% confidence interval of the accuracy is between 0.6927 and 0.7262. The most important features found by the model are ranked in Figure 5.

Figure 5 - Variable importance of LogitBoost growth prediction model



The age of the driver, the date of taking the driving test, and the age at which the test was taken are the most relevant features in the model. The discussion of the positive or negative impact of each feature on growth go beyond the scope of this thesis, however would shed more light on their causal relationship. Partial dependency plots (not available for LogitBoost) are suggested for the evaluation and can be applied to other decision tree related methods.

III.5 Summary

III.5.1 Discussion and Conclusion

The findings of the study suggest that vehicle insurance policies of SMEs reveal insight about policyholders that can be exploited when conducting future business with the SME. Analyzing the age of the insured driver with the CAGR that is indicative for the annual growth shows that the older the drivers are, the lower the CAGR. The relationship between age and revenue, however, is inverted until the official retirement age. The older the driver—in many cases, the ages are identical to those of SME owners—the larger the revenues. When a driver/SME owner reaches retirement age, then average CAGR turns negative in many cases, and average revenues are much smaller compared to SMEs with drivers aged mid-50 to early-60.

Evaluating the relationship between age of the insured vehicle and average CAGR, the findings suggest that SMEs that own relatively new vehicles, grow more slowly than SMEs

that lease new vehicles. This may be the result of advanced cash-flow management, "saving" liquidity, which can be used for growth-generating activities instead of paying for vehicles. Analyzing the age when the driver test was taken results in controversial findings. The average CAGR of SMEs with driver/owner who take the driver's license test at age 18 is much higher than for all the age groups up to the age of 22. Also, the average SME revenues decline stepwise from 18 to 26, plateauing at age 26. The consistency in the reduction of the CAGR and average revenue is remarkable and not intuitive at first glance. Taking the driving test is a costly endeavor. Many people taking the test either plan to buy a car of have access to a family-owned car. Therefore, the socioeconomic status of the person taking the test is at least partially related to the age at which somebody takes the test. Assuming that entrepreneurs with a higher economic status also have advantages to grow their businesses, the relationship between business growth and test-taking age could be explained to some extent. Access to liquidity at a young age could also suggest a higher propensity to build a more capital-intensive business and also explain the higher revenues. A competing theory that may explain this effect could argue with the help of behavioral decision-making theory. It postulates that time discounting influences decision choices. Therefore, entrepreneurs discount the future costs of procrastinating and therefore choose to do other, more enjoyable things. Differences in the steepness of discounting can consequently differentiate between procrastinators and nonprocrastinators (König and Kleinmann, 2004). Speculating, one could assume that the entrepreneurs that do not procrastinate when taking the driver's test are also more eager to achieve timely results for their SME and are better capitalized. Because of both, they achieve higher growth.

The choice of the vehicle insured in respect to CAGR showed that most insured vehicles are in the lower middle class or are SUVs. Some SMEs have vehicles insured that generally would be considered of recreational nature. With the exception of SMEs with ATVs/quads, which one finds also in nature/forest/winery related industries, SMEs with recreational vehicles insured have, on average, a negative CAGR, indicating that their revenues are declining over time. Similar to the observation of SMEs with drivers/owners post the official retirement age, one could assume that the SME purpose is at least partially a hobby and the pursuits of revenue growth subordinated compared to other SME cohorts.

The fuel type of the vehicles did reveal a noticeable CAGR difference between the common fuel types, which are regular gas and diesel. Leaded gas, however, an indicator of

rather old vehicles, is attended by a negative growth. Further segmenting the fuel types and CAGR evaluation by leased and owned vehicles, however, shows bigger growth in all major categories (regular gas and diesel) in CAGR. Over longer periods of time, SMEs with owned vehicles had a higher growth. This could again be related to the socioeconomic status of the SME owner. The ability to pay by cash (or have access to very good credit terms) may be indicative for the capacity to invest in the SME, hence partially explain the differences in CAGR.

Examining the more frequent of the total of 357 vehicle brands in the data sample revealed that the highest CAGR have SMEs insuring a Subaru, Fiat, Toyota, or Peugeot. Similarly, SMEs with a Mercedes-Benz or an Audi had, on average, a lower CAGR. At the same time, the SMEs with the later brands, however, had the highest average revenues, which may be of equal importance for insurance companies. Other vehicle brands (see Appendix 20), especially when they are of low frequency compared to the whole sample, showed very high stand deviations and partially economically unsustainable double-digit CAGR rates caused by data-heavy points and low counts in cohort observations.

This study was partially motived by social identity and brand personality theory, and it was expected to find strong signals for groups of brands, such as the example of high revenue, low CAGR in the group containing Mercedes-Benz and Audi. It is exciting to see the data patterns; however, their role for the prediction seems questionable. The reader has to assume that also other covariates cause the choice of a brand at different stages of the entrepreneurial cycle.

The growth signals that may or may not exist in the vehicle brand attribute are challenging to describe holistically. The LogitBoost model I used to evaluate the prediction capacity of vehicle insurance policies also revealed the relatively low importance of the attribute brand for the overall model performance. The aforementioned attributes, such as age of the driver, age at taking the driving test, and other attributes, contribute more of the models' classification performance. With an overall accuracy of 70.96%, the model was able to discover 88.47% observations of SMEs with positive CAGR. This is solely based on attributes derived from vehicle and general liability insurance policies. The findings itself are new; however, in a bigger context, they relate to those of other authors investigating the potential of data mining in the insurance context. The nature of most studies relating to the suitability of insurance data for predictions of various domains differs substantially. Hence, a

quantitative comparison is difficult. From a qualitative perspective, the findings of most authors correspond with the view taken within this study: insurance data are extensive sources of insight given the ability to mine and analyze it for a specific research question.

III.5.2 Implications for Research and Practice

The results of this study provide implications for insurance firms as well. When implemented in practice, the presented approach of data mining, analyzing, feature engineering, and predicting SME growth enables insurance companies to identify SME customers that are most relevant for them. Hence, insurers can exploit insights about a probable future state of a customer efficiently to retain SMEs that otherwise terminate or modify their insurance policies. Thus, it helps insurance companies to prepare the necessary actions to improve loyalty and protect their customer base against competitors with aggressive growth strategies, which is critical to the survival of many insurance companies (McKinsey and Company, 2013). Based on the high number of correct classifications of the presented model, the portfolios of sales agents could be prioritized. There CRM systems should contain a ranking of potential growing and shrinking SMEs. High-ranking SMEs would thereby receive more attention/discounts and more personal coverage.

The implications for CRM professionals in the insurance sector to enrich customer data with additional insights is also evident. When SME customers are newly classified as growing SMEs, this could trigger CRM activities. Agents could increase the contact frequency, adopt their marketing strategy, and, assuming an SME is grown in the past, adapt the SME liability coverage to the higher revenues. Thus, insurance companies can use such opportunities to efficiently increase premiums paid and to strengthen the tie with their customers in an otherwise silent relationship.

III.5.3 Limitations and Future Work

Though the presented study provides practical and academic insights, it is also limited in certain respects. First, the study includes only data from the period of 2008 to 2016, with most of the observations from the year 2015. The sample evaluated contained SMEs with all insured vehicles, the vehicle attributes, and the registered drivers. The sample was reduced to those SMEs that updated their insured revenues on their policies, which could have introduced a selection bias. Also, changes in attributes or changes in drivers resulted in double counts of the respective SME. Despite the large sample size, this may have caused heavy data points,

which had impact on the descriptive results and may have caused the prediction to not reach its potential accuracy. In interpreting the findings, in many views, it was argued that the driver is also the entrepreneur. Before implementing or further claiming any strong link, the connection between driver and SME entrepreneur needs to be verified. Second, the results are based on data of one insurer only and could be biased by the specific structure of the organization and the firm's marketing strategy. Future studies could extend the focus of the current study by replicating the approach of this study with samples from other insurance products, such as building insurance or machine insurance. Finally, the context of this paper could be expanded to a field study to validate the applicability of the prediction model in a real business setting. This would provide a better understanding of the stability of the model and a critical review of the basic assumption that historic SME business performance measures are indicators for future SME business performance.

IV) Improving Growth Prediction with Patent Data

IV.1 Introduction

From the macroeconomic view, the long-term growth of an economy depends on the aggregate demand and aggregate supply (productive capacity). Aggregate supply can increase if capital is utilized appropriately or more people provide labor. For example, investments in new factories or in infrastructure such as roads, increases in the size of the working population—for example, through immigration, higher birth rates, or increased labor productivity—and better education and improved technology can also increase aggregate supply. Technological improvements and innovations are the most difficult to investigate. Researchers have shown that SMEs are pillars of innovation (Helfat, 2006) and have the capability of structural change and market disruption in any economy. From other studies (Chew and Yeung, 2001), it has been revealed that small firms spend almost twice as much of their R&D dollars on fundamental research compared to the large firms. Innovation in companies is difficult to measure, but one way to examine parts of it is through patenting behavior (Melo-Martı, 2013) of SMEs. The patenting of innovation leads to the urge to acquire intellectual property rights of the invention in order to benefit from the investment in knowledge creation. Patents grant temporary monopoly for exploitation of knowledge. Through this monopoly, commercial use of an invention can be prevented, thereby providing an opportunity of market dominance by selling the invention. Therefore, the impact of the patent system upon SMEs is of particular importance.

Researchers (Cefis and Orsenigo, 2001) have claimed that evaluating patterns of innovation in respect to sales growth is difficult at the empirical and theoretical levels. Research in this field is not extensive, but practitioners and academics alike have in their studies identified growth of revenue as a meaningful indicator of post-innovation performance (Coad and Rao, 2008). These studies are mainly done on data sets of large-scale, global companies and are industry specific (Mowery, 2017). Such studies are limited for SMEs as their revenue figures are not public and hence not easily available. Empirical research in

Western Europe has not been able to find any strong link between SME innovation and growth of revenue for SMEs. I aim to address this research gap by capturing the effect of innovation on growth of revenue for small- and medium-sized enterprises.

I will address these topics with this study of patent- and nonpatent-filing SMEs, all operating in Switzerland. This study applies data analysis and machine learning on a Swiss SME data set to explain the growth of companies with patent data and derived gender data of patent applications. If current growth numbers can be explained by the patent data or are correlated to patent data, future growth will eventually also have similar dependence on patents. By explaining the historic growth of revenues, I want to investigate whether patent data could be used as a feature for the prediction of future growth of companies.

IV.2 Literature Review and Hypothesis

Many researchers have investigated the relationship between innovation and business performance (Beck et al., 2017; Jiménez-Jiménez and Sanz-Valle, 2011; Lyon and Ferrier, 2002). It is also widely shown that innovative companies respond better and faster to challenges and changes and compete most effectively in the market (Eisenhardt, 2017). They can exploit new products and market opportunities better than noninnovating companies. Executives predominantly say that innovation is what their companies “need most for growth” (Loren, 2017). Further, authors (Jiménez-Jiménez and Sanz-Valle, 2011) showed that organizational size strengthens the positive relationship between innovation and performance within firms. Baker and Sinkula (2002) showed the effect of market orientation and business performance. They explain how capabilities lead to superior products through innovation capabilities. Mowery (Mowery, 2017) used a data set of U.S. manufacturing firms and found that R&D expenditures have a positive impact on a firm’s growth. Law's (Law, 2017) research utilized a data set from the steel and petroleum industry over a 40-year period and found that innovators grew quickly. Freel (2000), in its study of 209 leading UK firms, showed that innovating companies grew faster and had larger profits. However, these studies either exclusively address large companies or are specific to a particular industry. Previous literature investigating innovation and growth for SMEs relied on survey data and research innovation in terms of process and product release (Roper, 1997; Teng et al., 2011). This leads to a scope of ambiguities and reliance on answers from the companies.

Further, studies that examine corporate innovation in terms of patent data mainly address the appropriation strategies of small firms (Arundel, 2001; Arundel and Kabla, 1998; Levin et al., 2013). They assert patenting to an expensive process and confirm secrecy to be more important for small firms (Darroch and McNaughton, 2002). Studies establishing patents as success indicators in small firms are rare and mainly address the role of patents in venture capital funding and survival rate among software firms (Helmers and Rogers, 2011; Mann and Sager, 2007). Wagner and Cockburn (2010) showed in their research that companies with patents have 34% higher survival rate than companies without patents after 5 years of their foundation. Helmers and Rogers (2011) also measured the company's survival rate with respect to patents portfolios using a data set of 162 000 SMEs created in the United Kingdom. They found that companies with patents have a 16% higher survival rate than the companies without patents.

Existing studies that established the role of patents in growth of revenue of firms (Coad and Rao, 2008; Scherer, 1965) were based on data sets of large and global companies whose financial data could be acquired as big public companies are legally required to produce statements after every quarter. Coad and Rao (2008) used a quantile regression method over 2 113 large firms to prove that R&D expenditures and patents have a positive impact on a firm's sales. They quote in their study that "compared to the average firm, innovation is of great importance for the fastest-growing firms." Further, Scherer (1965) used a data set of the 365 largest US firms and found that patents have a positive impact on the sales of companies. They asserted an increase in profits of innovative companies by constant margins due to an increase in sales. One of the major complaints about using patents as innovation indicators is that not all inventions are patentable. In addition, it has been shown in various studies that to small firms, secrecy and lead time are more important than patents (Arundel, 2001; Wesley M. Cohen Richard R. Nelson John P., 2000). Moreover, small firms face a significant disadvantage in protecting their intellectual property (IP) rights due to high litigation risk (Lanjouw and Schankerman, 2004). Despite all these hurdles, some small companies do invest resources into protecting their inventions by patenting. Therefore, patents should capture continuous innovation and should reflect its existence through growth of SMEs.

Studies establishing the role of innovation in growth of sales for SMEs used a survey data set by questioning companies that resulted in a low turnaround (Freel, 2000; Lussier, 1996). Roper (Roper, 1997) used survey data of 2 721 SMEs to prove that innovative products

made a positive contribution to firm's revenues. I find that the literature addressing the role of patents in growth of small and medium companies is underdeveloped, mainly due to the paucity of factual data sets; more research is required.

In this chapter, I investigate the role of patents in SME growth. To study the impact of patents on SME business performance, I evaluate the following two research hypotheses:

H3.1: *There are difference in the growth rates of SMEs with and without patents.*

And secondly, assuming patents have some explanatory power for predicting the revenues of companies, I try to evaluate the following hypothesis:

H3.2: *Patents' application attributes are useful in predicting business performance of SMEs.*

This chapter aims to investigate the role of patents as a public data source to predict the future business performance of their SMEs. Investigating these research hypotheses answers to research questions three. The expected findings hold importance for business-to-business companies, especially for financial institutions that benefit from understanding which of their business customers will outperform the others, hence eventually directly or indirectly benefiting them. This is due to the fact that most financial institutions (i.e., banks and insurance companies) own revenues correlative to some extent with the volume of the business of their customers.

IV.3 Research Design

IV.3.1 Data Set and Record Linkage

I created the population sample from a set of general liability insurance policy data, combined with an operational database of the insurance companies, to extract the SMEs' names. The provided data set used in this research includes policy information from 2010-2016. The data set is very diverse in size, group, ownership, industry type, and location. For the integrity of the data, I have excluded firms that have multiple policies. I have removed the outliers in revenue numbers by using the quantile function to truncate values at the 1st and 99th percentiles in the data (Beck et al., 2017).

I then combined this data with data from the Swiss commercial register and data from the federal statistic office data sets of Switzerland (Federal Statistical Office, 2017), which I

will use as a baseline to evaluate the performance gain of patent features. Finally, I matched this data set with a patent database that covers all patents granted in Switzerland (the Lens (Cambia, 2017), the Data Set 2). As the linking attribute, I used the policy holder name from Data Set 1 and matched it with the applicant name from Data Set 2. By matching, I identified those Swiss SMEs that have filed at least one patent after 2009.

IV.3.2 Covariates for Prediction

Table 18, Table 19, Table 20 and Table 21 describe the covariates extracted from the selected data sources.

Table 18 - Insurance data set covariates

Attribute	Variable	Description
Name	Name	Legal name of the SME or entrepreneur
Canton	Canton	Canton in which the SME operates
Zip	Zip	Zip code of location
Industry	Industry_[Group]	Insurance internal classification of the business type (323 categorical values)
Employees	Maximum_Employees Minimum_Employees	Derived from the insurance internal estimated range of employees of the SME (i.e. 1-5, 6-10, 11-20, …)

Table 19 - Commercial register (HREG) covariates

Attribute	Variable	Description
SME name	Name	Legal name of the SME or entrepreneur
Incorporation year	Founding_year	Year the SME was incorporated
New / Old	New, Medium_Aged, Old	New if the SME is < 5 years old Medium_Aged if the SME is >5 but >20 years old Old, if the SME is > 20 years old
Type	Sole proprietorship, AG, GmbH,	Legal type of company (binary encoded for AG, GmbH, sole proprietorship as base)
Nominal capital	Nominal_Capital	Capital (payment at full) at foundation

Table 20 - Swiss federal statistical office (BFS) covariates

Attribute	Variable	Description
City	City	City (unique value)
Zip	Zip	Zip code of location
Prosperity	Area_Income	Standardized area income
Rural Index	City_Rural_Index	Number between 0 (=city) and 1 (=very rural)
City area	Small_large	Area associated with city (small or large)

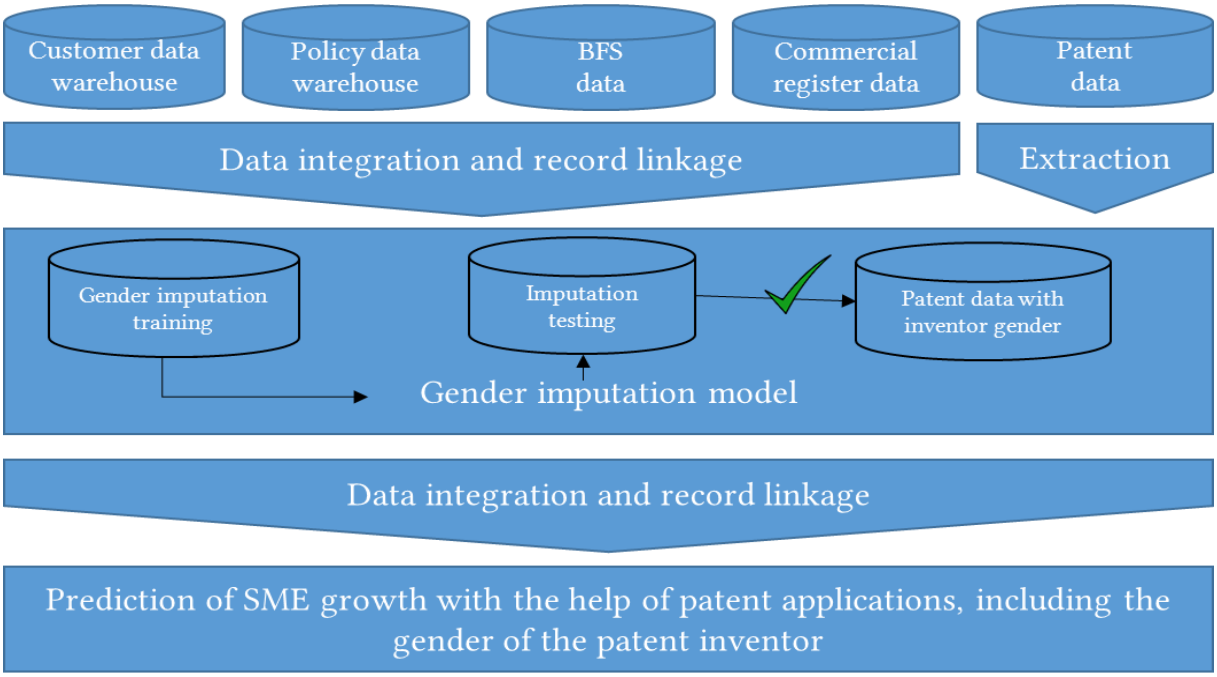
Table 21 - Patent database (LENS) covariates

Attribute	Variable	Description
SME name	Name	Legal name of the SME or entrepreneur, filing a patent
List of inventor names	Male_Inventors, Female Inventors	Derived number of male and female inventors
Gender of inventor	Gender	Imputed gender of inventor
Number of patents	One, Two, (Zero)	One, if the SME has 1-2 patens Two, if the SME has >2 patents, all others: Zero
Citations	Cited_Count	Number of citations of a patent

IV.3.3 Methodology

The process of this study follows the illustrated framework in Figure 6. First, the data sources for a general model (customer data warehouse, policy data, BFS data and commercial register data) are merged. Then, before merging the patent data, the names of the inventors on the patent applications are assigned a gender and counted. To automate the process of assigning a gender, I built a model that is trained and validated on data from the motor vehicle insurance data. The source was already introduced in Chapter III.3.2., and also contains the drivers' gender and its nationality. Both attributes however have not been used in the previous study. Merged together, the resulting data set is then fed to a prediction model.

Figure 6 - Framework of growth prediction study using BFS, HReg and patent data



Prediction model selection For the prediction, I choose random forests, a supervised learning algorithm that is known to be powerful and simple to use (Fernández-Delgado et al., 2014). Random forests work by creating an entire forest of random uncorrelated decision trees to arrive at the (near) best possible answer. The random feature sub-setting aims at diversifying individual trees and is at the same time improving the prediction performance. The algorithm turns down the fraction of the features that is considered at any given node (Breiman, 2001), which allows to work on large data sets with a large number of attributes.

The advantage of random forest is that it is correcting for simpler algorithms' habit of overfitting. This is because if there are enough trees in the random forest, the classifier will not learn from the residual variation (i.e. the noise) as if that variation represents an underlying model structure. Further, random forest run times are fast, it is able to handle unbalanced and missing data, both of which I encounter in the patent filing data set. Additionally, random forests requires almost no input preparation. It can handle binary, categorical, and numerical features without any need for scaling. This means the requirement for data preprocessing are minimal. Additionally, it allows to view the relative importance it assigns to the input features. Random forest is also considered as simple to use, because its

default hyperparameters often produce a good prediction result. The number of hyperparameters is small and they are straightforward to understand.

Although many studies show that complex models such as neural nets or boosting algorithm can beat the performance of random forest, the difference in performance is most often small and requires mostly data set specific hyperparameters' tuning (Caruana et al., 2008). Because the specification of these complex models requires more technical domain expertise, the complex models take longer to compute, require more input preparation and are expected to only lead to marginal potential outperformance compared to random forest, I choose the versatile and simple random forest classification algorithm. This choice allows to use the same hyperparameters to create a benchmark model, which is then re-trained and tested with additional attributes, everything else being the equal. This makes comparability easier, since the only differences that can cause a performance difference between the benchmark model and the model variation is the additional data input (patent filings).

Prediction application I have divided the merged data set in a training set (75%) and a test set (25%). The training set is used to build the random forest model. The random forest is trained with 500 trees, a maximum depth of the trees of 10 and the parameter $r = \sqrt{n}$, with n = Total features available. This process is performed on all the data sets using all available features and, in a second step, on the data set, excluding all patent-related features (LENS), resulting in two models that can be compared.

Prediction evaluation I compute the performance of the two models by computing RMSE and R^2 values on the test set. Both random forests had same hyper parameters except for the number of features, which cause the differences in the evaluation metrics. In order to compare the models' performance, I show their R^2 and RMSE values on the test set.

Imputation methodology In order to create the final data set for this analysis, the gender attribute has to be imputed correctly. To achieve this goal, the study interim goal is to investigate an improved approach to inferring gender from digital data sources that dynamically capture other user attributes to improve the categorization precision. I also discuss, depending on the application type of the gender mapping, whether a user might be interested only in the high-precision outcomes of the mapping or a "best guess," thereby providing a binary gender categorization with a high recall score for a complete list of names. Therefore, the quality of a mapping can only be evaluated in the context of the application.

Effectively building such a method with the goal of improving the data quality of a company's data records and the effects of precision and recall on each other will be discussed in the following subchapters. I provide an overview, and I review the current literature on how name-gender mappings have been created and where they have been applied with success. In addition, I illustrate how a mapping can be created and how recall and precision results can be improved.

The categorization of gender is a problem that has been approached in different research domains. Studies and applications of available gender-categorization tools can be found in the domains of marketing, humanities, information science, and census literature. Many researchers have performed gender categorization to uncover instances of gender inequity in a range of areas, from authorship of French literature to the disparities between attendees of the annual Digital Humanities Conference (Argamon et al., 2017). Others have contributed to identify attributes, which enable gender categorization and can be applied in reverse.

In marketing, gender is one of the most common forms of segmentation that is used by marketers in general and advertisers in particular (Darley et al., 2017). Males and females are likely to differ in terms of information processing and decision-making. Therefore, marketing researchers have, for example, tried to identify gender based on web users' perceptions of web advertising (Sultan and Uddin, 2011) or browsing patterns (Weiser, 2000). They found that genders make use of the web differently (Xue-wui et al., 2000). Hence, a user's gender can be identified, which presents opportunities for advertisers such as ad-placement targeting. A prerequisite of any gender segmentation is the representation of real-world entities to which they refer in a consistent, accurate, complete, timely, and unique way (Saha and Srivastava, 2014). The quality of the input data strongly influences the quality of the results (Sattler and Schallehn, 2001) (the "garbage in, garbage out" principle) and is essentially studied in two research communities: databases and management (Chen and Popovich, 2017). The first research community investigating the aspects of databases studies data quality from a technical point of view (e.g., Shasha et al., 1996), while the second is also concerned with other aspects or dimensions (e.g., accessibility, believability, relevancy, interpretability, objectivity) that are involved in data quality (Pipino et al., 2002). Hence, the completion and data quality improvement of a firm's records creates an opportunity for those enterprises that

engage in efforts to take advantage of best practices in data quality management from technical and dimensional points of view (Cai and Zhu, 2015).

In the domain of information science, scholars predicted, for example, gender based on people's Internet browsing histories (Hu et al., 2007). The authors' experimental results, which are based on click-through logs, showed that they were able to achieve 79.7% precision in gender categorization. Similarly, the authors of (Coltheart, 1981) found significant linguistic differences between men and women, which can be identified in written or spoken form. The author determined multiple linguistic features such as character usage, writing syntax, functional words, and word frequency that can be mapped to gender. Those and other features have been examined and are contained in the Media Research Center (MRC) Psycholinguistic database. However, they have not yet been used in reverse to categorize gender (Coltheart, 1981). Deitrick et al. (2012) applied a simple neural network to categorize gender on a sample that was extracted from the Enron email data set, which was provided by Carnegie Mellon University. The emails were labeled according to gender, and the authors' algorithm was able to achieve 95% precision using word-based features (Deitrick et al., 2012).

In the domain of the humanities, initiatives such as the Orlando Project, the Poetess Archive, and the Women Writers Project have evaluated the share of female authors and writers. In addition, authors such as Argamon et al. (2017) have studied the linguistic styles of male and female playwrights and representations of gendered bodies in European fairy tales. Authors Goldstone and Underwood (2017) uncovered the underlying trends across a century of academic articles in literary studies (Goldstone and Underwood, 2017) by mapping first name to gender. Moreover, Cook (2011) performed a mapping between names and gender based on historical African American naming practices and identified a set of "distinctively African American names." This mapping between name and gender has been used in multiple papers to evaluate the share of African Americans in a list of names that are attributed to patenting activity (Cook, 2011), for general gender categorization (Jung and Ejermo, 2014) and age (Jones, 2017) and income estimation (Celik, 2015). Most recently, Blevins and Mullen (2015) inferred gender from one of the most common features of humanities data sets: the personal names of authors. Mapping first names and gender over time, they investigated changes in naming conventions in the United States.

In census, scholars have examined naming conventions and changes in them. They found that, for example, the conventional genders for various names switched over the course

of a few decades. In 1900, approximately 92% of newborn babies who were named Leslie were male, while in 2000, approximately 96% of the Leslies born in that year in the United States were female (Blevins and Mullen, 2015). Changes in naming practices create a problem, especially for databases with records of people who were born in the "transition phase" of naming conventions. This is known as the "Leslie problem." As the average lifespan of humans is increasing, this problem is further intensifying. Today, the average European citizen reaches an age of more than 80 years (Ludwig et al., 2014), which is sufficient time for naming practices to change (Blevins and Mullen, 2015). In addition, applying a name-to-gender mapping enables the role of patents that are attributed to female inventors to be discussed in research about gender disparity in patenting (Sugimoto et al., 2015). In their research, Sugimoto et al. (2015) created their mapping based on universal and country-specific name lists of unknown origin, which they did not further elaborate upon, and applied it on a sample of 4.6 million utility patents that were granted by the United States Patent and Trade Office (USPTO). A study by Hunt et al. (2013) concluded that the number of academic and female innovators, which they identified from their mapping of name and gender, is a suitable metric for female innovation. Patent activity was also studied in terms of gender distribution in a recent study by Elsevier Analytical Services (Hunt et al., 2013). In this study, the authors mapped research performance, including patents that were granted, with the inventors' genders. The authors relied on social networking service data to calculate the probabilities. Hence, naming conventions from 1930-1950 may not be entirely representative for the current share of people who are alive. In this study, which uses social media profiles, the authors limited their mapping to names that appeared at least 5 times in their data set and had a probability of corresponding to a male or female of at least 85%.

Many studies that I found discuss differences in behavior between the genders or segment based on gender. To label by gender, some authors found creative ways to create name-gender mappings, whereas others relied on lists that were created by other researchers. Some identified the need to consider other variables to improve the mapping quality, and only very few (Blevins and Mullen, 2015) quantified their findings. The importance of these findings depends on the intended application of the name mapping, which determines whether high precision or high recall is needed.

To the best of my knowledge, there is no research on how changes in naming conventions and nationality can improve mappings. To this point, it is unclear what the

determinants of a mapping that outperforms in either high recall or high precision are. My literature review identified studies that address how additional attributes can enhance the mapping quality ("Leslie problem"); hence, I believe that incorporating additional features into a mapping can enhance the mapping quality, thus enabling me in a second step to correctly identify the gender of inventors, which will be discussed in Chapter IV.4.2.

In order to build this mapping, I collect a data set from policy data warehouse of the Swiss car insurer and extract the first name and gender for each available policy. In addition, I extract the nationality and date of birth. In the process, I evaluate whether the inclusion of those demographic characteristics increases the discriminative power and improves the categorization precision over basic first-name-gender mappings. Then, I apply the mapping to 20 000 names that the mapping algorithm has not seen and evaluate the mapping performance in two scenarios:

Mappings in 2-label and 3-label categorization setting First, by evaluating precision of the mapping outcome in categorizing male and female names, and second, by evaluating the performance of the mapping in a three-class scenario with female, male, and unisex labels.

Probabilistic frequency calculation for gender The models are based on the probabilistic occurrences of names and gender (see Tables 1, 2 and 3). Therefore, I count the occurrences of each first name being labeled as male and female. From the occurrence in each category, relative to the overall counts, I compute the probabilities of the first names being categorized as male (P(m)) and female (P(f)), as illustrated in Table 1 for the name "Peter."

Table 22 - Distribution of the name "Peter"

Name	Male (m)	Female (f)	Total (t=m+f)	P(m)	P(f)
Peter	67 759	52	67 811	.99	0 [4]

Then, I further divide the two categories according to birth decade. From the occurrence in each subgroup, relative to the overall counts, I compute the male and female probabilities of

[4] The true value of P(f) is calculated as 52/67,811 = 0.000767.

the first name being categorized as male and female, as illustrated in Table 2 for the name "Gabriele."

Table 23 - Distribution of the name "Gabriele" by decade

Name = Gabriele	Male (m)	Female (f)	Total (t=m+f)	P(m)	P(f)
Decade:					
1900	0	9	9	0	1
1910	3	12	20	.15	.85
…	…	…	…	…	…
1990	180	6	186	.96	.04

Finally, I distinguish nationalities for all the subgroups. Then, from the occurrences in each subgroup, relative to the overall counts, I compute the probabilities of each of the first names being categorized as male and female, as illustrated in Table 24 for the name "Andrea." I assume that adding decade and nationality will improve the mapping of name and gender for some of the naming exceptions.

Table 24 - Distribution of the name "Andrea" by nationality

Name = Andrea Decade = 1970-1979	Male (m)	Female (f)	Total (t=m+f)	P(m)	P(f)
Nationality					
CH	718	8 313	9 036	.07	.93
AT	1	67	68	.01	.99
…	…	…	…	…	…
IT	386	3	389	.99	.01

Probabilistic nationality imputation I also provide a solution for imputing nationality using a prediction model, which follows the procedure discussed by Batista and Monard (2003). Out of the 4 058 139 observations from the motor insurance, I extract 50 000 distinct entities, which I call the holdout sample. I use the remaining 4 008 139 records to construct a mapping between last name and probabilistic nationality. Most last names of high count have multiple nationalities assigned to them. Almost all names have observations with Swiss nationality (majority class) assigned to them, in addition to other nationalities. However, to construct a final mapping, I mapped each available name to the nationality that I observed in at least 50%

the respective name-nationality mappings. The final mapping consists of 780 876 name-nationality tuples. The most common names within the mapping are all assigned the Swiss nationality. For illustration, the five most frequent last names are *Müller, Meier, Schmid, Keller* and *Weber* (full list in Appendix 27). The most frequent names that are assigned to non-Swiss nationalities have much lower frequency counts. I found 139 942 last names that are predominantly carried by non-Swiss individuals (dominant names). However, most of these names only occur once (long-tailed distribution), and I did not further validate the spellings in the complete list of Swiss and non-Swiss names. Out of 139 942 dominant non-Swiss names, 6 072 were carried by 10 or more people. The summary of the distribution of those names by nationality is listed in Table 25.

Table 25 - Non-Swiss dominant names

Dominant Nationality	Count of names in data set by Dominant Nationality	Share of dominant non-Swiss names
Italian	1 440	24%
Portuguese	1 287	21%
German	679	11%
Others	2 666	44%

The most frequent unique dominant non-Swiss last names are Portuguese. These names occur more than 500 times. The mapping is used to impute the nationality of a person who is living in Switzerland when given only a last name. The mapping does not consider naming conventions outside of Switzerland, nor does it distinguish between last names with a high share of the total observations or names that are part of the mapping only by chance. The last name-nationality mapping is a compendium of Switzerland's immigration and naturalization standards and its inhabitants' marriage and naming inheritance conventions. However, its composition will not be elaborated in further detail and will be considered as given in the remainder of this study.

Mapping of name to gender I map the names to gender based on the male and female probabilities in Table 22, Table 23 and Table 24. First, I assign the binary labels of male and female to the names in the mapping, based on a probability of more than 50%. Second, I add another label to those, which shows dispersion of gender for each name. I assign an additional label that indicates the gender dispersion for each name. I choose to label all names that have

a probability of less than 0.95 of being either male or female as unisex names. For example, in the case of a binary categorization, “Gabriele” in row 2 of Table 23 is labeled as a female name because the probability of “Gabriele” being female is greater than that of being male according to the data set. However, in the case of three-category labeling, I assigned a unisex label because the probabilities of “Gabriele” being male and female are both less than 0.95 for some of the decades. I performed this labeling procedure for all entries of Table 22, Table 23 and Table 24. The procedure of mapping names from the holdout sample (test sample), which also includes names that are not found in the name-to-gender mapping, will be explained in Chapter IV.4.2

Testing of the gender categorization methodology I estimate the gender for the test set of 20 000 observations (see Test Sample 1) by mapping (1) first names, (2) first names and decades, and finally (3) first names, decades, and nationalities for the binary- and three-label categorizations. I compare the results against the ground truth for the binary- and three-label categorizations for Table 22, Table 23 and Table 24. However, names that are found in the test sample are not part of the mapping table and are not assigned a gender. The number of unmapped names is counted and used in the evaluation. For the binary categorization, I compute true positives (tp) as the number of males and females who are correctly identified as male and female, respectively, by the predicted labels (Colquhoun, 2014). For the 3-label categorization, I omit the names that are being labeled as unisex from the random sample and compute true positives (tp) as the number of males and females who are correctly identified as male and female, respectively, by the predicted labels. I assume that names with unisex labels cannot be predicted based on the information of name, age, or nationality. Therefore, omitting unisex labels will reduce the number of samples from the random set but should increase the precision of the model as I try to predict the genders of only those names whose probability of being associated with a gender is greater than 0.95. I further calculate precision and recall scores. Recall is defined as the ratio of the number of relevant observations that are retrieved to the total number of relevant observations (Raghavan et al., 1989). Precision is defined as the number of relevant observations that are retrieved divided by the total number of retrieved observations (Raghavan et al., 1989). Additionally, I calculate the F1-Scores for all three tables in both categorizations. These metrics enable the comparison of the results of the algorithms according to Equation 9, Equation 10 and

Equation 16 - F1-Score

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Using the binary categorization approach, I also apply the mapping procedure to categorize Test Sample 2 to estimate the share of inventors by gender. I use the calculated probabilities from Table 22 as attributes such as nationality and birth year are not published with patent data. As I do not know the actual genders of the inventors, I can only compare the mapping results with other studies using aggregated numbers. Hence, I do not calculate the evaluation metrics of precision, recall, and F1-Score.

Testing of the nationality imputation methodology To enable the categorization based on name, decade, and nationality as mentioned in Chapter IV.3.3, I provide a solution completing the nationality attribute when the attribute is incomplete. Therefore, I estimate the nationality of the holdout set of 50 000 observations by mapping last names to the most frequent nationality that is found with at least 50% probabilistic occurrence. I treat nationality, which is the attribute with missing data, as the class attribute and use the remaining attribute (last name) as input for the predictive model to predict nationality (Batista and Monard, 2003).

I compare the results against the ground truth. I use the ground truth as reference data, which is critical for data validation (Gudivada et al., 2017). I compute true positives (tp) as the number of (name, nationalities) tuples that are predicted correctly. For the names that are found in the holdout sample but not in the mapping, I show the results in two ways.

First, I predict each record that is not found in the mapping as Swiss, since more than 50% of all observations that were used to create the mapping are Swiss. Hence, I expect this naive imputation to improve the recall score. I call this No Information replacement of nonmapped names.

Second, I omit records that are not found in the mapping. This will reduce the number of samples from the test set. However, I expect the precision of the model to increase.

Imputation evaluation I calculate the precision, recall, and F1-Score using the caret-package in R. I do this for both approaches with respect to the majority class (Swiss names, 85.78%), which enables the comparison of the results of the name-to-nationality mapping according to Equation 8-10. In addition, I evaluate the overall mapping quality by examining the accuracy, which is calculated using Equation 8. Additionally, I provide 95% confidence intervals of the

Accuracy and the No Information Rate (NIR), which allow to evaluate the quality of the mapping since my data set has a predominant majority class. Finally, I show the P-Value of the Accuracy is greater than that of NIR.

IV.4 Results

I will first show how SMEs with patents show a different growth pattern than SMEs without patents. Then, as an interim step to augment the patent data set with the gender attribute, I visually present some of the findings when analyzing the data, which give supporting evidence that detailed information about nationality and decade of birth can improve the name-to-gender mapping. Then the results of the imputation are shown and evaluated. Based on the imputation findings, the correct name-to-gender mapping is selected and applied. Then, the augmented data set is analyzed and used to predict differences in SME growth.

IV.4.1 Descriptive Statistics

Patenting SMEs The SMEs for the final analysis have been reduced to the industries in which SMEs with patents were found. This is done for comparability reasons, mitigating a potential industry selection bias. The down-sampled SME data set contains approximately 60% SMEs without patents and 40% SMEs with patents. The distribution of each category is therefore not representative for the insurance data set or the Swiss economy. The data show that 1 625 SMEs belong to the service industry, 1 067 are related to real estate, 530 to trade, and 881 to manufacturing industry.

In order to investigate hypothesis H3.1, I separated those SMEs with patents from those without. To each of the groups, I assigned a class according to the following rule:

SME with 0 patents	as class zero (green line in Figure 7)
1-2 patents	as class one (black line in Figure 7)
>2 patents	as class two (orange line in Figure 7)

The final data set has 4 103 companies, of which 3 618 belong to Class 0, 376 belong to Class 1, and 109 belong to Class 2. I call the categorization the SME class. I further calculate patent count, number of cites, count of male and female inventors as features from patent data to investigate their significance in predicting business growth between 2010 and 2016.

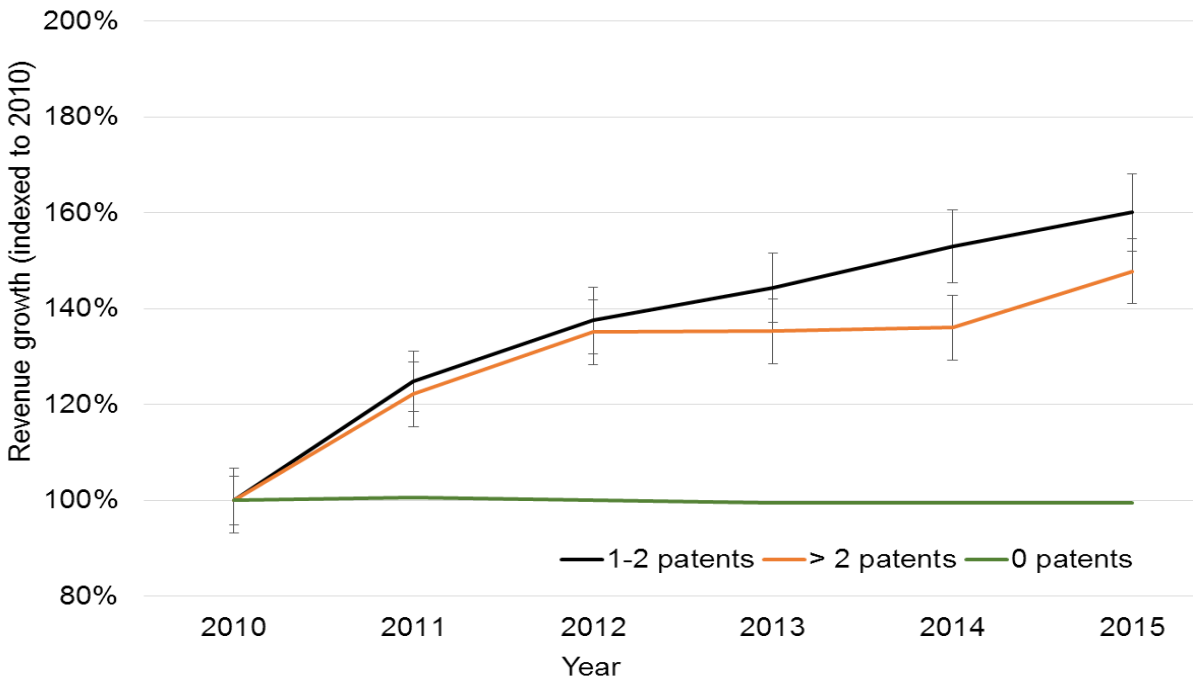
According to the insurance company's internal classification, the SMEs can be grouped by sector. Table 26 summarizes the sector's count of each patent class.

Table 26 - Patent and SME counts of industry groups

Industry	SMEs with patents	Total SMEs	Class zero	Class one	Class two
Services (misc. Services)	797	1 625	1 490	104	31
Manufacturing and engineering	441	1 067	897	126	44
Real Estate related (construction, modification, renting, real estate based services)	206	881	800	67	14
Trade (selling of any kinds of goods)	253	530	431	79	20
Total	1 697	4 103	3 618	376	109

In order to investigate the importance of patents, I plot the change in revenue, indexed to 1 for the year 2010 of SMEs and the consecutive 5 years by SME class.

Figure 7 - SME growth by patent count



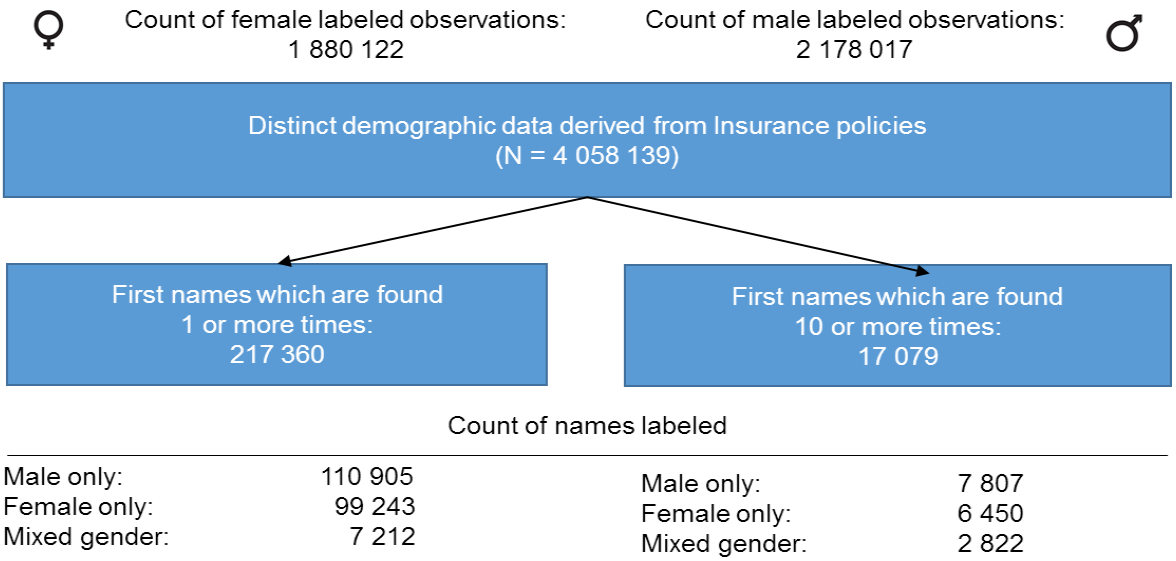
n = 4 103

Figure 7 shows the growth of SME with (black, *n* = 376 and orange, *n* = 109) and without patents (green, *n* = 3 618) in the sample. The bars indicate the standard error of the mean for each year. I find that companies with one or at least two patents (top line) experienced on average 60% and 48% growth accumulated over five years. On the other hand, the groups zero, representing SMEs without patent applications, experience no growth (-0.53% accumulated

over five years, indicated by flat green line). In order to validate the statistical significance of the average growth values of the three classes, I performed a *t*-test on all three possible combinations as well as an analysis of variance (ANOVA, see Appendix 28-31). ANOVA is a hypothesis-testing statistical technique used to compare the means of more than two groups. I have used a confidence interval of 0.05 for the test. Both tests show a significant difference between the patent groups and the nonpatent groups. Applying a *t*-test to all three possible combinations of SME classes, I find that the *t*-tests between Class 0 and Classes 1 and 2 show a *p*-value less than 0.05; hence, I conclude that there is a difference in SME growth between SMEs with patents and SMEs without. This supports H3.1.

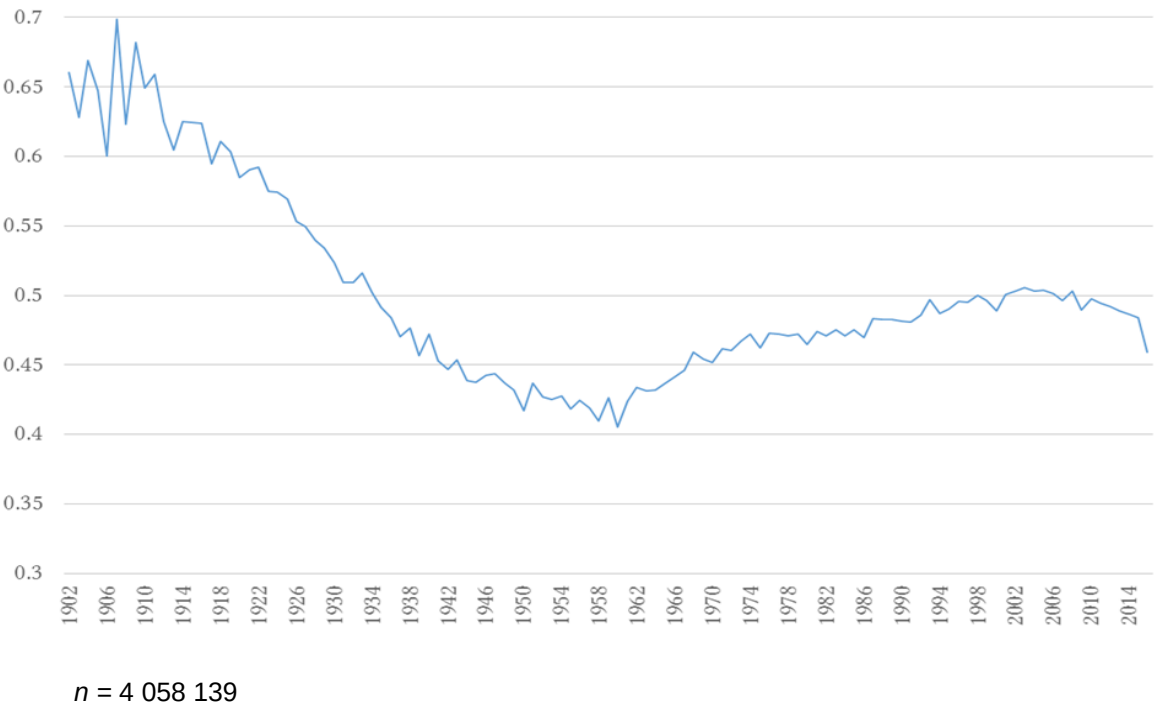
Name and gender distribution in the motor insurance data The data set derived from the motor insurance information is comprised of 4 058 139 first and last names and their corresponding genders, nationalities, and dates of birth. Out of the 4 058 139 names, 46.3% (1 880 122) are labeled as females and 53.6% (2 178 017) as males. The data are comprised of 217 360 unique first names, of which 17 079 unique names occur more than 10 times. Out of the 217 360 unique names, a total of 99 243 are uniquely labeled as females, whereas 110,905 are uniquely labeled as males. The remaining 7 212 names have various frequencies of being male and female. Of the 17 079 names that occur more than 10 times in the data set, a total of 7 807 are uniquely male names, whereas 6 450 are uniquely female names, and the remaining 2 822 names of the samples have instances of male and female first names in the data. Figure 8 illustrates the composition of the source data with respect to female and male observations and highlights the distributions between high- and low-frequency counts of first names by gender.

Figure 8 - Data description and first name distribution



From the list of first names and their gender, I compute the probabilities of first names being male or female. Sorting this list by date of birth, number of observations, and gender, as illustrated in Figure 9, shows the distribution of female and male insured over time.

Figure 9 - Percentage of females insured to total

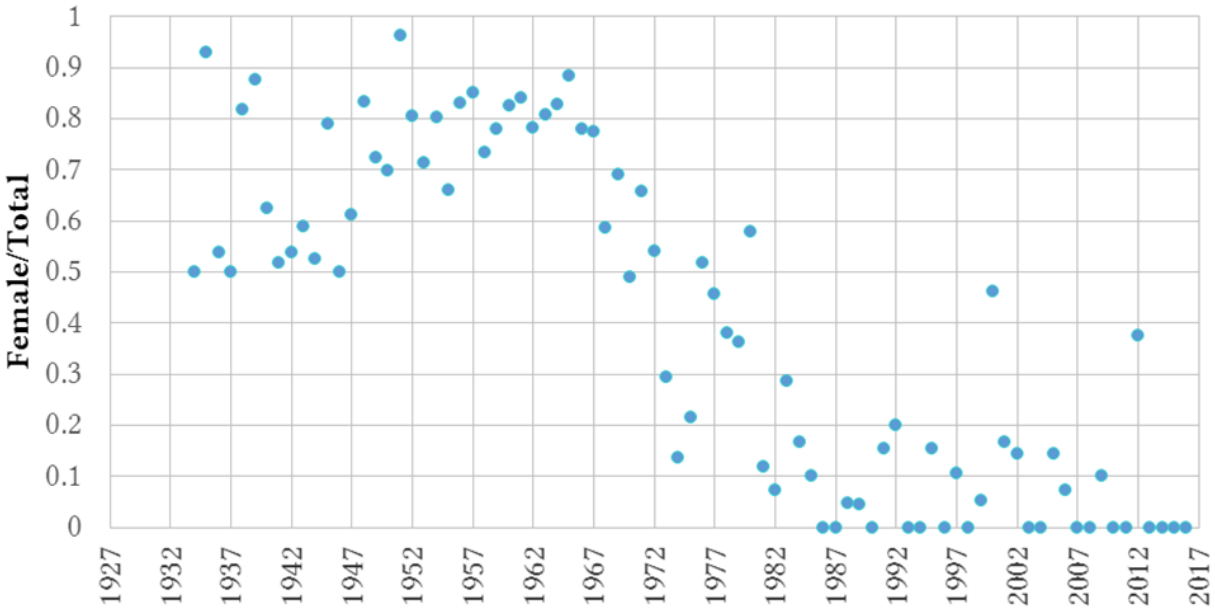


n = 4 058 139

The data show a high share of females for the year at the beginning of the 20th century. The female population as a share of total customers peaked in 1907 at 7 out of 10 customers who were born that year. The overrepresentation of women who were born in the beginning of the 20th century in the sample, which was derived from insurance policies, may be associated with gender-biased intrahousehold financial decision-making (Ashraf, 2009) and the consequences of World War II.

To further explore the data set, I test whether naming conventions have changed over time. Naming practices were mostly consistent, with the exception of a few first names such as "Gabriele," "Michele," "Dominique," "Deniz," "Isa," or "Kim." These names occur frequently as male or female names. I find that some of these naming conventions for gender have changed over the years, even within a single decade. Figure 10 shows the share of females who were named "Gabriele" in the database relative to the total number of people who were named "Gabriele" and born in the same year. As indicated by Figure 10, the name "Gabriele" was carried predominantly by females in the early 1900s.

Figure 10 - Female share of the name "Gabriele"



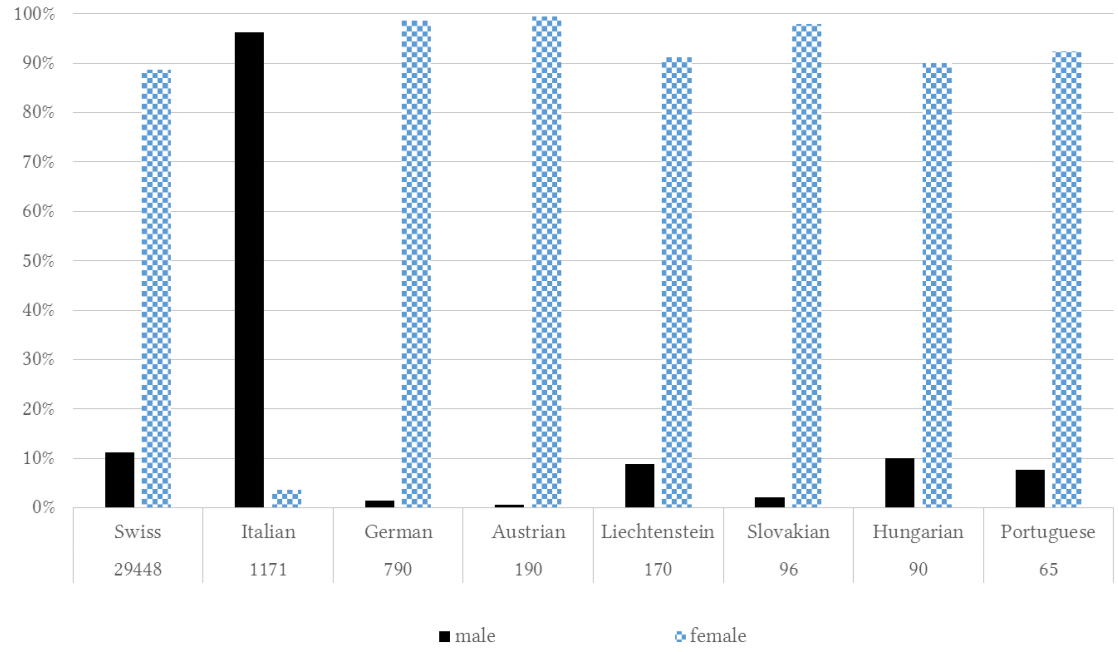
n = 2 136

From 1930-1940, "Gabriele" was a unisex name, while from 1960s onwards, it became a predominantly male name. Besides changes in naming conventions, the data also shows a

sharp rise and decline of popularity of the name "Gabriele" as a female name around 1965. From 1985 onward, the data only show rare instances of "Gabriele" as a female name in the data.

Finally, I explore differences in naming conventions across the 37 nationalities that are prevalent in Data Set 1. I found that a few names, such as Andrea, Luca, and Nicola, showed particular differences in conventions and had sufficiently large counts (>100) for evaluation. For illustration, Figure 11 shows the male and female percentage shares of the name "Andrea" in the countries Switzerland, Italy, Germany, Austria, Liechtenstein, Slovakia, Hungary, and Portugal, with counts of 29 448, 1 171, 790, 190, 170, 96, 90 and 65. In Italy, people who are named "Andrea" are predominantly males, whereas in Switzerland and all other countries from Data Set 1, the name is female. Figure 11 shows the nationality differences for the name "Andrea".

Figure 11 - Nationality differences for the name "Andrea"



n = 32 020

IV.4.2 Imputation Results

As illustrated, naming conventions have changed over time. Further, different nationalities have different naming conventions. Therefore, I assume that an imputation of first name-to-gender can be improved by recognizing the differences in naming conventions. To test this

assumption, I follow the train-test procedure. After the mapping has been created, I apply the gender-mapping procedure to test set for validation. I apply the mappings in 2-label and 3-label categorization settings (see IV.3.3). The attributes that I initially use for the mapping are:

- Imp1: name,
- Imp2: name and decade,
- Imp3: name, decade and nationality

Additionally, I apply nationality imputation to name and decade mapping to demonstrate a scenario in which I do not have full information about the nationalities of all observations (as in the patent data example). Since not all last names of the test set are found in the nationality imputation table, I perform the imputation in two ways: (1) I omit the names that are not found in the table, and (2) I assign the majority class (Swiss nationality) to all names that are not found in the nationality imputation table.

Therefore, the additional mapping attributes that I use are:

- Imp4: name, decade, and imputed nationality (omitting all names that are not found in the nationality imputation table)
- Imp5: name, decade, and imputed nationality (No information (NI), assigning all names that are not found in the nationality imputation table to the majority class, which is the Swiss nationality)

Successfully imputing the nationality is a prerequisite to expanding the mapping models with the mentioned two attribute sets. Hence, I first state the results of the nationality imputation before continuing with the name-to-gender mapping based on all five iterations of the mapping attributes. The best suitable mapping is then selected and applied to the patent data.

Results of nationality imputation I use the last-name-to-nationality mapping to predict the nationalities of the last names from the holdout sample. Table 27 shows the results for the majority class, which has 42 891 entries.

Table 27 - Nationality imputation with no information replacement of non-mapped names

	N=50 000	**Actual class**	
		Swiss	Not-Swiss
Predicted class	Swiss	41 974 (tp)	5 230 (fp)
	Not-Swiss	917 (fn)	1 879 (tn)
Precision (%)	88.92		
Recall (%)	97.80		
F1-Score:	93.18		

However, since not all observations in the holdout sample are found in the mapping, several last names were assigned using a no-information approach to the majority class of the mapping. To show the actual performance of the mapping, I also show the results of the mapping for the majority class, limited to entries that were found in the mapping (37 732 = 36 815 + 917) in Table 28.

Table 28 - Nationality imputation without non-mapped names

	N=42 522	**Actual class**	
		Swiss	Not-Swiss
Predicted class	Swiss	36 815 (tp)	2 966 (fp)
	Not-Swiss	917 (fn)	1 824 (tn)
Precision (%)	92.54		
Recall (%)	97.57		
F1-Score	94.99		

To compare the last-name-to-nationality mapping for all classes (not only the majority class, which is the Swiss nationality), I provide the mapping results for both approaches. Table 29 shows the mapping accuracy and enables a comparison with the No Information Rate.

Table 29 - Overall nationality mapping evaluation

Approach procedure	**Approach 1** 1.) Apply mapping 2.) Assign nationality “Swiss” to unmapped cases (7 478)	**Approach 2** 1.) Apply mapping 2.) Omit all unmapped cases
N	50 000	42 522

Accuracy (%)	87.18	90.26
95% CI:	(0.8688, 0.8747)	(0.8997, 0.9054)
No Information Rate (%)	85.78	88.62
P-Value [Acc > NIR]	< 2.2e-16	< 2.2e-16

Both mapping approaches show significance and a high accuracy. Using the mapping derived from the motor insurance data, in particular the last name of a person, the imputation of nationality is possible at a high accuracy using both approaches. Therefore, I include both of them in the evaluation of a suitable imputation approach for the name-to-gender mapping, augmented by attributes such as (imputed) nationality and decade of birth.

Results of 2-label gender categorization The 2-label (binary) gender categorization maps each first name found in the mappings to its most frequent gender (male, female). As a result, I matched 18 994 names from Table 22. I matched 18 357 name-decade combinations from Table 23 and 16 716 name-decade-nationality combinations of the test set from Table 24. Having no information about nationality and imputing the attribute yields 17 169 and 17 902 name-decade-imputed-nationality combinations. Imputing nationality and replacing those not found in the mapping with the majority class (Swiss) yields 17 902 name-decade-nationality combinations. Out of the 20 000 instances in test set 1, 10 762 were male and 9 238 were female. The results of the binary categorization can be found in Table 30. With only first names, a total of 10 072 out of 10 762 (93.59%) males were identified correctly. Thus, the precision increased from 98.55% to 98.67% by adding the decade to the first name and increased to 98.78% by adding the nationality to the name and the decade. Imputing nationality yields improved precision results compared to only name or name-and-decade mappings. The F1-Score decreases with the addition of decade and nationality to name, which indicates that the name-to-gender mapping is not augmented by the decade and nationality.

Table 30 - 2-label model performance

Label	*Name*	*Name, Decade*	*Name, Decade, Nationality*	*Name, Decade, Imputed Nationality*[5]	*Name, Decade, Imputed Nationality (IN)*[6]
Abbreviation	***Imp1***	***Imp2***	***Imp3***	***Imp4***	***Imp5***
Males (tp)	10 072	9 753	8 893	9 148	9 519
Females (tp)	8 647	8 361	7 621	7 816	8 160
Total (tp)	18 719	18 114	16 514	16 964	17 679
Total (fp)	275	243	202	205	223
Mapped	18 994	18 357	16 716	17 169	17 902
Unmapped	1 006	1 643	3 284	2 831	2 098
Recall (%)	93.60	90.57	82.57	84.82	88.40
Precision (%)	98.55	98.67	98.79	98.80	98.75
F1-score	96.01	94.45%	89.95	91.28	93.29

Results of 3-label gender categorization The 3-label gender categorization maps each first name found in the mappings to its most frequent gender (male, female), and if the probability is less than 95% for one of the two genders, it maps it to unisex. As a result, I was able to match 19 124 names, 14 709 name-decade combinations, and 10 109 name-decade-nationality combinations out of 20 000 instances of the random sample with the data sets from Table 22, Table 23 and Table 24, respectively. For evaluation purposes, I only considered the male and female labels, which reduced the sample to 18 184 names, 13 232 name-decade combinations,

[5] Nationality was imputed using the mapping that was introduced in section IV.3.3 - Imputation methodology.

[6] Nationality was imputed using the mapping introduced in section IV.3.3 - Imputation methodology. Additionally, all names that could not be found in the mapping but occurred in the holdout set were assigned the majority class (Swiss nationality).

and 7 797 name-decade-nationality combinations. Imputing nationality increased the number of name-decade-nationality combinations to 15 563 and 16 160. With only first names, 8 410 females and 9 547 males out of 8 506 and 9 848, respectively, were identified correctly. The precision of the model was increased from 99.12% to 99.26% by adding decade and nationality to predict the first names. The F1-score also increased from left to right, as I observed in the binary categorization. The results are summarized in Table 31.

Table 31 - 3-label model performance

Label	***Name***	***Name, Decade***	***Name, Decade, Nationality***	***Name, Decade, Imputed Nationality***[7]	***Name, Decade, Imputed Nationality (IN)***[8]
Abbreviation	***Imp1***	***Imp2***	***Imp3***	***Imp4***	***Imp5***
Males (tp)	9 757	8 252	5 845	8 417	8 677
Females (tp)	8 267	4 880	1 894	7 019	7 257
Total (tp)	18 024	13 132	7 739	15 436	15 934
Total (fp)	160	100	58	127	136
Mapped	18 184	13 232	7 797	15 563	16 070
Unmapped	940	1 477	2 312	597	90
Recall (%)	94.25	89.28	76.55	95.52	98.60
Precision (%)	99.12	99.24	99.26	99.18	99.15
F1-score	96.62	94.00	86.44	97.32	98.87

[7] Nationality was imputed using the mapping that was introduced in section IV.3.3 - Imputation methodology

[8] Nationality was imputed using the mapping introduced in section IV.3.3 - Imputation methodology. Additionally, all names that could not be found in the mapping but occurred in the holdout set were assigned the majority class (Swiss nationality).

The precision of the 3-label categorization is higher than the precision of the binary categorization in all five cases. However, the absolute number of true positives (tp) is higher in the 2-label categorization, which is partially due to the eschewal of unisex labels. The F1-Score is highest for the 3-label categorization with Name, Decade, and Imputed Nationality (IN).

Imputation of gender of patent inventors The purpose of the gender imputation on the patent data set is to be able to categorize the patents by male, female, and mixed inventor teams. To assign the gender to many patents, the imputation required a high recall score and only permits a binary classification as there I choose not to further increase the patent groups by the introduction of unisex labels. Due to the relatively small difference in precision scores among Imp 1-3 in the 2-label gender categorization, I apply the method that yields the highest recall score. Applying Imp1 to the patent data set, I identified 1 829 male inventors and 229 female inventors. The method is able to identify the names of the inventors of 464 out of 496 SME by the name-matching algorithm. Out of those 464 companies, 41 had female-only inventors, 393 had male-only inventors, and 30 had mixed-gender inventors. Out of the 1 667 patents, 75 had only female inventors, 887 had male inventors, and 705 had mixed-gender inventors, as summarized in Table 32.

Table 32 - Counts of SMEs and patents by gender

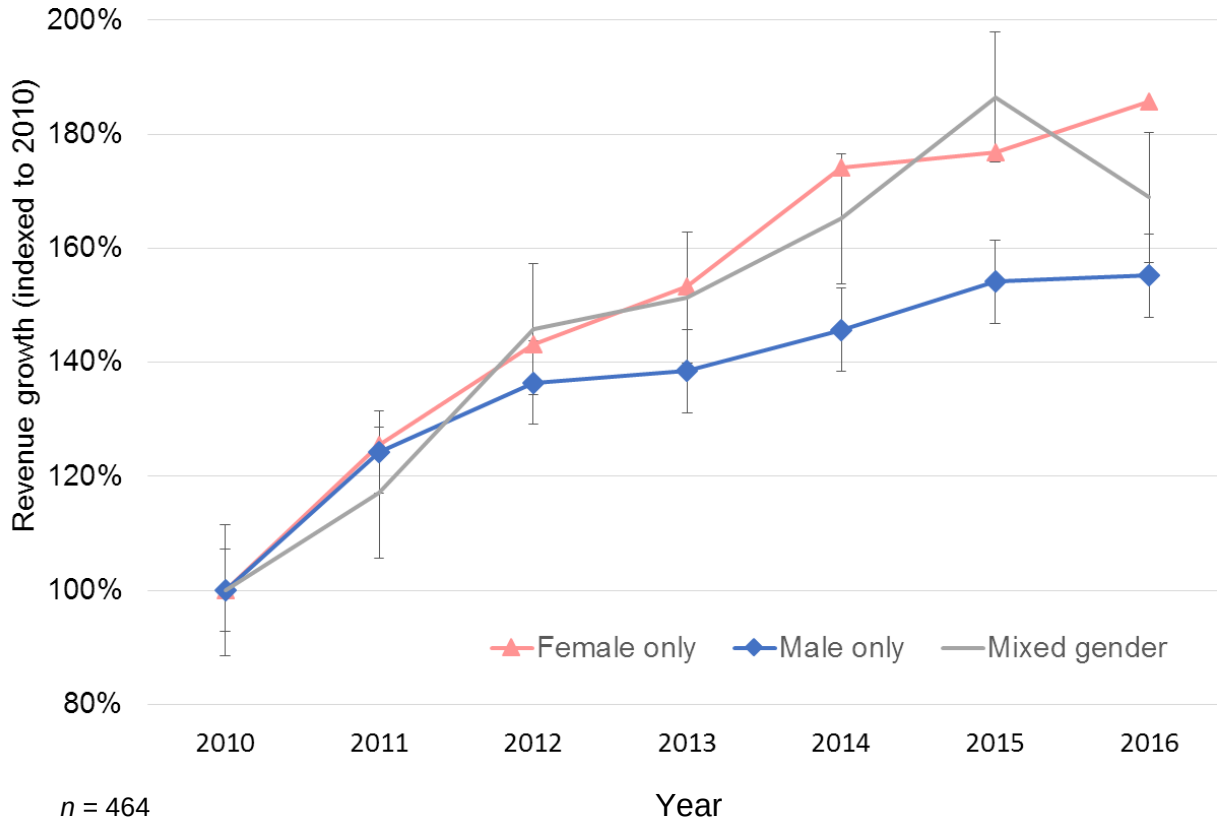
SMEs with patents	Count
Male only	393
Female only	41
Mixed gender	30

Patents	Count
Male only	887
Female only	75
Mixed gender	705

Counting the individual inventors by gender, I find that 11% of all inventors are females and 89% are males.

Applying the imputation categorization Further, itemizing the attributes of patent application, I group SMEs with patents by gender composition of the inventors mentioned on the application. I show the indexed performance of male only, female only, mixed teams as well as the sum of all patent-filing SMEs (grand total) for the years 2010-2015 in Figure 12.

Figure 12 - SME growth by patent application gender



I validated the findings for statistical significance of the three groups over all the years and found that there was no significant difference between the female-only inventor SMEs (n = 41) and mixed-gender companies (n = 30) in any year (pink and grey line). There was significant difference between mixed-gender and male-inventor companies (n = 393) in the years 2014 (p value = 0.0381, T value = 2.764) and 2015 (p value = 0.0491, T value = 2.034, see Appendix 32-34). However, there was also no significant difference in the growth of male-only and female-only patent inventors. I conclude, given a sufficiently long observation period, that SMEs with mixed-gender invention teams outperform most other patent-filing SMEs in terms of revenue growth.

IV.4.3 Prediction Results

To evaluate the predictive power of patents, I train a prediction model on 3 104 companies with and without the patent-application-derived features. For the prediction, I selected a random forest as a suitable algorithm and fed the test set to the model. The model's performance on the test set is summarized in Table 33.

Table 33 - Random forest prediction results with and without patent attributes

Results	Random Forest (w/o patent features)	Random Forest (with patent features)	Improvement
R^2	.527	.838	.311
RMSE	.26	.18	-.08

The model with patent features has a better R^2 and a lower RMSE (residual mean squared error). This supports H3.2. To evaluate which features from the patent applications contribute most to the model's performance, I further plot the features importance to the (improved second) random forest model. I find that the IncNodePurity in random forests, which is the total decrease in node impurities from splitting on the variable, averaged over all trees has a group-like structure of features. The most important model feature is Area_income and expresses the average income per person in CHF in each Zip code. Also important are City_Rural_Index, a numerical value that separates rural and metropolitan areas and Nominal_capital, which relates to the initial funding when a company was set up. From the patent-related features, I find Cited_Count, Patent_Count, Male_Inventors, Female_Inventors and Zero (SMEs with 0 patents) to be important.

Due to brevity, I do not comment on whether the value of the features for the trees has a positive or negative impact. However, I rank the features. The IncNodePurity of features from model with and without patents can be found in Figure 13 and Figure 14. The patent related features are highlighted with a yellow star.

Figure 13 - Random forest feature importance without patent data

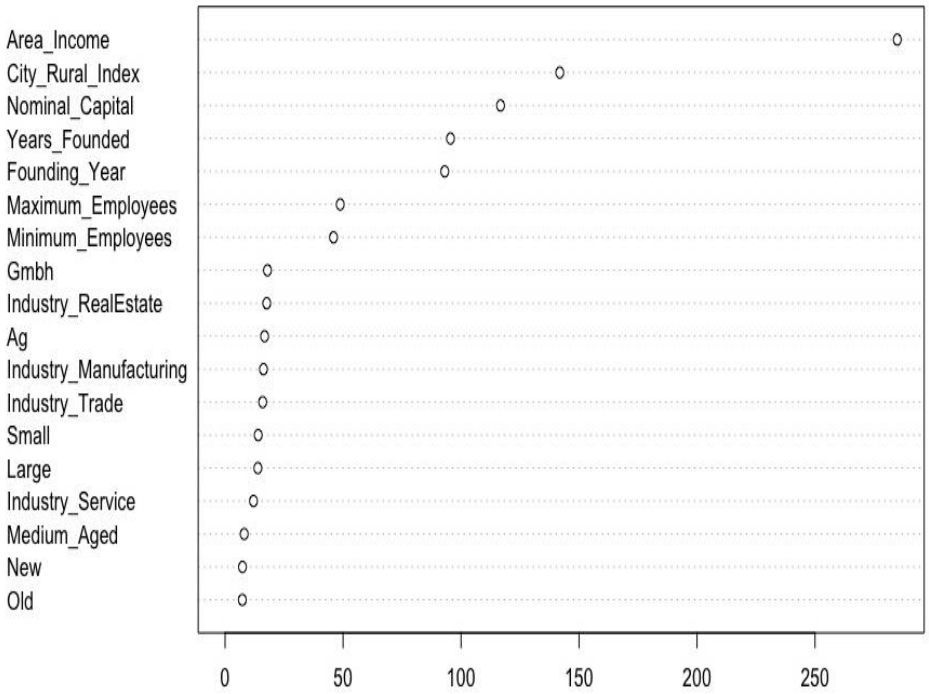
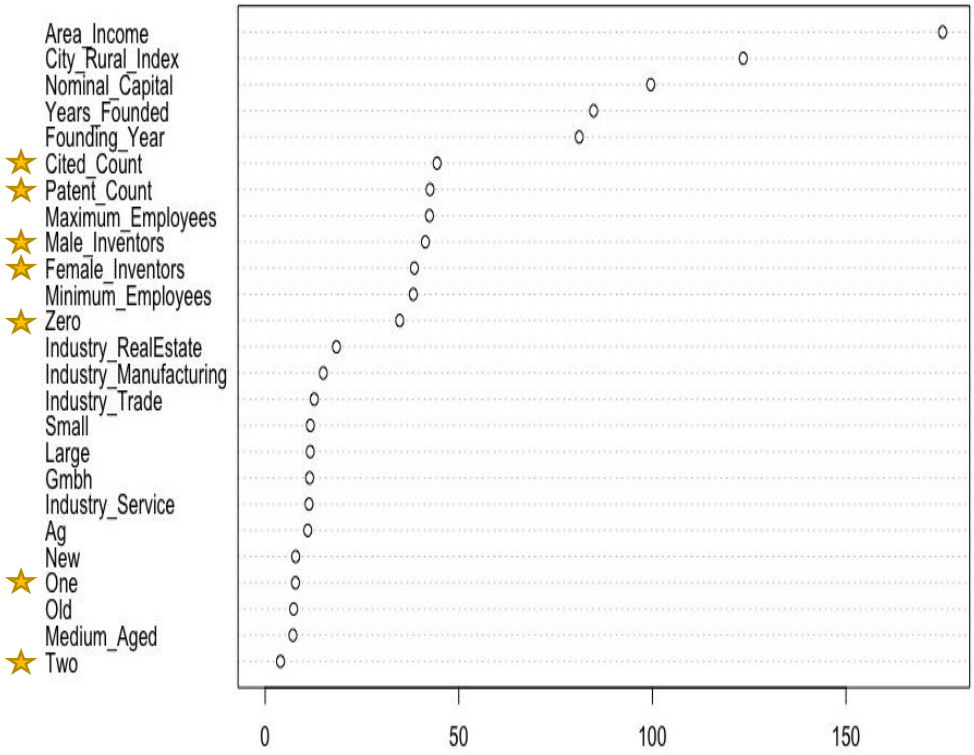


Figure 14 - Random forest feature importance with patent data



IV.5 Summary

IV.5.1 Discussion and Conclusion

The findings of this study reveal that, based on the provided data sample, out of 106 000 Swiss SMEs in business between 2010 and 2015, 496 SMEs (0.47%) had filed a patent in Switzerland. Adjusted for industry composition, the patent-filing SMEs grew significantly faster within the observation period. Itemizing the patent applications, the study further finds that SMEs with mixed-inventor teams (females and males) grow annual revenues significantly faster than male-only inventor teams.

Further, a prediction of the indexed growth of SMEs based on attributes provided by the insurance company, the commercial register, and Swiss federal office of statistics yields a R^2 value of .527 and a RMSE value of .26 on the test set. Additionally, using the augmented patent data, retraining the model, and reevaluating it with the help of a test set improves the R^2 value of .838 (+.311) and reduces the RMSE to .18 (-.08). The study results reveal that attributes found in patent applications allow to infer the growth pattern of patent filing SMEs. Adding information about the patents of an SME can greatly enhance the prediction capacity of a model due to the prediction capacity of patents. The data source is publicly available for an insurance company, and its contents can be collected without any reservation.

IV.5.2 Implications for Research and Practice

The study's findings indicate that a large population of SMEs do not file patents. This is because most SMEs do not have intellectual property that requires protection or because once they do, they decide against patents as a strategy for appropriation. Those SMEs that do file patents grow much faster compared to those who do not file patents. Therefore, the filing of a patent can be used as a trigger signal of SME growth, which can be captured form public data. Implemented into a company's CRM system, such a trigger may be of interest for executives of business-to-business companies such as insurance companies, which benefit by knowing which of their customers are high-growing companies and which are low-growing ones. Correctly estimating expected growth rates could eventually help them prioritize among their customers and divert resources to the most promising ones (i.e., the ones with the most growth).

IV.5.3 Limitations and Future Work

Despite the novelty and width of the study, it has limitations. I investigated the importance of patent-related attributes to predicting SME revenue growth. I compared the averages of growth of revenues in a span of 6 years (2010-2015) for three groups of companies that are companies without patents, companies with one or two patents, and companies with more than two patents. As the selection of the SMEs in the sample was limited to resampling the industries, ex-ante found in the patent filing SMEs; the study findings are not generalizable to all SMEs. The same is true for the evaluation metric of the random forest prediction. The features used in the model, which included size, number of employees, and jurisdictional form, mitigate the section bias with respect to the size, type, and industry of the firm. This was a major concern when constructing the study to investigate SMEs that file patents; however, it also created additional limitations. One example is the grouping of the industries, which has been performed to the best of my knowledge, but will remain subjective to some extent.

With respect to the extensive process of gender (and nationality) imputation, its application is greatly dependent on the availability of vehicle insurance policies or other products that provide historic records of attributes such as names, gender, and nationality. Only very few organizations have a sizable amount that can be used to build mappings that go beyond the commonly used name-gender relationship.

In respect to the gender attribute, I only had access to a small and possibly biased data set of 41 SMEs with female-only inventor teams and 30 SMEs with mixed-inventor teams. I therefore encourage other researchers to replicate the study using a larger data set for other countries. Eventually, further studies could also perform a correlation analysis of the time of patent application and the commercial success of a product or service related to the application that would provide a more causal explanation for the effect.

V) Growth Prediction with Tourism Data

V.1 Introduction

One of the goals of management research is to provide ideas, tools, and processes for current and future business leaders to make evidence-based decisions (Torres et al., 2015). For tourism business managers, the identification of emerging determinants of business performance remains therefore one of the most critical activities for those concerned with the planning and optimization of their organizations. Permanent change, caused by technological turbulence in the external environment, existing market competition, new challengers emerging from the sharing economy (Cheng, 2016), niche competitors such as pop-up restaurants and food trucks (Schwarzkopf, 2016), and more discerning customers make it difficult for businesses to constantly reappraise their performance measurements and the effectiveness of their competitive strategies (Phillips et al., 2015). Among the existing academic studies of novel determinants of tourism business performance is a study by Ye et al. (2009) in which a mathematical model was developed to explain the impact of user-generated comments, known as electronic word-of-mouth (eWOM), on hospitality sales and profitability. In addition to hospitality, the topic of eWOM as a determinant of business performance has garnered significant practitioners (Blal and Sturman, 2014) and academic attention in gastronomy (Torres et al., 2015) and a few industries (Blal and Sturman, 2014; Gretzel et al., 2007; Litvin et al., 2008; Melián-González and Bulchand-Gidumal, 2016; Ye et al., 2011). These investigations are spurred by the increasingly notable role of the eWOM in consumers' lives and its economic role in the tourism industry.

The importance of the eWOM model and its relationship with business performance has been shown by Litvin et al. (2008), who stated that 60% of tourists considered the Internet an effective tool for restaurant selection before their current vacation. Given that travelers, in addition to using information available through search engines, are relying heavily on eWOM webpages to evaluate and buy hospitality and gastronomy services (Park and Nicolau, 2015; Zhang et al., 2010), eWOM has inevitably changed the process of information gathering, the

accessibility of recent information, and subsequently the consumers' knowledge and perception of various service offerings (Litvin et al., 2008). The dominant form of eWOM is found on a few dominant websites (Anderson, 2012; Ghose et al., 2012; Litvin et al., 2008; O'Conor et al., 2008). However, more recently, new incumbents, such as Google and the business chains themselves, have also started to host reviews (Torres et al., 2015). The magnitude and wealth of information available from the various websites in both industries is reflected in a February 2018 press release by TripAdvisor. According to the company's own data, they find that user reviews and opinions grew 29% year-over-year and reached 600 million on December 31, 2017, covering (among others) approximately 1.2 million hospitality businesses and 4.6 million restaurants. The average monthly number of unique visitors of TripAdvisor grew 17% in Q4 2017 and grew to 455 million during the 2017 peak summer travel season. The average monthly number of unique hotel shoppers grew 3% in Q4 2017 and grew 7% in the full year 2017 (TripAdvisor, 2018). The historically permanent increase in online reviews structurally relates to the growth of electronic sales in many countries (Scaglione et al., 2009), including Switzerland (Schenk, 2018). The increasing economic relevance of eWOM for the tourism industry justifies the rise in interest by practitioners and academics alike and, hence, explains prior efforts to capture it systematically and to evaluate and discuss it in a scholarly manner in respect to various dimensions, including as a determinant of performance measures.

Despite the existing studies, more research is needed to demonstrate the effects of eWOM on the performance of hotels and restaurants (H&R) (Torres et al., 2015). Advancements in eWOM understanding are important since a comprehensive understanding of customer feedback can help H&R managers gain a lead in the market in terms of strategic planning, marketing, and product development (Wilkins, 2010). In addition, given the nature of the tourism industry and the efforts that companies put forth toward maintaining their competitiveness, it would be prudent to examine the potential benefits from a better understanding of the relationship between eWOM and business performance (Torres et al., 2015). Although research has started to examine financial performance measures as a result of eWOM online feedback, many of these studies use proxy data to estimate actual financial measures (Torres et al., 2015). While this can be helpful in drawing attention to the research problem, I propose that by utilizing actual annual revenues, disaggregated at the individual H&R level, this study can add value. I do this by comprehensively assessing eWOM data from

different sources as well as other publicly available data. I consider H&R in different locations, taking into account a variety of types of businesses in terms of size, quality, and service offerings. Then, I combine actual H&R performance data with restaurants and hotel profiles from Trip Advisor, Google Ranking, SwissHotel and data from Swiss Federal Statistical Office (BFS). I measure the business performance of H&Rs, expressed as positive or negative growth. In addition, I suggest a new approach to estimate the business performance of H&R with the help of eWOM data. I propose the use of the boosted logistic regression as a method for analyzing the determinants of H&Rs as well as for making predictions using data the model has not seen. Studies using prediction models that incorporate the use of open data for tourism purposes are still limited (Ba and Pavlou, 2016; Klein, 2011; Mariani et al., 2014). Thus, this research represents one of the first attempts to explore the usage of eWOM to predict the business performance of H&R in terms of revenue growth. To this end, my efforts cater to the needs of hospitality and gastronomy managers, investors, banks, insurances, merger and acquisition advisors or management consulting firms that will eventually consider the issue of critical importance and might adopt open data analysis to make better predictions about the attractiveness of a certain businesses (Pantano et al., 2017). With this study, I try to provide value to the mentioned stakeholders and answer whether I am able to predict the business performance of H&R using publicly available data. The structure of this study is as follows. First, the concept of user eWOM and its use in the gastronomy and hospitality industry are discussed. Second, eWOM and business performance studies are reviewed. Third, the data sample is introduced, followed by a presentation of the method of the study. Fourth, I show how the user-generated content can be used to predict the business performance of H&R and evaluate the prediction results. Finally, I discuss opportunities the suggested approach creates for practitioners and provide conclusions and implications.

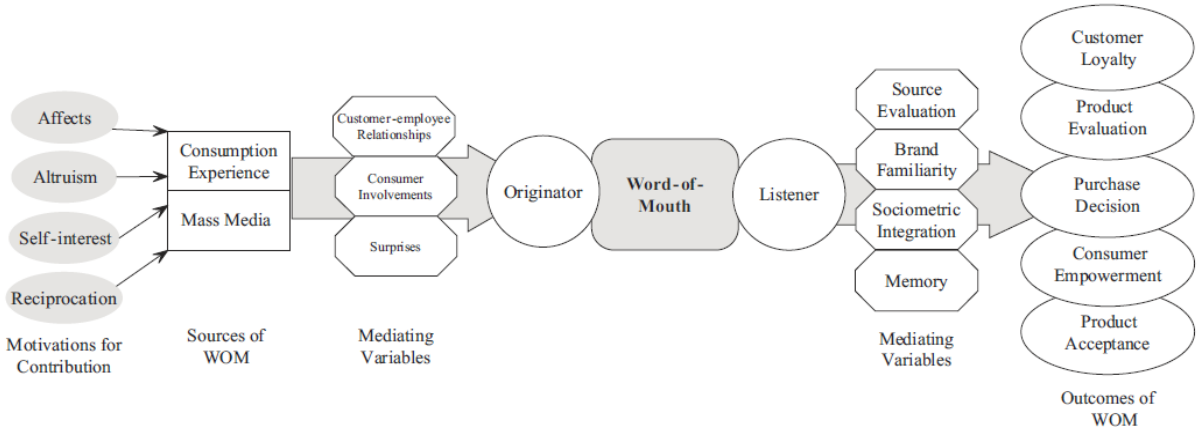
V.2 Literature Review and Hypothesis

V.2.1 Electronic Word-of-Mouth

Word-of-mouth (WOM) is defined as the process that allows consumers to share information and opinions that direct other consumers toward or away from products, brands, and services (Hawkins and Mothersbaugh, 2012). Over time, WOM definitions have evolved (Carl, 2006; Litvin et al., 2008). The essence of the construct, however, has remained stable as consumers imitate each other following a social or vicarious learning paradigm (Hawkins et al., 2014).

Other authors noted that consumer's affective elements of satisfaction, pleasure, and sadness motivate them to share experiences with each other (Dichter, 1966; Klein, 2011; Prashanth, 1997). A conceptual framework of WOM and its upstream originators and downstream listeners has been developed by Litvin et al. (2008) and is illustrated in Figure 15.

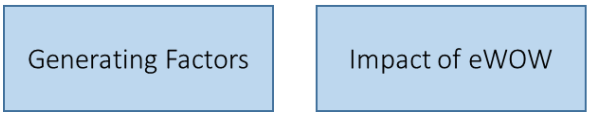
Figure 15 - The conceptual model of word-of-mouth by Litvin et al. (2008)



While every aspect of WOM illustrated has been discussed in marketing research dating back to the 1960s (Arndt, 1967; Dichter, 1966; Engel et al., 1969), this study only discusses the consequences of WOM on the purchase decision (illustrated far right) and its meaning for the present and near future. Since the emergence of the commercial Internet, consumers had have the opportunity to share their experiences by writing online reviews about their subjective level of satisfaction (Liu, 2006). The term electronic word-of-mouth (eWOM) was first mentioned in the context of "internet customer communications" by Stauss (2000) and was later defined by Hennig-Thurau et al. (2004) as "any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet." The authors further add that motivation, sources, mediating variables, and outcomes remain the same, regardless whether the consumer is in the physical and online space. It is a part of human nature that consumers rely on other individuals' experiences to make inferences about the quality of goods and services (Wirtz and Chew, 1992), regardless of the medium. Different from the ephemeral nature of traditional WOM, eWOM exists in an online "space" that can be accessed, linked, and searched. Studies have investigated relevant factors that consumers evaluate before purchase or consumption, including review valence (Duverger, 2013; von Wangenheim and Bayon, 2004; Ye et al., 2009), product rankings (Ghose et al., 2014, 2012), perceived usefulness

(Racherla and Friske, 2012), expert reviews (Qin et al., 2007; Zhang et al., 2010) trust in consumer reviews (Ayeh et al., 2011; O'Conor et al., 2008; Racherla and Friske, 2012), and management responses to consumer reviews (Senecal and Nantel, 2004a; Torres et al., 2015; Zhang et al., 2010). The rapid growth of eWOM websites has led to an enormous amount of consumer-generated online reviews (Tuominen, 2011). As consumers post their recommendations and opinions about a product on social media, they attempt to persuade other consumers to see their point of view and thus influence their decision-making (Benzing et al., 2009), which in turn impacts the company offering a product or service. As one of the first, Kirkpatrick (2005) proposed the need to manage eWOM for purposes of revenue generation. This can be grouped with other recent studies related to the action taken by companies in respect to eWOM and the impact on the company. Serra Cantallops and Salvi (2014), for example, distinguish between the review generating factors of eWOM and the impacts of eWOM, as illustrated in Figure 16.

Figure 16 - eWOM research streams: perspective 1/4

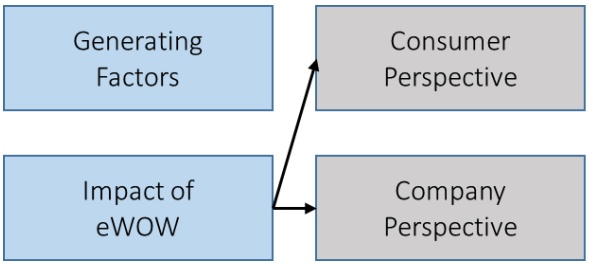


Review-generating factors eWOM Studies of eWOM that relate to the review-generating process analyze factors such as motivation, demographics of the review writer, cognitive and psychological aspects, satisfaction/dissatisfaction, group influence, sense of community belonging, and elements related to service quality. Thus, the researchers in this domain investigate if there is any identifiable set of factors that contributes to generating and publishing reviews (Craig et al., 2014; Filieri and McLeay, 2014; Rensink, 2013; Yoo and Gretzel, 2011).

Impacts of eWOM Studies that relate to impacts of eWOM analyze the effects that eWOM has on the consumer and on the company itself. Studies in the consumer perspective have identified factors related to positive or negative reviews, including gender differences, reliability, confidence, different behaviors depending on valuation ratios, content, ease of accessing the reviews, product acceptance, and media (blogs and virtual communities, e-mails, websites, product review sites) (Serra Cantallops and Salvi, 2014). Factors related to influence of purchase, decision models, repurchase intention, and loyalty, among others, have also been studied (Black and Kelley, 2009; Sparks and Browning, 2011). The impact of eWOM can be

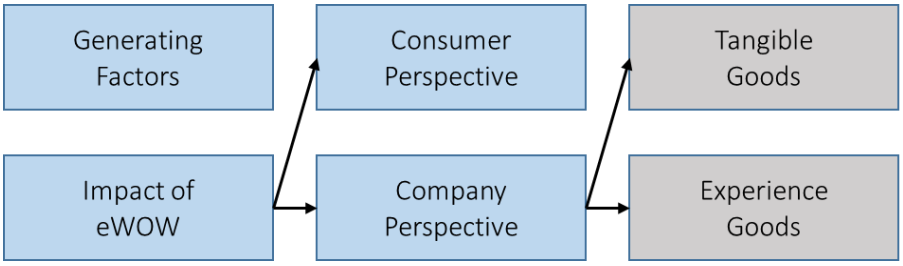
direct, affecting a potential customer, or indirect, affecting the review subject, as illustrated in Figure 17.

Figure 17 - eWOM research streams: perspective 2/4



Studies of eWOM from the company perspective identify the consequences of user-generated content on revenue, and how the companies can influence the content. Authors such as (Anderson and Lawrence, 2014; Chen and Xie, 2008; Pavlou and Dimoka, 2006) discuss the possibility of generating price premium, specific marketing strategies, and corporate reputation (Dickinger, 2010; Yacouel and Fleischer, 2012; Ye et al., 2009). In this study, I further distinguish between eWOM in the context of tangibles, mainly physical product, and experience goods (intangibles), as illustrated in Figure 18.

Figure 18 - eWOM research streams: perspective 3/4



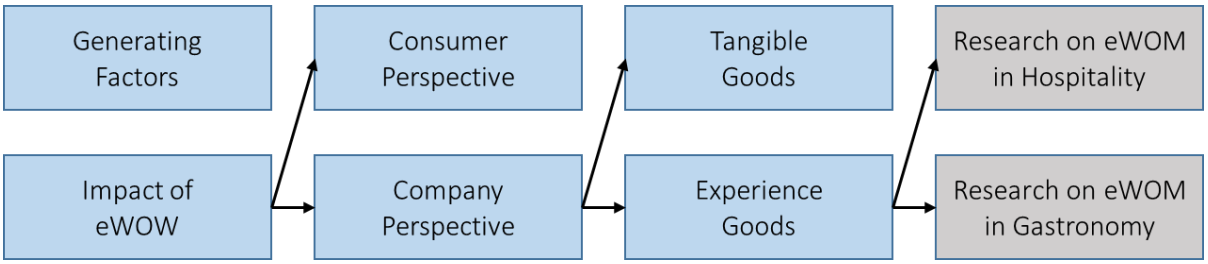
Tangible goods Motivated to find out whether the number of consumer reviews may signal the popularity of a product (Dellarocas et al., 2013), many studies provide support to the positive relationship between volume of eWOM and product sales (Chen et al., 2004; Chevalier and Mayzlin, 2006; Gu et al., 2012). They find that the more consumer reviews are written for a specific product, the more consumers tend to be aware of this product (Chevalier and Mayzlin, 2006; Dellarocas et al., 2013; Li and Sun, 2010) find that online consumer ratings significantly influence book sales and that a high average rating on Amazon.com drives book sales. Further studies had shown that online user-generated reviews could significantly influence the sales of products like books and CDs (Chevalier and Mayzlin, 2006; Ghose and

Ipeirotis, 2006; Hu et al., 2006; Zhu and Zhang, 2010). Another example is provided by authors such as (Schlieker, 2014; Ye et al., 2009), who studied the relationship of eWOM and business performance in consumer electronics. They compared the impact of eWOM sites such as Epinions, DpReview, and Cnet with Amazon and found that despite its popularity, Amazon eWOM had the lowest impact on business performance, measured by eWOM providing consumers' influence on other consumers (Gu et al., 2012). Another area of interest is the valence or customer rating level (Cheung and Thadani, 2012; Duan et al., 2008a), which describes the heterogeneity of rating scores given by consumers. Authors such as (Blal and Sturman, 2014; Cheung and Thadani, 2012; Duan et al., 2008a) find that a consistency in rating value provided by consumers has a positive relationship with business performance. Focusing on the content of reviews, authors of (Xie et al., 2014) examined the effects of eWOM communication on product judgments, showing that negative eWOM tends to be more diagnostic or informative than positive or neutral eWOM. They further find that negative attributes strongly imply products of low quality, whereas positive or neutral eWOM is ambiguous and associated with products of mixed quality.

Experience goods Comparing tangible and intangibles, the work of Cui et al. (2012) indicates that the volume of reviews has a greater impact on sales of experience products (intangibles) than on the sales of tangible products. This is confirmed by Senecal and Nantel (2004b), who added that eWOM would be more influential to purchasers of experiential products compared to tangible products. Studies on the influence of eWOM find the difference in effects is especially strong comparing products with experience goods as the quality of intangibles is often unknown before consumption (Klein, 1998). In a study investigating the strength of WOM influence in service provider switching von Wangenheim and Bayon (2004) find that subject expertise and WOM author similarity with the receiver have an effect on the perceived influence and consumer decision making. In another study investigating experience goods Duan et al. (2008b) find in the context of eWOM that movie sales are significantly influenced by the volume of online postings. In a study, also investigating WOM and movies, Liu (2006) finds that as variety in ratings increases, consumers dissatisfied with the movie have a stronger incentive to distribute negative WOM, whereas consumers satisfied with their experience have a stronger incentive to distribute positive feedback, affecting future customers. The impact of changes in ratings has been shown to have an impact on movie sales in several regional contexts (Dellarocas et al., 2013). Many other examples of experience

goods can be found in hospitality and gastronomy; intangible offerings can be found in nature. In general, tourism-related services cannot be evaluated before the consumer experiences the service and are seen as high-risk purchases. Therefore, an evaluation prior to consumers' own consumption is an important aspect in the decision-making process (Lewis and Chambers, 2000). Hospitality and gastronomy perfectly match the nature of the experience goods, and tourism business performance is expected to be influenced by eWOM (Blal and Sturman, 2014). Going forward, I narrow down the research focus to these two industries.

Figure 19 - eWOM research streams: perspective 4/4



V.2.2 Studies of eWOM in Hospitality and Gastronomy

Xie et al. (2014) examined the effect of consumer reviews on hotel performance. They were able to investigate their research questions by collecting a time series of consumer reviews and management responses of 843 hotels in the United States on TripAdvisor.com on a daily basis. The authors matched the data with the business performance of these hotels. They found that management responses moderate the relationship between consumer reviews and hotel performance.

The study's findings suggest that managers need to effectively manage online ratings to affect the business performance of hotels and that executives can use ratings as an effective measurement metric for an internal and external analyses of their operations. A study investigating the content of eWOM and positive ratings found that the most valuable hotel characteristics that are prominently displayed on eWOM websites significantly affect a traveler's decision-making. These characteristics include location, price, facilities, and cleanliness (Lockyer, 2005). Other characteristics, such as the room size and type of building, quality of service, and a quiet environment, are important to certain demographics only (Merlo and de Souza João, 2011). Some studies (Sohrabi et al., 2012) present another list of important hotel features such as promenades, comfort, security, network, pleasure, news,

recreational information, expenditure, room facilities, and the convenience of parking. A study by (Ye et al., 2011) analyzed the impact of online reviews on hotel bookings and found the variance or polarity of eWOM for the reviews of a hotel had a negative impact on online sales. Previous studies (Lu et al., 2012) suggest that a higher customer rating significantly increases the online sales of hotels (Xie et al., 2014). Research by other authors demonstrated a positive relationship between consumer ratings and the lodging establishment's average room price (Öğüt and Onur Taş, 2012). A study (Torres et al., 2015) examined the relationship between the hotels' average revenue per booking and publicly available TripAdvisor data. The authors found that there is a significant relationship between a hotel's overall rating (i.e., 1, 2, 3 or 4 stars) on TripAdvisor and the hotel's average revenues from online transactions. Other research (Ye et al., 2011) finds that positive reviews can significantly increase the number of bookings, and a further study (Kim et al., 2017) reports that customer ratings have a strong positive effect on the customers' willingness to book a hotel online. A study (Ye et al., 2009) in which the authors developed a mathematical model to explain the impact of eWOM content on hotel business performance was expressed in sales numbers and profitability. The authors found that a 10% improvement in rating scores leads to a 4.4% increase in revenue. Similarly, the travel company Expedia shared that they observed that a 1-point increase in a review score (on a 1 to 5 scale) equates to a 9% increase in the average daily rate, calculated as the lodging revenue divided by the number of sold rooms (Lu et al., 2012). Additionally, a research study (Blal and Sturman, 2014) found a significant positive impact of ratings on the revenues per available room. Blal and Sturman (2014) confirmed the effect of increased sales using a sample of hotels in London, and another study (Torres et al., 2015) confirmed the effect existed by using a cross section of hotels in the United States. For China, the authors of a research study (Tuominen, 2011) made an initial attempt to investigate the impact of eWOM on hotel bookings, using data collected from a major travel website.

While much smaller in size, a nascent stream of literature exists to explore the role of eWOM online feedback on business performance in gastronomy. In a study by Klein (2011), this relationship was empirically tested for restaurants, and it was confirmed that there is a positive relationship between good reviews at a third-party site and traffic to a restaurant proprietary website. Furthermore, authors (Blal and Sturman, 2014; Klein, 2011) find that good reviews can lead to greater website traffic for restaurants, indirectly impacting revenue changes.

Considering the analysis performed on eWOM related to the gastronomy and hospitality industry, I and others observe that there are ample opportunities for future research to extend the level of knowledge with respect to the impact of eWOM from a companies' perspective (Serra Cantallops and Salvi, 2014). Despite the availability of open (free) data and its limited usage by hospitality and gastronomy managers (Pantano et al., 2017), empirical research investigating the economic value of consumer reviews to hotel businesses still lags in the literature (Duverger, 2013). In 2011, Tuominen claimed that the impact of online consumer-generated reviews on the performance of hospitality businesses has been overlooked by researchers (Tuominen, 2011) and called for more research into the relationship between online reviews and business performance (Serra Cantallops and Salvi, 2014; Ye et al., 2011). A recent study by Melián-González and Bulchand-Gidumal (2016) further proposes that the research field needs further evidence between the relation between number of reviews, review score, and the specific monetary impact of reviews. This is partially due to the lack of actual financial data used in most studies (Öğüt and Onur Taş, 2012). Many have used the number of reviews as a proxy for sales (Ye et al., 2011), and none, to the best of my knowledge, had actual annual revenues of businesses. In a study (Serra Cantallops and Salvi, 2014), the authors ask whether eWOM generates different impacts on hotels dependent on the location.

In this context, this study provides answers to some of the questions raised by authors requesting more research on eWOM's impact on the company level. I chose to investigate this relationship by using the example of the hospitality and gastronomy industry. I explore the availability of data sources in the tourism context and compare them. I then hypothesize that:

H4.1: Revenue growth of hotels can be predicted with the help of eWOW data.
H4.2: Revenue growth of restaurants can be predicted with the help of eWOM data.

Evaluating these hypotheses, this study shows the predictive power of eWOM on two different data sets. I include business attributes, which are provided by the business owners, and attributes derived from guest reviews. Joining the findings from H4.1 and H4.2 answers to research questions four.

Methodologically, the study contributes to the existing tourism literature by suggesting a new approach to predict the business performance of restaurants and hotels with

the help of eWOM. I collect granular eWOM data on the company level and calculate several features. I test the directly collected and calculated variables for their predictive power. As I build a multiyear data set, I also test attributes to determine whether they are early indicators.

Finally, I expect that this study's results will provide information on the relevance and importance of online reviews for organizations' performance and, as such, can be used in the strategic management of H&R businesses. If I succeed, the findings will make a meaningful theoretical and practical contribution to the Swiss tourism sector, which eventually can help business executives to enhance the level of economic and social benefits (Phillips et al., 2015).

V.3 Research Design

V.3.1 Data Set and Context of Study

Tourism activities of domestic and international tourists contribute notably to Switzerland's economic activities. The sector represents 2.6% of Switzerland's total gross valued added and employs 5% of all working people in the country. In absolute numbers, the sector provided a gross value added of CHF 16 billion in 2015, of which 4 billion (25%) came from accommodation and 2.1 (13%) from gastronomy (GastroSuisse Verband für Hotellerie und Restauration, 2018, 2017; Schweizer Tourismusverband, 2016; Weber, 2007). The country is an example of a mature Western tourist destination, a pioneer in winter sports tourism and includes a worldwide recognized offering of hospitality and gastronomy businesses. The economic crises of 2008, as well as the Euro-Crisis in 2011, followed by changes in exchange rates between the Swiss Franc and several other currencies, had a strong impact on the business performance of many H&Rs. Aggregated for all H&R in Switzerland, Table 34 shows the annual changes in total gross value added between 2007 and 2016.

Table 34 - Annual changes in gastronomy hospitality revenues in Switzerland

Revenue changes	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Change to previous year Restaurants	6.7%	8.5%	-5.2%	-1.1%	-1.1%	-3.5%	1.5%	-0.7%	-3.2%	0.2%
Change to previous year Hotels	13.5%	6.2%	-6.3%	0.5%	-0.9%	-1.3%	1.5%	0.6%	-2.0%	0.5%

Source: Schwarzkopf and Gujan (2017)

Gastronomy in Switzerland According to GastroSuisse, there are 27 000 restaurants operating in Switzerland employing 240 000 people, of which approximately 38 600 work full time. Half of all revenues come from domestic and international tourists. The relevance of gastronomy differs by regions. Only in some alpine touristic areas, gastronomy plays a major part for the local economy (Weber, 2007). Based on the numbers provided by the Swiss Tourism Federation (2010), the sample represents 18% of all restaurants.

Hospitality in Switzerland There are 5 055 hotels operating in Switzerland employing 78 000 people. All hotels in Switzerland offer tourists 273 510 beds in 141 019 rooms. The relevance of hospitality also differs by regions and is characterized by seasonality. Based on the numbers provided by the Swiss Tourism Federation (2010), the sample represents 24% of all hotels in Switzerland.

With the economic crises in Europe, followed by turbulences in the exchange rates and the reaction of the Swiss National Bank, the Swiss tourism sector saw a decline in demand, and room nights decreased to below 35 million in 2012 (GastroSuisse Verband für Hotellerie und Restauration, 2017). This downturn was more pronounced in the alpine leisure resorts compared to the dynamic markets in the urban centers. Hotelleriesuisse, the Swiss hotel association (known as boards), calculates for each year the revenue per available room. I collected their annual reports and extracted various values to evaluate whether the data sample of hotels is representative. The reports, summarized in Table 35, show the time series of the hotels grouped by their star classification. Negative changes in the weighted average revenue per available room for all star-rated hotels occurred, according to HESTA statistic of the Swiss Federal Institute, in the years 2009, 2011, 2012, and 2015 (Federal Statistical Office, 2017). The weighted average revenue per available rooms today has not reached the peak of 2007, as illustrated in Table 35.

Table 35 - Average revenue per room night available for hotels in Switzerland

Revenue per Room in CHF	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Members, misc	49	57	51	54	53	54	55	47	67	53
Non-members	31	32	31	31	32	34	37	38	44	40
1-star	58	61	55	62	79	77	77	75	69	78
2-stars	54	61	58	58	59	58	65	68	68	65
3-stars	72	79	76	78	76	74	74	76	80	73

4-stars	121	131	119	119	117	113	118	116	125	112
5-stars	281	300	262	261	252	249	234	231	255	227
Weighted average all	83	88	82	82	81	80	81	82	80	81
Average of star-rated hotels	109	118	107	109	107	105	105	108	105	105
Average of non-star-rated hotels	31	32	31	31	32	34	37	38	44	40

Data collection and overview To obtain and analyze the data needed for this study, I teamed up with a domestic insurance company. The insurance company sells (among other products) general liability insurance for businesses, including H&R, all over Switzerland. The company provided the authors with a data set that reflects the policy information of businesses from 2010-2016. When taking out general liability insurance, companies need to reveal their annual revenues to the insurance companies in order to determine the price of the premium. The data set is limited to H&R and includes information such as company name, unique id, industry type, incorporation year, and annual revenue changes, which will serve as the ground truth. The data set is very diverse in size, group, ownership, and location. The hospitality and gastronomy establishments represent a cross-section of different cities and the primary focus group (i.e., business, leisure, convention), which I will show in further detail.

V.3.2 Covariates for Prediction

Hotel and restaurant annual revenue data Exploring the data set (in) provided by the insurance company, I was able to extract 667 unique entries for hotels and 2 556 restaurants from the insurance company database, containing name, address, owner, and revenue of the business for at least two consecutive years between 2013 and 2016. A total of 105 hotels and 559 restaurants reported a relevant revenue change (as defined by the terms of service of the insurance) in at least one of the years. I manually looked up these businesses on TripAdvisor.com and google.com. In addition, I also looked up the hotels on swisshoteldata.ch. For the businesses I found on these websites, I noted down their URL. I was not able to find many of the businesses, which reduces the final data set. The counts of businesses for which I found more attributes are summarized in Table 36.

Table 36 - Description of additional data

Attribute	Description	Number of Hotels	Number of Restaurants
in_street	Street name	667	2 556
in_Sn	Street number	535	2 552
in_zip	Zip code of venue	667	2 540
in_city	City name	667	2 556
	of which reported a relevant revenue change in period	105	559
sh_url	URL to businesses' TripAdvisor listing	88	Not applicable
ta_url	URL to businesses' TripAdvisor listing	78	520
go_url	URL to businesses' Google Places listing	77	427

Merging all the data sources together, I had a data overlap for 76 unique hotels and 407 restaurants, having

- a listing for all applicable properties;
- two or more different revenue entries for the years between 2014 and 2016, enabling to calculate growth; and
- at least 5 reviews in the year proceeding the revenue change period.

I will provide details of the attributes that were collected and are available for the H&R, grouped by data source.

Amenity description for hotels Swisshotel is a database managed by hotelleriesuisse – the Swiss Hotel Association. According to the self-description, all hotels in Switzerland are listed on this platform with a portrait. It displays many amenities for hotels, which the hotel managers provided at the time they registered a business. In addition to a basic overview, I find information about the hotel infrastructure, management, local infrastructure in a 10 km radius, hotel classification, group/chain information, accepted payment methods, and many more attributes. I found all hotels in the merged data set in the database and collected all available attributes. The collected swisshotel data set (sh) has 254 different attributes. Each hotel has only a few attributes; hence, the sh data set is sparse. Table 37 summarizes the collected attributes that were available for the hotels.

Table 37 - Description of swisshotel data

Attribute	Description
sh_swisshotel	URL to the website of the hotel
sh_street	Street name and number
sh_zip	Zip code of the city
sh_city	Name of the city
sh_rooms	Number of rooms
sh_name	Name of the hotel
sh_stars	Number of stars and category (Superior/Garni)
sh_check_out	Check out time
sh_meeting_room	Minimum and maximum size of meeting room
sh_trust_you	URL of the associated trust me website
sh_managers	Name(s) of the managers
sh_banquet_room	Minimum and maximum size of the banquet room
sh_check_in	Check in time
sh_beds	Number of beds in hotel
sh_telephone	Telephone number of the hotel
sh_google_name	Name of the hotel on Google
sh_google_ratingvalue	Rating value on Google
sh_google_reviewcount	Number of reviews on Google
sh_x_coordinate	Longitudinal GPS coordinate
sh_y_coordinate	Latitudinal GPS coordinate
sh_max_meeting_room_size	Maximum size of the meeting room
sh_max_banquet_room_size	Maximum size of the banquet room
sh_nb_stars	Numbers of stars as an integer
sh_nb_managers	Numbers of managers
sh_[feature_name]	25 out of 254 possible attributes (payment method, Wi-Fi, tv, pets welcome, spa, elevator, in-house bar or –(vegetarian) restaurant, nearby public transport, hiking, golf, tennis, …)
sh_NrFeatures	Total number of the 254 features associated with a hotel (254 choices, multiple features possible)

The swisshotel data set reflects the knowledge of the website owners about the listed hotels at the time of the data collection in January 2018. I assume that the attributes provided are all static and have not changed in the past. To mitigate the curse of dimensionality when

predicting the revenue growth of hotels, I limited the individual features to 25 instead of 254. I selected the features that, in my view, best segment the individual hotels and showed a low correlation among each other. Some examples of high correlation features were the individual payment options, which almost always occurred simultaneously: Eurocard, Visa, Mastercard, and American Express.

Economic data The Swiss Federal Statistical Office (BFS) collects several data points each year in respect to tourism. It includes data from the tourist accommodation statistics (HESTA) The data is publicly available and downloadable in a structured data format. Table 38 describes the collected attributes that I extracted.

Table 38 - Description of Swiss Federal Office of Statistics data

Attribute	Description
ed_zip	Zip code of the commune
ed_arrivals_per commune_[2012-2016]	Number of annual arrivals (from visitors, guests, …) in commune (aggregated, grouped by year)
ed_room_stays_[2012-2016]	Number of days visitors spend in commune (aggregated, grouped by year)
ed_rooms_per_commune_[2012-2016]	Number of available rooms in commune (aggregated, grouped by year)
ed_hotels_per_commune_[2012-2016]	Number of hospitality businesses operating in commune (grouped by year)
ed_rooms_per_commune_[2012-2016]	Accumulated number of hotel rooms of operating businesses in commune (grouped by year)
ed_average_hotelrooms_[2012-2016]	Calculated average rooms per hotels in commune (grouped by year)
ed_occupancy_per commune_[2012-2016]	Average annual occupancy of hospitality businesses in commune (grouped by year)
ed_gva_per_annum_[2012-2016]	Gross value added by hotels and restaurants, measured in CHF (grouped by year)
ed_tourismReg_[class]	Classification (Factor group 1 to 9) of regional areas in Switzerland

The economic tourism data set (ed) is time dependent and publicly available from 2001 on. Each attribute refers to a specific year. I collected the attributes for the years 2012-2016. Some of the attributes provided change annually, which allows to calculate relative (change) and absolute changes (delta change). When merging this data set with the insurance data,

including the annual revenue, I merged only information that refers to a time period earlier than the revenue year.

eWOM To evaluate different user content as well as business descriptions, I choose to collect data from popular review and business listing pages on which I was able to find the business entries of the provided companies.

TripAdvisor TripAdvisor is an international travel and restaurant website company providing H&R reviews and other tourism-related content. The TripAdvisor data are included next to the identification properties of each business in addition to the rating values, which go from 1 to 5 in 0.5 increments. The rating values are rounded up or down from the actual value calculated from the reviews. Due to the availability of historic data, I can calculate the number of reviews and ratings at the end of year. The historic data of TripAdvisor reaches back to 2008. However, due to the sparsity of the data set, I exclude the years 2008 to 2011. I further collected 21 general attributes per hotel and 24 per restaurants. Table 39 describes the collected attributes if available.

Table 39 - Description of TripAdvisor data

Attribute	Description
ta_url	URL for the TripAdvisor page of the business
ta_reviewcount_[2012-2016]	Counted number of reviews on Jan. 1st in the years 2012 to 2016 (grouped by year)
ta_reviewcount_[1-5]_[2012-2016]	Counted number of reviews in the years 2012 to 2016 with a value 1 to 5 (grouped by year and rating)
ta_reviewcount_var_[2012-2016]	Calculated variance of rating distribution at time of data collection (grouped by year)
ta_reviewcount_csum_[1-5]_[2012-2016]	Calculated accumulated reviews (grouped by year and rating)
ta_ratingvalue_[2012-2016]	Calculated rating value on Jan. 1st in the year 2012 to 2016 (grouped by year)
ta_stars	Star value of hotel
ta_rooms	Number of rooms at the hotel
ta_openTime	Opening hours of business, calculated hours of operations (in h)
ta_classification_[classification]	Hotel, Pension, Hotel Garni, other (single choice)
ta_language_[language]	Language spoken at business (28 choices)

ta_NrLanguages	Total number of languages spoken at business (multiple languages possible)
ta_cuisine_[cuisine]	Cuisine of restaurants (113 choices, multiple cuisines possible)
ta_NrCuisines	Total number of cuisines offered at restaurant
ta_meals_[meal]	6 meal types offered at restaurant (breakfast, brunch, lunch, dinner, dessert, buffet)
ta_NrMeals	Total number of meals offered at restaurant (6 choices, multiple meals possible)
ta_goodFor_[preference]	9 different possible attributes (business meeting, romantic, pet lovers, …)
ta_NrGoodFor	Total number of “good for” attributes (multiple features possible)
ta_price	Interval of average min. and average max. price provided, calculated as (Min+Max)/2
ta_feature_[feature]	26 different possible attributes (Payment options, serviced restaurant, served alcohol, Wi-Fi, wheelchair accessible, …)
ta_NrFeatures	Total number of features (26 choices, count of features)
ta_url	URL for the TripAdvisor page of the business
ta_reviewcount_[2012-2016]	Counted number of reviews on Jan. 1st in the years 2012 to 2016 (grouped by year)
ta_reviewcount_[1-5]_[2012-2016]	Counted number of reviews in the years 2012 to 2016 with a value 1 to 5 (grouped by year and rating)
ta_reviewcount_var_[2012-2016]	Calculated variance of rating distribution at time of data collection (grouped by year)
ta_reviewcount_csum_[1-5]_[2012-2016]	Calculated accumulated reviews (grouped by year and rating)
ta_ratingvalue_[2012-2016]	Calculated rating value on Jan. 1st in the year 2012 to 2016 (grouped by year)
ta_stars	Star value of hotel
ta_rooms	Number of rooms at the hotel
ta_openTime	Opening hours of business, calculated hours of operations (in h)
ta_classification_[classification]	Hotel, Pension, Hotel Garni, other (single choice)
ta_language_[language]	Language spoken at business (28 choices)
ta_NrLanguages	Total number of languages spoken at business (multiple languages possible)

The TripAdvisor.com data set (ta) is partially time dependent. Each attribute, which I indicated with a time stamp [year], refers to a specific year (2012-2016). Hence, these attributes are all dynamic and change annually, which allow to calculate relative (rel) and absolute previous year (1prev) and two and three-previous-year changes (2prev, 3prev). Derived from the previous and two-previous-year changes, I also calculated variances and changes in variances. Further, I calculated the share of reviews grouped by rating and previous year changes of those shares. In addition, I calculated the ratio of the shares (i.e., the share of annual excellent ratings compared to the share of poor ratings and changes thereof). When merging this data set with the insurance data, including the annual revenue, I merged only information that refers to a time period earlier than the revenue year. Other TripAdvisor attributes, which I assume to be static, include, for example, the number of rooms of a hotel or the local ranking of a place among similar hotels or restaurants.

Google.com Integrated into Google Search and Maps, Google allows people to review businesses such as H&R. Table 40 summarized the collected attributes.

Table 40 Description of Google Reviews data

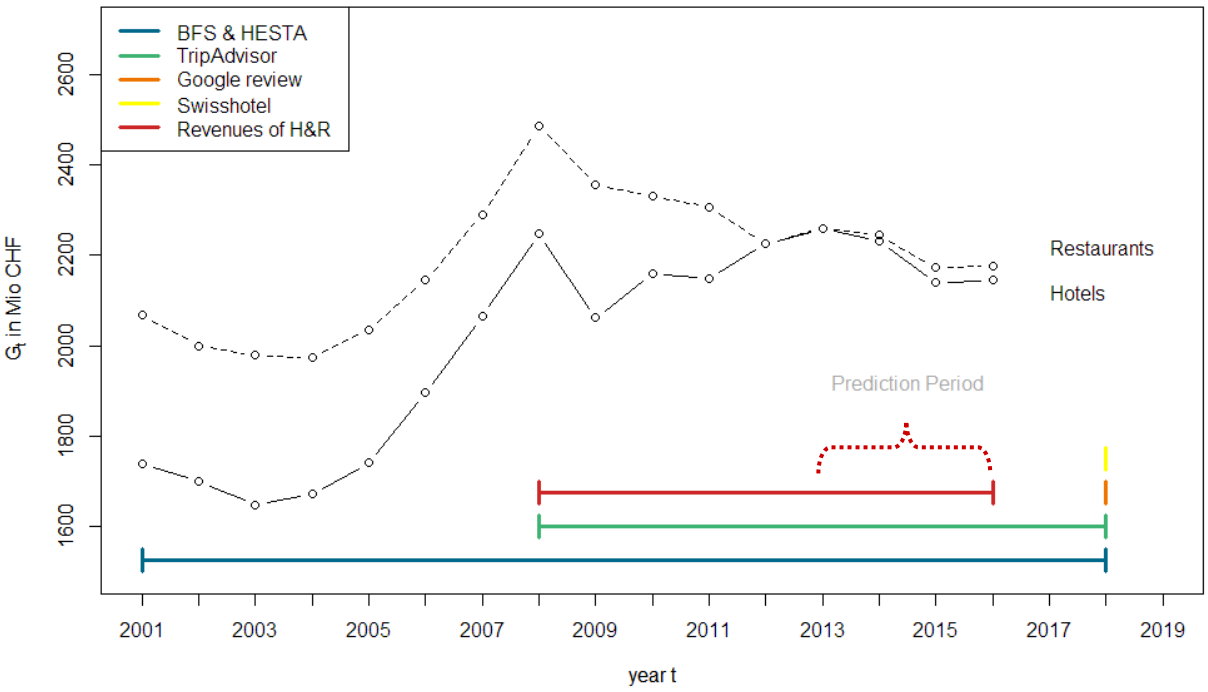
Attribute	Description
go_url	URL for the Google place page of the business
go_name	The name of the hotel
go_street	Street name and number
go_postalcode	Zip code
go_city	Name of the city
go_type	418 different business types
go_x_coordinate	Latitudinal GPS coordinate
go_y_coordinate	Longitudinal GPS coordinate
go_url	URL for the Google place page of the business
go_name	The name of the hotel
go_street	Street name and number

The Google.com data set (go) is time independent. I collected a snapshot of the information available at the time of the data collection in January 2018. I assume that the attributes provided are all static.

In respect to the rigor of the predictive modeling, I include attributes that I assume to remain stable over time, even if they were derived from a data source that was collected chronologically at a later date with respect to the prediction horizon. Such attributes include, for example, the location of a business, the number of rooms in a hotel, or the availability of an elevator. However, for the derived attributes, I store only those that relate to an earlier date than the temporal ground truth (annual revenue of H&R from the Insurance data). This means, for example, that TripAdvisor rating changes of H&R after 2016 are not considered for the predictive model, as I may not predict 2015 to 2016 revenue growth with the help of rating changes that occurred between 2016 and 2017.

For better illustration of the data sources and the time period of the available data, I plotted the available sources in Figure 20. Additionally, I show the gross value added (GVA) for H&Rs, which helps to put the study's findings in perspective.

Figure 20 - Data sources overview and gross value added Hotels and Restaurants



As Figure 20 illustrates, part of the TripAdvisor (green), the Google (orange) and the swisshotel attributes are a snapshot of the current listing at the time of the writing of this section in January 2018. The other data sources show periodicity. Periodicity allows me to take changes of, for example, ratings and ranking or changes of business occupancy over time

into account, especially when training the growth models. In respect to the rigor of the predictive modeling, I include attributes that I assume to remain stable over time, even if they were derived from a data source that was collected, chronologically at a later date with respect to the prediction horizon. Such attributes include, for example, the location of a business, the number of rooms in a hotel, or the availability of an elevator. However, for the derived attributes, I only store those that relate to an earlier date than the temporal ground truth (annual revenue of H&R from the Insurance data). This means, for example, that TripAdvisor rating changes of H&R after 2016 are not considered for the predictive model, as I may not predict 2016 revenue growth with the help of rating changes that occurred between 2016 and 2017.

V.3.3 Methodology

Predictive model and data split It is this study's goal to accurately predict positive or negative changes in the revenues of H&R companies using a subset of the attributes I collected. I evaluated different methods used in similar contexts (Tsai and Lu, 2010) and selected the Boosted Logistic Regression as the appropriate statistical and predictive methodology for this study (also compare III.3.3 – Prediction model selection and Fernández-Delgado et al., 2014). I try to predict whether the business had positive or negative changes in revenues from the years 2014 to 2015 and 2015 to 2016. Businesses with positive changes in any particular year (growing H&R) were labeled with the class "1," and those with negative changes (shrinking H&R) in any particular year were labeled with the class "-1." To train the Boosted Logistic Classification model, I selected businesses that had changes in revenues in the years 2014 to 2015 or from 2015 to 2016. I split the merged data set for H&R into a 60% (train) and 40% (test) subset.

V.4 Results

To determine the relationship among the variables of interest, I use boosted logistic regression. I choose the sign of the annual changes in H&R revenues as the independent variable (growing or shrinking). The objective was to try several parameters and topologies in order to obtain the best possible classification result. The features identified as most important will be of interest to academics and practitioners, as they can be considered unique predictors or determinants of H&R performance. Moreover, I also show the feature

importance. The models for the growth prediction each use the same hyperparameters except for the number of features and data sources. I cross validate the model 10 times and evaluate the model's performance on the test sets of 52 hotel and 235 restaurant observations, which had not been previously tested in the model.

V.4.1 Hotel Revenue Growth

To evaluate the performance of the classification models, the average prediction accuracy and the no-information rate, which enable the interpretation of the models' performance given unbalanced classes, are usually examined in related literature (Tsai and Lu, 2010). Accuracy is defined as the percentage of records that are correctly predicted by the model (Kuzey et al., 2014). Table 40 shows the confusion matrix for the test set I used for obtaining the performance measures.

Table 40 - Study results hotel growth prediction

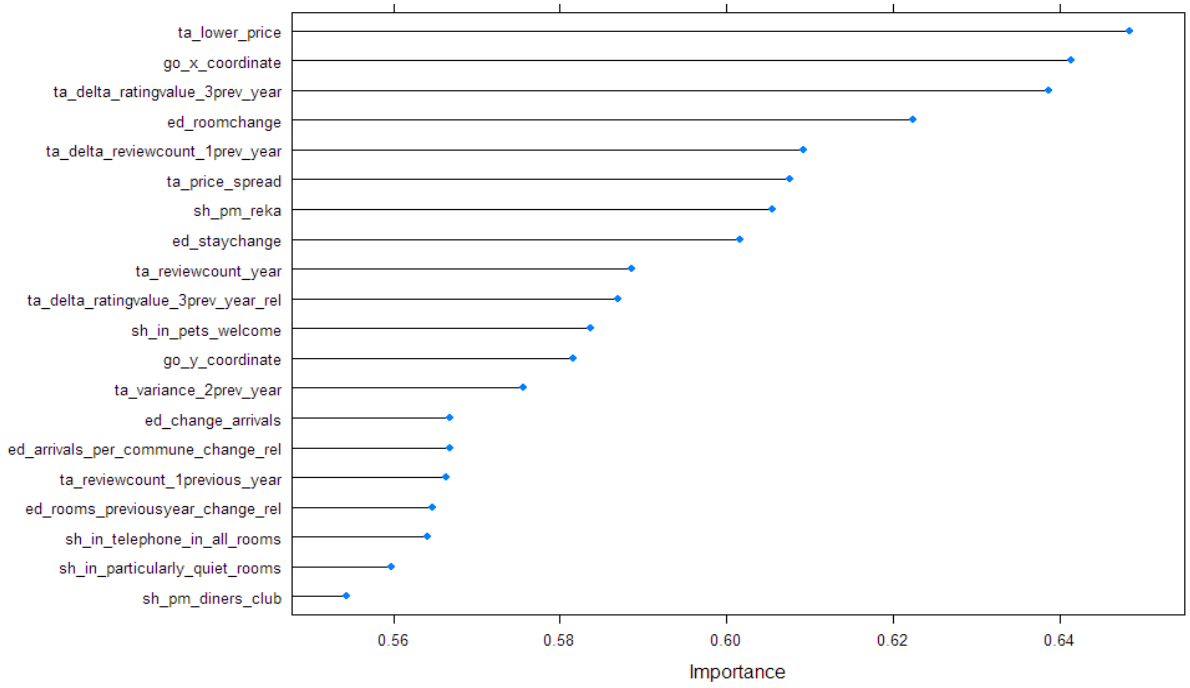
Actual/Predicted	**Growing**	**Shrinking**	**Total**	**Recall**	
Growing SMEs	**19**	8	27	70.37%	Pos. Pred. Value
Shrinking SMEs	5	**12**	17	70.59%	Neg. Pred. Value
Total	24	20			
Precision	79.17%	60.00%			
	Sensitivity	Specificity			
Accuracy	70.45%				
No Information Rate	54.55%				
P-Value [Acc > NIR]	0.02312 *		*** P ≤ 0.001	** P ≤ 0.01	* P ≤ 0.05
Balanced Accuracy	69.58%				

I computationally optimized the hotel model by having the machine choose the parameters that produced the best accuracy for the models. The final values used for the model were nIter = 11. The two classes predicted were 'growing' and 'shrinking', which were resampled using a 10-fold cross validation with sample sizes of (60, 59, 60, 60, 59, 59,...) resampled across the tuning parameters.

The model achieved an accuracy of 70.5% (balanced accuracy of 69.6%). The algorithm identified 31 out 44 hotel observations in the test set correctly. The 95% accuracy confidence

interval lies at 0.5480 to 0.8324. Compared with the no-information rate (NIR) of 0.5455, the classification performance is significant (P-Value [Acc > NIR] < 0.05) and supports H4.1. In addition, for the classification, I show the model feature importance.

Figure 21 - Hotel growth prediction feature importance



I find that the average low-end room price (ta_lower_price) and the latitude (go_x_coordinate), which separates hotels in the north from hotels in the south of Switzerland, to be the most important variables (see Appendix 35 for GPS plot). Further, I find the change in rating values (ta_delta_ratingvalue_3prev_year) compared to ratings three years ago, as well as the one-year change (ta_delta_reviewcount_1prev_year) in the number of reviews, to be important. Additionally, I found the change in the supply of hotel rooms (ed_roomchange) in a certain commune, expressed as the change in rooms, to be among the most important variables for the model. The feature price spread (ta_price_spread), which expresses the ratio between the average low and average high price, was also found to be important. In addition, some of the attributes from the swisshotel database, such as the Swiss payment option “Reka-Check” (sh_pm_reka) or “Pets welcome” (sh_in_pets_welcome), were found to be relevant for the model. Additionally, some of the economic data attributes (i.e., the change in the average length of trip compared to the previous year (ed_staychange) and

the changes in visitors that arrive compared to arrivals for the previous year (ed_change_arrivals) were found to be relevant.

V.4.2 Restaurant Revenue Growth

Table 41 shows a confusion matrix used for obtaining the performance measures stated below for the restaurant-growth classification problem.

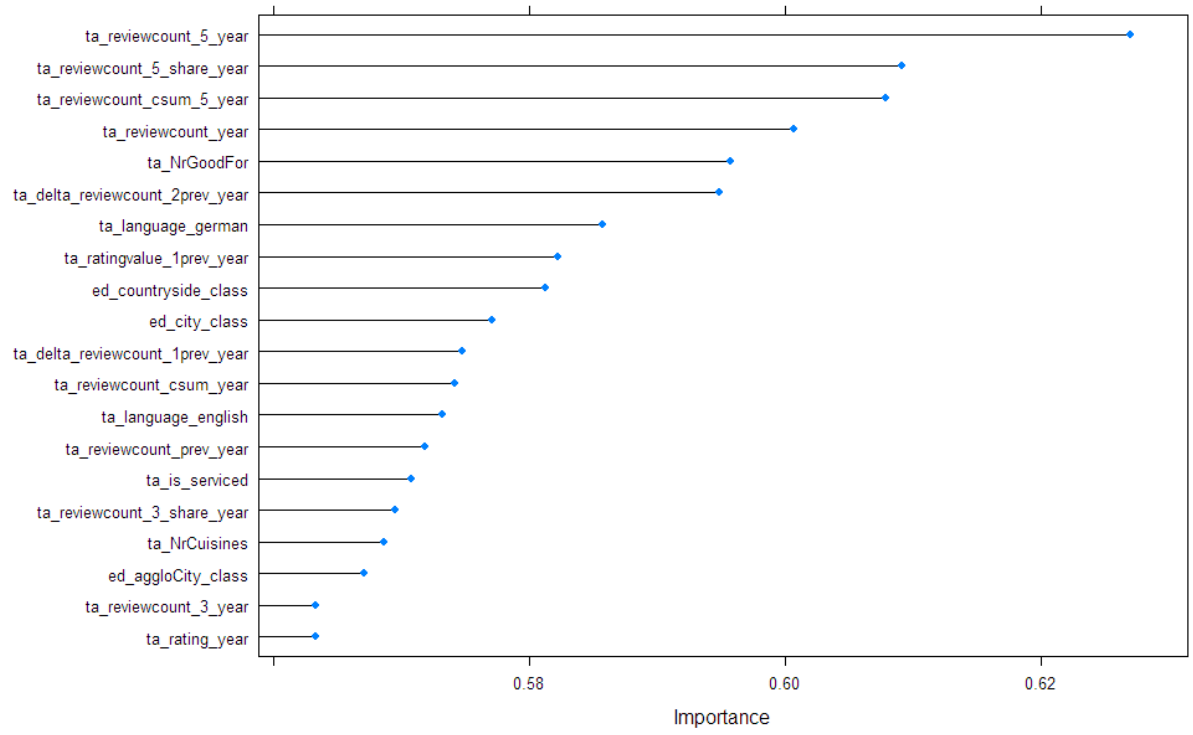
Table 41 - Study results restaurant growth prediction

Actual/Predicted	Growing	Shrinking	Total	Recall	
Growing SMEs	**81**	42	123	65.85%	Pos. Pred. Value
Shrinking SMEs	46	**64**	110	58.18%	Neg. Pred. Value
Total	127	106			
Precision	63.78%	60.38%			
	Sensitivity	Specificity			
Accuracy	62.23%				
No Information Rate	54.51%				
P-Value [Acc > NIR]	0.01032 *		*** P ≤ 0.001 ** P ≤ 0.01 * P ≤ 0.05		
Balanced Accuracy	62.08%				

I computationally optimized the hotel model by having the machine choose the parameters that produced the best accuracy for the models. The final values used for the model were nIter = 11. The two classes predicted were "growing" and "shrinking," which were resampled using a 10-fold cross validation with sample sizes of (318, 319, 319, 318, 319, 318,...) resampled across the tuning parameters.

The model achieved an accuracy of 62.2% (balanced accuracy of 62.1%). The algorithm identified 145 out 233 restaurant observations in the test set correctly. The 95% accuracy confidence interval lies at 0.5567 to 0.6848. Compared with the no-information rate of 0.5451, the classification performance is significant (P-Value [Acc > NIR] < 0.05) and supports H4.2. I further show the restaurant-growth model feature importance in Figure 22.

Figure 22 - Restaurant growth prediction feature importance



I find the number of excellent ratings (ta_reviewcount_5_year), the share of excellent ratings compared to the overall amount of ratings (ta_reviewcount_5_share_year) and the total number of excellent ratings (ta_reviewcount_csum_5_year) to be the model's most important features. Further, the absolute number of reviews (ta_reviewcount_year) and the difference in review counts (ta_reviewcount_year) between the current and two years ago are important. In addition, the number of "good for" attributes (ta_NrGoodFor), such as special offerings of Sunday brunch or special offerings for families, is found to be important. Additionally, the location of the business, expressed as the rural and city character (ed_countryside_class, ed_city_class) of a restaurant location, is important for the model.

The discussion of the positive or negative impact of each feature on growth go beyond the scope of this thesis, however would shed more light on their causal relationship. Partial dependency plots (not available for LogitBoost) are suggested for the evaluation and can be applied to other decision tree related methods.

V.5 Summary

V.5.1 Discussion and Conclusion

This study's findings suggest that a computational growth prediction based on data that is a combination of eWOW and economic data is possible and performs significantly better than a random guess, even when considering the knowledge about the distribution of growing and shrinking businesses. As the model is able to correctly classify 79.2% of all growing hotels and 63.7% of all growing restaurants and achieved balanced accuracies of 69.6% and 62.1%, respectively, I believe that through the combination and analysis of historic and present data sources, the future growth of H&R can be predicted by researchers and practitioners that are able to collect publicly available data sources, combine them, and apply algorithmic procedures to them.

The quantitative nature of the present study allows to further evaluate the relationships among eWOM attributes, some of which have been found to explain revenue per available room and have been questioned by Torres et al. (2015). This emphasizes the need for future research examining other financial measures, such as actual annual revenues and revenues outside the United States. The study, which took place in Switzerland, used actual changes in the revenue numbers of businesses over several years. As such, I believe that this is the first study inferring and predicting economic changes with the help of eWOM attributes, which has been pointed out by Pantano et al. (2017) to be a blind spot in current research. I find that the collection of eWOM attributes enables the classification of business growth. However, I did not find the strong link that other researchers have found between changes in rating value and revenue growth. Moreover, I did not find that a strong relative increase in reviews predominantly enables the prediction of growth. The feature importance figures show other features to be the models' most relevant features and show that a change in rating and in counts only partially contributes to the model's performance.

Still, this research supports the assumption of Torres et al. (2015) that increasing the number of reviews can be beneficial for a lodging establishment in various ways, as I find that the changes in review counts contribute to the models' prediction capability of growing and shrinking hotels. I also find similar features important for restaurants, especially "excellent reviews" and their share compared to all reviews. I cannot comment on their positive or negative impact on business performance. However, I find that growing and shrinking

restaurants can, ceteris paribus, be distinguished using review counts; hence, I find no evidence contradicting the findings of Torres et al. (2015) and assume this phenomenon to be valid across tourism industries.

V.5.2 Implications

The exploration of models that are based on an open-data analysis to make better predictions about the (future) attractiveness of a certain business is valuable for hospitality and gastronomy managers, investors, banks, insurances, merger and acquisition advisors, management consulting firms, as well as marketers of regions or tourism boards. As such, the study results generate ideas for practitioners to create their own performance indicators that could be linked to managers' incentives. Ideas for attributes that help explain revenues can be found in the feature importance ranking in the revenue models. In addition, ideas for (early) indicators of growth can be found in the feature importance ranking of the growth classification. Combining the findings in the growth predictions and assuming knowledge about the average business profitability, the process of valuing businesses can be augmented and eventually be automated.

V.5.3 Limitations and Future Work

Despite the contribution that this study makes to both theory and practice, there are a number of limitations. First, the study uses data from a time period that is likely to have been impacted by the European debt crisis as well as the rare events of the Swiss National Bank, heavily influencing exchange rates. Because of both, I assume that the relationship between performance determinants and business performance was severely affected. I further assume that as I was not able to include many other factors, such as overall demand, new forms of supply (e.g., AirBnB, retail gastronomy, food trucks), and service levels, I cannot claim that this study's findings will hold for other countries and different time periods. The main focus of the present study was limited to only online reviews for businesses that are insured with one single insurance provider. Hence, I cannot postulate that I would have found the same effects using a sample of companies from a different provider.

I have attempted to answer the research questions with the highest care given the data-availability limitation. As I considered only a small period of time and excluded the businesses with no (reported) revenue change for the growth prediction, I restrict the

generalization of the findings to real-world applications. While I find evidence regarding the predictive capabilities of the data collected, the growth prediction model itself was evaluated in only a three-year context of classifying growing and shrinking businesses. Additionally, especially for hotels, the sample size of 110 observations is limited. Due to the large number of features, I eventually encounter the curse of dimensionality, which I tried to mitigate by systematically reducing the categorical features from the swisshotel database. The prevalent high dimensionality of the data sets still limits the generalization of some of the features that were found to be important. Consequently, these limitations provide research ideas that future studies might try to address.

I encourage more research replicating this study's approach using longitudinal data over much longer periods of time with more observations. This would provide an opportunity to examine the data sources I introduced and shed more light on which features can be practically used to predict the business performance of H&R companies.

Additionally, I believe it would be worthwhile to include other potential variables that may influence business performance, such as increases in marketing budgets or the characteristics of the managers. Future research leading to a refinement of the prediction model, as well as studies in other geographies and different observation periods, are required to improve the generalization of the findings. Scholars can continue to expand this emerging field of knowledge by proposing theoretically and economically sound research iterations, which can be implemented and tested in joint efforts between practitioners and academia. In addition to research on changes in revenues, joint studies between H&R could further shift focus to not only evaluate the impact of eWOM on revenue and revenue proxies but also to consider profit. Working toward a better understanding of the relationship between revenue, profits, and eWOW data, other researchers might also reveal additional relationships or confounding variables that have not yet been discovered.

V.5.4 Conclusion

For academics and practitioners alike, the identification of emerging determinants of tourism business performance is one of the most critical activities when engaging in the optimization of organizations. Due to the large-scale and searchable nature of the Internet, cost-effectively capturing and analyzing eWOW and other tourism related public data sources becomes possible. Doing so yields great potential as an early indicator for the growth of H&R. In this

study, I linked the revenue of H&R with attributes from TripAdvisor, GoogleReview, and the Swiss Federal Statistical office for both H&R. In addition, I linked attributes from hotelssuisse.ch with the records of hotels only. By combining actual H&R business performance data, I contribute to the tourism literature in that I evaluate the relationship between eWOM data sources, business descriptions from travel and eWOM websites, and the actual business performance of a wide range of H&R. I find that the combination of the collected attributes enables the prediction of revenue changes of H&R using boosted logistic regression. The results are partially consistent with prior studies in similar fields; however, this study's findings may put the strength of previously found effects into perspective. I conclude that the potential of user-generated content and the information stored at eWOM websites can have a discriminating effect on tourism performance, and its collection has value when predicting the business performance of H&R. Online reviews are an important source of information and can crucially shape the business performance of H&Rs. Arguably, the results are not a proof of a direct causal relationship. However, the evidence and effect I find suggest that practitioners should continue to pay attention to their businesses' online listings.

VI) Discussion and Conclusion

This chapter discusses the findings of this thesis and reflects the results within the insurance context. It starts with a summary and discussion of principal findings in relation to initially posed research questions. Then, it highlights the most relevant results of the work and discusses their implications for research and practical applications in the insurance sector. Further, limitations of this thesis are presented and a future outlook for data mining with respect to growth prediction in the insurance sector is provided.

VI.1 Summary of Key Findings

With the increasing awareness of the value of operational data and availability of public data, that can be combined with internal data sources, firms are exploring opportunities how to leverage digital resources to improve their own business capabilities. Recent business intelligence studies have formulated the term of data-rich environments referring to the ability of companies to amass large amounts of customer data and process it within analytical models (e.g., Wedel and Kannan, 2016). This strategic orientation finds applications in many industries, including insurance. However, the questions remain for insurance operators which of these various sources are generating insights and how they are to be processed. On the example of business insurance, this thesis identifies internal and external data sources that can be combined and used to identify the economic future of insurance customers. As such, it provides value by demonstrating how data analytics can enable insurance companies to help prioritizing their client portfolio, price insurance policies more competitively and increase premiums by reducing underinsurance of SME clients. Within this thesis, the value of insurance internal and external data sources in respect to forecasting the revenue growth of insurance companies' SME clients are explored with the help of four case studies. This investigation was guided by the overall research question:

> *How can insurance companies predict the revenue growth of small and medium-sized enterprises with the help of internal and external data sources?*

Based on this general problem, four research questions, which relate to (1) the model applicability, (2) the data sources, (3) the prediction methodology, and (4) have application in

the insurance sector, are derived. These questions are addressed in four case studies, which are based on the operational data from a Swiss insurer. For the purpose of this thesis, the operational data sources are enriched with data from other insurance data warehouses and external data such as patent data, federal statistical data, online reviews and tourism databases. These collected data are analyzed and used to fit statistical models to infer the growth of SMEs. The composition of the thesis's findings that show the value of developed methodology and internal and external data sources in the insurance context is summarized in the following sections. Moreover, Table 42 (see page 129) lists the key findings of the research questions on a single page.

VI.2 Answers to research questions

Due to the competitive nature of the commercial insurance market, margins are under pressure. Customers are bargaining their insurance premiums; hence, insurance companies have to consider the expected economic future of a client when pricing policies and investing firm resources into the relationship with a business client. Further, customers that are underinsured pay insufficient premiums which puts a burden on the business relationship in case of a claim. Competitive policy pricing and avoiding underinsurance are both goals of insurance companies that can be achieved, if revenue growth of SMEs can be predicted accurately.

Research Question 1

The starting point of this thesis was the question for what kind of SME clients it is possible and promising to predict future revenue growth and how insurance companies can prioritize among their SME clients. This led to the first research question of the case study:

RQ1: For which SMEs should insurance companies predict revenue growth?

Insurance companies prefer to have the capability of correctly estimating the future economic situation of all of their clients. The gained insights could then be used to price policies competitively, give discounts, and reach out to their clients if insured revenues do not match actual revenues, thus increasing premiums collected and avoiding underinsurance. In practice, the holistic prediction of all of their SME clients' economic future at a high accuracy is difficult to achieve; hence, the customer portfolio has to be grouped by cohorts, suitable for prediction. By answering RQ1, I suggest a top-down approach, based on suitability criteria an

insurance company can establish before engaging in any prediction efforts. As such, it is an attempt to find a more suitable data basis compared to building one scaled predictive model for all SMEs. I suggest that based on the identified cohorts, the insurance company then builds a community of micromodels, which as an ensemble operates over the whole cohort population. The cohorts, in this case SME industries, are derived from the operational and policy data warehouse of a Swiss insurer. A sample of SMEs with their corresponding revenues over several years was extracted and analyzed. I provide four different sample evaluation dimensions that are relevant to rank industries by prediction suitability. The results of the statistical analysis reveal that (1) restaurants and hair dressers are the most frequent SMEs in the sample, (2) the aggregated industry revenue of motor vehicle garages and restaurants contribute most to the portfolio of insured revenues, (3) revenues of dental practices and pharmacies are most similar to the mean revenue (low CV) of the industry, whereas business consulting and commercial agencies' revenues spread around the mean (high CV), and (4) the fastest-growing industries[9] were breweries and florists shops. Multiplying these ranks provides anecdotal evidence how insurances can select industry companies that are most relevant and suitable for a prediction. The results suggest that industries such as motor vehicle garages, restaurants, and consultancies are suitable industries for revenue growth predictions.

Research Question 2

The second case study provides an example of how a cross-industry prediction of SME revenue growth can be performed without the effort of predetermining a cohort and costs involved with collecting external data. Therefore, RQ2 is stated as:

> RQ2: *How can insurance companies predict the future revenue growth of SMEs with vehicle insurance data, and what are the benefits of such models?*

Contingent of the product offering of an insurance, other internal data warehouses contain information about some of their SME customers (bottom-up). On the example of motor insurance policies, the prediction capacity of this data source is evaluated. For the study in

[9] With more than 50 observations.

Chapter III, it is hypothesized that choices of motor vehicles of SME owners, inferred social status, values, and its desired perception are related to SME performance. The data to investigate the relationship has been gathered from the operational and policy data warehouse of a Swiss insurer. The sample includes data from the two nonlife insurance products: general business insurance and automobile insurance. Moreover, the data set includes the general covariates of customers and further insurance-product specific covariates. To evaluate the influence of motor vehicle (related) choices and SME growth, several analyses have been conducted.

The results of the statistical analysis reveal that the demographics of the driver, often assumed to be the manager, owner, and founder of an SME, help to explain the growth, expressed as CAGR of an SME. The results suggest that SMEs with younger driver mentioned in the vehicle policy, grow faster but have lower absolute revenues. Further, SMEs with older vehicles insured tend to grow faster. SMEs that lease their vehicles in the first three years seem to grow faster than those SMEs that own new or up to three-year-old vehicles. However, overall, SMEs with leased vehicles grow slower than SMEs that own their vehicles. Surprisingly, the data also suggests that SMEs with drivers that took the driving test at age 18 grow faster and have higher revenues than those taking the test at a later age. For those years investigated (18-27 years), this effect becomes stronger the later the driving test was taken. In addition, investigating the kind of vehicle insured, the study finds that sport and luxury cars indicate negative growth. Fuel type and vehicle brand (or models) did not indicate any strong difference in SME growth. As a result of the statistical analysis of the data sample, I find that several cohorts of SMEs, grouped by the attributes of the motor insurance policies, suggest there are differences between the cohorts.

Training a prediction model with attributes from the general business and motor insurance, I find that SME business growth can be predicted with the help of these covariates. The model was able to recognize some of the differences between the cohorts and was therefore able to identify 1 619 (72.15%) out of 2 244 growing SMEs. The model results are significant and achieved 70.96% accuracy. Among the most important covariates for the model are the age of the driver, kilometers driven, the date, and the age at which the driving test were taken.

Overall, the findings indicate that differences in growth exist between SME insuring different motor vehicles with different vehicle attributes. With the help of these attributes,

growing SMEs can be identified. The study provides empirical evidence that by combining already collected information about SME customers, insurance companies can generate further insights, which they can leverage to improve their understanding of the economic status of SME clients. In the examined data sample, the overlap between the general liability insurance and the vehicle insurance was 10.1%[10] of all SMEs. This is the largest cohort and has 56.06% more observations than any industry-related cohort, hence it provides sufficient data for the algorithm to learn. Due to the applicability of this large cohort and the derived model accuracy, the introduced methodology ought to be implemented into the CRM systems of all insurances that offer both general liability and commercial motor insurance.

Research Question 3

With regard to the third research question (RQ3), an external data source is introduced, analyzed, and used to predict SME growth. Assuming an insurance company's internal data sources are already exploited, such publicly available sources can be collected and tested for their prediction capacity. Chapter IV investigates research question 3.

> *RQ3: How can insurance companies improve revenue prediction models with the help of patent filings and how can these models be applied?*

The main objective of the study is to validate whether data related to SME innovation activity, expressed through patent filing, improves prediction results. The study is motivated by current studies such as Freel (2000), who showed that innovating companies grew faster, and Wagner and Cockburn (2010), who showed in their research that companies with patents have 34% higher survival rate; hence, I hypothesize that the presence of patents should indicate a higher revenue growth of SMEs. Therefore, enriching a growth prediction model with SME patent information should improve the prediction outcome.

The findings of the study suggest that SMEs with patents grow faster than SMEs without patents. Furthermore, I show that with the help of motor vehicle insurance data, the gender of SME patent inventors can be imputed with a high accuracy. Applying the imputation to the patent filings enabled the investigation of the gender composition of patent-

[10] Calculated as the number of SMEs in the motor vehicle sample (11 235) divided by the number of SMEs in the general liability sample (111 236)

filing SME inventor teams. The findings of this investigation suggest that SMEs with mixed gender teams grow faster than SMEs with male-only teams.

Feeding the augmented patent information, including the gender attribute, into a prediction model, which is based on insurance internal and publicly available data, improves the prediction results, expressed by an increase in R^2 from 0.527 to 0.838. Hence, I conclude that with the help of patent data, even though only applicable for a small share of SMEs (~0.5%), general models of SME growth prediction can be improved. In the insurance portfolio context, the patent filing of an SME could therefore be used as a trigger event and should be implemented into insurances' CRM systems. As such, it would notify insurance agents whenever a patent is filed and indicate the increased future customer importance.

Research Question 4

With regard to the fourth research question (RQ4), following the findings of RQ1, an industry cohort, combined of hotels and restaurants, has been investigated. Through the combination, a cohort that is high in observations and relevant in respect to insured portfolio revenues has been created and is studied to answer the following question:

> *RQ4: How can insurance companies predict the future revenue growth of hotels and restaurants?*

The importance of an industry cohort to an insurance company is determined through the multiplication metric, developed as an answer to RQ1. By creating a tourism cohort, as the combination of restaurants, hotels and similar businesses, a much larger and more important (sum of the insured revenues) joint-cohort could be created. Restaurants are the third most frequent businesses and hotels are the fifteenth most frequent businesses in the sample; hence, the relevancy for the insurance is provided, yet the sample can be augmented with external data sources applicable to all tourism related SMEs. I was not able to link any other larger industry cohorts found by answering RQ1 (i.e., consultancies, motor vehicle garages) to an external data source, containing potential relevant information. The focus of the analysis is therefore to investigate an external data sources that can reveal signals related to growth in the tourism industry. The study is motivated by the understanding in existing marketing research that the tourism businesses with high ratings on electronic word-of-mouth websites (such as TripAdvisor.com) have higher average revenues (Torres et al., 2015). Therefore, I

hypothesize that past increases in rating scores also lead to increases in revenues. The data for the study span from a period from 2010 to 2015 and are extracted from the operational data warehouse of a Swiss insurer. The external data sample was collected in 2018; however, for the analysis, only data points valid in the years before 2015 were considered. When combined, the sample contains traditional covariates of business customers (i.e., company name, location, industry). Further, the annual insured revenues are extracted from the general business insurance of the same hotels and restaurants. The external data points include number of reviews and ratings at specific points in time. The external data have been collected by looking up the tourism businesses online on several eWOM websites. If available, the historic reviews and ratings were collected and analyzed. Feeding the collected information into a prediction model suggests that the business growth of hotels and restaurants can be predicted with the help of eWOM data.

The hotel model was able recognize the importance of features such as price and the location of a hotel. The model was able to identify 70.5% (69.6% balanced accuracy) of all SMEs correctly. The model results are significant. The restaurant model was able recognize the importance of features such as the number of “excellent” ratings and the share of “excellent” ratings to total ratings it has received during the year. The model was able to identify 62.2% (62.1% balanced accuracy) of all SMEs correctly. The restaurant model results are also significant.

Overall, the findings indicate that differences in growth exist between individual hotels and restaurants that can be inferred from eWOM websites. With the help of these attributes, which can be collected from these websites, growing SMEs can be identified. The study provides anecdotal evidence that by combining several eWOM data sources with insurance internal information, insurances can leverage their own data to improve the internal understanding of the economic status of tourism SMEs. In the examined data sample, the overlap between the general liability insurance and a single eWOM websites were 36.0%. This is the largest cohort between the sample and an external data source investigated. Due to the applicability of this relatively large industry cohort and the derived model accuracy, the introduced methodology ought to be implemented in any insurance CRM system that has a sufficiently large tourism business customer base.

Table 42 - Summary of research questions and findings

	Research question	Findings
RQ1:	*For which SMEs should insurance companies predict revenue growth?*	Industries such as restaurants, motor vehicle garages, or consultancies are most promising cohorts for revenue prediction models. Meaningful prediction models have to be applicable to a high number of SMEs and contribute considerably to the insured portfolio revenue. In addition, the cohort members should have heterogeneous revenues, and have increased revenues in the past.
RQ2:	*How can insurance companies predict the future revenue growth of SMEs with vehicle insurance data, and what are the benefits of such models?*	With data from motor vehicle insurance policies, insurance companies can predict future SME revenues. Age of the driver, age at which the driving test was taken, and kilometers driven are the most relevant attributes for the model. The model's main advantages are accuracy and applicability to a large cohort (>10% of the examined insurance portfolio).
RQ3:	*How can insurance companies improve revenue prediction models with the help of patent filings and how can these models be applied?*	Patent filings are signals of innovation. It can be collected at scale and linked with the SME portfolio. Fed to a model, the availability of the patent attributes improves the prediction capacity of a model. Only 0.5% of all SMEs have filed a patent; hence, the low applicability precludes general prediction modeling for miscellaneous SME portfolios. However, used as a trigger event, patent filings signal increased future customer importance for an insurance.
RQ4:	*How can insurance companies predict the future revenue growth of hotels and restaurants?*	Data of the SME cohort of all restaurants, joined with hotels, to form a larger, more relevant cohort, can be augmented with eWOM data and fed to a prediction model. Such a model can predict revenue growth but is limited in applicability because many businesses are not listed on eWOM websites and have low counts in reviews, and legal constraints limit the data collection.

VI.3 Implication

VI.3.1 Implications for Practice

This thesis and the research project taken as basis have been conducted in cooperation with a Swiss insurance company, and thus the four research questions address topics that inherit a practical relevance. Further, the empirical findings of my work are based on operational data of this insurer and provide tangible insights for the insurance professionals in the area of business intelligence, insurance policy pricing, underwriting, and customer relationship management of business clients. The presented case studies offer examples and conceptual guidance how to benefit from the usage of a data-rich environment when analyzing the growth differences of SME customers.

The approach taken in the analysis that is related to research question one (RQ1) develops a methodology how an insurance company can use only data from the general liability policies to guide the selection of industries suitable for growth prediction. It also introduces the reader to the characteristics of data sample, which is diverse in industries and SME size. The findings of case study one imply that cohort size and its economic importance should be taken into consideration when building revenue-growth prediction models. In order to find a meaningful cohort for prediction, the cohort itself should not be too homogeneous and should show growth variation compared to other cohorts. Further, by combining cohorts that may share similarities, the importance of a cohort to an insurance building growth models can further be increased.

Following up the results described above, research question two investigates an alternative to the top-down approach. For this study, only SMEs that have a vehicle insured with the insurance were considered. The results of the study show that with the help of internal insurance data describing the characteristics of a large share of SME portfolio customers, their growth can be predicted. The detailed results of the study provide a guideline for practitioners that plan to leverage the internal information available to them. The knowledge about another insurance product an SME customer has (such as a motor vehicle policy) enables insures to leverage existing digital resources and direct their client relationship efforts to growing SMEs. Specifically, based on the effect of the different vehicle attributes, simple managerial rules could be derived and implemented to prioritize the contacting of SMEs with higher growth expectations. To these SMEs, the insurance could

further allow a discount when negotiating premiums, instead of losing the customer otherwise.

Besides internal data sources, external sources can also reveal promising insights about an SME customer. This topic has been investigated in research question three (and four). Used as a trigger event, a patent filing could be implemented into an insurance CRM, with the goal of an agent reaching out to the customer and discuss the economic future. The filing of a (first) patent, information that is publicly available, is an important milestone for most SMEs; hence, besides the economic signaling effect, it may also serve as a good topic to discuss from a relationship point of view. Patent data are collectable without much effort as they are stored and continuously updated in a structured format. In summary (see Chapter IV) the presented approach, to merge traditional customer data from the operational data warehouse with external data from innovation data sources such as patents, shows value by indicating a strong effect between SMEs with and without patents. Only very few SMEs file patents; hence, the application is limited in the portfolio context. Due to the effect strength and availability, however, it served practitioners well and could be implemented without much effort as the patent data collection can be automated and the information itself is free of charge.

A second example of an external data source that can be captured systematically and could reveal promising insights about SMEs is discussed in RQ4. Data sources, to which the literature refers to as electronic word-of-mouth, are plentiful and reveal many economically relevant data points. A hotel's or restaurant's (change in) number of reviews, rating score, or rank, in comparison to a peer group, intuitively conveys information about the economic future of the business. The collection of eWOM data, as concluded by the answer to RQ4, enables the prediction of future growth. The potential existence of confounding and lagging variables, technical, legal, and availability challenges make the collection and evaluation of eWOM data at scale a tedious process for an insurance, given its limited applicability for only a small share of the portfolio. Within the examined data sample, only 76 hotels and 407 restaurants had a listing on the eWOM pages and reported change in revenues. This represents less than 1% of all insured businesses in the total sample. The approach itself is valuable, however, will most likely find more application for tourism itself. For example, the prediction of the attractiveness of a certain business is valuable for hospitality and gastronomy managers, investors, merger and acquisition advisors, and management consulting firms as well as marketers of regions or tourism boards.

Overall, the four studies contribute to the question how the collection and analysis of insurance internal and external data can generate valuable insights for insurance firms. Further, they provide tangible implications for how insurance companies can build capabilities to predict the future revenue growth of their commercial business customers and use such deepened understanding for operational implementations. The economic benefits lie in agents' prioritizing growing SME within their portfolio and foster the personal relationship with those SMEs. Further, the prediction of SME revenue growth may provide suggestions for which SMEs' insured revenues are not consistent with actual revenues (underinsurance), hence enabling the insurance company to reach out to growing SMEs, update the insured revenues, and thereby increase the premiums collected from the SMEs.

Besides the benefits that explicitly support insurance implementation and application of, for example, the cross-product insights from the motor vehicle insurance or the trigger events from patent filings, there are also concerns, which may limit the application of the findings discussed. The usage of such additional data sources for internal analytical purposes, at least, or as a conversational starting point with a customer at most, must be carried out with utmost discretion by any firm. But financial service providers in particular, because they invest heavily to build and maintain a trustworthy relationship with their customers, are required to act with caution. A customer's perception of a breach of data privacy alone may damage the relationship. As the recent Facebook privacy scandal revealed (DSK, 2018), consumers see a responsibility with those who store data to explain to them what is done with their information in plain terms. The mundane agreement of terms of conditions is not enough, and insurance data-derived economic gains may dissolve when trust is lost in the long term.

Based on four cases studies, this thesis presents approaches for insurance firms on how to benefit from data-mining activities such as combining internal and external data sources, analyzing the source, and building prediction models to recognize growing SME customers. The described processes and the evaluation of novel data sources that arise from this development carry relevance for practitioners, who may consider to adapt and further develop the approaches presented in this thesis.

VI.3.2 Implications for Research

Although the presented work has a strong practical and managerial focus, it provides novel empirical insights for researchers in the domains of information science and insurance management as well. Insurance-related research on how to utilize internal and novel external data sources of business customers is sparse, yet the topic has been raised in other domains, such as auto insurances or banking, which shares the contractual setting commercial insurances find themselves in. This thesis aims to contribute to the research direction and is structured to work toward filling remaining research gaps. At the same time, some of the findings, especially the statistical distinctive attributes highlighted and discussed in Chapters III.4.1 and IV.4.1 and 4, also raise new questions.

The focus of research question one (RQ1) in this thesis is to investigate how and which cohorts of SMEs can be selected for a meaningful top-down growth prediction without prior additional data source investigation. As such, it serves as a methodological guideline how to combine economical and statistical considerations when selecting cohorts for revenue growth prediction.

The motivation to investigate research question two (RQ2) builds on the assumption that before external data are acquired, an insurance preference lies in the use of internal data that are already available. The study's findings provide novel evidence in the insurance context that the value of cross-product internal data is not yet fully recognized by insurances. Motivated by social identity theory and brand identity theory that suggest that consumer choices such as a vehicle purchase are indicative for status signaling, lifestyle and value, the information related to the insured vehicles mentioned in the SME policies can be used to predict SME growth patterns.

For research question three (RQ3) the usage of patent data is discussed. As one of the first empirical studies, it takes a large real-world sample of SMEs with patents and compares it with an industry-adjusted peer group. As such, it supports the finding of Freel (2000) and provides evidence that revenue growth can be inferred from patent application of SMEs and not just from patent applications of the much more studied large corporations. Methodologically, the study showed how a data source such as motor insurance contains information, which in aggregate can be used for data imputation, thereby increasing the data quality. As such, the methodological part answered to Blevins and Mullen (2015), asking for a method for gender categorization that takes demographic changes over time into account.

The results related to research question four (RQ4) show the relevance and availability of eWOM websites for hotels and restaurants in Switzerland. The study findings answer to Serra Cantallops and Salvi (2014) and Ye et al. (2011), who called for more research on the relationship between online reviews and business performance. Further, evidence from the study contributes toward shedding more lights on the relationship between the number of reviews, rating score, and specific monetary impact on revenues as requested by Melián-González and Bulchand-Gidumal (2016) and Senecal and Nantel (2004a).

Overall, these results contribute to how the usage of insurance internal and external data could (1) reveal more detailed insights to describe the growth of SME insurance customers and (2) improve prediction approaches to identify which of the SMEs in an insurance portfolio are growing with high accuracy.

VI.4 Limitations and Future Research

The contributions of this thesis are associated with a number of limitations with regard to the scope of the thesis as well as the characteristics of the applied research methods. Further, as this thesis provides valuable insights for practitioners, applicability and effect relevance in a portfolio context have to be reflected critically. This section summarizes and discusses these limitations as well as formulates future research ideas. For each of the four case studies, limitations exist that also need a critical reflection with respect to the overall goal of the thesis. At first, the data for all empirical studies of this thesis are obtained from a single insurance firm, solely operating in Switzerland and Liechtenstein. Although the insurance company is one of the market leaders in the nonlife insurance sector and serves businesses and households in all regions of the country, the results of the case studies are influenced by the customer composition of the company. The company is historically strong in rural areas, which may have affected the data sample. The dependent variable, the SMEs' revenues are taken from general liability insurance policies. The reported revenue numbers that only cover a limited revenue period (2009-2016) underwent a basic plausibility check by the agents at initiation; however, no systematic validation by the insurance took place from then on. This explains the lacking actualization behavior of many SMEs, that, according to the findings of case study one, only report large revenue changes at infrequent intervals. A more fine-grained reporting standard, matching insured revenues with, for example, tax-compliant revenue numbers, would certainly allow the increase of sample sizes and tune prediction models to recognize smaller changes in revenues. Due to the short observation period, restricted by the data availability, and a historically abnormal currency exchange rate volatility, the study's findings generalization is limited, especially for the findings in case-study four, because tourism was more severely affected by the aforementioned economic irregularities. Furthermore, the nature of anecdotal evidence of case studies restrict the generalization and must not be interpreted as a proof for a causal relationship. The thesis provides a foundation for researchers and firms to further explore internal and external data sources for analytical purposes to understand and manage the special relationship between them and their customers. Therefore, the opportunities and needs for future research are plentiful. Study iterations most meaningful in the context of the work performed by this thesis could gather other non-self-reported, timely revenue data and eventually even profits that would broaden the research focus and provide additional relevance to the topic. Further, a longitudinal study

covering several business cycles in more than one country could be performed. Such findings would increase the robustness of the results discussed within this thesis. Studies with the aim of finding generalizable cross-industry relevant covariates could also investigate in much further depth which macroeconomic covariates influence general business growth and hence be considered when predicting future revenues of companies. In terms of applicability, the study findings suggest several strategies for implementation into insurance companies' customer-relationship management systems. What has not been critically reflected is the user acceptance as well the economic benefit-versus-cost consideration of building a prediction model for SME revenue growth in the insurance context. Both are of the highest priority before implementing the suggested approaches to recognizing SME growth. For example, a field test that would implement the findings of case study two could be implemented without much effort. Whenever an SME files a patent, assuming the SME is a client of the insurance, the responsible agent could receive an e-mail notification with guidance on how to proceed. Structured as a randomized controlled trial, this would allow study of the agents' acceptance of the information delivery channel as well as the SMEs' reaction to such an approach as well as the economic outcome.

Moreover, the linkage of distinct sources of customer data with internal and external data, as suggested within this thesis, must be viewed critically from a data privacy perspective. With the General Data Protection Regulation (GDPR Regulation) for the European Union (2016/679), which came into effect in May 2018, the topic is of high relevance. Despite the potential absence of legal constraints, firms ought to recognize that business intelligence stretching to a level considered intrusive from customers bears reputational cost, which may outweigh the benefits desired. Further, as corporate data ownership has become a challenging issue in a data-rich environment such as insurance companies find themselves, they have to find suitable governance structures for the collection and use of data. Recent events like Wikileaks, Edward Snowden, and Cambridge Analytica (Lin, 2018; Williams et al., 2018) have helped to raise the level of discussion that is providing insight into the dangers of aggregating data centrally (Zhou et al., 2014). For these reasons, the usage of data and the inference of upselling opportunities or triggers of outbound sales activities, which could harm the relationship between business customer and the company, should be investigated thoroughly.

References

Accenture, 2011. The Path to High Performance in Insurance, Transforming Distribution and Marketing with Predictive Analytics.

Albrecht, J.P., 2016. How the GDPR Will Change the World. Eur. Data Prot. Law Rev. 2, 287–289. https://doi.org/10.3366/ajicl.2011.0005

Altowim, Y., Kalashnikov, D. V., Mehrotra, S., 2014. Progressive approach to relational entity resolution. Proc. VLDB Endow. 7, 999–1010. https://doi.org/10.14778/2732967.2732975

Ammann, K., Reusser, K., 2017. Welches sind die häufigsten Schweizer Nachnamen? [WWW Document]. URL https://www.swissinfo.ch/ger/wirtschaft/umstrittene-aliasnamen-in-callcentern_welches-sind-die-haeufigsten-schweizer-nachnamen/43291670

Anderson, C.K., Lawrence, B., 2014. The Influence of Online Reputation and Product Heterogeneity on Service Firm Financial Performance.

Argamon, S., Goulain, J., Horton, R., Olsen, M., 2017. DHQ : Digital Humanities Quarterly Vive la Différence ! Text Mining Gender Difference in French Literature 3, 1–11.

Arndt, J., 1967. Role of Product-Related Conversations in the Diffusion of a New Product. J. Mark. Res. 4, 291. https://doi.org/10.2307/3149462

Arundel, A., 2001. The relative effectiveness of patents and secrecy for appropriation. Res. Policy 30, 611–624. https://doi.org/10.1016/S0048-7333(00)00100-1

Arundel, A., Kabla, I., 1998. What Percentage of Innovations are Patented? Empirical Estimates for European firms. Res. Policy 27, 127–141. https://doi.org/10.1016/S0048-7333(98)00033-X

Ashraf, N., 2009. American Economic Association Spousal Control and Intra-Household Decision Making : An Experimental Study in the Philippines Spousal An Experimental Study in the Philippines Control and Intra-Household Decision Making : Aer 99, 1245–1277. https://doi.org/10.1257/aer.99.4.1245

Auge-Dickhut, S., Koye, B., Liebetrau, A., 2016. Customer Value Generation in Banking.

Ayeh, J.K., Au, N., Law, R., 2011. Journal of Travel Research. J. Travel Res. 1–3. https://doi.org/10.1177/0047287510382296

Baker, W., Sinkula, J., 2002. Market orientation, learning orientation and product innovation: delving into the organization's black box. J. Mark. Manag. 2002 5, 5–23.

Batista, G., Monard, M.C., 2003. An analysis of four missing data treatment methods for supervised learning. Appl. Artif. Intell. 17, 519–533. https://doi.org/10.1080/713827181

Beck, T., Cull, R., Jerome, A., 2017. Bank Privatization and Performance: Empirical Evidence from Nigeria 1–2.

Ben-David, S., Shalev-Shwartz, S., 2014. Understanding Machine Learning: From Theory to Algorithms, Understanding Machine Learning: From Theory to Algorithms. https://doi.org/10.1017/CBO9781107298019

Benzing, C., Chu, H.M., Kara, O., 2009. Entrepreneurs in Turkey: A factor analysis of motivations, success factors, and problems. J. Small Bus. Manag. 47, 58–91. https://doi.org/10.1111/j.1540-627X.2008.00262.x

Bharadwaj, A., El Sawy, O.A., Pavlou, P.A., Venkatraman, N., 2013. Digital Business Strategy:

Toward a Next Generation of Insights. MIS Q. 37, 471–482. https://doi.org/10.25300/MISQ/2013/37:2.3

Black, H.G., Kelley, S.W., 2009. A storytelling perspective on online customer reviews reporting service failure and recovery. J. Travel Tour. Mark. 26, 169–179. https://doi.org/10.1080/10548400902864768

Blal, I., Sturman, M.C., 2014. The Differential Effects of the Quality and Quantity of Online Reviews on Hotel Room Sales. Cornell Hosp. Q. 55, 365–375. https://doi.org/10.1177/1938965514533419

Blevins, C., Mullen, L., 2015. Jane, John … Leslie? A Historical Method for Algorithmic Gender Prediction. Digit. Humanit. Q. 9, 1–19.

Brachman, R.J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., Simoudis, E., 1996. Mining business databases. Commun. ACM 39, 42–48. https://doi.org/10.1145/240455.240468

Brandstätter, H., 2011. Personality aspects of entrepreneurship: A look at five meta-analyses. Pers. Individ. Dif. 51, 222–230. https://doi.org/10.1016/j.paid.2010.07.007

Breiman, L., 2001. Statistical Modeling: The Two Cultures. Stat. Sci. 16, 199–215. https://doi.org/10.2307/2676681

Cai, L., Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Sci. J. 14, 2. https://doi.org/10.5334/dsj-2015-002

Cai, Y.D., Feng, K.Y., Lu, W.C., Chou, K.C., 2006. Using LogitBoost classifier to predict protein structural classes. J. Theor. Biol. 238, 172–176. https://doi.org/10.1016/j.jtbi.2005.05.034

Cambia, 2017. Lens patent database [WWW Document]. URL https://www.lens.org/lens/

Capon, N., Farley, J.U., Hoenig, S., 1990. Determinants of Financial Performance: A Meta-Analysis. Manage. Sci. 36, 1143–1159.

Carl, W.J., 2006. What's all the buzz about?: Everyday Communication and the Relational Basis of Word-of-Mouth and Buzz Marketing Practices. Manag. Commun. Q. 19, 601–634. https://doi.org/10.1177/0893318905284763

Caruana, R., Karampatziakis, N., Yessenalina, A., 2008. An empirical evaluation of supervised learning in high dimensions. Proc. 25th Int. Conf. Mach. Learn. 96–103. https://doi.org/10.1145/1390156.1390169

Cătălin, M.C., Andreea, P., 2014. Brands as a Mean of Consumer Self-expression and Desired Personal Lifestyle. Procedia - Soc. Behav. Sci. 109, 103–107. https://doi.org/10.1016/j.sbspro.2013.12.427

Cefis, E., Orsenigo, L., 2001. The persistence of innovative activities: A cross-countries and cross-sectors comparative analysis. Res. Policy 30, 1139–1158. https://doi.org/10.1016/S0048-7333(00)00139-6

Celik, M.A., 2015. Does the Cream Always Rise to the Top ? The Misallocation of Talent in Innovation.

Chen, H., Storey, V.C., 2012. Business Intelligence and analytics: From big data to big impact.

Chen, I., Popovich, K., 2017. Understanding customer relationship management (CRM): People, process and technology. Bus. Process Manag. J. 21, 191–206. https://doi.org/10.1108/MBE-09-2016-0047

Chen, P.-Y., Wu, S., Yoon, J., 2004. The Impact of Online Recommendation and Consumer Feedback on Sales. Proceeding Int. Conf. Inf. Syst. Paper 58, 711–724.

https://doi.org/http://aisel.aisnet.org/icis2004/58

Chen, Y., Xie, J., 2008. Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix. Manage. Sci. 54, 477–491. https://doi.org/10.1287/mnsc.1070.0810

Cheung, C.M.K., Thadani, D.R., 2012. The impact of electronic word-of-mouth communication: A literature analysis and integrative model. Decis. Support Syst. 54, 461–470. https://doi.org/10.1016/j.dss.2012.06.008

Chevalier, J.A., Mayzlin, D., 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. J. Mark. Res. 43, 345–354. https://doi.org/10.1509/jmkr.43.3.345

Chew, Y.-T., Yeung, H.W.-C., 2001. The SME Advantage: Adding Local Touch to Foreign Transnational Corporations in Singapore. Reg. Stud. 35, 431–448. https://doi.org/Article

Coad, A., Rao, R., 2008. Innovation and firm growth in high-tech sectors: A quantile regression approach. Res. Policy 37, 633–648. https://doi.org/10.1016/j.respol.2008.01.003

Colquhoun, D., 2014. An investigation of the false discovery rate and the misinterpretation of P values. R. Soc. Open Sci. 1–15. https://doi.org/10.1098/rsos.140216

Coltheart, M., 1981. MRC Psycholinguistic Database. Q. J. Exp. Psychol. 33, 497–505.

Comin, D., 2008. Total Factor Productivity. New Palgrave Dict. Econ. 1088–92. https://doi.org/10.1057/9781137336583.1849

Cook, L.D., 2011. Explorations in Economic History Inventing social capital: Evidence from African American inventors from 1843 – 1930. 48, 507–518. https://doi.org/10.1016/j.eeh.2011.05.003

Covin, J.G., Miller, D., 2014. International Entrepreneurial Orientation: Conceptual Considerations, Research Themes, Measurement Issues, and Future Research Directions. Entrep. Theory Pract. 38, 11–44. https://doi.org/10.1111/etap.12027

Covin, J.G., Wales, W.J., 2012. The Measurement of Entrepreneurial Orientation. Entrep. Theory Pract. 36, 677–702. https://doi.org/10.1111/j.1540-6520.2010.00432.x

Craig, J.B., Dibrell, C., Garrett, R., 2014. Examining relationships among family influence, family culture, flexible planning systems, innovativeness and firm performance. J. Fam. Bus. Strateg. 5, 229–238. https://doi.org/10.1016/j.jfbs.2013.09.002

Cusick, K., 2016. Voice of the Small Business Consumer, in: Deloitte CAGNY Fall 2016 Meeting. pp. 1–15.

Darley, W.K., Smith, R.E., Darley, W.K., Smith, R.E., 2017. Gender Differences in Information Processing Strategies : An Empirical Test of the Selectivity Model in Advertising Response Gender 3367. https://doi.org/10.1080/00913367.1995.10673467

Darroch, J., McNaughton, R., 2002. Examining the link between knowledge management practices and types of innovation. J. Intellect. Cap. 3, 210–222. https://doi.org/10.1108/14691930210435570

Davison, A.C., 2003. Statistical models. https://doi.org/10.4236/health.2010.27098

Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., Hu, W., 2012. Author Gender Prediction in an Email Stream Using Neural Networks. J. Intell. Learn. Syst. Appl. 4, 169–175. https://doi.org/10.4236/jilsa.2012.43017

Dellarocas, C., Zhang, X., Awad, N., 2013. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. Acad. Mark. Stud. J. 17, 119–132. https://doi.org/10.1002/dir

Deloitte, 2015. Small-business insurance in transition: Agents difficult to displace, but direct sellers

challenge status quo.

Dichter, E., 1966. How Word-of-Mouth Advertising Works. Harv. Bus. Rev. 44, 147–160.

Dickinger, A., 2010. The Trustworthiness of Online Channels for Experience- and Goal- Directed Search Tasks. J. Travel Res. https://doi.org/10.1177/0047287510382296

Dobbin, K.K., Simon, R.M., 2011. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med. Genomics 4, 31. https://doi.org/10.1186/1755-8794-4-31

DSK, 2018. Facebook Privacy Scandal – Enforcing the New Data Protection Law within Social Network Services, in: Entschließung Der Konferenz Der Unabhängigen Datenschutzbehörden Des Bundes Und Der Länder. pp. 1–2.

Duan, W., Gu, B., Whinston, A.B., 2008a. The dynamics of online word-of-mouth and product sales-An empirical investigation of the movie industry. J. Retail. 84, 233–242. https://doi.org/10.1016/j.jretai.2008.04.005

Duan, W., Gu, B., Whinston, A.B., 2008b. Do online reviews matter? - An empirical investigation of panel data. Decis. Support Syst. 45, 1007–1016. https://doi.org/10.1016/j.dss.2008.04.001

Duggal, R., Soni, A., 2016. Reducing Risk in KYC (Know Your Customer) for large Indian banks using Big Data Analytics. Int. J. Comput. Appl. https://doi.org/10.5120/17039-7347

Duverger, P., 2013. Curvilinear Effects of User-Generated Content on Hotels' Market Share: A Dynamic Panel-Data Analysis. J. Travel Res. 52, 465–478. https://doi.org/10.1177/0047287513478498

Eidgenössisches Departement für Wirtschaft Bildung und Forschung WBF, 2013. Die KMU-Politik der Schweiz, Eidgenössisches Departement für Wirtschaft Bildung und Forschung. Bern.

Eisenhardt, K.M., 2017. Product Development: Past Research, Present Findings, and Future Directions 20, 343–378.

Engel, J.F., Kegerreis, R.J., Blackwell, R.D., 1969. Word-of-mouth communication by the innovator. J. Mark. 33, 15–19. https://doi.org/10.2307/1248475

Erevelles, S., Fukawa, N., Swayne, L., 2016. Big Data consumer analytics and the transformation of marketing. J. Bus. Res. 69, 897–904. https://doi.org/10.1016/j.jbusres.2015.07.001

Federal Statistical Office, S., 2017. Federal Statistical Office - Look for Statistics [WWW Document]. URL https://www.bfs.admin.ch/bfs/en/home/statistics.html (accessed 3.2.17).

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., Amorim Fernández-Delgado, D., 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? J. Mach. Learn. Res. 15, 3133–3181. https://doi.org/10.1016/j.csda.2008.10.033

Filieri, R., McLeay, F., 2014. E-WOM and Accommodation: An Analysis of the Factors That Influence Travelers' Adoption of Information from Online Reviews. J. Travel Res. 53, 44–57. https://doi.org/10.1177/0047287513481274

Fisher, C., Lauria, E., Chengular-Smith, S., 2012. Introduction to Information Quality. AuthorHouse. https://doi.org/10.1007/978-3-319-24106-7_1

Fraz, M.M., Remagnino, P., Hoppe, A., Uyyanonvara, B., Rudnicka, A.R., Owen, C.G., Barman, S.A., 2012. An ensemble classification-based approach applied to retinal blood vessel segmentation. IEEE Trans. Biomed. Eng. 59, 2538–2548. https://doi.org/10.1109/TBME.2012.2205687

Freel, M.S., 2000. Do small innovating firms outperform non-innovators? Small Bus. Econ. 14, 195–210. https://doi.org/10.1023/a:1008100206266

Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. Ann. Stat. https://doi.org/10.1214/aos/1016218223

GastroSuisse Verband für Hotellerie und Restauration, 2018. Konjunktur / KOF Pulsmesser der Gastronomie und Hotellerie 2–5.

GastroSuisse Verband für Hotellerie und Restauration, 2017. Branchenspiegel 2017.

Ghose, A., Ipeirotis, P.G., 2006. Designing Ranking Systems for Consumer Reviews: The Impact of Review Subjectivity on Product Sales and Review Quality. Proc. Int. Converence Decis. Support Syst. 1–25. https://doi.org/10.1145/1282100.1282158

Ghose, A., Ipeirotis, P.G., Li, B., 2014. Examining the Impact of Ranking on Consumer Behavior and Search Engine Revenue. Manage. Sci. 60, 1632–1654. https://doi.org/10.1287/mnsc.2013.1828

Ghose, A., Ipeirotis, P.G., Li, B., 2012. Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowdsourced Content. Mark. Sci. 31, 493–520. https://doi.org/10.1287/mksc.1110.0700

Goldstone, A., Underwood, T., 2017. The Quiet Transformations of Literary Studies : What Thirteen Thousand Scholars Could Tell Us. New Lit. Hist. 9–10.

Gordini, N., 2014. A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy. Expert Syst. Appl. 41, 6433–6445. https://doi.org/10.1016/j.eswa.2014.04.026

Gretzel, U., Yoo, K.H., Purifoy, M., 2007. Online Travel Review Study: Role & Impact of Online Travel Reviews. Lab. Intell. Syst. Tour. 1–70. https://doi.org/10.1300/J052v11n03_03

Gu, B., Park, J., Konana, P., 2012. The Impact of External Word - of - Mouth Sources on Retailer Sales of High - Involvement Products. Inf. Syst. Res. 23, 182–196. https://doi.org/10.2307/23207880

Gu, L., Baxter, R., Vickers, D., Rainsford, C., 2003. Record Linkage: Current Practice and Future Directions. Tech. Rep.

Gudivada, V., Apon, A., Ding, J., 2017. Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations. Int. J. Adv. Softw. 10, 1–20.

Gupta, A.K., Gupta, C., 2010. Analyzing Customer Behavior using Data Mining Techniques: Optimizing Relationships with Customer. Manag. Insight VI, 92–98.

Hawkins, D.I., Best, R.J., Coney, K.A., 2014. Comportamiento del consumidor: repercusiones en la estrategía de marketing. Reper. en la estrategía Mark. 1–3.

Hawkins, D.I., Mothersbaugh, D.L., 2012. Consumer Behavior: Building Marketing Strategy, 12th Editi. ed. McGraw-Hill Education.

Helfat, C.E., 2006. Open Innovation: The New Imperative for Creating and Profiting from Technology. Acad. Manag. Perspect. 20, 86–88. https://doi.org/10.5465/AMP.2006.20591014

Helmers, C., Rogers, M., 2011. Does patenting help high-tech start-ups? Res. Policy 40, 1016–1027. https://doi.org/10.1016/j.respol.2011.05.003

Hennig-Thurau, T., Gwinner, K.P., Walsh, G., Gremler, D.D., Lis, B., Korchmar, S., 2004. Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? J. Interact. Mark. 18, 38–52. https://doi.org/10.1002/dir.10073

Henry, O., 2016. Commercial General Liability Insurance and Coverage : A Theoretical Review 5, 509–517.

Herschel, M., 2008. Space and Time Scalability of Duplicate Detection in Graph Data.

Hobey, E., 2017. Next Insurance Introduces New Portal for SMEs, Promises Transparency [WWW Document]. URL https://www.crowdfundinsider.com/2017/09/121835-next-insurance-introduces-new-portal-smes-promises-transparency/ (accessed 8.1.17).

Hu, J., Zeng, H., Li, H., Niu, C., Chen, Z., 2007. Demographic Prediction Based on User's Browsing Behavior, in: WWW 2006, Data Mining. pp. 151–160.

Hu, N., Pavlou, P., Zhang, J., 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. Proc. 7th ACM Conf. Electron. Commer. 324–330. https://doi.org/10.1145/1134707.1134743

Hunt, J., Garant, J., Herman, H., Munroe, D.J., 2013. Why are women underrepresented amongst patentees? Res. Policy 42, 831–843. https://doi.org/10.1016/j.respol.2012.11.004

Hussain Naqvi, S.W., 2011. Critical Success and Failure Factors of Entrepreneurial Organizations: Study of SMEs in Bahawalpur. J. Public Adm. Gov. 1, 96–100. https://doi.org/10.5296/jpag.v1i2.824

Hutner, M., 2018. Die beliebtesten schweizer Nachnamen [WWW Document]. URL https://flugzentrale.de/1905/blog/swiss-surnames

Jardin, P. du, Séverin, E., 2010. Dynamic analysis of the business failure process: a study of bankruptcy trajectories. Munich Pers. RePEc.

Jiménez-Jiménez, D., Sanz-Valle, R., 2011. Innovation, organizational learning, and performance. J. Bus. Res. 64, 408–417. https://doi.org/10.1016/j.jbusres.2010.09.010

Johnson, M.P., Midgley, G., Chichirau, G., 2017. Emerging trends and new frontiers in community operational research. Eur. J. Oper. Res. 268, 1178–1191. https://doi.org/10.1016/j.ejor.2017.11.032

Jones, B.F., 2017. The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder? Rev. Econ. Stud. 76, 283–317.

Jordan, M.I., Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science (80-.). 349, 255–260. https://doi.org/10.1126/science.aaa8415

Jung, T., Ejermo, O., 2014. Technological Forecasting & Social Change Demographic patterns and trends in patenting: Gender, age, and education of inventors. Technol. Forecast. Soc. Chang. 86, 110–124. https://doi.org/10.1016/j.techfore.2013.08.023

Kirkpatrick, D., 2005. Why There's no escaping the Blog. January 10–12.

Klein, L., 2011. Evaluating the Potential of Interactive Media through a New Lens: Search versus Experience Goods. J. Travel Res. 23, 1–3. https://doi.org/10.1287/mksc.1040.0071

Klein, L.R., 1998. Evaluating the Potential of Interactive Media through a New Lens: Search versus Experience Goods. J. Bus. Res. 41, 195–203. https://doi.org/10.1016/S0148-2963(97)00062-3

König, C.J., Kleinmann, M., 2004. Business before pleasure: No strategy for procrastinators? Pers. Individ. Dif. 37, 1045–1057. https://doi.org/10.1016/j.paid.2003.11.013

Kuenzel, S., Halliday, S.V., 2010. The chain of effects from reputation and brand personality congruence to brand loyalty: The role of brand identification. J. Targeting, Meas. Anal. Mark. 18, 167–176. https://doi.org/10.1057/jt.2010.15

Kurt, I., Ture, M., Kurum, A.T., 2008. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. Expert Syst. Appl. 34, 366–374. https://doi.org/10.1016/j.eswa.2006.09.004

Kuzu, M., Kantarcioglu, M., Durham, E., Malin, B., 2011. A constraint satisfaction cryptanalysis of

bloom filters in private record linkage. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 6794 LNCS, 226–245. https://doi.org/10.1007/978-3-642-22263-4_13

Landwehr, N., Hall, M., Frank, E., 2005. Logistic model trees. Mach. Learn. 59, 161–205. https://doi.org/10.1007/s10994-005-0466-3

Lanjouw, J.O., Schankerman, M., 2004. Protecting Intellectual Property Rights: Are Small Firms Handicapped? J. Law Econ. 47, 45–74. https://doi.org/10.1086/380476

Law, G., 2017. Entry, Gibrat's Law, Innovation, and the Growth of Firms 2–3.

Lever, J., Krzywinski, M., Altman, N., 2016. Classification evaluation. Nat. Publ. Gr. 13, 603–605. https://doi.org/10.1038/nmeth.3945

Levin, R.C., Klevorick, A.K., Nelson, R.R., Winter, S.G., 2013. Appropriating the returns from industrial research and development. Compet. Policy Int. 9, 160–196. https://doi.org/10.2307/2534454

Lewis, R.C., Chambers, R.E., 2000. Marketing Leadership in Hospitality: Foundations and Practices, 3rd Revise. ed. John Wiley & Sons.

Li, H., Sun, J., 2010. Forecasting Business Failure in China Using Case-Based Reasoning with Hybrid Case Representation. J. Forecast. 29, 486–501. https://doi.org/10.1002/for

Li, P., 2012. Robust logitboost and adaptive base class (abc) logitboost. arXiv Prepr. arXiv1203.3491 1–30.

Lin, Y.Y., 2018. #DeleteFacebook is still feeding the beast – but there are ways to overcome surveillance capitalism. Conversat. 1–4.

Litvin, S.W., Goldsmith, R.E., Pan, B., 2008. Electronic word-of-mouth in hospitality and tourism management. Tour. Manag. 29, 458–468. https://doi.org/10.1016/j.tourman.2007.05.011

Liu, Y., 2006. Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue. J. Mark. 70, 74–89. https://doi.org/10.1509/jmkg.70.3.74

Loren, A., 2017. What Global Executives Think About Growth & Risk, D&B's, McKinsey Quarterly 2005.

Ludwig, C., Cavalli, S., Oris, M., 2014. "Vivre / Leben / Vivere": An interdisciplinary survey addressing progress and inequalities of aging over the past 30 years in Switzerland. Arch. Gerontol. Geriatr. 59, 240–248. https://doi.org/10.1016/j.archger.2014.04.004

Lussier, R., 1996. A business success versus failure prediction model for service industries. J. Bus. Entrep. 8, 23–37.

Lyon, D.W., Ferrier, W.J., 2002. Enhancing performance with product-market innovation: The influence of the top management team. J. Manag. Issues 14, 452–469. https://doi.org/Article

Mann, R.J., Sager, T.W., 2007. Patents, venture capital, and software start-ups. Res. Policy 36, 193–208. https://doi.org/10.1016/j.respol.2006.10.002

Marc Sumner, E.F. and M.H., 2005. Speeding Up Logistic Model Tree Induction. Lect. Notes Comput. Sci. 3721, 675–683. https://doi.org/10.1007/11564126_72

Marom, S., Lussier, R.N., 2014. A Business Success Versus Failure Prediction Model for Small Businesses in Israel. Bus. Econ. Res. 4, 63. https://doi.org/10.5296/ber.v4i2.5997

Mau, S., Mueller, D., Cvijikj, I., Wagner, J., 2016. Anticipating insurance customers' next likely purchase events, in: IPAG Business School. pp. 1–9.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, The American Statistician. https://doi.org/10.1080/00031305.1994.10476073

McKinsey and Company, 2016. Small Commercial Insurance: A Bright Spot In the U.S. Property-Casualty Market.

McKinsey and Company, 2013. Winning Share and Customer Loyalty in Auto Insurance.

Melián-González, S., Bulchand-Gidumal, J., 2016. A model that connects information technology and hotel performance. Tour. Manag. 53, 30–37. https://doi.org/10.1016/j.tourman.2015.09.005

Melo-Martı, I. de, 2013. Patenting and the Gender Gap: Should Women Be Encouraged to Patent More? Sci Eng Ethics 19, 491–504. https://doi.org/10.1007/s11948-011-9344-5

Morgan, N.A., 2012. Marketing and business performance. J. Acad. Mark. Sci. 40, 102–119. https://doi.org/10.1007/s11747-011-0279-9

Morgan, R.E., Strong, C.A., 2003. Business Performance and Dimensions of Strategic Orientation. J. Bus. Res. 56, 163–176.

Mowery, D.C., 2017. Industrial Research and Firm Size, Survival, and Growth in American Manufacturing 43, 953–980.

Nyffenegger, R., 2018. Statistiken über das Telefonbuch der Schweiz [WWW Document]. URL http://www.adp-gmbh.ch/misc/tel_book_ch.html

O'Conor, P., Öpken, W., Gretzel, U., 2008. Information and Communication Technologies in Tourism 2008, Information and Communication Technologies in Tourism 2008. https://doi.org/10.1017/CBO9781107415324.004

Park, S., Nicolau, J.L., 2015. Asymmetric effects of online consumer reviews. Ann. Tour. Res. https://doi.org/10.1016/j.annals.2014.10.007

Pavlou, P.A., Dimoka, A., 2006. The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums and seller differentiation. Inf. Syst. Res. 17, 392–414. https://doi.org/10.1287/isre.1060.0106

Phillips, C., 2003. How do consumers express their identity through the choice of products that they buy ? Manag. Work. Pap. Ser. 1–20.

Pipino, L.L., Lee, Y.W., Wang, R.Y., Lowell Yang Lee, M.W., Yang, R.Y., 2002. Data Quality Assessment. Commun. ACM 45, 211. https://doi.org/10.1145/505248.506010

Pome, P.P., Bilderbeek, J., 2014. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. J. Bus. Ventur. 34, 127–142. https://doi.org/10.1016/j.eswa.2006.09.004

Prajogo, D., Toy, J., Bhattacharya, A., Oke, A., Cheng, T.C.E., 2018. The relationships between information management, process management and operational performance: Internal and external contexts. Int. J. Prod. Econ. 199, 95–103. https://doi.org/10.1016/j.ijpe.2018.02.019

Prashanth, N., 1997. A Study of the Relationships between Cognitive Appraisals and Consumption Emotions. J. Acad. Mark. Sci. 2–3. https://doi.org/10.1177/009207039001800305

Qin, Y., Zhang, S., Zhu, X., Zhang, J., Zhang, C., 2007. Semi-parametric optimization for missing data imputation. Appl. Intell. 27, 79–88. https://doi.org/10.1007/s10489-006-0032-0

Racherla, P., Friske, W., 2012. Perceived "usefulness" of online consumer reviews: An exploratory investigation across three services categories. Electron. Commer. Res. Appl. 11, 548–559. https://doi.org/10.1016/j.elerap.2012.06.003

Raghavan, V., Bollmann, P., Jung, G.S., 1989. A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans. Inf. Syst. 7, 205–229. https://doi.org/10.1145/65943.65945

Reimer, K., Becker, J.U., 2015. What customer information should companies use for customer relationship management? Practical insights from empirical research, Management Review Quarterly. Springer Berlin Heidelberg. https://doi.org/10.1007/s11301-014-0110-z

Rensink, J.M., 2013. What motivates people to write online reviews and which role does personality play ? University of Twente.

Roper, S., 1997. Product Innovation and Small Business Growth: A Comparison of the Strategies of German, U. K. and Irish Companies. Small Bus. Econ. 523–537. https://doi.org/10.1023/A:1007963604397

Rust, R.T., Moorman, C., van Beuningen, J., 2016. Quality mental model convergence and business performance. Int. J. Res. Mark. 33, 155–171. https://doi.org/10.1016/j.ijresmar.2015.07.005

Saha, B., Srivastava, D., 2014. Data quality: The other face of Big Data. Proc. - Int. Conf. Data Eng. 1294–1297. https://doi.org/10.1109/ICDE.2014.6816764

Salavou, H., Baltas, G., Lioukas, S., 2004. Organisational innovation in SMEs. Eur. J. Mark. 38, 1091–1112. https://doi.org/10.1016/B978-0-12-375678-7.01110-X

Samitsch, C., 2015. Data quality and its impacts on decision-making: How managers can benefit from good data. Data Qual. its Impacts Decis. How Manag. can Benefit from Good Data 1–59. https://doi.org/10.1007/978-3-658-08200-0

Sattler, K.-U., Schallehn, E., 2001. A Data Preparation Framework based on a Multidatabase Language 219–228.

Scherer, F.M., 1965. Corporate Inventive Output, Profits, and Growth. J. Polit. Econ.

Schlieker, A., 2014. Das Wettrennen um die Gunst des neuen Kunden.

Schwarzkopf, S., Gujan, S., 2017. Branchenspiegel 2017. Zürich.

Schweizer Tourismusverband, 2016. Schweizer Tourismus in Zahlen 2013 Struktur- und Branchendaten.

Schweizerische Mobiliar Holding AG, 2016. Geschäftsbericht 2015 - Die Mobiliar.

Senecal, S., Nantel, J., 2004a. The influence of online product recommendations on consumers' online choices. J. Retail. 159–169. https://doi.org/10.1017/CBO9781107415324.004

Senecal, S., Nantel, J., 2004b. The influence of online product recommendations on consumers' online choices. J. Retail. 80, 159–169. https://doi.org/10.1016/j.jretai.2004.04.001

Serra Cantallops, A., Salvi, F., 2014. New consumer behavior: A review of research on eWOM and hotels. Int. J. Hosp. Manag. 36, 41–51. https://doi.org/10.1016/j.ijhm.2013.08.007

Shasha, D., Galhardas, H., Saita, C., Simon, E., Rocquencourt, I., 1996. Improving Data Cleaning Quality using a Data Lineage Facility. Informatica 1–13.

Shmueli, G., Koppius, O., 2010. Predictive analytics in information systems research. Robert H. Smith Sch. Res. Pap. No. RHS 35, 6–138.

Sidik, I.G., 2012. Conceptual Framework of Factors Affecting SME Development: Mediating Factors on the Relationship of Entrepreneur Traits and SME Performance. Procedia Econ. Financ. 4, 373–383. https://doi.org/10.1016/S2212-5671(12)00351-6

Simpson, M., Padmore, J., Newman, N., 2013. Towards a new model of success and performance in SMEs. Int. J. Entrep. Behav. Res. 18, 264–285. https://doi.org/http://dx.doi.org/10.1108/MRR-09-2015-0216

Smith, K.A., Willis, R.J., Brooks, M., 2000. An analysis of customer retention and insurance claim patterns using data mining: A case study. J. Oper. Res. Soc. 51, 532–541. https://doi.org/10.1057/palgrave.jors.2600941

Sparks, B.A., Browning, V., 2011. The impact of online reviews on hotel booking intentions and perception of trust. Tour. Manag. 32, 1310–1323. https://doi.org/10.1016/j.tourman.2010.12.011

Stauss, B., 2000. Using New Media for Customer Interaction: A Challenge for Relationship Marketing. Relatsh. Mark. 233–253. https://doi.org/10.1017/CBO9781107415324.004

Stone, M., 1974. Cross-validation and multinomial prediction. Biometrika 61, 509–515. https://doi.org/10.1093/biomet/61.3.509

Sugimoto, C.R., Ni, C., West, J.D., Larivi, V., 2015. The Academic Advantage : Gender Disparities in Patenting 1–10. https://doi.org/10.1371/journal.pone.0128000

Sultan, M.U., Uddin, N., 2011. Consumers' Attitude towards Online Shopping.

Teng, H.S.S., Bhatia, G.S., Anwar, S., 2011. A success versus failure prediction model for small businesses in Singapore. Am. J. Bus. 26, 50–64. https://doi.org/10.1108/19355181111124106

Torres, E.N., Singh, D., Robertson-Ring, A., 2015. Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry. Int. J. Hosp. Manag. 50, 77–83. https://doi.org/10.1016/j.ijhm.2015.07.012

Tuominen, P., 2011. The influence of tripadvisor consumer-generated travel reviews on hotel performance. 19th Annu. Front. Serv. Conf. 1–11.

Van Doorn, S., Jansen, J.J.P., Van Den Bosch, F.A.J., Volberda, H.W., 2013. Entrepreneurial orientation and firm performance: Drawing attention to the senior team. J. Prod. Innov. Manag. 30, 821–836. https://doi.org/10.1111/jpim.12032

Venter, P., Wright, A., Dibb, S., 2015. Performing market segmentation: a performative perspective. J. Mark. Manag. 31, 62–83. https://doi.org/10.1080/0267257X.2014.980437

Verhoef, P.C., Venkatesan, R., McAlister, L., Malthouse, E.C., Krafft, M., Ganesan, S., 2010. CRM in data-rich multichannel retailing environments: A review and future research directions. J. Interact. Mark. 24, 121–137. https://doi.org/10.1016/j.intmar.2010.02.009

von Wangenheim, F., Bayon, T., 2004. The effect of word of mouth on services switching. Measurement and moderating variables. Eur. J. Mark. 9/10, 1173–1185. https://doi.org/http://dx.doi.org/10.1108/MRR-09-2015-0216

Vyncke, P., 2002. Lifestyle segmentation: From attitudes, interests and opinions, to values, aesthetic styles, life visions and media preferences. Eur. J. Commun. 17, 445–463. https://doi.org/10.1177/0267323102017004030 1

Wagner, S., Cockburn, I., 2010. Patents and the survival of Internet-related IPOs. Res. Policy 39, 214–228. https://doi.org/10.1016/j.respol.2009.12.003

Wang, R.Y., Strong, D.M., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manag. Inf. Syst. 12, 5–33. https://doi.org/10.1080/07421222.1996.11518099

Weber, K., 2007. Das Schweizer Gastgewerbe : Eine Branche im Wandel 42–46.

Wedel, M., Kannan, P.K., 2016. Marketing Analytics for Data-Rich Environments. J. Mark. 80, 97–121.

https://doi.org/10.1509/jm.15.0413

Weinstein, A., 2017. Handbook of Market Research. https://doi.org/10.1007/978-3-319-05542-8

Weiser, B.E., 2000. Gender Differences in Internet Use Patterns and Internet Application Preferences: A Two-Sample Comparison. CyberPsychology Behav. https://doi.org/10.1089/109493100316012

Wesley M. Cohen Richard R. Nelson John P., 2000. Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firmspatent (or Not). Nber Work. Pap. Ser.

Westhead, P., Wright, M., Ucbasaran, D., 2001. The internationalization of new and small firms: a resource-based view. J. Bus. Ventur. 16, 333–358. https://doi.org/10.1016/S0883-9026(99)00063-4

Williams, M.L., Levi, M., Burnap, P., Gundur, R. V, Williams, M.L., Levi, M., Burnap, P., Under, R.V.G., 2018. Under the Corporate Radar : Examining Insider Business Cybercrime Victimization through an Application of Routine Activities Theory. Deviant Behav. 0, 1–13. https://doi.org/10.1080/01639625.2018.1461786

Williams, N., Zander, S., Armitage, G., 2006. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification. ACM SIGCOMM Comput. Commun. Rev. 36, 5. https://doi.org/10.1145/1163593.1163596

Wind, Y., 1978. Issues and Advances in Segmentation Research. J. Mark. Res. 15, 317. https://doi.org/10.2307/3150580

Wirtz, J., Chew, P., 1992. The effect of incentives, deal proneess, satisfaction and tie strenght on word-of-mouth behaviour. Int. J. Serv. Ind. Manag. 3, 57–69. https://doi.org/10.1108/09564230310474165 Hierarchical

Wu, C.-H., Kao, S.-C., Su, Y.-Y., Wu, C.-C., 2005. Targeting customers via discovery knowledge for the insurance industry. Expert Syst. Appl. 29, 291–299. https://doi.org/10.1016/j.eswa.2005.04.002

Xie, K.L., Zhang, Z., Zhang, Z., 2014. The business value of online consumer reviews and management response to hotel performance. Int. J. Hosp. Manag. 43, 1–12. https://doi.org/10.1016/j.ijhm.2014.07.007

Xue-wui, S., Gui-hua, N.I.E., Ling, S., 2000. Gender-Based Differences in the Effect of Web Advertising in E-business.

Yacouel, N., Fleischer, A., 2012. The Role of Cybermediaries in Reputation Building and Price Premiums in the Online Hotel Market. J. Travel Res. 2–3. https://doi.org/10.1177/0047287510382296

Ye, Q., Law, R., Gu, B., 2009. The impact of online user reviews on hotel room sales. Int. J. Hosp. Manag. 28, 180–182. https://doi.org/10.1016/j.ijhm.2008.06.011

Ye, Q., Law, R., Gu, B., Chen, W., 2011. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. Comput. Human Behav. 27, 634–639. https://doi.org/10.1016/j.chb.2010.04.014

Yoo, K.H., Gretzel, U., 2011. Influence of personality on travel-related consumer-generated media creation. Comput. Human Behav. 27, 609–621. https://doi.org/10.1016/j.chb.2010.05.002

Zhang, Z., Ye, Q., Law, R., Li, Y., 2010. The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. Int. J. Hosp. Manag. 29, 694–700. https://doi.org/10.1016/j.ijhm.2010.02.002

Zhou, Z., Chawla, N. V, Jin, Y., Williams, G.J., Office, A.T., 2014. Big Data Opportunities and

Challenges: Discussions from Data Analytics Perspectives 62–74.

Zhu, F., Zhang, X., 2010. Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. J. Mark. 74, 133–148. https://doi.org/10.1509/jmkg.74.2.133

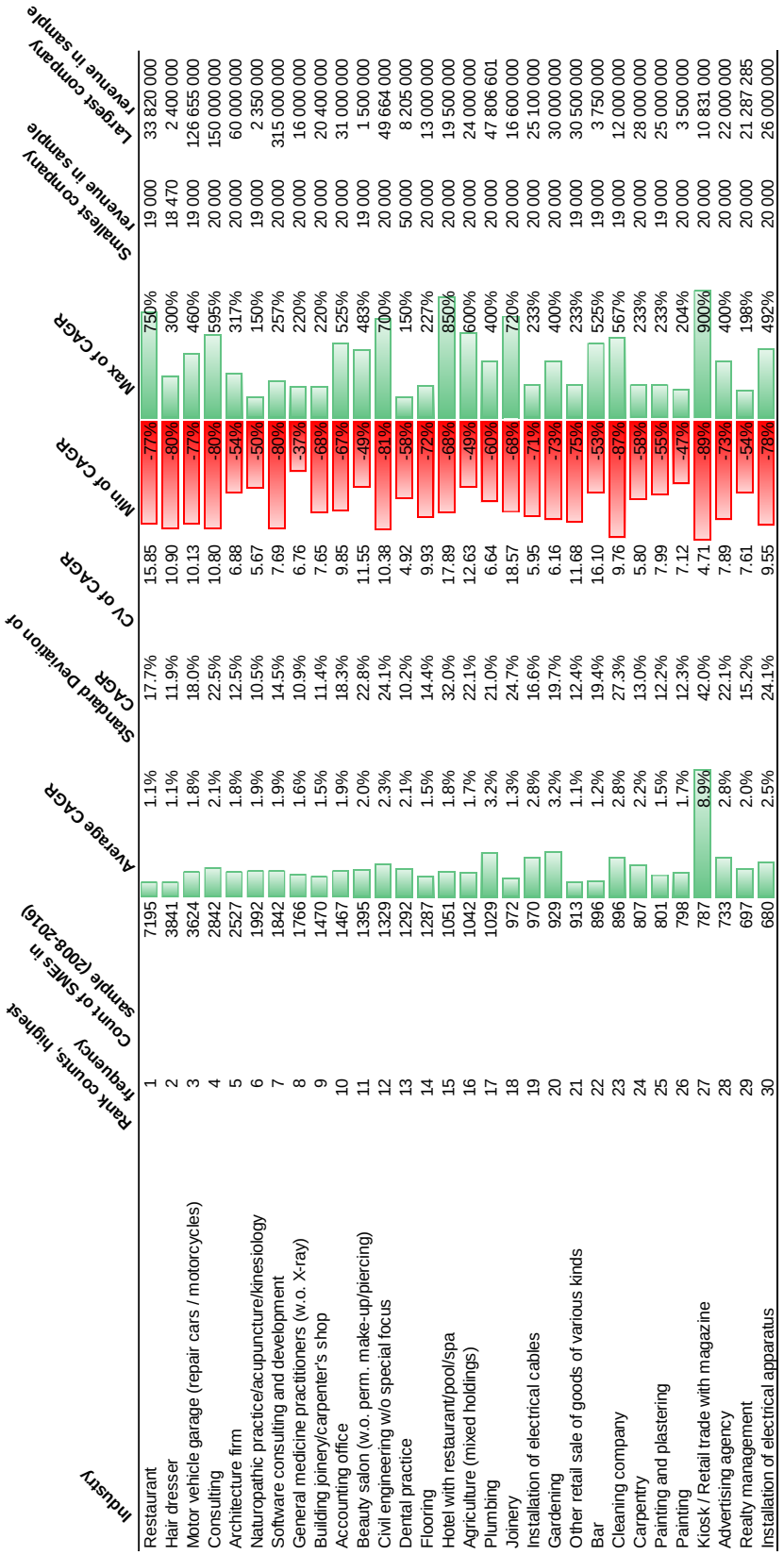
Appendix

Appendix 1 – Employment in Switzerland (2013)

Size	Numer of SMEs	%	Employees	% of total Employees
1 employee	291 293	50.05%	291 293	6.64%
2 employees	84 751	14.56%	169 502	3.87%
3-4 employees	82 410	14.16%	278 990	6.36%
5-9 employees	63 926	10.98%	413 572	9.43%
10-19 employees	31 633	5.44%	422 247	9.63%
20-49 employees	17 497	3.01%	523 451	11.94%
50-99 employees	5 689	0.98%	391 923	8.94%
100-199 employees	2 659	0.46%	366 906	8.37%
200-249 employees	533	0.09%	118 870	2.71%
Subtotal of SMEs (1-249)	**580 391**	**99.73%**	**2 976 754**	**67.90%**
250-499 employees	894	0.15%	302 486	6.90%
500-999 employees	383	0.07%	263 904	6.02%
>1000 employees	286	0.05%	841 077	19.18%
Total of SMEs	**581 954**	**100.00%**	**4 384 221**	**100%**
Micro enterprises (1-9)	522 380	89.76%	1 153 357	26.31%
Small enterprices (10-49)	49 130	8.44%	945 698	21.57%
Medium enterprices (50-249)	8 881	1.53%	877 699	20.02%
Large enterprises (>250)	1 563	0.27%	1 407 467	32.10%
Total of SMEs	**581 954**	**100.00%**	**4 384 221**	**100%**

Appendix 2 – Most common SMEs by industry

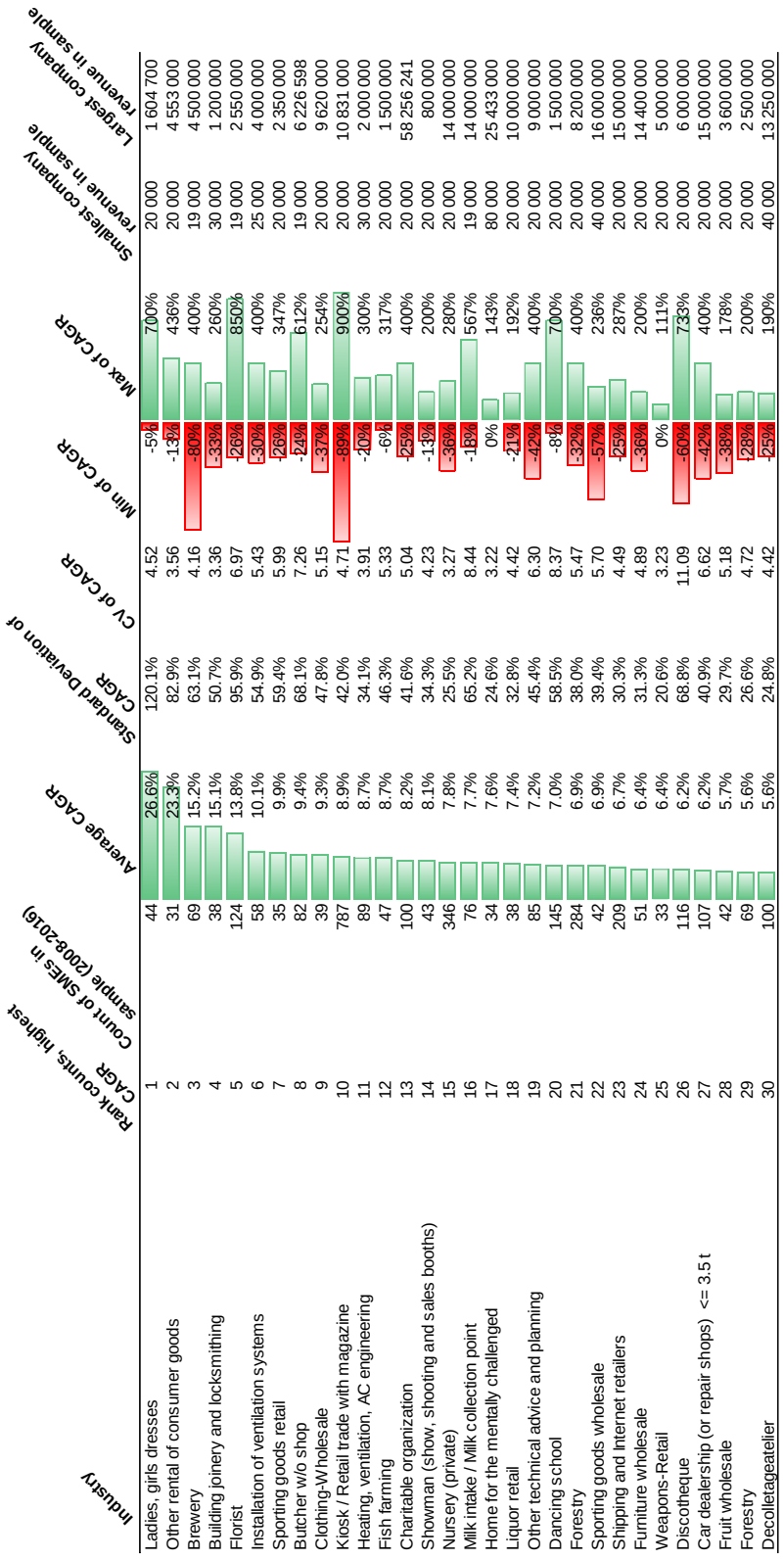
(2008-2016)



Industry	Rank counts, highest frequency	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR	CV of CAGR	Min of CAGR	Max of CAGR	Smallest company revenue in sample	Largest company revenue in sample
Restaurant	1	7195	1.1%	17.7%	15.85	-77%	750%	19 000	33 820 000
Hair dresser	2	3841	1.1%	11.9%	10.90	-80%	300%	18 470	2 400 000
Motor vehicle garage (repair cars / motorcycles)	3	3624	1.8%	18.0%	10.13	-77%	460%	19 000	126 655 000
Consulting	4	2842	2.1%	22.5%	10.80	-80%	595%	20 000	150 000 000
Architecture firm	5	2527	1.8%	12.5%	6.88	-54%	317%	20 000	60 000 000
Naturopathic practice/acupuncture/kinesiology	6	1992	1.9%	10.5%	5.67	-50%	150%	19 000	2 350 000
Software consulting and development	7	1842	1.9%	14.5%	7.69	-80%	257%	20 000	315 000 000
General medicine practitioners (w.o. X-ray)	8	1766	1.6%	10.9%	6.76	-37%	220%	20 000	16 000 000
Building joinery/carpenter's shop	9	1470	1.5%	11.4%	7.65	-68%	220%	20 000	20 400 000
Accounting office	10	1467	1.9%	18.3%	9.85	-67%	525%	20 000	31 000 000
Beauty salon (w.o. perm. make-up/piercing)	11	1395	2.0%	22.8%	11.55	-49%	483%	19 000	1 500 000
Civil engineering w/o special focus	12	1329	2.3%	24.1%	10.38	-81%	700%	20 000	49 664 000
Dental practice	13	1292	2.1%	10.2%	4.92	-58%	150%	50 000	8 205 000
Flooring	14	1287	1.5%	14.4%	9.93	-72%	227%	20 000	13 000 000
Hotel with restaurant/pool/spa	15	1051	1.8%	32.0%	17.89	-68%	850%	20 000	19 500 000
Agriculture (mixed holdings)	16	1042	1.7%	22.1%	12.63	-49%	600%	20 000	24 000 000
Plumbing	17	1029	3.2%	21.0%	6.64	-60%	400%	20 000	47 806 601
Joinery	18	972	1.3%	24.7%	18.57	-68%	720%	20 000	16 600 000
Installation of electrical cables	19	970	2.8%	16.6%	5.95	-71%	233%	20 000	25 100 000
Gardening	20	929	3.2%	19.7%	6.16	-73%	400%	20 000	30 000 000
Other retail sale of goods of various kinds	21	913	1.1%	12.4%	11.68	-75%	233%	19 000	30 500 000
Bar	22	896	1.2%	19.4%	16.10	-53%	525%	19 000	3 750 000
Cleaning company	23	896	2.8%	27.3%	9.76	-87%	567%	19 000	12 000 000
Carpentry	24	807	2.2%	13.0%	5.80	-58%	233%	20 000	28 000 000
Painting and plastering	25	801	1.5%	12.2%	7.99	-55%	233%	19 000	25 000 000
Painting	26	798	1.7%	12.3%	7.12	-47%	204%	20 000	3 500 000
Kiosk / Retail trade with magazine	27	787	8.9%	42.0%	4.71	-89%	900%	20 000	10 831 000
Advertising agency	28	733	2.8%	22.1%	7.89	-73%	400%	20 000	22 000 000
Realty management	29	697	2.0%	15.2%	7.61	-54%	198%	20 000	21 287 285
Installation of electrical apparatus	30	680	2.5%	24.1%	9.55	-78%	492%	20 000	26 000 000

n = 111 236

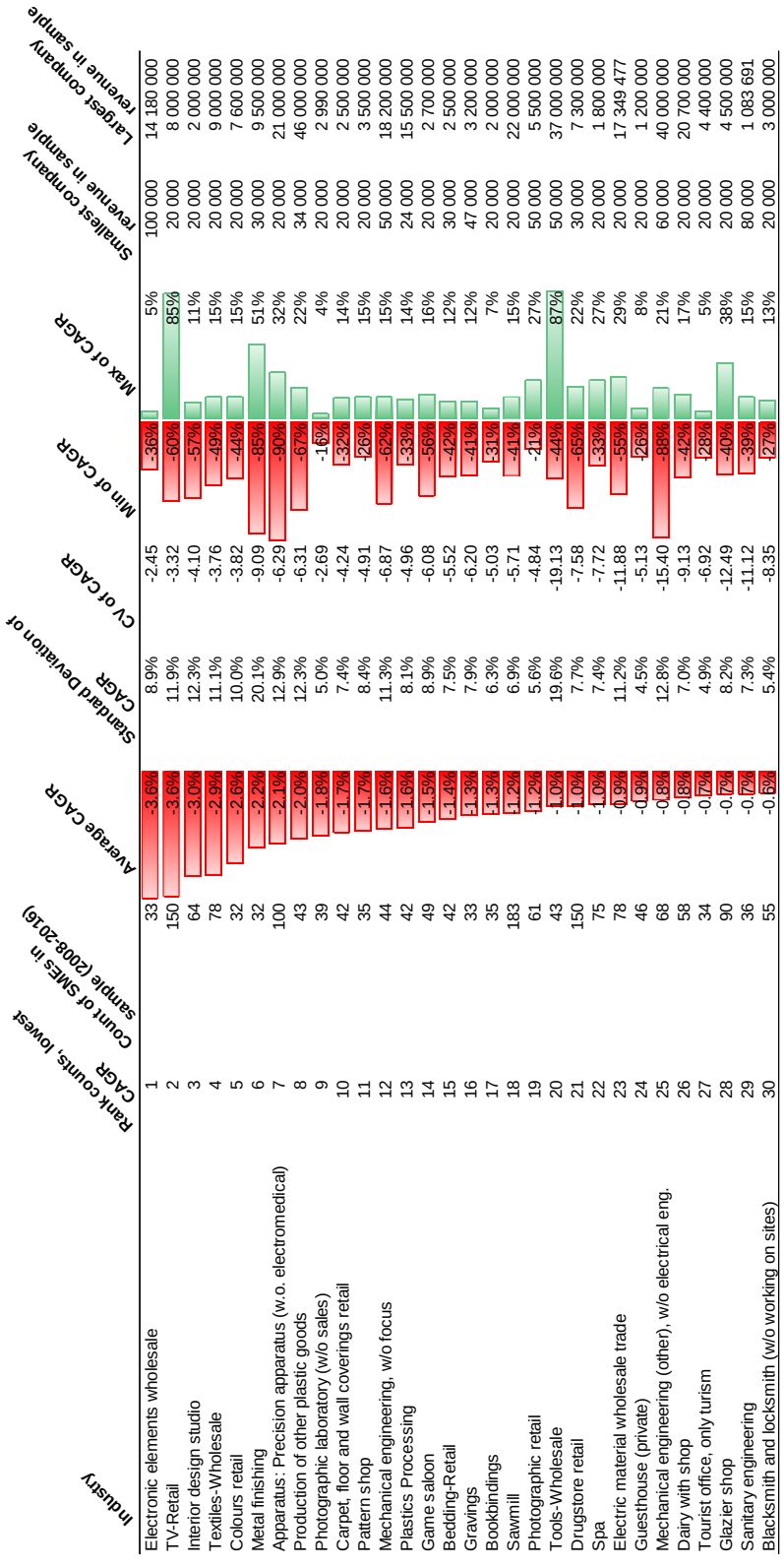
Appendix 3 – Highest average CAGR by industry with more than 30 observations (2008-2016)



Industry	Rank counts, highest CAGR	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR	CV of CAGR	Min of CAGR	Max of CAGR	Smallest company revenue in sample	Largest company revenue in sample
Ladies, girls dresses	1	44	26.6%	120.1%	4.52	-5%	700%	20 000	1 604 700
Other rental of consumer goods	2	31	23.3%	82.9%	3.56	-13%	436%	20 000	4 553 000
Brewery	3	69	15.2%	63.1%	4.16	-80%	400%	19 000	4 500 000
Building joinery and locksmithing	4	38	15.1%	50.7%	3.36	-33%	260%	30 000	1 200 000
Florist	5	124	13.8%	95.9%	6.97	-26%	850%	19 000	2 550 000
Installation of ventilation systems	6	58	10.1%	54.9%	5.43	-30%	400%	25 000	4 000 000
Sporting goods retail	7	35	9.9%	59.4%	5.99	-26%	347%	20 000	2 350 000
Butcher w/o shop	8	82	9.4%	68.1%	7.26	-21%	612%	19 000	6 226 598
Clothing-Wholesale	9	39	9.3%	47.8%	5.15	-37%	254%	20 000	9 620 000
Kiosk / Retail trade with magazine	10	787	8.9%	42.0%	4.71	-89%	900%	20 000	10 831 000
Heating, ventilation, AC engineering	11	89	8.7%	34.1%	3.91	-20%	300%	30 000	2 000 000
Fish farming	12	47	8.7%	46.3%	5.33	-6%	317%	20 000	1 500 000
Charitable organization	13	100	8.2%	41.6%	5.04	-25%	400%	20 000	58 256 241
Showman (show, shooting and sales booths)	14	43	8.1%	34.3%	4.23	-13%	200%	20 000	800 000
Nursery (private)	15	346	7.8%	25.5%	3.27	-36%	280%	20 000	14 000 000
Milk intake / Milk collection point	16	76	7.7%	65.2%	8.44	-18%	567%	19 000	14 000 000
Home for the mentally challenged	17	34	7.6%	24.6%	3.22	0%	143%	80 000	25 433 000
Liquor retail	18	38	7.4%	32.8%	4.42	-21%	192%	20 000	10 000 000
Other technical advice and planning	19	85	7.2%	45.4%	6.30	-42%	400%	20 000	9 000 000
Dancing school	20	145	7.0%	58.5%	8.37	-8%	700%	20 000	1 500 000
Forestry	21	284	6.9%	38.0%	5.47	-32%	400%	20 000	8 200 000
Sporting goods wholesale	22	42	6.9%	39.4%	5.70	-57%	236%	40 000	16 000 000
Shipping and Internet retailers	23	209	6.7%	30.3%	4.49	-25%	287%	20 000	15 000 000
Furniture wholesale	24	51	6.4%	31.3%	4.89	-36%	200%	20 000	14 400 000
Weapons-Retail	25	33	6.4%	20.6%	3.23	0%	111%	20 000	5 000 000
Discotheque	26	116	6.2%	68.8%	11.09	-60%	733%	20 000	6 000 000
Car dealership (or repair shops) <= 3.5 t	27	107	6.2%	40.9%	6.62	-42%	400%	20 000	15 000 000
Fruit wholesale	28	42	5.7%	29.7%	5.18	-38%	178%	20 000	3 600 000
Forestry	29	69	5.6%	26.6%	4.72	-28%	200%	20 000	2 500 000
Decolletageatelier	30	100	5.6%	24.8%	4.42	-25%	190%	40 000	13 250 000

n = 111 236

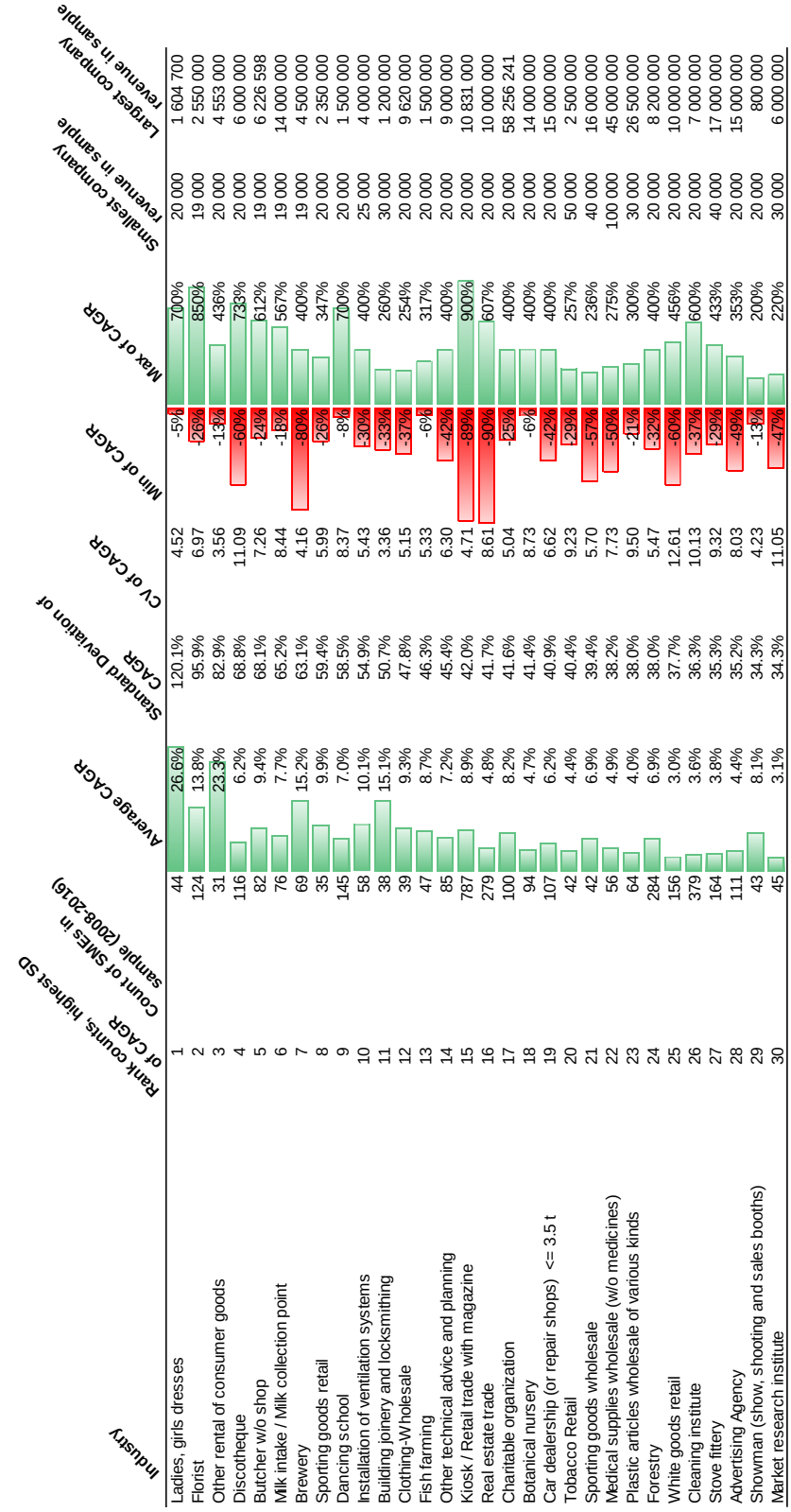
Appendix 4 – Lowest average CAGR by industry with more than 30 observations (2008-2016)



Industry	Rank counts, lowest CAGR	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR	CV of CAGR	Min of CAGR	Max of CAGR	Smallest company revenue in sample	Largest company revenue in sample
Electronic elements wholesale	1	33	-3.6%	8.9%	-2.45	-36%	5%	100 000	14 180 000
TV-Retail	2	150	-3.6%	11.9%	-3.32	-60%	85%	20 000	8 000 000
Interior design studio	3	64	-3.0%	12.3%	-4.10	-57%	11%	20 000	2 000 000
Textiles-Wholesale	4	78	-2.9%	11.1%	-3.76	-49%	15%	20 000	9 000 000
Colours retail	5	32	-2.6%	10.0%	-3.82	-44%	15%	20 000	7 600 000
Metal finishing	6	32	-2.2%	20.1%	-9.09	-85%	51%	30 000	9 500 000
Apparatus: Precision apparatus (w.o. electromedical)	7	100	-2.1%	12.9%	-6.29	-90%	32%	20 000	21 000 000
Production of other plastic goods	8	43	-2.0%	12.3%	-6.31	-67%	22%	34 000	46 000 000
Photographic laboratory (w/o sales)	9	39	-1.8%	5.0%	-2.69	-16%	4%	20 000	2 990 000
Carpet, floor and wall coverings retail	10	42	-1.7%	7.4%	-4.24	-32%	14%	20 000	2 500 000
Pattern shop	11	35	-1.7%	8.4%	-4.91	-26%	15%	20 000	3 500 000
Mechanical engineering, w/o focus	12	44	-1.6%	11.3%	-6.87	-62%	15%	50 000	18 200 000
Plastics Processing	13	42	-1.6%	8.1%	-4.96	-33%	14%	24 000	15 500 000
Game saloon	14	49	-1.5%	8.9%	-6.08	-56%	16%	20 000	2 700 000
Bedding-Retail	15	42	-1.4%	7.5%	-5.52	-42%	12%	30 000	2 500 000
Gravings	16	33	-1.3%	7.9%	-6.20	-41%	12%	47 000	3 200 000
Bookbindings	17	35	-1.3%	6.3%	-5.03	-31%	7%	20 000	2 000 000
Sawmill	18	183	-1.2%	6.9%	-5.71	-41%	15%	20 000	22 000 000
Photographic retail	19	61	-1.2%	5.6%	-4.84	-21%	27%	50 000	5 500 000
Tools-Wholesale	20	43	-1.0%	19.6%	-19.13	-44%	87%	50 000	37 000 000
Drugstore retail	21	150	-1.0%	7.7%	-7.58	-65%	22%	30 000	7 300 000
Spa	22	75	-1.0%	7.4%	-7.72	-33%	27%	20 000	1 800 000
Electric material wholesale trade	23	78	-0.9%	11.2%	-11.88	-55%	29%	20 000	17 349 477
Guesthouse (private)	24	46	-0.9%	4.5%	-5.13	-26%	8%	20 000	1 200 000
Mechanical engineering (other), w/o electrical eng.	25	68	-0.8%	12.8%	-15.40	-88%	21%	60 000	40 000 000
Dairy with shop	26	58	-0.8%	7.0%	-9.13	-42%	17%	20 000	20 700 000
Tourist office, only turism	27	34	-0.7%	4.9%	-6.92	-28%	5%	20 000	4 400 000
Glazier shop	28	90	-0.7%	8.2%	-12.49	-40%	38%	20 000	4 500 000
Sanitary engineering	29	36	-0.7%	7.3%	-11.12	-39%	15%	80 000	1 083 691
Blacksmith and locksmith (w/o working on sites)	30	55	-0.6%	5.4%	-8.35	-27%	13%	20 000	3 000 000

n = 111 236

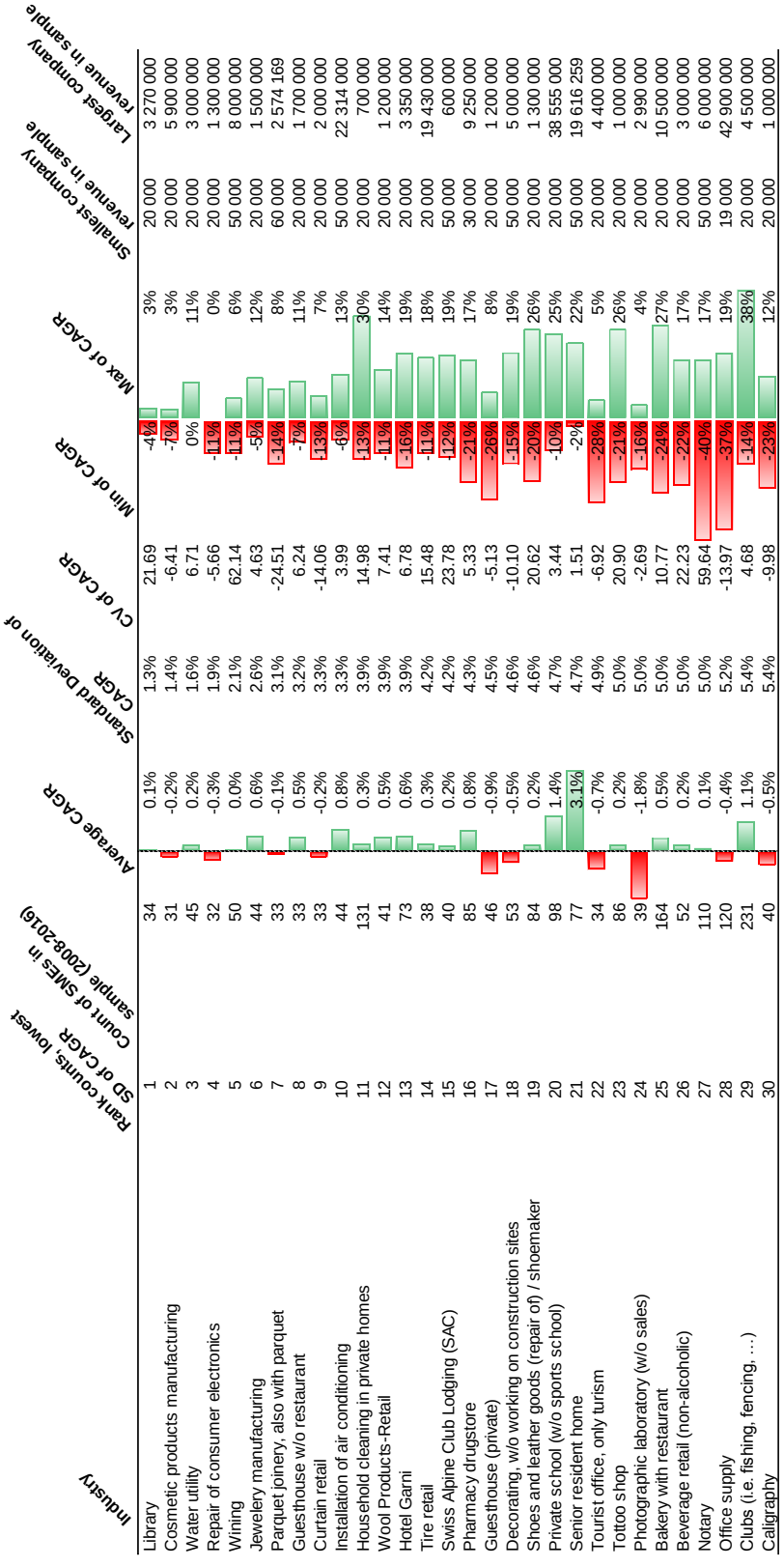
Appendix 5 – Highest average standard deviation with more than 30 observations (2008-2016)



Industry	Rank counts. highest SD of CAGR	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR	CV of CAGR	Min of CAGR	Max of CAGR	Smallest company revenue in sample	Largest company revenue in sample
Ladies, girls dresses	1	44	26.6%	120.1%	4.52	-5%	700%	20 000	1 604 700
Florist	2	124	13.8%	95.9%	6.97	-26%	850%	19 000	2 550 000
Other rental of consumer goods	3	31	23.3%	82.9%	3.56	-13%	436%	20 000	4 553 000
Discotheque	4	116	6.2%	68.8%	11.09	-60%	733%	20 000	6 000 000
Butcher w/o shop	5	82	9.4%	68.1%	7.26	-24%	612%	19 000	6 226 598
Milk intake / Milk collection point	6	76	7.7%	65.2%	8.44	-18%	567%	19 000	14 000 000
Brewery	7	69	15.2%	63.1%	4.16	-80%	400%	19 000	4 500 000
Sporting goods retail	8	35	9.9%	59.4%	5.99	-26%	347%	20 000	2 350 000
Dancing school	9	145	7.0%	58.5%	8.37	-8%	700%	20 000	1 500 000
Installation of ventilation systems	10	58	10.1%	54.9%	5.43	-30%	400%	25 000	4 000 000
Building joinery and locksmithing	11	38	15.1%	50.7%	3.36	-33%	260%	30 000	1 200 000
Clothing-Wholesale	12	39	9.3%	47.8%	5.15	-37%	254%	20 000	9 620 000
Fish farming	13	47	8.7%	46.3%	5.33	-6%	317%	20 000	1 500 000
Other technical advice and planning	14	85	7.2%	45.4%	6.30	-42%	400%	20 000	9 000 000
Kiosk / Retail trade with magazine	15	787	8.9%	42.0%	4.71	-89%	900%	20 000	10 831 000
Real estate trade	16	279	4.8%	41.7%	8.61	-90%	607%	20 000	10 000 000
Charitable organization	17	100	8.2%	41.6%	5.04	-25%	400%	20 000	58 256 241
Botanical nursery	18	94	4.7%	41.4%	8.73	-6%	400%	20 000	14 000 000
Car dealership (or repair shops) <= 3.5 t	19	107	6.2%	40.9%	6.62	-42%	400%	20 000	15 000 000
Tobacco Retail	20	42	4.4%	40.4%	9.23	-29%	257%	50 000	2 500 000
Sporting goods wholesale	21	42	6.9%	39.4%	5.70	-57%	236%	40 000	16 000 000
Medical supplies wholesale (w/o medicines)	22	56	4.9%	38.2%	7.73	-50%	275%	100 000	45 000 000
Plastic articles wholesale of various kinds	23	64	4.0%	38.0%	9.50	-21%	300%	30 000	26 500 000
Forestry	24	284	6.9%	38.0%	5.47	-32%	400%	20 000	8 200 000
White goods retail	25	156	3.0%	37.7%	12.61	-60%	456%	20 000	10 000 000
Cleaning institute	26	379	3.6%	36.3%	10.13	-37%	600%	20 000	7 000 000
Stove fittery	27	164	3.8%	35.3%	9.32	-29%	433%	40 000	17 000 000
Advertising Agency	28	111	4.4%	35.2%	8.03	-49%	353%	20 000	15 000 000
Showman (show, shooting and sales booths)	29	43	8.1%	34.3%	4.23	-13%	200%	20 000	800 000
Market research institute	30	45	3.1%	34.3%	11.05	-47%	220%	30 000	6 000 000

n = 111 236

Appendix 6 – Lowest average standard deviation with more than 30 observations (2008-2016)



Industry	Rank counts, lowest SD of CAGR	Count of SMEs in sample (2008-2016)	Average CAGR	Standard Deviation of CAGR	CV of CAGR	Min of CAGR	Max of CAGR	Smallest company revenue in sample	Largest company revenue in sample
Library	1	34	0.1%	1.3%	21.69	-4%	3%	20 000	3 270 000
Cosmetic products manufacturing	2	31	-0.2%	1.4%	-6.41	-7%	3%	20 000	5 900 000
Water utility	3	45	0.2%	1.6%	6.71	0%	11%	20 000	3 000 000
Repair of consumer electronics	4	32	-0.3%	1.9%	-5.66	-11%	0%	20 000	1 300 000
Wining	5	50	0.0%	2.1%	62.14	-11%	6%	50 000	8 000 000
Jewelery manufacturing	6	44	0.6%	2.6%	4.63	-5%	12%	20 000	1 500 000
Parquet joinery, also with parquet	7	33	-0.1%	3.1%	-24.51	-14%	8%	60 000	2 574 169
Guesthouse w/o restaurant	8	33	0.5%	3.2%	6.24	-7%	11%	20 000	1 700 000
Curtain retail	9	33	-0.2%	3.3%	-14.06	-13%	7%	20 000	2 000 000
Installation of air conditioning	10	44	0.8%	3.3%	3.99	-6%	13%	50 000	22 314 000
Household cleaning in private homes	11	131	0.3%	3.9%	14.98	-13%	30%	20 000	700 000
Wool Products-Retail	12	41	0.5%	3.9%	7.41	-11%	14%	20 000	1 200 000
Hotel Garni	13	73	0.6%	3.9%	6.78	-16%	19%	20 000	3 350 000
Tire retail	14	38	0.3%	4.2%	15.48	-11%	18%	20 000	19 430 000
Swiss Alpine Club Lodging (SAC)	15	40	0.2%	4.2%	23.78	-12%	19%	50 000	600 000
Pharmacy drugstore	16	85	0.8%	4.3%	5.33	-21%	17%	30 000	9 250 000
Guesthouse (private)	17	46	-0.9%	4.5%	-5.13	-26%	8%	20 000	1 200 000
Decorating, w/o working on construction sites	18	53	-0.5%	4.6%	-10.10	-15%	19%	50 000	5 000 000
Shoes and leather goods (repair of) / shoemaker	19	84	0.2%	4.6%	20.62	-20%	26%	20 000	1 300 000
Private school (w/o sports school)	20	98	1.4%	4.7%	3.44	-10%	25%	20 000	38 555 000
Senior resident home	21	77	3.1%	4.7%	1.51	-2%	22%	50 000	19 616 259
Tourist office, only turism	22	34	-0.7%	4.9%	-6.92	-28%	5%	20 000	4 400 000
Tottoo shop	23	86	0.2%	5.0%	20.90	-21%	26%	20 000	1 000 000
Photographic laboratory (w/o sales)	24	39	-1.8%	5.0%	-2.69	-16%	4%	20 000	2 990 000
Bakery with restaurant	25	164	0.5%	5.0%	10.77	-24%	27%	20 000	10 500 000
Beverage retail (non-alcoholic)	26	52	0.2%	5.0%	22.23	-22%	17%	20 000	3 000 000
Notary	27	110	0.1%	5.0%	59.64	-40%	17%	50 000	6 000 000
Office supply	28	120	-0.4%	5.2%	-13.97	-37%	19%	19 000	42 900 000
Clubs (i.e. fishing, fencing, ...)	29	231	1.1%	5.4%	4.68	-14%	38%	20 000	4 500 000
Caligraphy	30	40	-0.5%	5.4%	-9.98	-23%	12%	20 000	1 000 000

n = 111 236

Appendix 7 – Gender distribution of high average CAGR industries

Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male Stddev of CAGR	Female Stddev of CAGR	Total Count	Total Average of CAGR	Total Stddev of CAGR
Photo studio (only photographs w/o laboratory)	1	87	32	0.6%	110.5%	3.0%	303.2%	166	32.5%	168.6%
Cleaning company	2	72	108	29.9%	0.9%	54.7%	4.5%	251	13.9%	39.2%
Forestry	3	557	7	8.5%	-7.0%	53.9%	0.0%	763	8.9%	55.2%
Windows, doors: fabrication and assembly	4	90		4.9%		21.7%		129	6.9%	25.2%
Butcher w/o shop	5	108		5.2%		11.6%		144	5.4%	11.8%
Showman (show, shooting and sales booths)	6	62	20	3.2%	0.0%	25.4%	0.0%	121	5.0%	31.2%
Taxi	7	82		5.1%		3.9%		113	4.9%	3.8%
Sanitary, gas, water installation	8	166		5.0%		13.8%		240	4.8%	13.6%
Comestibles-Retail	9	89	23	4.2%	0.0%	12.0%	0.0%	149	3.6%	11.2%
Forestry	10	95	3	3.6%	0.0%	10.5%	0.0%	141	3.5%	12.5%
Motor vehicle garage (repair cars / motorcycles)	11	1732	12	3.0%	0.0%	19.6%	0.0%	2384	3.2%	19.7%
Chimney sweep	12	401		2.8%		9.3%		544	2.9%	9.8%
Pig farming	13	156	8	3.1%	0.0%	12.6%	0.0%	230	2.9%	12.5%
Agriculture (mixed holdings)	14	4684	148	2.8%	-0.9%	24.7%	4.7%	6578	2.7%	24.3%
Pension / Guesthouse	15	73	9	2.8%	0.0%	6.0%	0.0%	113	2.6%	5.9%
Food stall, take away	16	67	16	2.8%	2.1%	11.4%	5.6%	116	2.6%	10.6%
Wholesale of beverages, misc.	17	76	8	3.0%	-3.6%	15.1%	0.0%	108	2.5%	14.7%
Landscaping	18	330	8	2.2%	-5.0%	6.2%	5.3%	471	2.5%	7.0%
Misc. Services	19	202	52	2.6%	3.2%	12.3%	11.1%	339	2.5%	11.6%
Restaurant	20	1849	453	2.6%	-0.6%	21.5%	4.4%	3222	2.2%	20.2%
Horse riding school	21	70	49	0.0%	6.5%	0.0%	9.8%	165	2.1%	6.2%
Wooden Flooring	22	105		2.1%		4.4%		154	2.0%	4.3%
Fitness Center / Gym	23	35	64	4.9%	0.6%	9.2%	2.3%	133	1.8%	5.7%
Paiting (Real estate)	24	267	16	1.8%	-1.4%	5.5%	2.1%	385	1.8%	5.7%
Caretaking	25	247	11	2.2%	0.0%	6.3%	0.0%	362	1.7%	5.9%
Consulting	26	209	60	0.8%	5.7%	8.7%	24.0%	390	1.7%	13.4%
Landscaping	27	309	8	1.6%	0.0%	3.8%	0.0%	411	1.5%	3.9%
Civil engineering w/o special focus	28	946		1.3%		9.4%		1304	1.4%	9.4%
Blacksmith	29	253	53	1.6%	0.0%	5.8%	0.0%	419	1.4%	5.4%
Breeding and fattening of cattle	30	970	51	1.6%	0.9%	7.6%	3.5%	1392	1.3%	6.8%

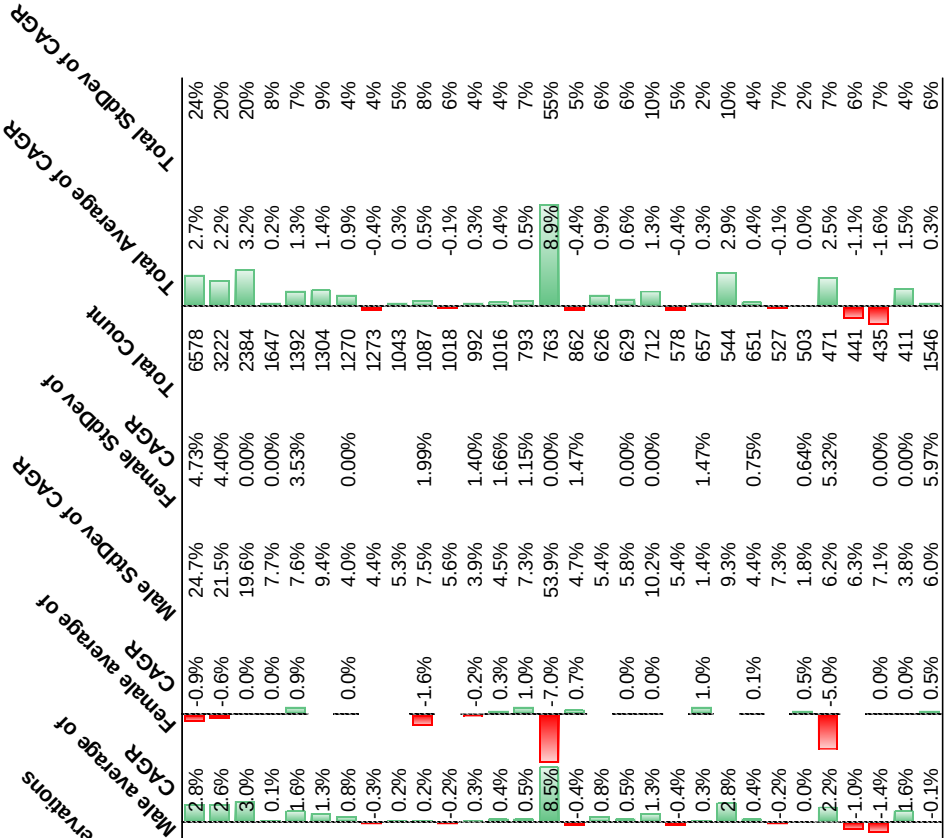
n = 66 227

Appendix 8 – Gender distribution of low average CAGR industries

Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male StdDev of CAGR	Female StdDev of CAGR	Total Count	Total Average of CAGR	Total StdDev of CAGR
Cleaning institute	1	87	99	-16.9%	0.3%	10.5%	1.5%	245	-4.9%	9.2%
Heating installation (w/o electric heaters)	2	147		-3.0%		10.2%		194	-3.3%	10.5%
Special Civil Engineering	3	174		-2.8%		8.7%		253	-2.8%	8.7%
Precision Engineering	4	98		-2.8%		7.5%		128	-2.4%	7.5%
Software consulting and development	5	95	7	-2.5%	0.0%	6.0%	0.0%	144	-2.1%	5.7%
Roofing	6	252		-2.4%		11.3%		344	-1.9%	11.1%
Hotel with restaurant, w/o swimming pool / wellness	7	115	11	-1.2%	-4.7%	7.4%	5.3%	180	-1.7%	7.8%
Installation of electrical apparatus	8	247	16	-1.6%	0.0%	10.0%	0.0%	351	-1.6%	9.7%
Motor vehicle industry, misc.	9	315	10	-1.4%	0.0%	7.1%	0.0%	435	-1.6%	7.1%
Locksmith and sanitary installations	10	163		-1.5%		7.9%		223	-1.6%	7.9%
Graphics	11	73	19	-1.5%	0.0%	5.5%	0.2%	130	-1.4%	5.4%
Antique market	12	63	38	-1.3%	-1.3%	5.2%	3.1%	139	-1.3%	4.7%
Dentist studio	13	114	21	-1.5%	2.5%	4.7%	2.4%	184	-1.3%	5.1%
Photographic studio (w/o sales office)	14	62	39	-2.2%	0.0%	7.6%	0.0%	145	-1.1%	5.5%
Installation of electrical cables	15	327		-1.0%		6.3%		441	-1.1%	6.4%
Bakery with groceries shop	16	119	4	-1.2%	0.0%	5.0%	0.0%	180	-1.1%	5.0%
Guesthouse with restaurant	17	291	57	-0.7%	-1.5%	3.1%	9.1%	475	-0.9%	4.7%
Other retail sale of goods of various kinds	18	84	49	-1.9%	-0.4%	13.0%	2.4%	179	-0.9%	10.1%
Trucking (driving only)	19	244	13	-0.8%	0.0%	7.1%	0.0%	325	-0.8%	6.9%
Mechanical workshops	20	266	3	-0.8%	0.0%	2.8%	0.0%	363	-0.7%	2.7%
Art painting studio	21	57	30	-0.5%	-1.0%	3.5%	3.0%	113	-0.7%	3.3%
Furniture retail	22	42	43	-1.3%	0.0%	1.6%	0.0%	120	-0.7%	1.4%
Machine repair	23	136		-0.3%		6.3%		173	-0.6%	6.6%
General building construction	24	282	11	-0.8%	6.3%	6.7%	0.0%	400	-0.5%	6.7%
Accounting office	25	179	72	-0.4%	-0.4%	3.6%	2.1%	352	-0.4%	3.5%
Furniture joinery (w/o works on sites)	26	187	2	-0.5%	0.0%	1.8%	0.0%	268	-0.4%	1.7%
Gardening (w/o horticulture)	27	220	8	-0.3%	-5.4%	2.3%	5.8%	326	-0.4%	2.8%
Locksmith w.o. sanitary installations	28	426		-0.4%		5.4%		578	-0.4%	5.4%
Carpentry	29	919		-0.3%		4.4%		1273	-0.4%	4.5%
Hotel with restaurant/pool/spa	30	538	58	-0.4%	0.7%	4.7%	1.5%	862	-0.4%	4.7%

n = 66 227

Appendix 9 - Most insured male drivers by industry



Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male StdDev of CAGR	Female StdDev of CAGR	Total Count	Total Average of CAGR	Total StdDev of CAGR
Agriculture (mixed holdings)	1	4684	148	2.8%	-0.9%	24.7%	4.73%	6578	2.7%	24%
Restaurant	2	1849	453	2.6%	-0.6%	21.5%	4.40%	3222	2.2%	20%
Motor vehicle garage (repair cars / motorcycles)	3	1732	12	3.0%	0.0%	19.6%	0.00%	2384	3.2%	20%
Building joinery/carpenter's shop	4	1201	3	0.1%	0.0%	7.7%	0.00%	1647	0.2%	8%
Breeding and fattening of cattle	5	970	51	1.6%	0.9%	7.6%	3.53%	1392	1.3%	7%
Civil engineering w/o special focus	6	946		1.3%		9.4%		1304	1.4%	9%
Poultry farm	7	929	17	0.8%	0.0%	4.0%	0.00%	1270	0.9%	4%
Carpentry	8	919		-0.3%		4.4%		1273	-0.4%	4%
Flooring	9	765		0.2%		5.3%		1043	0.3%	5%
Gardening	10	755	36	0.2%	-1.6%	7.5%	1.99%	1087	0.5%	8%
Joinery	11	739		-0.2%		5.6%		1018	-0.1%	6%
Architecture firm	12	661	51	0.3%	-0.2%	3.9%	1.40%	992	0.3%	4%
General medicine practitioners (w.o. X-ray)	13	647	74	0.4%	0.3%	4.5%	1.66%	1016	0.4%	4%
Plumbing	14	561	17	0.5%	1.0%	7.3%	1.15%	793	0.5%	7%
Forestry	15	557	7	8.5%	-7.0%	53.9%	0.00%	763	8.9%	55%
Hotel with restaurant/pool/spa	16	538	58	-0.4%	0.7%	4.7%	1.47%	862	-0.4%	5%
Excavation/earth moving	17	473		0.8%		5.4%		626	0.9%	6%
Painting and plastering	18	469	2	0.5%	0.0%	5.8%	0.00%	629	0.6%	6%
Painting	19	454	23	1.3%	0.0%	10.2%	0.00%	712	1.3%	10%
Locksmith w.o. sanitary installations	20	426		-0.4%		5.4%		578	-0.4%	5%
Car transportion / trucking	21	412	47	0.3%	1.0%	1.4%	1.47%	657	0.3%	2%
Chimney sweep	22	401		2.8%		9.3%		544	2.9%	10%
Dental practice	23	398	40	0.4%	0.1%	4.4%	0.75%	651	0.4%	4%
Repair of other vehicles	24	378		-0.2%		7.3%		527	-0.1%	7%
Viniculture	25	374	2	0.0%	0.5%	1.8%	0.64%	503	0.0%	2%
Landscaping	26	330	8	2.2%	-5.0%	6.2%	5.32%	471	2.5%	7%
Installation of electrical cables	27	327		-1.0%		6.3%		441	-1.1%	6%
Motor vehicle industry, misc.	28	315	10	-1.4%	0.0%	7.1%	0.00%	435	-1.6%	7%
Landscaping	29	309	8	1.6%	0.0%	3.8%	0.00%	411	1.5%	4%
Hair dresser	30	301	806	-0.1%	0.5%	6.0%	5.97%	1546	0.3%	6%

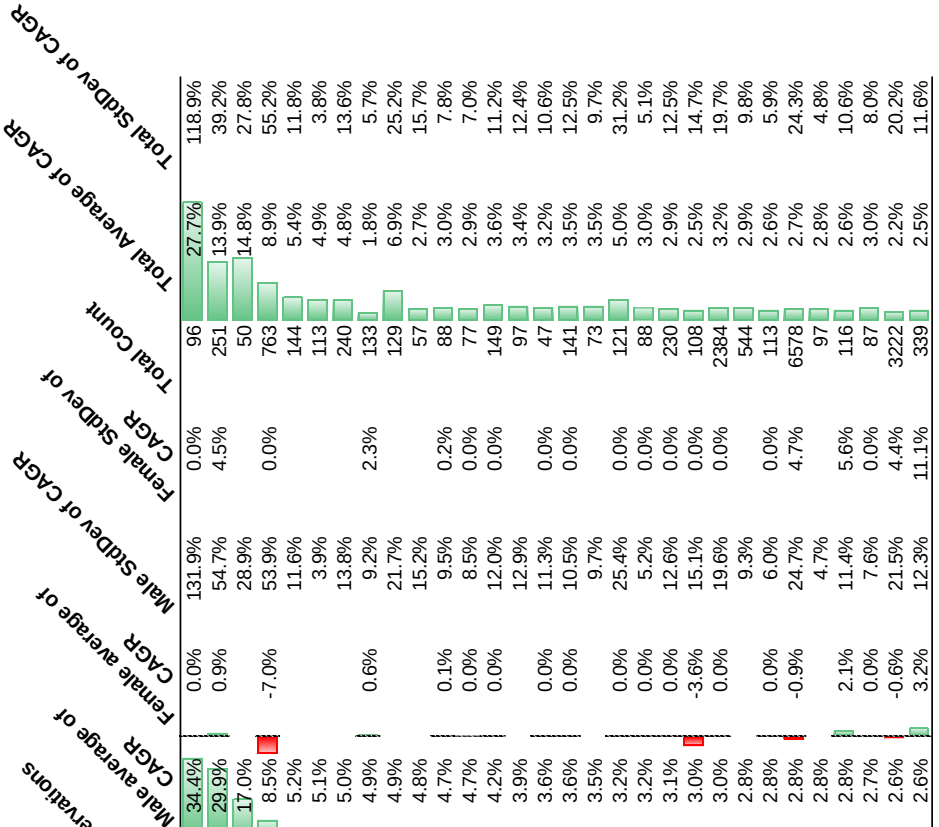
n = 66 227

Appendix 10 - Most insured female drivers by industry

Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male Stddev of CAGR	Female Stddev of CAGR	Total Count	Total Average of CAGR	Total Stddev of CAGR
Hair dresser	1	301	806	-0.1%	0.5%	6.0%	6.0%	1546	0.3%	5.8%
Naturopathic practice/acupuncture/kinesiology	2	231	467	0.6%	0.7%	3.2%	3.7%	973	0.7%	3.6%
Restaurant	3	1849	453	2.6%	-0.6%	21.5%	4.4%	3222	2.2%	20.2%
Beauty salon (w.o. perm. make-up/piercing)	4	1	331	0.0%	0.2%		4.2%	460	0.2%	4.0%
Pedicure	5		223		-0.3%		10.4%	313	0.0%	11.1%
Agriculture (mixed holdings)	6	4684	148	2.8%	-0.9%	24.7%	4.7%	6578	2.7%	24.3%
Massage	7	46	146	-2.1%	0.8%	10.5%	3.3%	276	0.3%	6.1%
Animal shelter	8	40	144	0.0%	0.0%	0.0%	2.6%	248	0.2%	3.1%
Cleaning company	9	72	108	29.9%	0.9%	54.7%	4.5%	251	13.9%	39.2%
Manicure	10		100		-0.2%		6.1%	136	-0.1%	6.0%
Cleaning institute	11	87	99	-10.9%	0.3%	10.5%	1.5%	245	-4.9%	9.2%
Household cleaning in private homes	12	1	83	0.0%	0.0%		0.6%	118	0.0%	0.6%
Kiosk / Retail trade with magazine	13	68	77	-0.2%	-0.1%	0.9%	0.5%	198	-0.2%	0.7%
General medicine practitioners (w.o. X-ray)	14	647	74	0.4%	0.3%	4.5%	1.7%	1016	0.4%	3.7%
Accounting office	15	179	72	-0.4%	-0.4%	3.6%	2.1%	352	-0.4%	3.5%
Beauty salon (w/o perm. make-up, w/o piercing)	16	2	65	0.0%	-0.7%	0.0%	2.6%	100	-0.7%	2.6%
Riding stables, horse boarding	17	239	64	0.5%	0.7%	1.7%	8.3%	412	0.4%	4.1%
Fitness Center / Gym	18	35	64	4.9%	0.6%	9.2%	2.3%	133	1.8%	5.7%
Consulting	19	209	60	0.8%	5.7%	8.7%	24.0%	390	1.7%	13.4%
Hotel with restaurant/pool/spa	20	538	58	-0.4%	0.7%	4.7%	1.5%	862	-0.4%	4.7%
Guesthouse with restaurant	21	291	57	-0.7%	-1.5%	3.1%	9.1%	475	-0.9%	4.7%
Beauty Institute (w/o perm. make-up, w/o piercing)	22	2	55	0.0%	0.3%	0.0%	1.1%	72	0.2%	1.0%
Blacksmith	23	253	53	1.6%	0.0%	5.8%	0.0%	419	1.4%	5.4%
Misc. Services	24	202	52	2.6%	3.2%	12.3%	11.1%	339	2.5%	11.6%
Breeding and fattening of cattle	25	970	51	1.6%	0.9%	7.6%	3.5%	1392	1.3%	6.8%
Architecture firm	26	661	51	0.3%	-0.2%	3.9%	1.4%	992	0.3%	4.3%
Other retail sale of goods of various kinds	27	84	49	-1.9%	-0.4%	13.0%	2.4%	179	-0.9%	10.1%
Horse riding school	28	70	49	0.0%	6.5%	0.0%	9.8%	165	2.1%	6.2%
Car transportion / trucking	29	412	47	0.3%	1.0%	1.4%	1.5%	657	0.3%	2.0%
Flower and plant retail	30	3	45	0.0%	0.2%	0.0%	5.5%	71	0.8%	5.9%

n = 66 227

Appendix 11 – Highest average CAGR of male insured drivers by industry



Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male Stddev of CAGR	Female Stddev of CAGR	Total Count	Total Average of CAGR	Total Stddev of CAGR
Notary	1	62	3	34.4%	0.0%	131.9%	0.0%	96	27.7%	118.9%
Cleaning company	2	72	108	29.9%	0.9%	54.7%	4.5%	251	13.9%	39.2%
Forge and sanitary installations	3	41		17.0%		28.9%		50	14.8%	27.8%
Forestry	4	557	7	8.5%	-7.0%	53.9%	0.0%	763	8.9%	55.2%
Butcher w/o shop	5	108		5.2%		11.6%		144	5.4%	11.8%
Taxi	6	82		5.1%		3.9%		113	4.9%	3.8%
Sanitary, gas, water installation	7	166		5.0%		13.8%		240	4.8%	13.6%
Fitness Center / Gym	8	35	64	4.9%	0.6%	9.2%	2.3%	133	1.8%	5.7%
Windows, doors: fabrication and assembly	9	90		4.9%		21.7%		129	6.9%	25.2%
Firewood (w/o working on construction sites)	10	50		4.8%		15.2%		57	2.7%	15.7%
Campsite administration	11	44	21	4.7%	0.1%	9.5%	0.2%	88	3.0%	7.8%
Toys-Retail	12	33	23	4.7%	0.0%	8.5%	0.0%	77	2.9%	7.0%
Comestibles-Retail	13	89	23	4.2%	0.0%	12.0%	0.0%	149	3.6%	11.2%
Movie recording studio	14	71		3.9%		12.9%		97	3.4%	12.4%
Elektroingenieurb¿ro	15	31	2	3.6%	0.0%	11.3%	0.0%	47	3.2%	10.6%
Forestry	16	95	3	3.6%	0.0%	10.5%	0.0%	141	3.5%	12.5%
Agriculture	17	49		3.5%		9.7%		73	3.5%	9.7%
Showman (show, shooting and sales booths)	18	62	20	3.2%	0.0%	25.4%	0.0%	121	5.0%	31.2%
Motorcycles (repair of)	19	65	3	3.2%	0.0%	5.2%	0.0%	88	3.0%	5.1%
Pig farming	20	156	8	3.1%	0.0%	12.6%	0.0%	230	2.9%	12.5%
Wholesale of beverages, misc.	21	76	8	3.0%	-3.6%	15.1%	0.0%	108	2.5%	14.7%
Motor vehicle garage (repair cars / motorcycles)	22	1732	12	3.0%	0.0%	19.6%	0.0%	2384	3.2%	19.7%
Chimney sweep	23	401		2.8%		9.3%		544	2.9%	9.8%
Pension / Guesthouse	24	73	9	2.8%	0.0%	6.0%	0.0%	113	2.6%	5.9%
Agriculture (mixed holdings)	25	4684	148	2.8%	-0.9%	24.7%	4.7%	6578	2.7%	24.3%
Joinery work in interior design	26	73		2.8%		4.7%		97	2.8%	4.8%
Food stall, take away	27	67	16	2.8%	2.1%	11.4%	5.6%	116	2.6%	10.6%
Botanical nursery	28	59	4	2.7%	0.0%	7.6%	0.0%	87	3.0%	8.0%
Restaurant	29	1849	453	2.6%	-0.6%	21.5%	4.4%	3222	2.2%	20.2%
Misc. Services	30	202	52	2.6%	3.2%	12.3%	11.1%	339	2.5%	11.6%

n = 66 227

Appendix 12 – Lowest average CAGR of male insured drivers by industry

Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male Stddev of CAGR	Female Stddev of CAGR	Total Count	Total Average of CAGR	Total Stddev of CAGR
Cleaning institute	1	87	99	-10.9%	0.3%	10.5%	1.5%	245	-4.9%	9.2%
Decolletageatelier	2	33		-6.8%		8.9%		42	-6.4%	8.8%
Glazier shop	3	41		-6.4%		9.0%		64	-6.0%	8.8%
Machine Retail	4	63	3	-3.7%	0.0%	8.3%	0.0%	88	-2.8%	7.2%
Nursery	5	60		-3.2%		5.8%		86	-3.3%	5.9%
Dairy with shop	6	31	3	-3.1%	0.0%	6.3%	0.0%	48	-3.0%	6.1%
Heating installation (w/o electric heaters)	7	147		-3.0%		10.2%		194	-3.3%	10.5%
Precision Engineering	8	98		-2.8%		7.5%		128	-2.4%	7.5%
Special Civil Engineering	9	174		-2.8%		8.7%		253	-2.8%	8.7%
Forestry with nurseries	10	74		-2.8%		3.3%		97	-2.6%	3.5%
Drugstore retail	11	34	4	-2.6%	0.0%	3.5%	0.0%	54	-2.2%	3.3%
Software consulting and development	12	95	7	-2.5%	0.0%	6.0%	0.0%	144	-2.1%	5.7%
TV-Retail	13	35		-2.5%		5.8%		50	-2.5%	5.6%
Roofing	14	252		-2.4%		11.3%		344	-1.9%	11.1%
Dredging of various kinds	15	31		-2.2%		10.9%		38	-3.2%	11.8%
Photographic studio (w/o sales office)	16	62	39	-2.2%	0.0%	7.6%	0.0%	145	-1.1%	5.5%
Massage	17	46	146	-2.1%	0.8%	10.5%	3.3%	276	0.3%	6.1%
Electrical material-Retail	18	38		-1.9%		8.1%		51	-2.1%	8.5%
Saddlery	19	50	20	-1.9%	0.7%	3.5%	1.3%	94	-1.0%	3.7%
Other retail sale of goods of various kinds	20	84	49	-1.9%	-0.4%	13.0%	2.4%	179	-0.9%	10.1%
Bicycles (repair of)	21	38		-1.7%		2.5%		48	-1.4%	2.3%
Installation of electrical apparatus	22	247	16	-1.6%	0.0%	10.0%	0.0%	351	-1.6%	9.7%
Graphics	23	73	19	-1.5%	0.0%	5.5%	0.2%	130	-1.4%	5.4%
Dentist studio	24	114	21	-1.5%	2.5%	4.7%	2.4%	184	-1.3%	5.1%
Metal deformation of various kinds	25	45		-1.5%		4.9%		70	-1.6%	5.0%
Realty management	26	65	3	-1.5%	0.0%	6.8%	0.0%	97	-1.1%	6.6%
Locksmith and sanitary installations	27	163		-1.5%		7.9%		223	-1.6%	7.9%
Motor vehicle industry, misc.	28	315	10	-1.4%	0.0%	7.1%	0.0%	435	-1.6%	7.1%
Advertising agency	29	43	11	-1.4%	-2.5%	12.3%	5.6%	74	-2.2%	11.7%
Antique market	30	63	38	-1.3%	-1.3%	5.2%	3.1%	139	-1.3%	4.7%

n = 66 227

Appendix 13 – Highest average CAGR of female insured drivers by industry

Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male StdDev of CAGR	Female StdDev of CAGR	Total Count	Total Average of CAGR	Total StdDev of CAGR
Photo studio (only photographs w/o laboratory)	1	87	32	0.6%	110.5%	3.0%	303.2%	166	32.5%	168.6%
Horse riding school	2	70	49	0.0%	6.5%	0.0%	9.8%	165	2.1%	6.2%
Consulting	3	209	60	0.8%	5.7%	8.7%	24.0%	390	1.7%	13.4%
Bar	4	85	42	-0.2%	5.6%	0.9%	17.0%	169	1.2%	8.9%
Misc. Services	5	202	52	2.6%	3.2%	12.3%	11.1%	339	2.5%	11.6%
Car transportion / trucking	6	412	47	0.3%	1.0%	1.4%	1.5%	657	0.3%	2.0%
Breeding and fattening of cattle	7	970	51	1.6%	0.9%	7.6%	3.5%	1392	1.3%	6.8%
Cleaning company	8	72	108	29.9%	0.9%	54.7%	4.5%	251	13.9%	39.2%
Massage	9	46	146	-2.1%	0.8%	10.5%	3.3%	276	0.3%	6.1%
Naturopathic practice/acupuncture/kinesiology	10	231	467	0.6%	0.7%	3.2%	3.7%	973	0.7%	3.6%
Pet Grooming	11		44		0.7%		3.1%	58	0.8%	3.3%
Hotel with restaurant/pool/spa	12	538	58	-0.4%	0.7%	4.7%	1.5%	862	-0.4%	4.7%
Riding stables, horse boarding	13	239	64	0.5%	0.7%	1.7%	8.3%	412	0.4%	4.1%
Fitness Center / Gym	14	35	64	4.9%	0.6%	9.2%	2.3%	133	1.8%	5.7%
Hair dresser	15	301	806	-0.1%	0.5%	6.0%	6.0%	1546	0.3%	5.8%
Cleaning institute	16	87	99	-10.9%	0.3%	10.5%	1.5%	245	-4.9%	9.2%
General medicine practitioners (w.o. X-ray)	17	647	74	0.4%	0.3%	4.5%	1.7%	1016	0.4%	3.7%
Beauty Institute (w/o perm. make-up, w/o piercing)	18	2	55	0.0%	0.3%	0.0%	1.1%	72	0.2%	1.0%
Flower and plant retail	19	3	45	0.0%	0.2%	0.0%	5.5%	71	0.8%	5.9%
Beauty salon (w.o. perm. make-up/piercing)	20	1	331	0.0%	0.2%		4.2%	460	0.2%	4.0%
Dental practice	21	398	40	0.4%	0.1%	4.4%	0.7%	651	0.4%	4.4%
Animal shelter	22	40	144	0.0%	0.0%	0.0%	2.6%	248	0.2%	3.1%
Household cleaning in private homes	23	1	83	0.0%	0.0%		0.6%	118	0.0%	0.6%
Photographic studio (w/o sales office)	24	62	39	-2.2%	0.0%	7.6%	0.0%	145	-1.1%	5.5%
Furniture retail	25	42	43	-1.3%	0.0%	1.6%	0.0%	120	-0.7%	1.4%
Veterinary practice / vet. affairs	26	78	33	-0.1%	0.0%	0.7%	0.0%	163	0.0%	0.6%
Special retail trade, other	27	103	36	1.3%	0.0%	3.5%	0.0%	189	0.9%	2.9%
Blacksmith	28	253	53	1.6%	0.0%	5.8%	0.0%	419	1.4%	5.4%
Kiosk / Retail trade with magazine	29	68	77	-0.2%	-0.1%	0.9%	0.5%	198	-0.2%	0.7%
Architecture firm	30	661	51	0.3%	-0.2%	3.9%	1.4%	992	0.3%	4.3%

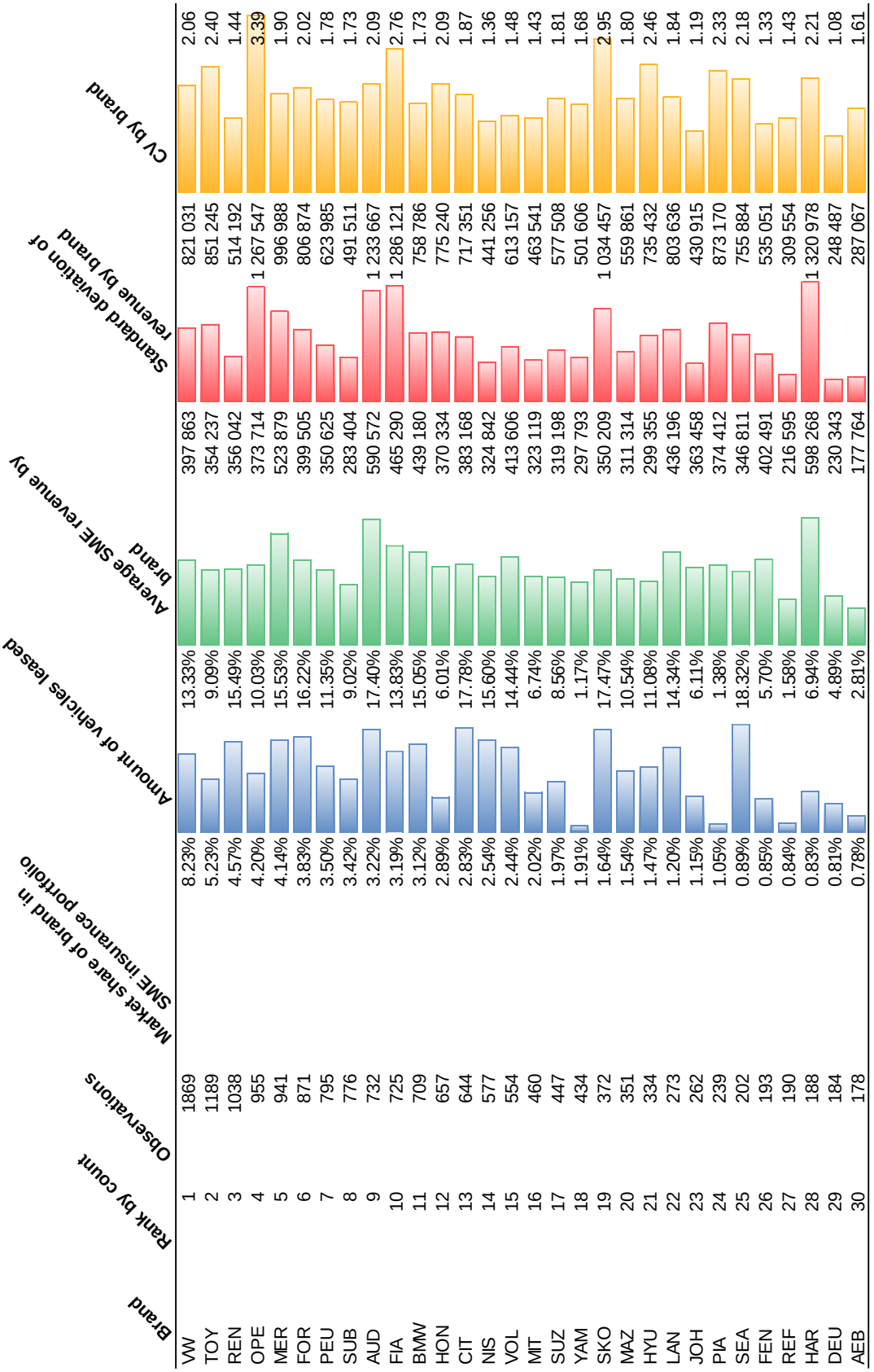
n = 66 227

Appendix 14 - Lowest average CAGR of female insured drivers by industry

Industry	Rank	Male observations	Female observations	Male average of CAGR	Female average of CAGR	Male Stddev of CAGR	Female Stddev of CAGR	Total Count	Total Average of CAGR	Total Stddev of CAGR
Tailoring	1	3	41	0.0%	-2.4%	0.0%	5.5%	64	-2.5%	5.2%
Gardening	2	755	36	0.2%	-1.6%	7.5%	2.0%	1087	0.5%	7.7%
Guesthouse with restaurant	3	291	57	-0.7%	-1.5%	3.1%	9.1%	475	-0.9%	4.7%
Antique market	4	63	38	-1.3%	-1.3%	5.2%	3.1%	139	-1.3%	4.7%
Artistic studio	5	39	39	2.4%	-1.1%	5.7%	4.6%	97	0.3%	5.3%
Agriculture (mixed holdings)	6	4684	148	2.8%	-0.9%	24.7%	4.7%	6578	2.7%	24.3%
Beauty salon (w/o perm. make-up, w/o piercing)	7	2	65	0.0%	-0.7%	0.0%	2.6%	100	-0.7%	2.6%
Restaurant	8	1849	453	2.6%	-0.6%	21.5%	4.4%	3222	2.2%	20.2%
Accounting office	9	179	72	-0.4%	-0.4%	3.6%	2.1%	352	-0.4%	3.5%
Other retail sale of goods of various kinds	10	84	49	-1.9%	-0.4%	13.0%	2.4%	179	-0.9%	10.1%
Pedicure	11		223		-0.3%		10.4%	313	0.0%	11.1%
Physiotherapy	12	18	39	0.0%	-0.2%	0.0%	2.4%	84	-0.1%	2.0%
Manicure	13		100		-0.2%		6.1%	136	-0.1%	6.0%
Architecture firm	14	661	51	0.3%	-0.2%	3.9%	1.4%	992	0.3%	4.3%
Kiosk / Retail trade with magazine	15	68	77	-0.2%	-0.1%	0.9%	0.5%	198	-0.2%	0.7%
Photographic studio (w/o sales office)	16	62	39	2.2%	0.0%	7.6%	0.0%	145	1.1%	5.5%
Furniture retail	17	42	43	-1.3%	0.0%	1.6%	0.0%	120	0.7%	1.4%
Veterinary practice / vet. affairs	18	78	33	-0.1%	0.0%	0.7%	0.0%	163	0.0%	0.6%
Special retail trade, other	19	103	36	1.3%	0.0%	3.5%	0.0%	189	0.9%	2.9%
Blacksmith	20	253	53	1.6%	0.0%	5.8%	0.0%	419	1.4%	5.4%
Household cleaning in private homes	21	1	83	0.0%	0.0%		0.6%	118	0.0%	0.6%
Animal shelter	22	40	144	0.0%	0.0%	0.0%	2.6%	248	0.2%	3.1%
Dental practice	23	398	40	0.4%	0.1%	4.4%	0.7%	651	0.4%	4.4%
Beauty salon (w.o. perm. make-up/piercing)	24	1	331	0.0%	0.2%		4.2%	460	0.2%	4.0%
Flower and plant retail	25	3	45	0.0%	0.2%	0.0%	5.5%	71	0.8%	5.9%
Beauty Institute (w/o perm. make-up, w/o piercing)	26	2	55	0.0%	0.3%	0.0%	1.1%	72	0.2%	1.0%
General medicine practitioners (w.o. X-ray)	27	647	74	0.4%	0.3%	4.5%	1.7%	1016	0.4%	3.7%
Cleaning institute	28	87	99	-10.9%	0.3%	10.5%	1.5%	245	-4.9%	9.2%
Hair dresser	29	301	806	-0.1%	0.5%	6.0%	6.0%	1546	0.3%	5.8%
Fitness Center / Gym	30	35	64	4.9%	0.6%	9.2%	2.3%	133	1.8%	5.7%

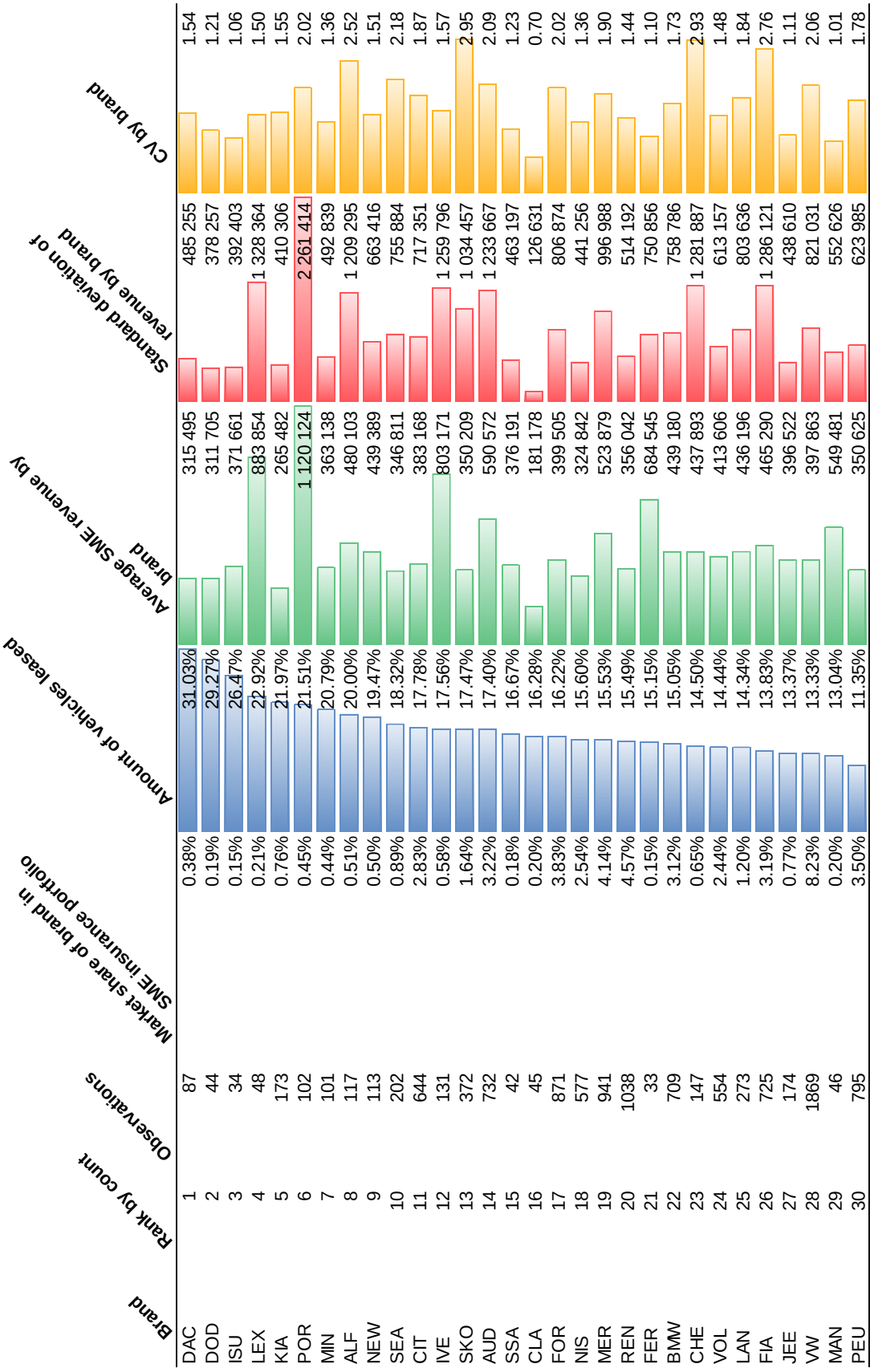
n = 66 227

Appendix 15 – Most common vehicle brands with SMEs



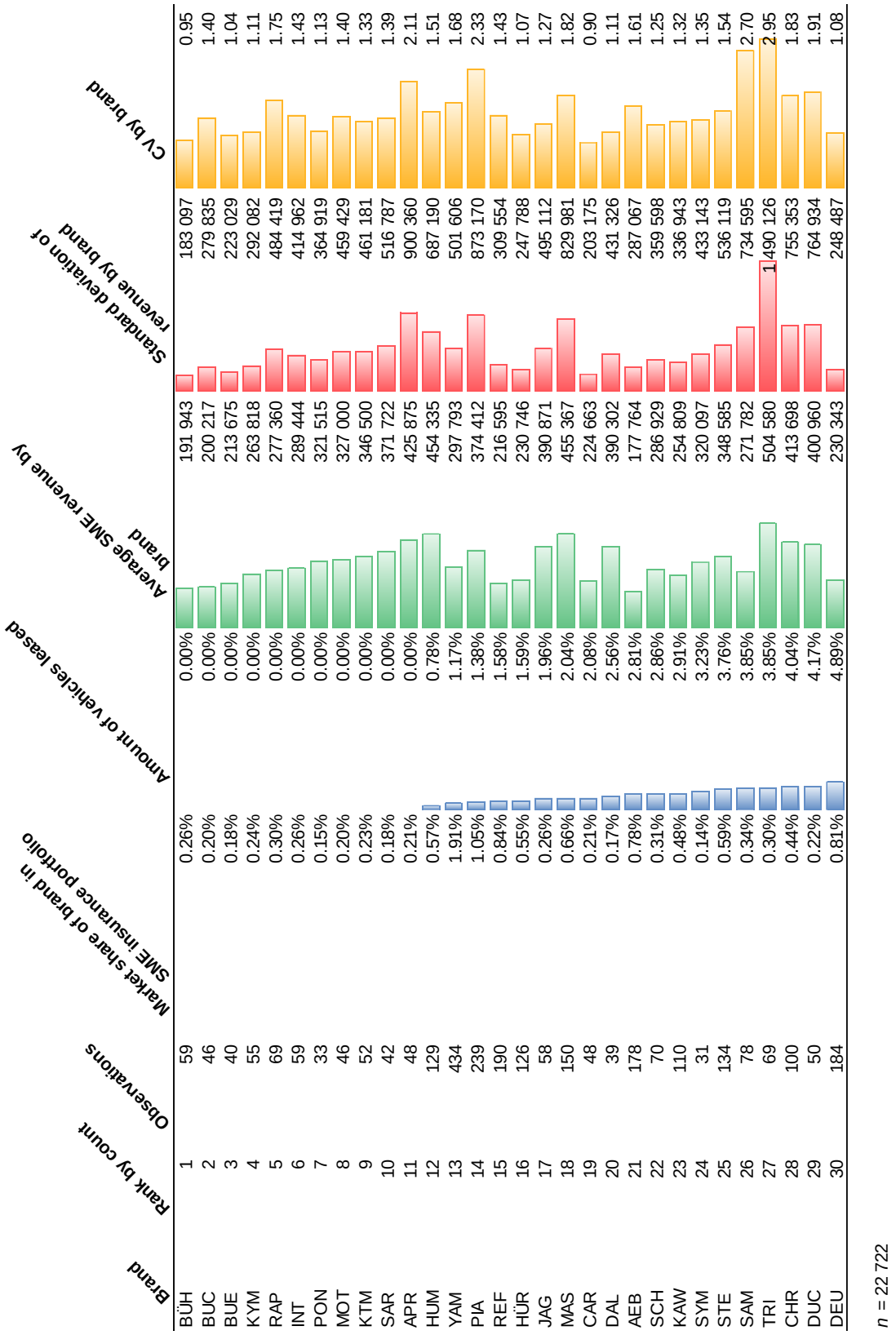
n = 22 722

Appendix 16 – Most common leasing vehicles brands

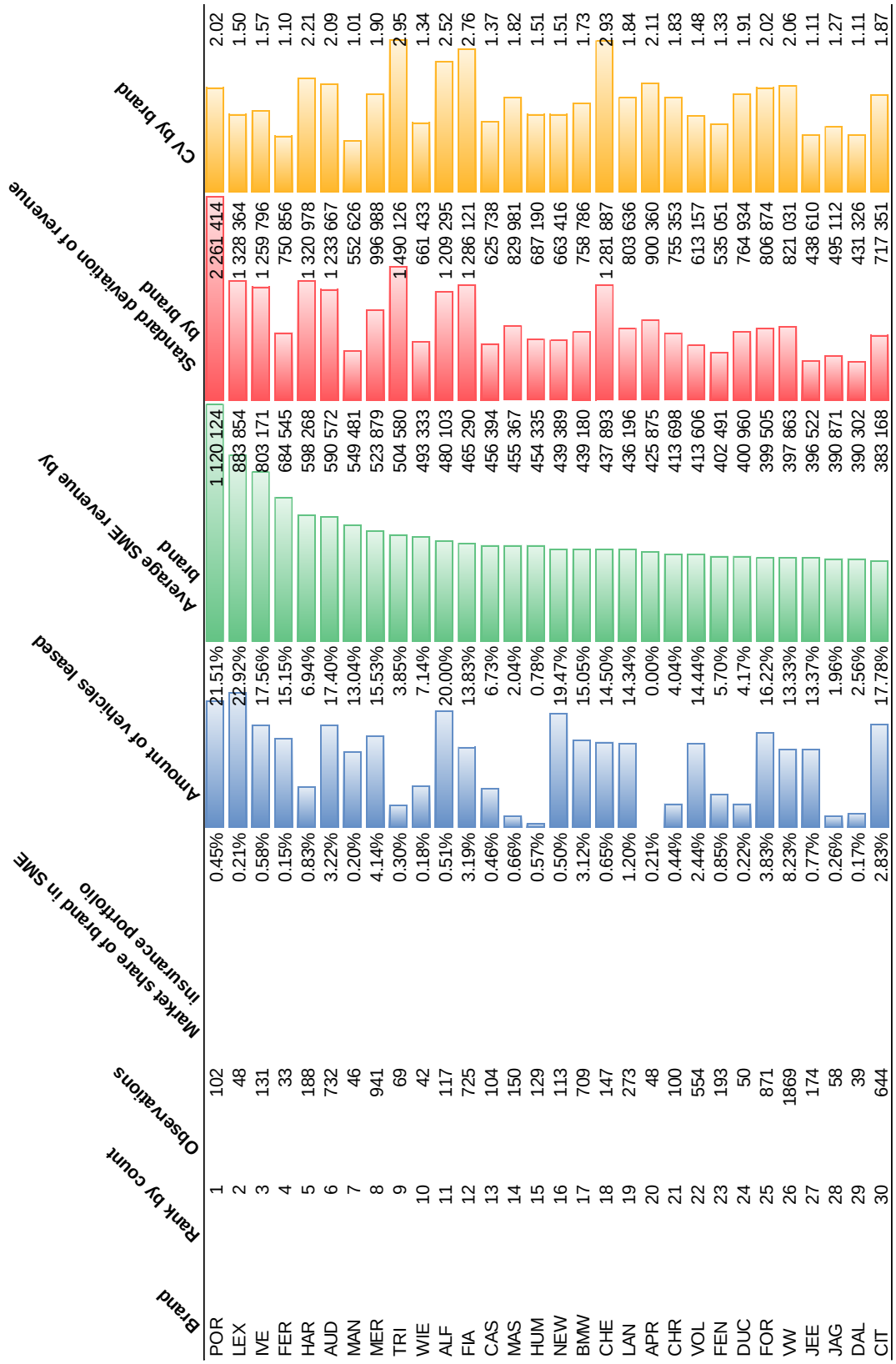


n = 22 722

Appendix 17 - Most common owned vehicles brands



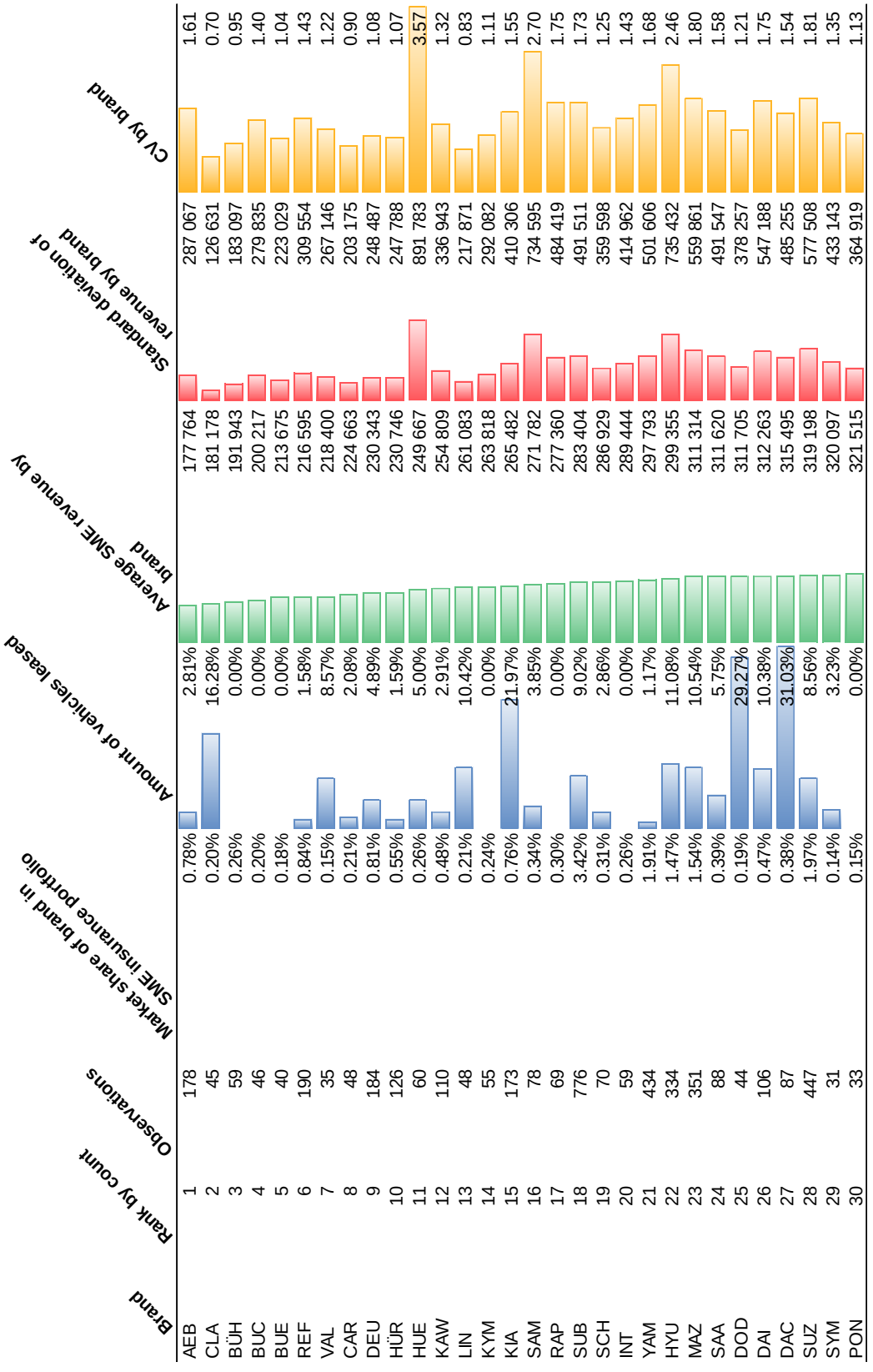
Appendix 18 – Highest average SME revenues by vehicle brand



Brand	Rank by count	Observations	Market share of brand in SME insurance portfolio	Amount of vehicles leased	Average SME revenue by brand	Standard deviation of revenue by brand	CV by brand
POR	1	102	0.45%	21.51%	1 120 124	2 261 414	2.02
LEX	2	48	0.21%	22.92%	883 854	1 328 364	1.50
IVE	3	131	0.58%	17.56%	803 171	1 259 796	1.57
FER	4	33	0.15%	15.15%	684 545	750 856	1.10
HAR	5	188	0.83%	6.94%	598 268	1 320 978	2.21
AUD	6	732	3.22%	17.40%	590 572	1 233 667	2.09
MAN	7	46	0.20%	13.04%	549 481	552 626	1.01
MER	8	941	4.14%	15.53%	523 879	996 988	1.90
TRI	9	69	0.30%	3.85%	504 580	1490 126	2.95
WIE	10	42	0.18%	7.14%	493 333	661 433	1.34
ALF	11	117	0.51%	20.00%	480 103	1 209 295	2.52
FIA	12	725	3.19%	13.83%	465 290	1 286 121	2.76
CAS	13	104	0.46%	6.73%	456 394	625 738	1.37
MAS	14	150	0.66%	2.04%	455 367	829 981	1.82
HUM	15	129	0.57%	0.78%	454 335	687 190	1.51
NEW	16	113	0.50%	19.47%	439 389	663 416	1.51
BMW	17	709	3.12%	15.05%	439 180	758 786	1.73
CHE	18	147	0.65%	14.50%	437 893	1 281 887	2.93
LAN	19	273	1.20%	14.34%	436 196	803 636	1.84
APR	20	48	0.21%	0.00%	425 875	900 360	2.11
CHR	21	100	0.44%	4.04%	413 698	755 353	1.83
VOL	22	554	2.44%	14.44%	413 606	613 157	1.48
FEN	23	193	0.85%	5.70%	402 491	535 051	1.33
DUC	24	50	0.22%	4.17%	400 960	764 934	1.91
FOR	25	871	3.83%	16.22%	399 505	806 874	2.02
VW	26	1869	8.23%	13.33%	397 863	821 031	2.06
JEE	27	174	0.77%	13.37%	396 522	438 610	1.11
JAG	28	58	0.26%	1.96%	390 871	495 112	1.27
DAL	29	39	0.17%	2.56%	390 302	431 326	1.11
CIT	30	644	2.83%	17.78%	383 168	717 351	1.87

n = 22 722

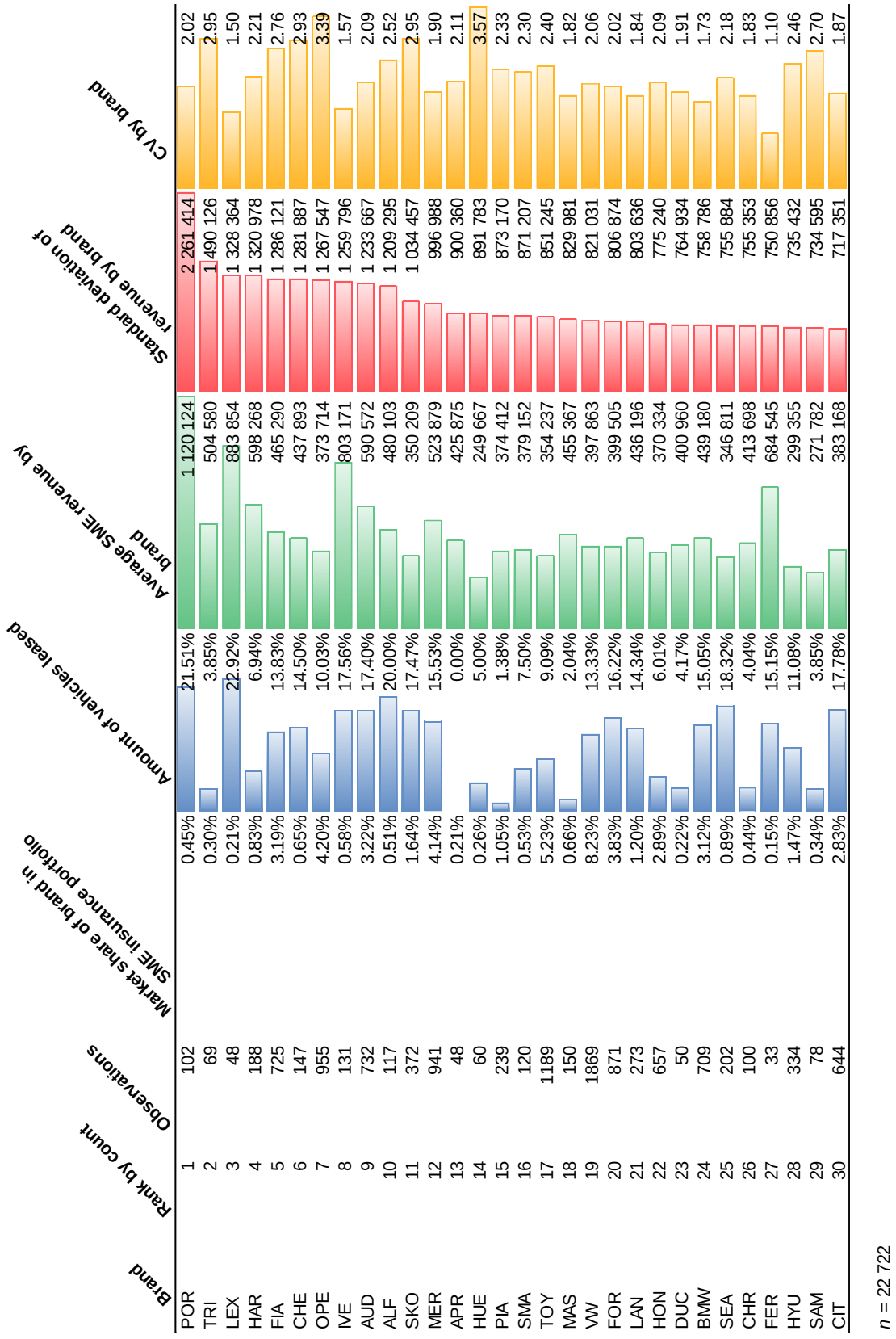
Appendix 19 – Lowest average SME revenues by vehicle brand



Brand	Rank by count	Observations	Market share of brand in SME insurance portfolio	Amount of vehicles leased	Average SME revenue by brand	Standard deviation of revenue by brand	CV by brand
AEB	1	178	0.78%	2.81%	177 764	287 067	1.61
CLA	2	45	0.20%	16.28%	181 178	126 631	0.70
BÜH	3	59	0.26%	0.00%	191 943	183 097	0.95
BUC	4	46	0.20%	0.00%	200 217	279 835	1.40
BUE	5	40	0.18%	0.00%	213 675	223 029	1.04
REF	6	190	0.84%	1.58%	216 595	309 554	1.43
VAL	7	35	0.15%	8.57%	218 400	267 146	1.22
CAR	8	48	0.21%	2.08%	224 663	203 175	0.90
DEU	9	184	0.81%	4.89%	230 343	248 487	1.08
HÜR	10	126	0.55%	1.59%	230 746	247 788	1.07
HUE	11	60	0.26%	5.00%	249 667	891 783	3.57
KAW	12	110	0.48%	2.91%	254 809	336 943	1.32
LIN	13	48	0.21%	10.42%	261 083	217 871	0.83
KYM	14	55	0.24%	0.00%	263 818	292 082	1.11
KIA	15	173	0.76%	21.97%	265 482	410 306	1.55
SAM	16	78	0.34%	3.85%	271 782	734 595	2.70
RAP	17	69	0.30%	0.00%	277 360	484 419	1.75
SUB	18	776	3.42%	9.02%	283 404	491 511	1.73
SCH	19	70	0.31%	2.86%	286 929	359 598	1.25
INT	20	59	0.26%	0.00%	289 444	414 962	1.43
YAM	21	434	1.91%	1.17%	297 793	501 606	1.68
HYU	22	334	1.47%	11.08%	299 355	735 432	2.46
MAZ	23	351	1.54%	10.54%	311 314	559 861	1.80
SAA	24	88	0.39%	5.75%	311 620	491 547	1.58
DOD	25	44	0.19%	29.27%	311 705	378 257	1.21
DAI	26	106	0.47%	10.38%	312 263	547 188	1.75
DAC	27	87	0.38%	31.03%	315 495	485 255	1.54
SUZ	28	447	1.97%	8.56%	319 198	577 508	1.81
SYM	29	31	0.14%	3.23%	320 097	433 143	1.35
PON	30	33	0.15%	0.00%	321 515	364 919	1.13

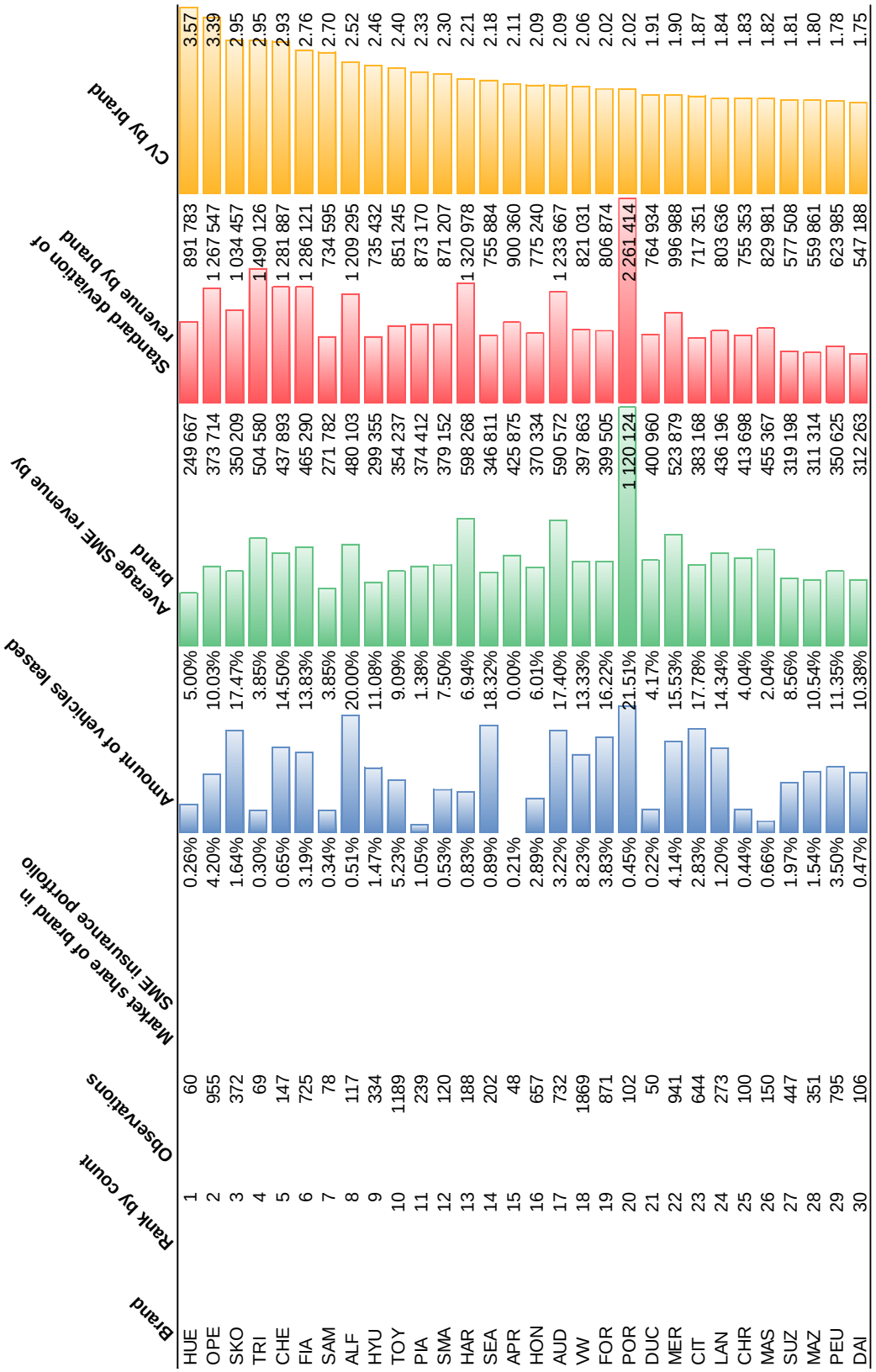
n = 22 722

Appendix 20 – Highest standard deviation by brand



n = 22 722

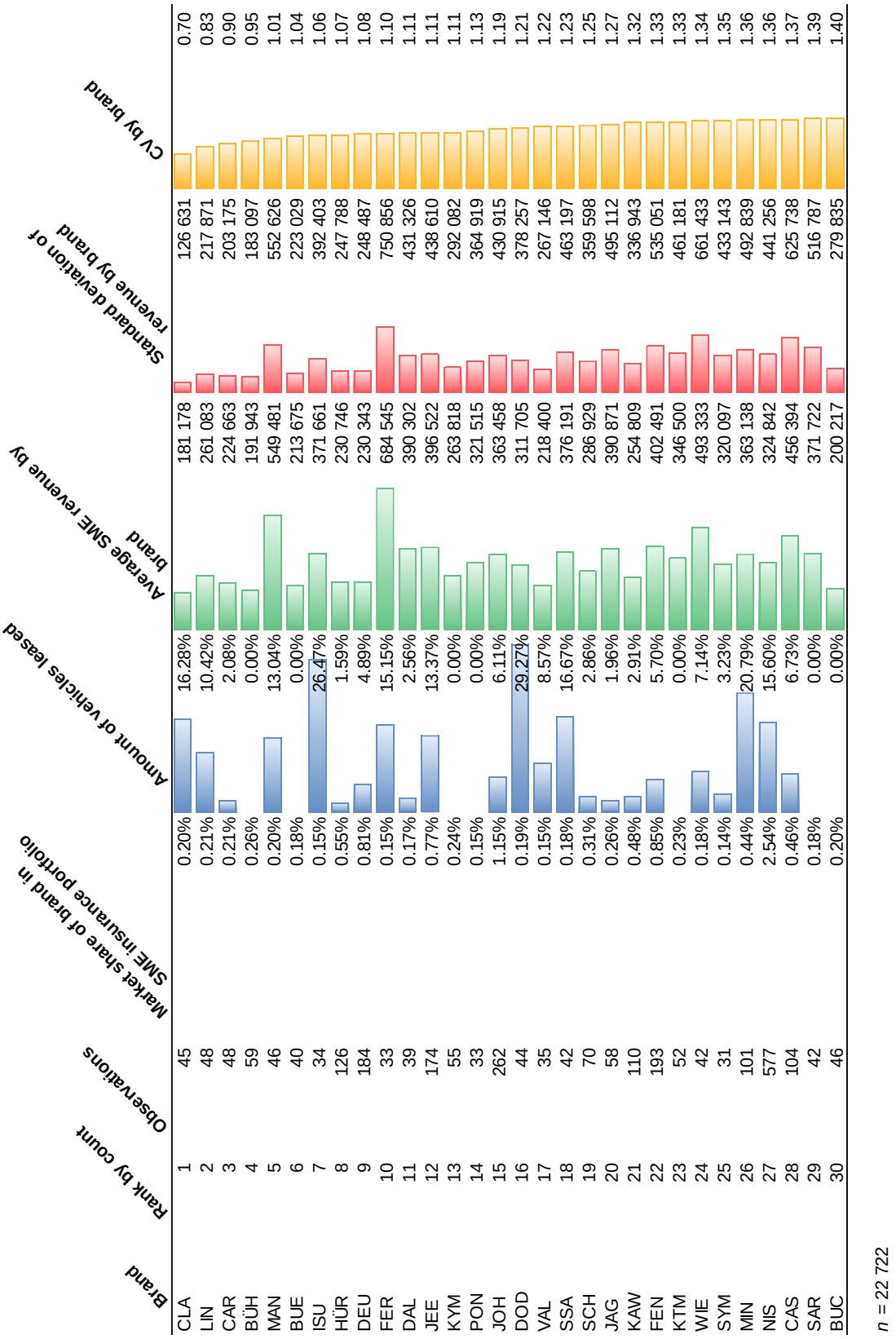
Appendix 21 – Highest CV by brand



Brand	Rank by count	Observations	Market Share of brand in SME insurance portfolio	Amount of vehicles leased	Average SME revenue by brand	Standard deviation of revenue by brand	CV by brand
HUE	1	60	0.26%	5.00%	249 667	891 783	3.57
OPE	2	955	4.20%	10.03%	373 714	1 267 547	3.39
SKO	3	372	1.64%	17.47%	350 209	1 034 457	2.95
TRI	4	69	0.30%	3.85%	504 580	1490 126	2.95
CHE	5	147	0.65%	14.50%	437 893	1 281 887	2.93
FIA	6	725	3.19%	13.83%	465 290	1 286 121	2.76
SAM	7	78	0.34%	3.85%	271 782	734 595	2.70
ALF	8	117	0.51%	20.00%	480 103	1 209 295	2.52
HYU	9	334	1.47%	11.08%	299 355	735 432	2.46
TOY	10	1189	5.23%	9.09%	354 237	851 245	2.40
PIA	11	239	1.05%	1.38%	374 412	873 170	2.33
SMA	12	120	0.53%	7.50%	379 152	871 207	2.30
HAR	13	188	0.83%	6.94%	598 268	1 320 978	2.21
SEA	14	202	0.89%	18.32%	346 811	755 884	2.18
APR	15	48	0.21%	0.00%	425 875	900 360	2.11
HON	16	657	2.89%	6.01%	370 334	775 240	2.09
AUD	17	732	3.22%	17.40%	590 572	1 233 667	2.09
VW	18	1869	8.23%	13.33%	397 863	821 031	2.06
FOR	19	871	3.83%	16.22%	399 505	806 874	2.02
POR	20	102	0.45%	21.51%	1 120 124	2 261 414	2.02
DUC	21	50	0.22%	4.17%	400 960	764 934	1.91
MER	22	941	4.14%	15.53%	523 879	996 988	1.90
CIT	23	644	2.83%	17.78%	383 168	717 351	1.87
LAN	24	273	1.20%	14.34%	436 196	803 636	1.84
CHR	25	100	0.44%	4.04%	413 698	755 353	1.83
MAS	26	150	0.66%	2.04%	455 367	829 981	1.82
SUZ	27	447	1.97%	8.56%	319 198	577 508	1.81
MAZ	28	351	1.54%	10.54%	311 314	559 861	1.80
PEU	29	795	3.50%	11.35%	350 625	623 985	1.78
DAI	30	106	0.47%	10.38%	312 263	547 188	1.75

n = 22 722

Appendix 22 – Lowest CV by brand



n = 22 722

Appendix 23 – Average number of vehicles per SME by industry

Industry	Rank average vehicles observations in industry	Average amount of vehicles per SME	SMEs in industry	Total vehicle observations	Standard deviation of vehicles per SME
Poultry farm	1	14.43	88	1270	9.45
Agriculture (mixed holdings)	2	13.63	486	6623	9.63
Car transportion / trucking	3	12.63	52	657	11.60
Breeding and fattening of cattle	4	12.54	111	1392	8.19
Forestry	5	12.21	63	769	10.39
Excavation/earth moving	6	11.65	55	641	9.51
Carpentry	7	9.95	128	1273	10.96
Gardening	8	9.66	111	1072	8.70
Motor vehicle garage (repair cars / motorcycles)	9	9.33	259	2417	10.89
Civil engineering w/o special focus	10	9.31	140	1304	7.61
Locksmith w.o. sanitary installations	11	8.69	67	582	5.91
Plumbing	12	8.49	94	798	7.23
Landscaping	13	8.26	57	471	6.94
Flooring	14	8.16	126	1028	5.76
Hotel with restaurant/pool/spa	15	7.37	117	862	6.35
Painting and plastering	16	7.34	87	639	5.39
Blacksmith	17	7.22	58	419	6.51
Mechanical workshops	18	7.04	52	366	5.19
Building joinery/carpenter's shop	19	7.00	235	1646	4.97
Joinery	20	6.93	147	1018	4.81
Caretaking	21	6.80	54	367	4.87
Paiting (Real estate)	22	6.75	57	385	4.07
Installation of electrical cables	23	6.64	66	438	5.04
Painting	24	6.62	106	702	6.16
Furniture joinery	25	6.25	55	344	3.88
Restaurant	26	6.23	517	3220	5.49
Dental practice	27	5.71	114	651	4.00
Guesthouse with restaurant	28	5.65	84	475	4.66
Architecture firm	29	5.62	177	994	3.44
Accounting office	30	5.50	64	352	5.37

n = 66 227

Appendix 24 – Most common vehicle brands by industry

Industry	Most common brand	Km	Leasing	Fuel type
Agriculture (mixed holdings)	John Deere	U	no	regular fuel
Restaurant	VW	>=7'000	no	regular fuel
Motor vehicle garage (repair cars / motorcycles)	Honda	>=7'000	no	regular fuel
Building joinery/carpenter's shop	Iveco	U	no	regular fuel
Hair dresser	Opel	>=7'000	no	regular fuel
Breeding and fattening of cattle	John Deere	U	no	regular fuel
Civil engineering w/o special focus	VW	U	no	regular fuel
Carpentry	VW	U	no	regular fuel
Poultry farm	Ford	U	no	regular fuel
Gardening	VW	U	no	regular fuel
Flooring	Fiat	U	no	regular fuel
Joinery	VW	>=7'000	no	regular fuel
General medicine practitioners (w.o. X-ray)	Peugeot	>=7'000	no	regular fuel
Architecture firm	Audi	>=7'000	no	regular fuel
Naturopathic practice/acupuncture/kinesiology	VW	>=7'000	no	regular fuel
Hotel with restaurant/pool/spa	BMW	>=7'000	no	regular fuel
Plumbing	Toyota	U	no	regular fuel
Forestry	Fendt	U	no	regular fuel
Painting	Peugeot	>=7'000	no	regular fuel
Car transportion / trucking	MAN	U	no	regular fuel
Dental practice	Mercedes	>=7'000	no	regular fuel
Painting and plastering	VW	U	no	regular fuel
Excavation/earth moving	Unimog	U	no	regular fuel
Locksmith w.o. sanitary installations	VW	>=7'000	no	regular fuel
Chimney sweep	Mitsubishi	U	no	regular fuel
Repair of other vehicles	Opel	U	no	regular fuel
Viniculture	Ford	U	no	regular fuel
Guesthouse with restaurant	John Deere	>=7'000	no	regular fuel
Landscaping	Nissan	U	no	regular fuel
Installation of electrical cables	Opel	U	no	regular fuel

n = 22 507

Appendix 25 – Revenue growth by leasing status, all years

	SMEs with leased cars				SMEs with owned cars			
Age of Car	Share of SMEs growing	Share of SMEs shrinking	Sample growing	Sample shrinking	Share of SMEs growing	Share of SMEs shrinking	Sample growing	Sample shrinking
0	5%	0%	11	1	3%	4%	15	16
1	10%	6%	40	24	6%	3%	46	23
2	11%	5%	54	25	7%	5%	64	52
3	15%	10%	99	63	9%	5%	126	63
4	12%	10%	107	86	12%	8%	254	161
5	15%	10%	148	104	13%	7%	311	176
6	16%	10%	135	83	15%	7%	338	156
7	19%	9%	125	62	14%	10%	332	234
8	16%	12%	89	63	10%	9%	257	219
9	26%	13%	110	58	14%	10%	395	272
10	26%	16%	84	51	12%	7%	358	203
11	11%	8%	20	14	14%	8%	398	225
12	13%	4%	15	5	12%	8%	348	223
13	8%	12%	7	10	15%	10%	421	293
14	22%	15%	13	9	13%	7%	373	215
15	9%	14%	4	6	14%	9%	408	265
16	54%		7		14%	9%	372	225
17					13%	8%	323	195
18					14%	9%	280	175
19					17%	9%	260	149
20					13%	11%	211	172
21					11%	9%	138	106
22					9%	6%	91	66
23					9%	10%	81	93
24					11%	10%	82	75
25					14%	8%	104	59
26					11%	8%	81	60
27					14%	13%	90	81
28					16%	11%	87	62
29					17%	4%	86	21
30					9%	15%	46	77

n = 66 227

Appendix 26 – CAGR by age of car and leasing status

Age of Car	CAGR leasing	CAGR owning	Observations leasing	Observations owning
new	0.54%	0.09%	207	438
1	0.61%	0.14%	381	821
2	1.01%	0.19%	477	948
3	0.38%	0.86%	661	1 339
4	0.71%	0.52%	896	2 111
5	0.76%	0.42%	1 002	2 471
6	0.11%	1.04%	819	2 307
7	1.42%	0.42%	657	2 444
8	0.53%	1.48%	543	2 545
9	1.35%	1.12%	431	2 733
10	1.24%	1.64%	326	2 891
11	1.67%	1.01%	180	2 847
12	0.21%	1.92%	118	2 968

n = 33 561

Appendix 27 – Most common last name

Position	Insurance Data	Airline Bookings	Phone-book (1)	Phone-book (2)
1	Müller	Müller	Müller	Müller
2	Meier	Meier	Meier	Schmid
3	Schmid	Schmid	Schmid	Meier
4	Keller	Keller	Keller	Keller
5	Weber	Weber	Weber	Huber
6	Schneider	Huber	Huber	
7	Huber	Meyer	Schneider	
8	Steiner	Schneider	Meyer	
9	Meyer	Steiner	Steiner	
10	Gerber	Fischer	Fischer	
Author:	Own	(Hutner, 2018)	(Nyffenegger, 2018)	(Ammann and Reusser, 2017)

Appendix 28 – ANOVA single factor analysis of SME by patent count

Year	Variance of class 0	Variance of class 1	Variance of class 2	F value	p value
2015	0.088	3.909	1.363	153.212	0.000*
2014	0.070	1.972	1.296	195.659	0.000*
2013	0.051	2.547	0.885	132.327	0.000*
2012	0.031	1.426	0.870	161.131	0.000*
2011	0.012	1.109	0.351	99.250	0.000*

F critical value=2.998

Appendix 29 - Paired two test results between revenues of class zero and class one

Year	T value	p value	mean(one)	mean(zero)
2015	5.935	3.31683E-09**	1.600	0.994
2014	7.377	5.15092E-13**	1.530	.995
2013	5.464	4.22044E-08**	1.444	.994
2012	6.064	1.60691E-09**	1.375	1.001
2011	4.470	5.1764E-06**	1.248	1.005

Appendix 30 - Paired two test results between revenues of class one and class two

Year	T value	p value	mean(one)	mean(two)
2015	.806	.210	1.600	1.478
2014	1.298	.090	1.530	1.360
2013	.748	.220	1.444	1.353
2012	.222	.411	1.375	1.351
2011	.337	.368	1.248	1.221

Appendix 31 Paired two test results between revenues of class two and class zero

Year	T value	p value	mean(two)	mean(one)
2015	4.321	1.7274E-05**	1.478	.994
2014	3.346	0.000562704**	1.360	.995
2013	3.978	6.29819E-05**	1.353	.994
2012	3.913	7.98244E-05**	1.351	1.001
2011	3.812	0.00011446**	1.221	1.005

All assuming unequal variances.

Appendix 32 - Paired two test results between revenue growth of male inventing and female inventing SMEs

Year	T value	p value	mean(female)	mean(male)
2015	0.944	0.175	1.769	1.541
2014	1.262	0.107	1.740	1.457
2013	1.110	0.135	1.621	1.384
2012	0.329	0.371	1.430	1.364
2011	0.078	0.468	1.254	1.242

Appendix 33 - Paired two test results between revenue growth of male inventing and mixed gender SMEs

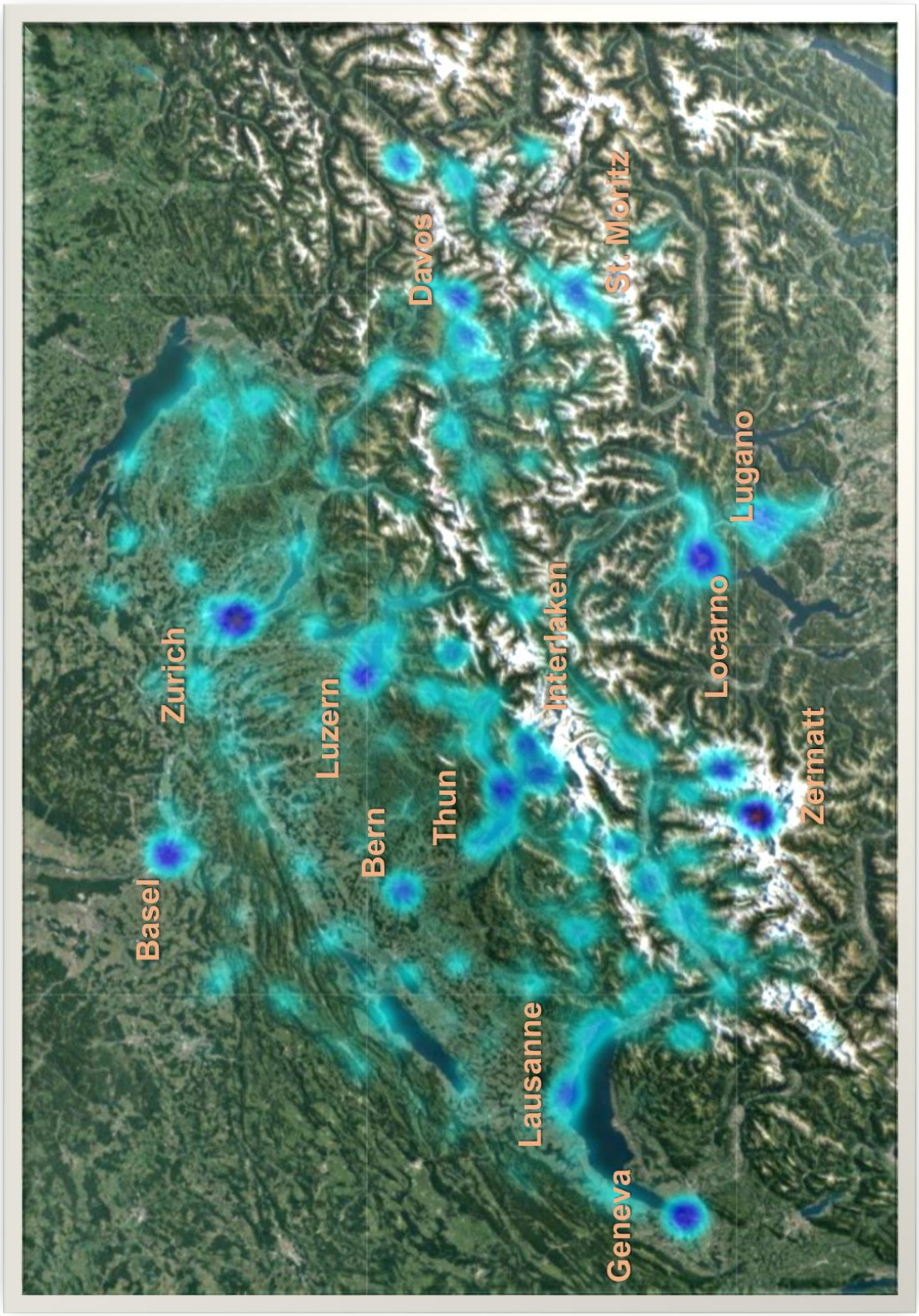
Year	T value	p value	mean(mixed)	mean(male)
2015	2.034	0.049*	1.865	1.541
2014	2.764	0.038*	1.652	1.457
2013	0.518	0.304	1.513	1.384
2012	0.383	0.352	1.458	1.364
2011	-0.593	0.278	1.171	1.242

Appendix 34 - Paired two test results between revenue growth of mixed gender inventing and female inventing SMEs

Year	T value	p value	mean(female)	mean(mixed)
2015	-0.246	0.403	1.769	1.865
2014	0.226	0.411	1.740	1.652
2013	0.349	0.364	1.621	1.513
2012	0.464	1.670	1.430	1.458
2011	0.325	1.668	1.255	1.171

All assuming unequal variances.

Appendix 35 – GPS mapping of Google Review hotel density in data sample



Curriculum Vitae

Personal Information

Name	Daniel Sebastian Müller
Date of birth	21.11.1985
Place of birth	St. Wendel (DE)
Nationality	German

Education

12/2014 – 06/2018	ETH Zurich	Zurich (CH)
	Doctoral studies at the Department of Management, Technology and Economics, Institute of Information Management	
09/2010 – 11/2014	University of St. Gallen and	St. Gallen (CH)
	Copenhagen Business School and	Copenhagen (DK)
	Stanford University	Stanford (US)
	Master of Arts (HSG) in Banking and Finance	
	Master in International Management CEMS MIM	
09/2006 – 06/2009	University of St. Gallen and	St. Gallen (CH)
	Singapore Management University (SMU)	Singapore (SG)
	Bachelor of Arts (HSG) in Business Administration	

Professional Experience

06/2012 – 09/2012	Keiretsu Forum Angel Investor Network Internship, pre-deal screening for acc. investors	San Francisco (US)
08/2009 – 08/2010	Royal Bank of Scotland Analyst, Sales and Trading	London (UK)
11/2008 – 02/2009	ABB Middle East Assistant to the country CEO	Doha (QA)
06/2008 – 08/2008	Accenture Internship, Banking Division	Zurich (CH)