

Big Data Veracity Assessment

Improving risk assessment by adding high veracity data to existing contents
insurance models

Robert Crone
1506870

Delft University of Technology
Faculty of Technology, Policy and Management
Management of Technology

Graduation committee

Chair	Prof. Dr. Ir. Marijn Janssen	TU Delft
1st Supervisor	Assistant professor André Teixeira	TU Delft
2nd Supervisor	Assistant professor Hadi Asghari	TU Delft
External supervisor	Dennis Willems	Aegon

December 5, 2016

Contents

Executive Summary	v
Preface	x
1 Introduction	1
1.1 Big Data	1
1.2 Contents Insurance	3
1.3 Research Questions	4
1.4 Practical Contributions	7
1.5 Academic Contributions	7
1.6 Ethical Considerations	7
1.7 Thesis Structure	9
2 Research Methodology	10
2.1 Research Overview	11
2.2 Action Design Research	11
2.3 Data Collection and Preparation	14
2.3.1 Internal Data	14
2.3.2 Data Brokers	14
2.3.3 Web-scraping a House Trading Platform	14
2.3.4 Social Media – why it is excluded	15
2.4 Modeling Approach, Variable Selection, and Performance Measurement	16
2.4.1 Modeling Approach	16
2.4.2 Variable Selection and Performance Measurement	18
2.5 Summary of the Research Methodology	19
3 Theoretical Background	20
3.1 Veracity	20
3.1.1 Veracity before Big Data	21
3.1.2 Veracity within Big Data	22
3.1.3 Defining Veracity for Big Data research	22
3.2 Big Data Analytics	24

3.3	Data Quality	25
3.3.1	Overview of current methodologies	27
3.3.2	Conclusions on the data quality methodologies	30
3.4	Ethical Big Data Analytics in Insurance	31
3.4.1	Critical questions for Big Data	31
3.4.2	The data supply chain	32
3.4.3	Beyond informed consent	33
3.4.4	The institute of business ethics	33
3.4.5	The insurance industry	33
3.5	Summary of the Theoretical Background	34
4	The Veracity Assessment Framework	36
4.1	Goal and Success Measurement	36
4.2	Context – Organization, People, and Technology	37
4.2.1	Organization: strategies, structure and culture	37
4.2.2	People: roles and capabilities	38
4.2.3	Technology: existing artifacts and processes	38
4.3	Knowledge Base – Foundations and Methodologies	38
4.4	Activity Selection	39
4.4.1	Activity group 1 - Data Quality Assessment	39
4.4.2	Activity group 2 - Entity Resolution	41
4.4.3	Activity group 3 - Model improvement	42
4.5	Application and Evaluation	42
4.6	Summary of Creating the Framework	44
5	Analysis & Results	45
5.1	Internal Data Overview and Baseline Models	45
5.1.1	Data cleaning	45
5.1.2	Internal Data Inspection	49
5.1.3	Baseline Models	53
5.2	Case 1 – Data Broker	55
5.2.1	Data Quality Measurement	55
5.2.2	Entity Resolution	57
5.2.3	Model Improvement	57
5.2.4	Conclusions on the broker’s data	60
5.3	Case 2 – House Trading Platform	64
5.3.1	Data Quality Measurement	64
5.3.2	Entity Resolution	68
5.3.3	Model Improvement	68
5.3.4	Conclusions on the House trading platform data	71
5.4	Evaluation of Case Results	73
5.4.1	Decision-maker and data analyst understanding	73
5.4.2	Evaluation of metrics	73
5.5	Consequences of Advancements in Data Analytics	77
5.5.1	The data collection process	77
5.5.2	The current process versus ethical boundaries	79
5.5.3	The industry perspective	82

6	Conclusions & Discussion	83
6.1	General Conclusions & Discussion	83
6.1.1	Practical Contributions	84
6.1.2	Academic Contributions	84
6.2	Discussion of Sub-questions	86
6.2.1	1 – Criteria for assessing claim risk	86
6.2.2	2 – Data sources that can potentially improve the risk assessment	87
6.2.3	3 – Key characteristics for data sources	87
6.2.4	4 – The value of data quality research to veracity assessment	87
6.2.5	5 – Contributions to Big Data research	88
6.2.6	6 – Consequences of increased data analytics capabilities	88
6.3	Limitations	89
6.3.1	Study design limitations	89
6.3.2	Impact limitations	89
6.4	Recommendations	90
6.4.1	Recommendations to the Organization	90
6.4.2	Recommendations for Future Research	91
	References	93
	List of Figures	97
	List of Tables	100
	Appendices	101
A	Data Definitions	102
B	Modeling Output	103
B.1	Base coefficients	103
B.2	Data broker’s data FA1	105
B.3	Data broker’s data FB1	110
B.4	Data broker’s data SA1	112
B.5	Data broker’s data SB1	117
B.6	House trader’s data H1	119
B.7	House trader’s data H6	123

Executive Summary

Companies and scholars around the world are figuring out how to utilize the rapid growing availability of data for new insight and business value. In both mainstream media and academic literature, Big Data analytics is considered to have huge potential. New challenges are introduced through the nature of Big Data. The amount of data makes traditional storage and analysis techniques inadequate. The speed with which new data is created enables near real time analysis, but only with the appropriate infrastructure. Most data is not created with the purpose of analysis, but as a product of online activity or collected by the sensors we surround ourselves with. This data needs to be structured in such a way that correctly captures the meaning of the original data. Furthermore, the quality, useability, and reliability, also known as the veracity of the data, are often uncertain.

According to literature reviewed in Chapter 3, four steps can be identified in the Big Data analytics chain of a company - data collection, data preparation, data analysis, and decision making. In each of these steps challenges arise due to the Big Data characteristics described. In this thesis the focus is on veracity in the data collection step, because this is where companies currently struggle to evaluate new data. In data collection knowing the veracity of the data is crucial, yet little research has been done into measurement techniques. Data analysts have to combine data from a variety of sources with different origins and estimating the veracity early can prevent time wasted on analysis and potentially bad decision making. This leads us to the main research questions of this thesis.

Research Question – How to assess the veracity of Big Data for the use in insurance risk assessment models?

Evaluating new data sources is a critical skill for contemporary data analysts. Researching how this is currently done and providing a solution firmly grounded in academic theory is both an academic contribution to Big Data research and a practical contribution to any organization looking to advance their Big Data analytics capabilities. This thesis is written from an insurance industry perspective, but most of these challenges apply to other industries as well.

The ethical perspective cannot be overlooked when dealing with personal data on this scale. Some fear that Big Data can lead us into a dystopian future where our free will is lost to companies knowing exactly how to manipulate us and everyone is under permanent surveillance, but the tremendous positive potential of this technology is also widely acknowledged. How a company should navigate into the era of Big Data ethically is also discussed within this thesis.

Methodology For the purpose of creating the veracity assessment framework the action design research methodology is chosen. This methodology combines aspects from both action research and design research. The key characteristics are that it entails development of a technological artifact in an organizational setting, solving a practical problem using academic theory. The academic contributions consist of the artifact itself and a reflection on the theories used to guide future research. The artifact is an example of exaptation, applying known solutions to new problems.

The academic theory used is from the fields of data quality assessment and entity resolution. In these two fields metrics can be found for assessing data that allow data analysts to judge its merits. Several data quality assessment methodologies are evaluated for their applicability to veracity assessment. These metrics are then used to evaluate different types of external data sources.

Different types of data sources are evaluated to test the framework for practicality and usefulness. Two cases are done where one data source is a structured dataset acquired from a data broker and the second data source is semi-structured data from a web based house trading platform. While being semi-structured, the data was converted to a structured dataset before analysis.

The strategic advice for navigating into the era of Big Data ethically follows from a literature review combined with semi-structured interviews with key decision-makers in the insurance company. The internal data acquisition processes are described and analyzed to find whether ethical issues found in the literature are taking into account. Then the impact of advancements of data analytics in the insurance is evaluated and recommendations are given to company on how to deal with the potential consequences.

Results Since no clear definition of veracity exists in Big Data literature, the first step in the research was creating this definition. The definition was made based on a collection of definitions found in academic literature with the focus on Big Data science. A distinction between Big Data science and other disciplines seems to be that the former focus mostly on truthfulness, while usefulness and data quality play a larger role in Big Data literature. In the context of the previous three V's – volume, variety, and velocity – the ability to trust and use the data for decision-making could be represented by veracity. Therefore the following definition was made.

Veracity

The ability of the data to support a decision making process by being appropriate, useful, and of sufficient quality in the context in which it is analyzed.

Collecting and preparing the data the company itself used for making their risk assessment models proved to be difficult due to a recent system migration, causing several iterations between preparation and collection to be necessary before the data was ready for analysis. Data acquired from data brokers was available within the company. Web-scraping as a form of data collection successfully yielded a dataset containing data of over 100.000 houses in The Netherlands. However, gathering data from social media posed new challenges of interpreting semi-structured data among other things that are outside the scope of this thesis.

The veracity assessment framework consists of three parts. The first step is measuring the data quality using the data quality assessment metrics (1) accuracy, (2) completeness, (3) consistency, and (4) timeliness. The second step involves matching the dataset to the existing customer dataset, measuring (5) precision and (6) recall. The third step is recreating the model that must be improved with the originally used variables and then checking if adding variables from the new data can actually improve (7) AIC and (8) BIC scores. These are scores that measure model performance based on maximum likelihood estimation, the number of parameters, and the number of observations. The first step relates to the sufficient quality aspect of veracity, the second and third step to its usefulness.

From developing the framework it has become clear that while the general challenges apply to all companies focusing on data analytics, the context is very important in measuring veracity. The type of data used, the type of decisions that need to be made, and the type of models used are all factors in the current iteration of the framework. Creating a framework that is more general risks losing practical value and since this thesis was set up to develop a useful tool for the business, the framework ended up as specifically useful for evaluating new data for actuarial models. Despite this it is a step in veracity assessment and can be used as a starting point for future research.

The framework is applied to a dataset acquired from a data broker and data scraped from a house trading platform. Initially social media data was considered as well, but it proved to be a challenging target for data collection due to ethical and practical obstacles. Since the social media case was especially interesting to the company, a reflection on its exclusion is given instead of removing it entirely from the report. Converting semi-structured social media data to structured observations profiling the account owner is already a huge challenge. Determining which social media account belongs to which historical customer without a key variable is equally difficult. Finally, doing this in an ethically responsible manner would require explicit permission from each individual account owner, which is simply impractical. A more realistic way of accessing this data would be to acquire profiles from a data broker that has developed a way to deal with these issues. Then applying the framework to the structured profiles gives an impression of their veracity.

The application of the framework in the two cases led to the following conclusions. First of all, the goals of the framework are to be able to distinguish between high and low veracity data sources, in a way that decision-makers can easily interpret, and data analysts must be able to apply it. The scores of the cases show a clear distinction between the broker's data and the house trading platform data. The house trading platform has an overall lower score and this makes sense due to low sample size and the fact that its web-based data converted to a structured dataset. The low sample size means that during entity resolution only a low amount of customers could be described by the data, therefore a slightly looser method of linking the records was used to attempt to describe more customers. This may have reduced the accuracy of the data and may have led to variables being excluded while a richer dataset may have led to higher scores.

In terms of data analysts being able to apply the framework to new data sources and decision-makers being able to interpret the results, this has been checked throughout the research as part of the action design research methodology. The development of the framework under supervision of a data analyst with intermediate presentations to a decision-maker guided the development such that these requirements would be met. A final test is a presentation of the research to the company, but this is scheduled after the deadline for this report. The results will be included in the presentation during the defense.

Reflecting on the metrics used in the framework, the following conclusions are reached. The data quality metrics were especially suited to structured data. As a large part of Big Data consists of semi-structured data, developing metrics that can evaluate veracity based on the limited structure that is available would be useful. In general it could be trustworthiness of the source of the data, analyzing the processes by which it has been created. For textual statements this could be syntax if a relation is found to exist between syntax and veracity. Or it could be case that meta data describing the data contains information useful to assessing veracity. However, if a structured dataset is available, these data quality metrics can certainly provide insight into the data veracity.

Entity resolution has been relatively straightforward in both cases. In the first case the data could be linked based on a primary key which clearly identifies each data record. In the second case a slight less reliable key was used. When considering social media as a data source this phase was one of the main practical obstacles. The absence of a clear identifying variable requires development of algorithms to link the new data to the company data, which is a topic for research by itself. The development of these algorithms could open up new opportunities for data sources to be considered of high enough veracity for use.

In terms of the model performance indicators, using both the AIC and BIC has proven be insightful. In the first case the data could clearly improve the model, scoring high on both AIC and BIC improvement. In the second case, the BIC, the stricter measure of the two, did not improve upon adding new variables. This means that the first case had a higher veracity data than the second, an insight that would have been lost by just using the AIC. It also means that the second data source contained interesting variables, an insight that would have been lost by focusing solely on the BIC. As a final note, it is recommended to also evaluate the coefficient estimations of the variables. Even if a variable improves performance on these metrics, that does not immediately make it suitable to base pricing decisions on. An example of this is included in the conclusions of the first case and the corresponding modeling output can be found in Appendix B.

Overall, the framework can be considered a new perspective on Big Data veracity assessment and opportunities exist to build on it in order to make better assessments.

The ethical analysis includes recommendations on a company and industry level. Through the abundance of both academic literature as well as papers published by insurers associations it is clear that this currently is highly relevant topic. Next to this interviews were conducted with the data protection officer, public relations spokesperson, privacy project manager, customer intelligence manager, two procurement professionals, screening professional, and a legal professional specializing in Big Data. From the company perspective regulatory compliance seems to be the main concern. The spokesperson advocated taking a broader view, involving stakeholders from outside the company. Being able to comply with the new EU general data protection regulation probably coming into effect in the first half of 2018 is already challenging.

From the academic literature the following issues have been identified. During data collection biases may enter the data as a result of the way in which the data is collected, collecting data through mobile apps for example biases data towards young people. Obtaining meaningful consent is a potential future issue, especially after the new EU regulation making rules concerning consent more strict. Furthermore, the ethical practices of data suppliers are not formally checked and the company can be considered at least morally responsible for the way the data was originally collected. Then there are values other than privacy, such as the autonomy of people described, requiring that they have control over their data. Finally, an increased sense of surveillance can be considered a negative externality of a growing data industry, impacting our self-expression. While not all of these issues can be dealt with by the company itself. It is important to keep them in mind. The recommendation is to formalize the data collection process in a way that both includes veracity assessment and an ethical risk assessment.

Insurance companies seek to improve their position in the industry while ensuring no ethical or legal boundaries are crossed. Market forces push insurers to seek competitive advantages by becoming leaders in data analytics and the possibilities that the increasing volume of personal information offers can fundamentally change the insurance industry. For a sustainable insurance industry it is important for insurers to focus on the long term systematic effects of their new business models. Several negative scenarios and their consequences are described. Recognizing these issues, the industry can take a proactive stance in negotiations with public stakeholders and their industry peers to achieve a regulatory situation that allows for a healthy industry contributing to a successful society.

Preface

With pride I present to you my master thesis. This thesis has been written as part of completing the Management of Technology master program at the faculty of Technology, Policy and Management of Delft University of Technology. The research was done during an internship at the insurance branch of Aegon NL, a Dutch financial services company. The internship lasted 6 months, from June to November 2016.

The research concerns the challenge of veracity when wanting to use the potential of Big Data for insurance risk assessment. With a background in Aerospace Engineering and Management of Technology, this topic was relatively new to me at the start of this research. However, once I started reading about it, my fascination for both the potential of data analytics as well as its societal implications grew quickly. If you are reading this because you share this interest, feel free to contact me.

I would first of all like to thank my supervisors. Marijn Janssen, as chair of the committee, for guidance in finding an interesting topic, an internship, and his incredibly fast responses whenever I had a question. André Teixeira, my first supervisor, for his continuous support and ideas on how to improve my research. Hadi Asghari, my second supervisor, for offering unique perspectives and connecting me to researchers with similar interests. Finally, Dennis Willems, my supervisor at the company, for supporting me on a day to day basis and helping me understand the intricacies of the insurance industry. You have all contributed greatly to the quality of this research and also to my continued enjoyment during the process. I hope you enjoy reading the results.

Then I would like to thank my colleagues at Aegon, starting with Folkert van der Ploeg for giving me the opportunity to do this research at an internship. I would like to thank Rob Egelink for always being there to help me understand statistics, the colleagues in my department for supporting me in the entire process, and the rest of Aegon for always making time to discuss my research over coffee.

Finally I would like to thank my friends, housemates, and family for their continuous support. Especially, Esmé Fijn for proofreading and always motivating me to go for my best.

One final note to those reading this from the repository. Given the confidential nature of some of data used, there is a public and private version of this report. So you will find that some parts are blacked out.

CHAPTER 1

Introduction

The increasing availability of data has caused many companies to rethink their business models. Contemplating how their performance can be improved by tapping into this gigantic pool of potential knowledge. Challenges are plentiful. Large amounts of data need to be processed near real-time while ensuring sufficient data quality, meaning needs to be given to unstructured data, and ethical concerns surrounding values such as privacy need to be taken into account. This thesis focuses on discovering the value of Big Data in improving claim risk assessment methods at a Dutch insurance company and how to assess whether this data is fit for use.

First the concept of Big Data and its veracity are discussed in section 1. Section 2 describes the contents insurance industry. Section 3 contains the research question and subquestions, and in sections 4 and 5 the societal and scientific relevance is discussed. In section 6 a reflection on ethical use of data for this thesis can be found. Finally, in section 7 the structure of the rest of this thesis is clarified.

1.1 Big Data

Big data is a buzzword that has attracted the attention of scholars in the last few years. However, the term Big Data itself is an ambiguous concept (Power, 2014). Diverse groups of scholars, physicist, sociologist, computer scientists, mathematicians, and many others are looking for insights from this vast quantity of information produced by and about people, things, and their interactions (Boyd & Crawford, 2012). While an agreed upon definition of Big Data is not immediately apparent from the literature, most scholars focus on the 3 or more V's (Frizzo-Barker, Chow-White, Mozafari, & Ha, 2016; White, 2012). The Volume of data, the Velocity with which it is generated and the Variety in data are larger than ever before. The Veracity of the data is a point of concern though and assessing it is the focus of this thesis.

The idea of large volumes of data is not new. Throughout history there have been technological advances that allowed more information to be stored than ever before. For example books allowed 'mass data storage' in the library of Alexandria 3rd century BC. However, mass storage alone does not define Big Data as it is used today. Its usage became popular more recently, possibly in 1989, when Erik Larson wrote, "The keepers of Big Data say they are doing it for

the consumer's benefit, but data have a way of being used for purposes other than originally intended" (Marr, 2015). This statement seems more true now than ever, when collecting data has become a goal in and of itself, with the idea that a purpose will be found at some point, but it does not hint towards high velocity and variety. Thus, the first people using it with its current meaning in mind were probably at lunch-table conversations at Silicon Graphics Inc. mid-1990s, publishing the first articles at the end of the 90s to give birth to the Big Data discipline (Diebold, 2012). In 1999 the first scientific article to use the term Big Data is published in *Communications for the ACM* (Press, 2013).

The increasing availability of data provides an opportunity for companies to gain a competitive advantage. Companies that leverage the potential of these opportunities are shown to be 5% more productive and 6% more profitable than their competitors (McAfee & Brynjolfsson, 2012). The future of companies might depend on their ability to extract value from this pool of data. With developments such as the *Internet of Things*, where more and more of our devices starts collecting data, the possibilities for insight through data analytics are only going to become more plentiful in the near future, increasing the Velocity, Volume, and Variety of data that is available.

The V's of Big Data present new challenges. How can the important be distinguished from the noise in the large quantities of data? How to deal with the uncertainty and heterogeneity of the data? And all of this preferably in real time data processing applications. Next to finding the right talent needed to operationalize Big Data, data quality is mentioned by companies as one of their main concerns (Schroeck, Shockley, Smart, Romero-Morales, & Tufano, 2012). Especially when using public data, not originally created for the purpose it is now used for, data quality must be carefully examined. When quality is insufficient, the principle "garbage in, garbage out" applies.

Since veracity is integral to this thesis, this characteristic is discussed first. Veracity is defined differently by scholars, some do not even define the term at all, see Table 1.1, but in general it is related to data quality or the data's fitness for use. For this thesis veracity is seen as the fitness for use of the data, which implies a degree of confidence in results obtained from the analysis of this data. When, as is often the case with Big Data, the data is not created for the purpose that it is now used, low veracity of data is a risk and thus needs to be evaluated.

Poor data veracity can be very harmful to a company's performance. While the use of veracity is not yet widespread, its definition it should include some measure of data quality. Data quality itself is already known to be an issue in many areas. When data quality is not checked it might lead to a case of "garbage in, garbage out" (Stvilia, Gasser, Twidale, & Smith, 2007). High data veracity is an important requirement for better predictability in the e-commerce environment and seen as one of the main obstacles for successful use of Big Data (Schroeck et al., 2012). Shell VP of architecture says it is more important to get data quality right than to find a 'Big Data silver bullet' (Hall, 2013). Models that are relied upon to make important business decisions, such as the determining the risk associated with a customer, can cause substantial losses if predictions turn out to be unreliable. When harnessing the potential of Big Data analysis this risk becomes magnified, due to the uncertain nature of this data.

Despite discussions highlighting a need to examine veracity of Big Data very few attempts have been made to do so (Lukoianova & Rubin, 2013). Therefore in this thesis an analysis will be made to identify the risks and potential of Big Data for contents insurance.

Volume refers to sheer amount of data that is currently generated. With communication becoming increasingly digital and thus generating data, with sensors measuring more and more of our daily lives. Our smart phones already contain several sensors and the current interest in Internet of Things means that we will live in smart homes tracking our behavior in our private lives. This leads to a large volume of data being available.

Velocity, the speeds at which Big Data is being generated, is an important characteristic of

Definition	Paper
An indication of data integrity and the trust on this information to make decisions	Lopez, del Rio, Benitez, and Herrera (2015)
Veracity directly refers to inconsistency and data quality problems	Saha and Srivastava (2014)
Low veracity equals uncertainty	Camacho (2014)
Trustworthiness and accuracy of Big Data	Lu, Zhu, Liu, Liu, and Shao (2014)
We use Veracity interchangeably with trust, reliability and credibility	Vlassov, Lozano, Rosell, and Franke (2015)
Objectivity, trustworthiness, and credibility are the main dimensions of veracity	Lukoianova and Rubin (2013)
No definition	Shyr and Spisic (2014)

Table 1.1: Definitions of veracity

Big Data. From the moment we started recording data until 2003, 5 billion gigabytes of data was gathered. In 2011, this amount of data was generated every two days, in 2013, every 10 minutes, and in 2015, every 10 seconds (Zwitter, 2014). This increasing amount of data means that a more complete digital image of our world is being created. With increasing computing power this allows real time evaluation of, for example, customer profiles. This may allow insurers to analyze a customers risk profile at the moment the customer is requesting a quote and give them a price within seconds.

Variety describes the heterogeneity of the data. Sensor readings, text messages, images, all kinds of data is available. It can be automatically generated by technological artefacts or intentionally created by people trying to achieve a certain purpose. Three types of data are typically described by data quality scholars: structured, semi-structured, and unstructured data (Batini, Cappiello, Francalanci, & Maurino, 2009; Schroeck et al., 2012).

1.2 Contents Insurance

The context of the thesis is a model that predicts claim risk of customers of contents insurance based on a variety of categorical variables. Contents insurance is a type of insurance that covers the contents of a house in case of theft, fire, water, and other causes of damages or losses. So the predictions are based on variables such as type of house, family composition, and level of income. Collecting data and making these risk assessment is a key activity of insurers (Verbond van Verzekeraars, 2016). Right now this is estimated using data provided by the customers themselves combined with some commercial datasets that give an indication of additional characteristics.

Key themes in insurance innovation are technology, data and analytics (Freiling, 2015). The scale on which data can be collected these days provide opportunities for insurers to improve their business through several means. On the one hand better risk profiling will allow insurers to offer lower premiums to specific customers without lowering margins and on the other hand it allows for preventive measures that can reduce damages. It is estimated that preventive action can reduce damages by up to 40% of their current values in contents insurance (Hocking et al., 2014). Other opportunities include faster claim handling, better customer service, rewarding customer's safe behavior, and fraud prevention (Verbond van Verzekeraars, 2016). For a South-African insurer fraud losses accounted for 6 to 10 percent of the premium costs, identifying fraud patterns allowed it to both reduce these losses and speed up claim handling for low risk customers

(Schroeck et al., 2012). In conclusion, the finance and insurance industry is considered to have one of the highest Big Data value potentials (Manyika, 2011) and recent industry reports all include Big Data and data analytics as critical area of interest for insurance companies (Freiling, 2015; Hocking et al., 2014; Manyika, 2011; McAfee & Brynjolfsson, 2012; McGranahan, 2015).

The industry is pressured to take advantage of the opportunities provided by Big Data. Fueled by digitization, insurance is becoming increasingly commoditized. Through online portals both large insurance firms as well as intermediaries are able to easily reach customers and insurance comparison websites enable customers to quickly find where they get the lowest premium. This leads to a market with low switching costs where companies are pressured to reduce margins in order to secure market share (McGranahan, 2015). Insurers should be looking to find ways to serve customers more efficiently, encourage brand loyalty, and optimally utilize the available data (Freiling, 2015).

1.3 Research Questions

Given the need for companies to assess what Big Data can mean for them combined with challenges posed by a lack of research into the issue of veracity, the following research question is formulated.

How to assess the veracity of Big Data for the use in insurance risk assessment models?

In order to answer this question, the thesis is divided in the following sub-questions. The rationale behind these questions is shortly described here. The research methodology used to answer them is described in detail in Chapter 2. The answers to all questions can be found in their section of the conclusions chapter.

1. What criteria are currently used for assessing claim risk?
2. Which data sources can potentially improve the current risk assessment?
3. What are the key characteristics of data sources for insurance risk assessment?
4. How can data and information quality assessment methodologies support the assessment of the veracity of these data sources?
5. What can be contributed to Big Data research from the development of the veracity assessment framework?
6. What are the consequences of increased data analytics capabilities in the insurance industry?

Sub-questions

What criteria are currently used for assessing claim risk?

This question will provide a baseline for examining the veracity of the new data sources. The criteria used in the insurance industry are governed by their company policy, modeling technique, and requirements from the regulator. Answering this question is done by examining the method through which models are made, the data that is used as input, and the considerations during this process.

An overview is given of the current types of data used and this data is analyzed, cleaned, and transformed to enable modeling. The challenges here are obtaining the required cooperation

and trust from the company to get access to their data and then cleaning and transforming this data into a workable dataset. Finally a baseline model is made using the tools and techniques that will be used in the framework.

Which data sources can potentially improve the current risk assessment?

This is a key question in insurance. To answer this sub-question several sources of data have been chosen to be analyzed for potential added value. These are social media, an online house trading platform, and a commercially available data set. The use of social media data is highly debated and has ethical implications discussed in sub-question 6. The EU also has legislation that forbids the collection and processing of personal information unless the subject has "unambiguously given his or her consent" (EU, 1995). This means that the focus is on data from a house trading platform and commercially available data. Social media data was initially considered, but excluded due to both ethical and practical obstacles discussed in section 2.3.4.

Adding value to the current assessment means finding new data sources that allow the company to further differentiate risk levels between customers. It will become clear that as part of the framework this value of the new data can be assessed and that the sources evaluated in the two case studies contain valuable data.

This is a challenging question to answer since it requires web scraping, linking different datasets, ethical and legal knowledge, and expert knowledge about what data may be valuable for insurance.

What are the key characteristics of data sources for insurance risk assessment?

This question follows directly from the previous one. From the current practices requirements for new data sources are distilled. In answering this sub-question the foundation for the framework is created. These requirements determine what veracity means in the context of contents insurance. The experience from creating the baseline model with additional data combined with expert opinions from company professionals will be the basis for determining these characteristics.

It turns out that for the purpose of the framework it is most relevant whether the model improved. On top of that the link between the company's current data and the new data as well as the data quality is an important characteristic.

How can data and information quality assessment methodologies support the assessment of the veracity of these data sources?

In answering this question a framework to assess data veracity will be developed. The basis for this framework are methodologies used to assess data and information quality from that field of research. This will allow scoring the chosen data on dimensions on the qualities important for data to be used in contents insurance that were found in the previous subquestion. Concepts are taken from several methodologies summarized in the literature review.

The methodology to develop the framework is a mixture between action and design research, action design research (ADR)(Sein, Henfridsson, Purao, Rossi, & Lindgren, 2011). This methodology is explained in more detail in Chapter 2. The basic idea is developing an IT artifact by applying contemporary academically generated knowledge and adding to this knowledge by reflecting on the development afterward. Next to this the development is done inside an organization in collaboration with the customers. The role of the customers is to provide feedback fueling multiple design iterations.

The resulting framework is not expected to be completely done after the first time answering this question. The framework will be tested on different data sources so it can evolve to a tool

that can be used by company employees. The number of iterations is only limited by time constraints and further research could focus on doing additional iterations or adapting it for a different industry.

Multiple difficulties arise when answering this question. Assessing data veracity by itself is complex, it is context dependent. Requirements to data quality come from the intended use and the practices in the organization. Next to this creating the framework itself is challenging theoretically as well practically. It requires thorough examination of data quality literature to find usable concepts for this framework, adjusting them to measure the attributes relevant to insurance, and determine a composite score from these measurements that gives an accurate indication of veracity of the data. Practical challenges are identifying potential users, motivating them participate in testing, and obtaining useful feedback. Additionally, in terms of Big Data research there are very few attempts at the development such a framework so far. The ones that do exist focus on text interpretation (Ashwin, Kammarpally, & George, 2016; Lukoianova & Rubin, 2013).

What can be contributed to Big Data research from the development of the veracity assessment framework?

When concluding an ADR project a formalization of learning is required. When applying theory from both data and information quality and Big Data research it is expected that insights are developed that can be generalized to add to the body of knowledge on these topics.

This is usually one of the more difficult parts of an (action) design research. The point of view for designing the artifact is from A. R. Hevner (2007) who sees design research studies as creative projects where existing academic theory, artifacts, and actually any source of information can be used as source of inspiration. This notably excludes the need that all design research is grounded in descriptive theories.

Big Data science was the starting point for this research, but it turned out that this science did not include any artifacts similar to the one required here. A few frameworks were available, but they were focused on areas with little resemblance to the aim of this research. Therefore the creation of the artifact itself as well as its application in the two cases is a contribution to Big Data science.

What are the consequences of increased data analytics capabilities in the insurance industry?

Although this question does not follow directly from the main research question the ethical aspects should be discussed, the potential issues will become clear from the literature review and the analysis concerning this question. The increasing availability of personal data has made the use of this data a widely discussed topic and the insurance industry is one of the main industries where this data is valuable. However, the value of this personal data may come as a cost to customers and society.

A reflection is done on the recent debate surrounding the use of Big Data by insurance companies. This is combined with the perspective gained from working inside a large insurance company and the results from trying to improve existing risk assessment models. The achieved improvements provide an empirical indication on the usefulness of Big Data and can support strategic decision making surrounding the expansion of data analytics capabilities.

In the literature several points of view on this topic are found. The company processes defining the acquisition of new personal data are examined and an analysis is made that shows potential issues arising from these processes. It is shown that some issues are not taken into account currently by the company and recommendation on improving these processes are given.

1.4 Practical Contributions

The first practical contribution is the framework itself. This framework can be applied within the company to formalize the analysis of new data sources replacing the current ad hoc process. This will make it easier for decision-makers to evaluate and compare different data sources.

The cases will demonstrate the application of the framework. Both applications concern data the company considers using so the results are a practical contribution as well. Next to being data sources that the company would have evaluated themselves, they are also examples on how to apply the framework for their data analysts.

Furthermore the analysis of the company's current data acquisition process and the associated risks give the company an opportunity to improve it. Suggestion on how to improve will also be provided.

Finally, the insurance industry analysis can facilitate a discussion about the company's strategic direction. Knowing potential future scenarios allows the company to prepare and adapt ensuring continuation of its business.

1.5 Academic Contributions

In the knowledge contribution framework from (Gregor & Hevner, 2013) this research is considered of the type *Exaptation*. This means that it is an extension of known solutions to new problems. There are four academic contributions in this research. Firstly, the application of a relatively new research method, action design research (ADR) and the generalization of the lessons learned from this design process. Secondly, the development of a framework adding a new perspective on veracity assessment. Thirdly, defining veracity broadly. Finally, adding to the ethical debate concerning the use of personal data by insurance companies.

This is an application of the relatively newly proposed research method ADR (Sein et al., 2011). The main idea behind this methodology is the use of theory to develop a new IT artifact that solves a practical problem and generalizing the lessons learned from the design process to add to the existing theory. In this thesis contemporary theory on Big Data and data quality are used to develop a framework for veracity assessment.

Veracity is a relatively new characteristic of Big Data and assessing it is one of main challenges in the Big Data domain (Lukoianova & Rubin, 2013; Vlassov et al., 2015). Table 1.2 shows challenges for the Big Data field from a recent literature review. Veracity is related to inconsistent data, timeliness, data provenance, and trust. An effort to develop a framework that can assess the veracity of new data sources furthers the field of Big Data science, because a framework in this form does not exist. While assessing veracity is one of the current challenges in the field.

When searching for data external to the company that may identify individuals as having a higher or lower risk profile, it is apparent that a discussion on ethics has to be included. This discussion is split into a part concerning this specific study in the next section and a reflection after completing that is included as the last research question. The answer to question can help further the debate by providing empirical results showing the usefulness of personal data as well as provide a perspective on societal consequences from within the insurance industry.

1.6 Ethical Considerations

When collecting personal data and using personal data, there are some ethical considerations that need to be taken into account. Especially when combining this with the insurance industry where these kinds of activities could have negative societal impact. What is technically possible, what

<i>Challenges</i>	<i>Issues</i>							
	Scalability	Source heterogeneity	Inconsistent data	Latency	Timeliness	Processing complexity	Data provenance	Trust, privacy
Cleansing/ acquisition	✓	✓	✓	✓	✓	✓	✓	✓
Storage/ transfer	✓			✓	✓		✓	✓
Analysis/ re- results	✓	✓		✓	✓	✓	✓	✓
Ethical con- siderations		✓					✓	✓

Table 1.2: Challenges versus issues to deal within Big Data, reproduced with permission from (Anagnostopoulos et al., 2016)

an organization would like to do, and what is ethically correct do not always overlap (Chessell, 2014). This section describes the ethical considerations that apply to the study itself and not the consequences of the study. The latter is answered by the last subquestion.

Aspects related to the execution and use of this research that are taken into account are fair distribution of results, consent to data use, privacy, and the company’s data use policies.

Access, Fairness, and Result Ownership

The results of this research will be equally accessible by everyone who’s data was used for it as it will be made publicly available on the TU Delft repository after 1 year.

Consent & Privacy

The data used for this thesis comes from public sources, data brokers, and company employees. The first two create challenges regarding consent as in theory the data could be connected to the individuals through the primary key of postcode and house number combination. The data itself however contains generalizations based on geographical area and does not describe exact know attributes, but attributes inferred based for example neighborhood characteristics. Using this kind of data instead of directly profiling individuals based on their social media activity minimizes the potential harmful impact to these individuals.

Company related considerations

The company has several principles in place when it comes to dealing with data collection about their customers. One is using all customer data to serve the customer to their best abilities. Meaning that if they collect data that allows them to know about damages happening to their customer, they must proactively go to their customers offering to take care of the damages. Or if they find out that the customer’s situation has decreased in risk, they must offer a lower premium insurance. Currently systems to operationalize this are not in place, so the company is reluctant with data collection. Another principle is that they inform the customer of all data used for pricing. When adding open data to their risk assessment models this would mean that all customers have to be informed of this change.

This thesis is seen as exploratory and resulting models will not be used in pricing. The sole purpose is to examine the value of open data, but none of the data used or models created will be

used by the company. The data is solely collected by the author and not added to the company's databases, therefore the above issues are not applicable to data collected for this study.

1.7 Thesis Structure

After clarifying the background and context of this study an overview of the structure of this thesis is as follows.

Chapter 2 describes the research methodology. First an overview of the research is given, stating the general approach and characteristics of this study. Then the data selection, collection, and handling approaches are discussed. The follow section clarifies the process through which the required models are created and lastly the method to obtain the data veracity measurement framework is given.

Chapter 3 contains a literature review on Big Data, data and information quality, and the ethical debate. Three topics of research central to this study. The current state-of-art of research in these fields is determined, the knowledge gap in Big Data Veracity will become clear and concepts from data and information quality are introduced that support the search for answers to the research questions.

Chapter 4 is the chapter where the framework is developed. It starts with the goals, success measurement, context and background theory used and from there activities are defined that could measure the veracity of new data sources. Finally, a reflection on the development is done.

Chapter 5 includes the application of the framework to two cases. The framework is first applied to broker's data and then is to the house trading platform data. This is followed by a reflection on the results of these applications. Finally, combining the knowledge gained from the application and the literature concerning the consequences of increasing data analytics in insurance are used to analyze potential issues in the company processes and future industry scenarios.

Chapter 6 finalizes this thesis with conclusions and discussion of its results. It includes a reflection on the design process, the feedback part of the relevance and rigor cycles can be found here. Finally limitations of this research and recommendation to the company as well as for future research are given.

CHAPTER 2

Research Methodology

In this chapter the research methodology is discussed. The approach used is call action design research, a form of design research with the emphasis on application in an organization. The goal of this chapter is to clarify how this methodology is used during the research, to discuss the data collection approach used for the cases, and to clarify modeling techniques that are used in the organization.

This chapter provides part of the answers for sub-questions 1 and 2. When discussing data collection important characteristics and criteria for risk assessment are both topic that have to be taken into account. This also includes the ethical and practical obstacles that caused the exclusion of the social media case in this research. Furthermore, the modeling approach used for the framework is explained. This approach is similar to the one used by the company and therefore includes part of the current criteria.

The research methodology applied for this thesis is action design research (ADR) (Sein et al., 2011). This methodology is chosen to reflect the need of making a theoretical contribution while solving a practical problem. This type of research focuses on the creation of an IT artifact using theoretical knowledge and generalizing the lessons learned during the project to add to the existing theory.

<i>Property</i>	DR	AR	ADR
Artifact	Central	Peripheral	Central
Organizational impact	Peripheral	Central	Central
Subject participation in research design	Possible	Mandatory	Mandatory
Subject feedback	Discrete	Continuous	Continuous
Transferability	Explicit	Implicit	Explicit
Success measure	Quantifiable measure of artifact behavior	Organizational impact	Organizational learning and artifact generalizability

Table 2.1: Comparing DR, AR, and ADR. Reproduced with permission (Henfridsson, 2011)

This relatively novel research method is an extension of design research with elements from action research. In design research a technological artifact is developed using theoretical knowledge that must be abstracted to develop knowledge again. An issue is that the development and the evaluation of its utility are separated. On top of that the development is rarely in the setting where the artifact should be applied. This leads to a situation where the developed artifact might not serve the needs of its intended users sufficiently. In action research, theory is used to solve a practical problem. The focus is on producing a useful tool for an organization.

The characteristics of the research are defined in Section 1. Section 2 describes in detail the steps in Action Design Research. Section 3 deals with data collection & preparation. Section 4 describes the method behind making the risk assessment models.

2.1 Research Overview

The purpose of this study is threefold, first existing literature will be synthesized in order to create a framework that allows the assessment of veracity of data. The second part is applying this to data collected for the purpose of improving risk assessment in contents insurance. The third is reflecting on the knowledge gained through working in an insurance company, interviews with key people in the data analytics chain, and literature on big data ethics to give strategic advice to the insurance company on their positioning in the near future. An overview of all steps taken can be found in Table 2.3.

The setting in which the case studies are conducted is noncontrived since all data is either from historical records of real world claims or open data that is publicly available and not created for the purpose of this study, merely collected. The unit of analysis in the insurance risk models is the individual policy holder. The characteristics of the house and household are seen as describing the environment that shapes the risk profile attached to this individual. Data will consist of profiles that describe the person making the claim, the independent variables are characteristics connected to the person or the claim.

The veracity assessment framework is first created from the literature and is then applied to the data collected about the policy holders. This provides a clear setting that demonstrates the usefulness as well as allows an iteration on the design to be made with feedback from the organizational feedback during this application.

The time horizon for data collection is cross-sectional, data will only be collected once. The company dataset contains customer characteristics and claims from 2009 to 2014. The commercial dataset is from 2015 and the house trading platform data is from 2016.

2.2 Action Design Research

Action design research aims to introduce elements of action research into design research leading to the following principles in each stage (Sein et al., 2011), see Table 2.2. Principle 1 is that it is practice inspired research, field problems are seen as knowledge-creation opportunities. The goal is to generate knowledge that can help solve the class of problems the specific problem exemplifies. Taking it one step further than a specific solution as would be proposed by a consultant for example. Principle 2 is creating a theory-ingrained artifact. Academic theory is used as basis for designing the artifact, then input from organizational practice is used to reshape the artifact into its final practically useful shape. These two principles have been applied by taking a practical problem from the insurance industry, the need to evaluate new data sources, and using academic theory to create an artifact that solves this problem.

Stages	Principles
1: Problem Formulation	1: Practice Inspired Research
	2: Theory-Ingained Artifact
2: Building, Intervention, and Evaluation	3: Reciprocal Shaping
	4: Mutually Influential Roles
	5: Authentic and Concurrent Evaluation
3: Reflection and Learning	6: Guided Emergence
4: Formalization	7: Generalized Outcomes

Table 2.2: Stages and principles of Action Design Research

Principles 3-5 in the second stage are implemented by working in the company during the execution of this research. Making the artifact while working inside the company, together with its final users, provides immediate feedback and thereby allows the practitioners to influence the framework. Next to immediate feedback from the colleagues worked with on a daily basis, there are two presentation for a small panel to show progress and allow for formal feedback.

Iterations are made between stage 2 and 3. The guided emergence refers to the artifact not precisely following the preliminary design, but emerging from the interplay between organization and theory. This means that the starting point of this research, using data quality methodologies to create a framework with which new data sources can be evaluated, might not be the only piece of theory used in the final version. Requirements might change and other theory might be necessary to adjust the framework.

Finally, after completing stages 1 to 3 an overall reflection on the project is done. This phase is challenging, because the setting of the research is highly situational. The solution to the specific organizational problem is influenced by the type of department, organization, industry, and other situation specific factors. The goal is to generalize this using the following three levels of generalization (Sein et al., 2011). Directions for further research can be suggested based on the generalizations made in this phase.

1. Generalization of problem instance
2. Generalization of solution instance
3. Derivation of design principles from the design research outcomes

<p>1: Problem Formulation</p> <ul style="list-style-type: none"> - Identifying the potential value of external data and the increasing demand for new data in the organization - Identifying uncertainty about Veracity of new data as a challenge in the organization - Developing the general idea of a Veracity assessment framework to assist the organization in evaluating new data sources - Learning the current methodology used for risk assessment in contents insurance - Reviewing literature on assessing model quality and model improvements - Reviewing Big Data literature for attempts at defining and assessing Veracity - Reviewing Data Quality literature for quality metrics and assessment methodologies - Reviewing ethical research practices for projects involving personal data - Reviewing literature discussing societal and ethical implications of using Big Data in insurance - Interviewing employees involved with data acquisition to map the process and evaluate ethical risks
<p>2: Building, Intervention, and Evaluation</p> <ul style="list-style-type: none"> - Defining Veracity - Developing the three pillars of the framework: data quality assessment, entity resolution, and model improvement - Collecting the data currently used in the organization - Creating a baseline model - Evaluation session 1 with 5 employees involved in data analytics - Evaluating new data sources of interest and settling on three sources with different levels of structure. Social media as least structured, house trading platform web data as semistructured, and data broker's data as fully structured - Obtaining permission for use of several social media accounts and collecting data - Scraping data from the house trading platform - Collecting the data broker's data available inside the organization - Exploring issues with the use of social media data - Creating combined datasets with the house and broker's data - Applying the framework to both datasets - Evaluation session 2 within the organization - Developing the final version of the framework
<p>3: Reflection and Learning</p> <ul style="list-style-type: none"> - Constant reflection and learning through working with the actuarial department on a daily basis - Evaluation sessions with key stakeholders within the organization to allow more formal feedback - Reflecting on societal and insurance industry implications of using the increased data analytics capabilities
<p>4: Formalization</p> <ul style="list-style-type: none"> - Discussing potential useful definitions for Veracity in the Big Data field - Discussing generalized lessons about Veracity assessment - Discussing advances in Big Data and Data Quality research that can support Veracity assessment

Table 2.3: Workflow

2.3 Data Collection and Preparation

Data collection and preparation is generally considered to be the bulk of the work for data analysts. The challenge begins by selecting the sources that could be potentially valuable. Next is determine how to collect the data and dealing with potential legal and ethical issues. Practically this means obtaining permissions from various supporting departments within the company. When the data is collected it needs to be prepared for analysis. Difficulty in preparing the data differs greatly depending on the type of data collected and the goal of the analysis.

For this thesis data is collected from a variety of different sources. This naturally follows from the purpose of testing the value of different sources of data for the risk assessment models as well as determining the veracity of these data sources. All data will be put into a relational table and linked to the internal data. The data collection and preparation process is different for each source so they are described separately in the following subsections.

2.3.1 Internal Data

This data is required, because it allows assessment of the current situation. This set of internal is going to be enriched by new data sources in order to improve the original risk assessments made with this dataset. It is already structured and contains characteristics of customers including the frequency and severity of insurance claims. The variables and their coding can be found in Appendix A. The data is obtained from the actuarial department.

This data is expected to be of high quality. Although contents insurance is only a small part of the business, this data is the basis for the policy pricing and it is also potentially subject to scrutiny by DNB (The Dutch Bank), which supervises financial institutions. This combination should make maintaining this data a priority for the actuarial department and therefore ensure trustworthiness and reliability.

2.3.2 Data Brokers

One of main sources already available in the company is data acquired from data brokers. This data is already used for customer intelligence and marketing making it freely available inside the company. It is also expected that acquisition of data through brokers is increasing in the future and therefore this is one of the most important sources to evaluate.

There are no expectations concerning the quality of this data, but the variance in quality between vendors is assumed to be high. It is not expected that results for this dataset correlate highly with future results of other acquired data. The goal in using this data is to ensure that the framework can evaluate these kinds of sources. The acquired data will be provided in a structured format with information about its quality. An overview of the quality and coding can be found in Appendix A.

2.3.3 Web-scraping a House Trading Platform

Web-scraping is an alternative to traditional data collection methods. It is essentially taking semistructured data from a webpage and turning it into a structured form to be able to analyze it. For this thesis the focus is on a house trading platform. The underlying hypothesis is that house characteristics influence the frequency and severity of contents insurance claims. For example, people owning expensive houses may also own more expensive furniture, older houses may have more damages, and gardens might make it easier for houses to be burglarized. The reason why the correlation exists is not further researched.

The results from web-scraping depend highly on the source. If the site is well structured it will be easier to obtain high quality results. The data on the website is in a table-like format and therefore well suited for scraping.

Web-scraping is not illegal, but is in a moral and legal gray zone. The issue is that although the data is public, the target's servers are being used for activities that do not provide any value to the owners of the platform. Also, if selling and analyzing the data is critical to the platform's business model, then scraping it from their website would directly harm them. This is the case for LinkedIn which recently sued scrapers and has relatively strong anti scraping measures (Conger, 2016). The house trading platform scraped for this thesis allows one copy of the data to be saved for non-commercial use, allowing the use of its data for this thesis, but in case the insurance company wants to use it commercially afterward they would need to obtain permission (Funda, 2016). To minimize the harm to the target for this thesis, the scraping is done purposefully slow, respecting the preferences that the platform owners saved in their robots.txt, and only scraping data needed for this research. Note that robots.txt is a document website owners can publish to tell bots where they are allowed to go on the domain, see for example funda.nl/robots.txt.

2.3.4 Social Media – why it is excluded

At the start of this research three different social media sources were evaluated: Facebook, LinkedIn, and Twitter. In After a short evaluation the conclusion is that it is both impractical and unethical to use this data for the research. However, given the interest in the value of this data an exploratory evaluation of the difficulties in collecting and analyzing this data is included.

Each of these sources contains personal information that could potentially be linked to low-risk or high-risk behavior. Facebook contains a wide variety of data, for example likes can predict character attributes Kosinski, Stillwell, and Graepel (2013), someone's status updates might indicate that this person goes on holiday frequently, and the characteristics of a person's friends might be a good predictor of his or her characteristics. LinkedIn contains data about education and professional background. In health and car insurance companies exist that specifically target highly educated people because of their low risk, LinkedIn data could back these claims up and even allow new customers to be profiled using this data. Twitter posts may also include information about a person's hobbies and professional background.

Several challenges exist in collecting social media data. The first challenge is matching social media accounts to the customer database. Duplicate names may exist or social media account names might not even remotely match a real name. The second challenge is extracting the relevant data. Profile attributes are available in a structured way, but analyzing status updates or tweets requires interpretation of data of such a low structure that the results are bound to have inaccuracies. Another challenge is related to the ethics of collecting personal information. Obtaining informed consent from thousands upon thousands of people for analyzing their social media account is simply unpractical. Next to this Facebook and LinkedIn disallow anyone to use data on their website without explicit permission.

These challenges have led to the exclusion of social media data as a third test case. While the organization is interested in the potential of this data, the conclusion was drawn the combination of ethical and practical challenges made this data unfit for assessment at this time. The recommendation is that if this data is still desired after the conclusions of this research, that the company obtains profiles directly from the providers of social media platforms. This makes linking the data more feasible and puts the challenge of consent in the hands of the platform owner. Although as discussed in Section 5.5 on page 77, ethical supply chain responsibilities as viewed from a public perspective have to be taken into account.

2.4 Modeling Approach, Variable Selection, and Performance Measurement

Through scientific research we can obtain an increased understanding of the complex world around us. Models are simplifications or approximations of reality. They are a useful tool to systematically capture observed relationships in a way that is easy to understand and communicate or at least easy to apply to specific problems. The decision of which modeling approach is used for answering a specific research question has a large impact on the type of answer found. Since models cannot include the entire complexity of the universe and each approach is created to best represent a select number of cases, it is important to consider whether the approach matches the goal of the research.

Now that the importance of model selection is established, the goal of model in this thesis needs to be established. Actuarial risk prediction is in essence a search for attributes of customers that affect their likelihood to make insurance claims and the amount of damages claimed. Exactly predicting when, where, and to whom damages occur would require a model of impractical complexity. It depends on the weather, the moods of people involved, possibly animals crossing the road and the disposition of a driver towards that animal. Therefore insurers rely on dividing in people in groups based on attributes that can indicate a different claim risk. This study includes a search for these attributes in data external to the insurance company. Thus post-hoc data analysis leads to the formation of model, as opposed to a priori modeling where a model is made based on theory before looking at the data.

When doing post-hoc data analysis a few issues arise. Inferential properties of post-hoc versus a priori data analysis are very different (Burnham & Anderson, 1998). The main criticism of post-hoc data analysis is that there is high risk of over-fitting to the sample data provided. Furthermore, if a model is created to fit a certain dataset, standard inferential tests are meaningless, because the model is fit such that the variables are 'significant' and have a high precision. Therefore it is important that after creating a model, new data is available that allows testing the model performance in a seemingly a priori fashion.

In general there are two ways to obtain this new data. Gather new data after training your model that can be used to test its performance or, much more convenient, randomly split the original dataset in multiple parts of which one is not used in training the model and is thus equivalent to new data. One drawback of this method is that it will result in a smaller amount of data to train the model on, but in this case with the dataset analyzed having much more observations than coefficients to estimate, this is the preferable approach (Hastie, Tibshirani, & Friedman, 2008). A general rule on how large the samples should be does not exist.

2.4.1 Modeling Approach

Model analysis will be done using generalized linear modeling (GLM) (Nelder & Wedderburn, 1972). GLM is a generalization of traditional linear regression that generalizes in two directions important to actuarial problems. First, deviations from the mean are not limited to normal distributions, it allows to have any distribution from the exponential dispersion family. Secondly, the mean of the random variable may be linear on another scale, for example logarithmic. This means that the model can be multiplicative instead of additive, which better describes actuarial risk (Kaas, Goovaerts, Dhaene, & Denuit, 2008).

Usually risk prediction is therefore split in predicting frequency and severity. Frequency is the amount of claims a customer is expected to make in a given time period and severity is the average amount of damages claimed. So two different models are required. Frequency is best modeled using a Poisson distribution, given the discrete count nature of the variable. Severity is

modeled using the gamma distribution as this is appropriate for variables that are positive and are expected to have a fat right hand tail (Kaas et al., 2008), see Figure 2.1.

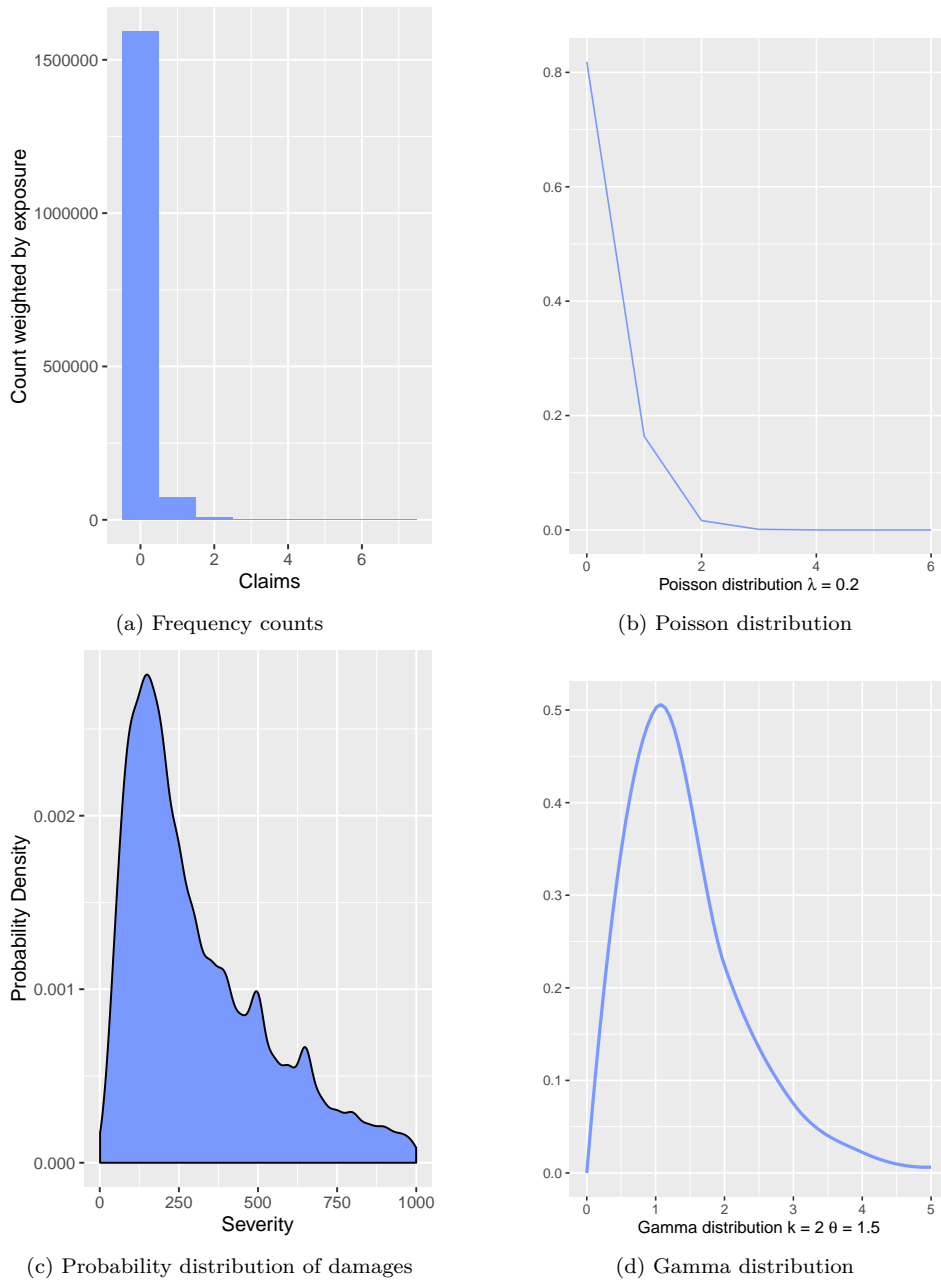


Figure 2.1: Comparing the response variable distribution to the Poisson and gamma distributions

2.4.2 Variable Selection and Performance Measurement

Using this modeling approach, the next challenge is determining how to find the candidate models. Each set of variables represents a candidate model, leaving many options. In some cases candidate modeling is done a priori based on a literature review with data being collected afterward to test the different models, but in this case the data is guiding the model creation. This means that every possible combination of independent variables is a potential model. Two general ways to approach this issue is to either start with the full model and dropping variables or start with an empty model and add variables, respectively stepwise backward or forward model selection. Another approach is to fit all possible combinations of variables and use a method to compare their performance.

While including all available variables will lead to a model with the highest explanatory performance, it also has the disadvantage of making the model complex. In some cases this is acceptable, but for the insurance industry parsimony is valued highly as this makes communicating the reasoning behind the insurance policy premiums more understandable for the customer. One way to make the model simple is to drop terms that do not meet a significance level as measured by a t-test or F-test, but with many tests needing to be done this poses the problem of selecting a relevant significance level (Harrel, 2001). Another approach involves the use information criteria (IC). Using a full IC selection strategy means fitting all possible combinations of variables and ranking them on their IC performance scores.

One of the most popular measures also used in the insurance industry is Akaike's information criterion (AIC) (Akaike, 1974), see Equation 2.1. In this equation k is number of independently adjusted parameters and L is the maximum likelihood. However, the AIC risks selecting over-fitted models, because the punishment for the inclusion of additional parameters is relatively small. To guard against over-fitting it is recommended to be thoughtful when selecting the models to be included in the test or to use a correction that accounts for small sizes (Burnham & Anderson, 1998). Since in this case no a priori models are formulated and the sample size is large compared to the number of estimated parameters, the Bayesian information criterion (BIC) is preferred, see Equation 2.2 (Schwarz, 1978). In the BIC n is additionally the amount of observations, so as long as $\ln(n) > 2$ the punishment for increasing k is larger than in the AIC. The AIC tends to choose models that are too complex as n becomes large (Hastie et al., 2008). Using the BIC should lead to a more parsimonious model, which will also better satisfy the requirement of being able to communicate the rationale behind the insurance policy cost. Both criteria can be used if computational resources and time allow it. Both are used in this thesis.

$$AIC = 2k - 2\ln(L) \tag{2.1}$$

$$BIC = k \cdot \ln(n) - 2\ln(L) \tag{2.2}$$

While in general the model with the lower score is considered to be better, how much better than the second best model is still unclear. Using Akaike weights the relative conditional probability of each proposed models can be calculated (Wagenmakers & Farrell, 2004). Given that we did find the models with the lowest scores, the relative probability of one model being superior the other can be calculated. Thus the strength of support for one model over another model can be determined by dividing their Akaike weights. For a more in depth explanation and examples see Wagenmakers and Farrell (2004). The relative likelihood L of a model i can be estimated through Equation 2.3. The Akaike weights are then calculated by normalizing the relative likelihoods, dividing them by the sum of all likelihoods of all models in consideration, as

per Equation 2.4.

$$L(M_i|data) \propto \exp \left\{ -\frac{1}{2} \Delta_i(AIC) \right\} \quad (2.3)$$

$$w_i(AIC) = \frac{\exp \left\{ -\frac{1}{2} \Delta_i(AIC) \right\}}{\sum_{k=1}^K \exp \left\{ -\frac{1}{2} \Delta_k(AIC) \right\}} \quad (2.4)$$

2.5 Summary of the Research Methodology

In summary, the action research design methodology that is going to be used for creating the artifact should be clear now. All the steps that were taken are outlined in Table 2.3.

Furthermore, the different data sources that will be used for the cases have been described. The base dataset is the internal company data, a dataset with a limited amount of characteristics of customers between 2009 and 2014, containing claim data as well. This dataset will be enriched with characteristics from a data broker's data and data from a house trading platform. Both of these 'new' datasets will be put through the veracity assessment framework as test cases. Finally, it has been explained why the social media data has been excluded as a third case.

Finally, in the previous section, the modeling approach used for the framework is described. The reasoning behind splitting the model in a frequency and severity should be clear. Next to this the chosen response variable distributions for the GLM are supported by Figure 2.1 and the logic behind using a log link function to obtain a multiplicative instead additive version of the GLM for insurance modeling is explained. This modeling technique is used for the two cases as it matches the technique used by the organization and thus can be applied by its analysts. The model performance of AIC and BIC and their differences have also been examined in depth and both will be used to judge model quality.

So in conclusion, we have treated the foundation for this research as well as started answering the first two sub-questions. For sub-question 1, it is clear that datasets should match the assumptions set by the GLM modeling technique. When using small sets of sample data these assumptions could be violated and the response variable distribution may not match anymore and the AIC becomes unreliable when the number of observations/number of variable ration becomes smaller (see AICc for a small sample size correction). For sub-question 2 it has become clear that a soft requirement is that it has to describe a large amount of historical customers. It is called soft, because there is no clear limit, but it is clear that reliability of the results is improved by increasing the amount observations available.

CHAPTER 3

Theoretical Background

Defining the academic background of the artifact is an essential step in design research. According to the ADR methodology it being theory ingrained artifact is required, see Table 2.2 on page 12. This chapter will describe the theoretical background. Four topics are examined. We start with veracity, defining the veracity characteristic so goals for the framework can be set. Secondly, Big Data analytics is examined. This forms part of the high level context in which the framework will be situated. Thirdly, data quality, from which theories will be used to create a method for assessing data veracity. Finally, basic ethical papers discussing the increasing use of personal and big data in the insurance industry.

This means that sub-questions 1, 2, 4 and 6 are partly treated in this chapter. The part big data analytics section provides insight into current company practices and thereby supports answering sub-question 1 concerning criteria for assessing claim risk. In terms of requirements of data sources for sub-question 2 the big data analytics and data quality chapter provide background needed to define them. Sub-question 4 is answered partly through the data quality assessment literature review, where concepts useful to veracity assessment are found. For starting to answer sub-question 6 the literature review on ethical big data analytics is provided.

3.1 Veracity

Veracity as a recent popular addition to the V's of Big Data does not have a uniform definition in academic literature so far. As stated in the introduction, every paper found uses a slightly different definition. Since Veracity is the central concept in this thesis, it is important to clearly define it. Attempting to find a definition that can be used in generally in Big Data research is also one of the academic contributions.

First prior use of the term in other academic literature is examined. Second is a search for the reason that this concept has been added to the Big Data characteristics. Third an overview is given of common definitions in recent academic literature. Finally, a definition is selected that can be useful to this thesis and related future research.

3.1.1 Veracity before Big Data

The concept of veracity has been in use in academic literature long before the rise in popularity of Big Data research. Examining the most cited articles using veracity on Scopus shows that the veracity of hypotheses and maxims is often discussed. Another common use is as the reliability of self-evaluations. The following section contains quotes from the most popular papers by number of citations using veracity in these ways. These are included to see its use in different contexts.

Examples of the use of Veracity

Hardy (1997)

It is argued that these genetic and molecular biological data provide strong support for the *veracity* of the 'amyloid cascade hypothesis' for disease pathogenesis, and that this hypothesis offers a coherent framework for drug discovery.

Langlois et al. (2000)

Results are used to evaluate social and fitness-related evolutionary theories and the *veracity* of maxims about beauty.

Kamel Boulos, Maramba, and Wheeler (2006)

While arguably, the very process of collaboration leads to a Darwinian type 'survival of the fittest' content within a Web page, the *veracity* of these resources can be assured through careful monitoring, moderation, and operation of the collaborationware in a closed and secure digital environment.

Landrum and Bone (2001)

Prospects for future research in the study of macular pigment require new initiatives that will probe more accurately into the localization of these carotenoids in the retina, identify possible transport proteins and mechanisms, and prove the *veracity* of the photoprotection hypothesis for the macular pigments.

Ellison, Heino, and Gibbs (2006)

Qualitative data analysis suggests that participants attended to small cues online, mediated the tension between impression management pressures and the desire to present an authentic sense of self through tactics such as creating a profile that reflected their "ideal self," and attempted to establish the *veracity* of their identity claims. This study provides empirical support for Social Information Processing theory in a naturalistic context while offering insight into the complicated way in which "honesty" is enacted online.

Ceci, Ross, and Toglia (1987)

Some of the apprehension about the *veracity* of children's recollections has arisen from a concern over the testimony provided by children during the Salem Witch Trials and been fueled further by research carried out around the turn of the century suggesting that children could not be trusted to accurately recount events. Today, all states have corroboration rules mandating that the testimony of a child be confirmed by another person prior to its being accepted as evidence in a court of law.

Earl, Martin, McCarthy, and Soule (2004)

Studying collective action with newspaper accounts of protest events, rare only 20 years ago, has become commonplace in the past decade. A critical literature has accompanied the growth of protest event analysis. The literature has focused on selection bias - particularly which subset of events are covered - and description bias - notably, the *veracity* of the coverage. The "hard news" of the event, if it is reported, tends to be relatively accurate.

Del Boca and Noll (2000)

This paper examines factors that influence the *veracity* of verbal self-report data in health services research, using a cognitive social-psychological model of the data-gathering process as an

organizing framework.

Liu, Wu, and Zidek (1997)

The basic condition on the error distribution required for the *veracity* of our asymptotic results is satisfied by any distribution with zero mean and a moment generating function (having bounded second derivative around zero).

3.1.2 Veracity within Big Data

The original 3 V's were originally defined by Laney (2001) and then an IBM employee was the first source found to coin Veracity as the fourth V (Snow, 2012). The reasoning being that trusted data in a Big Data setting cannot be compared to a traditional data that was purposefully collected. New processes and definitions are required to deal with the challenges that Big Data comes with.

After veracity being adopted by IBM it started appearing in Big Data research in 2013. In the Chapter 1 an overview of definitions used was already presented in Table 1.1. The amount of research of using this concept is still limited and every scholar has his own definition. A definition encompassing commonly used attributes and which will be used throughout this thesis is presented in the following section.

The attributes used in Big Data literature slightly differ from other definition in the sense that concepts such as usefulness accompany concepts such as trustworthiness and reliability. In the context of Big Data next to the other three V's – volume, variety, and velocity – veracity could indicate whether the data can improve decision-making. This makes definition slightly broader than its original meaning of truthfulness, but supports the development of practical measures focusing on supporting decision-making processes.

3.1.3 Defining Veracity for Big Data research

As a final source before determining how veracity can be defined, the dictionary definition can be consulted (Dictionary.com, 2016).

Veracity – noun, plural veracities

1. habitual observance of truth in speech or statement; truthfulness: He was not noted for his veracity.
2. conformity to truth or fact; accuracy: to question the veracity of his account.
3. correctness or accuracy, as of the senses or of a scientific instrument.
4. something veracious; a truth.

In conclusion data veracity describes a closeness to truth on a higher level than a measure such as accuracy does. Veracity is more comprehensive data quality measure. High veracity data is data that can be relied upon when making business decisions, data that meets minimal truthfulness requirements, data that describes the world in a way that is useful to the task at hand. Therefore the short definition of 'fitness for use' is proposed in this thesis. The full definition of veracity is proposed to be as follows.

Veracity

The ability of the data to support a decision making process by being appropriate, useful, and of sufficient quality in the context in which it is analyzed.

Historically veracity has focused mostly on closeness to truth. This definition is a slight departure from that focus by making it context dependent. In the definitions in Table 1.1 the idea to take the decision making into account is already found. This makes sense, because ultimately data analytics is a tool to improve decision making. This also creates a definition that is easily translated to requirements for its assessment.

In this definition the focus is on whether the data will be able to serve the goal it was collected for. Appropriate refers to selecting the data, there must a reasonable logical expectation by experts that the data can explain that which is analyzed. Useful is about whether the data actually allows the organization to make better decisions by using it. Finally, data quality should be a part of this, because this shows whether the data itself can be relied upon by decision makers.

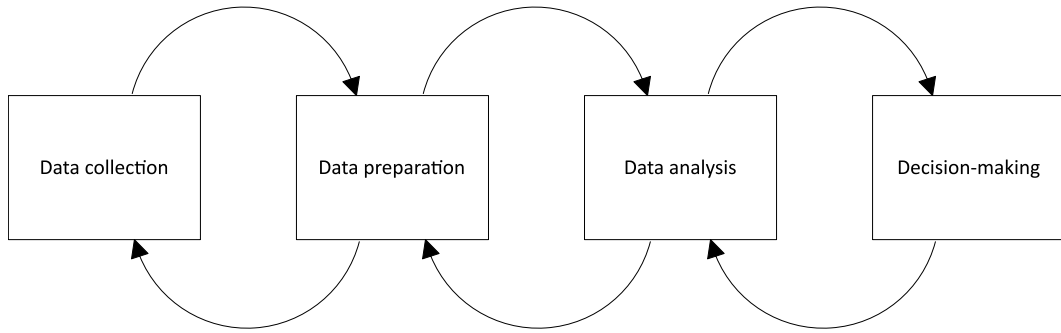


Figure 3.1: Steps and transfer points in the big data chain, reproduced from (Janssen et al., 2016).

3.2 Big Data Analytics

Combined with the rise of big data research is a growing interest in big data analytics. Analytics are essential to creating business value from the ever growing amount of data. Since big data and its characteristics have already been sufficiently introduced in the introduction, this section focuses on how data flows through a company from acquisition to the creation of value. The big data analysis chain from Janssen, van der Voort, and Wahyudi (2016) is used, see Figure 3.1.

Three types of data are collected by the insurance company. The first is data collected directly from customers when applying for an insurance policy, the second is data collected through customer activity on the company's online platform and the third is data acquired indirectly e.g. from other companies. Currently only data gathered directly from customers is used in pricing, but the challenges of using the other forms are explored in this thesis.

These other sources of data vary highly in veracity and therefore should be examined carefully before being used in decision-making. Third party data can be bought from data brokers or acquired directly from internet sources. Especially the latter can pose problems, because the source often contains unstructured or semistructured data. Uncertainty about the veracity of data can make decision makers reluctant to use the conclusions following from the analysis. The framework created in this thesis will help evaluate new data sources in a way that can provide clarity about the veracity and potential value of these new data sources.

The data collected is stored throughout the company with each department having access to their own datasets. Using data from other departments can be done on an ad hoc basis, through personal networks within the company. For using personal data special permissions are required.

Data preparation is done by the analysts themselves. Every analyst is responsible for creating a dataset that they want to use for analysis. While a data quality methodology is being implemented in several departments at this time, the project is still in its infancy and details regarding roles and responsibilities still need to be determined. In the next few years a central database will also be developed so that a centralized *truth* is available for analysis.

While currently the burden of data preparation rests mostly on data analysts themselves, which has also been a major time sink during this research, improvements are on the horizon.

Here it is important to distinguish between the different departments inside the company. The analysis and reporting standards differ. A large part of the data analysis, related to the analysis in this research, is actuarial. Actuaries have their own professional standards and best practices following from actuarial research. Becoming a certified actuary requires passing exams and keeping up with the current state-of-the-art through courses and conferences. As such the

analysis methods are restricted to current best practices and deviation from these norms would require thorough justification.

Then there is non-actuarial data analysis in departments such as marketing. Data analysts in these departments have much more freedom in experimenting with different types of analysis such as new machine learning techniques. Reporting the results naturally is linked to analysis techniques used and therefore differs per analysis.

Decision-making is distributed depending on type of decision. Small operational changes, like finding a specific type of customer that has much higher claims than projected and therefore stopping to accept new customer in that category can be done with the operational manager of the claims department. When price changes are proposed by the analysts the pricing board consisting of managers from operations, customer intelligence, and other departments has to approve it as well as the product approval board. The product approval board checks the effect of the proposed changes on customers, whether the amount of money returned to customers is fair and whether some customers are not impacted negatively in an unfair way.

3.3 Data Quality

In the analysis of data quality research there are two main objectives. The first is finding dimensions of data quality and their definitions so that an appropriate subset can be selected for the veracity assessment framework. The second is an examination of current data quality assessment methodologies to find how these can assist the development of the framework. The goal of this assessment is to determine fitness for use of the data.

One of the challenges of big data veracity assessment and improvement is that the volume and velocity of the data are so great that any methodology needs to be computationally efficient and scalable. This means that traditional data quality assessment methods might not be usable anymore. However, before creating such an approach, a theoretical framework is required that allows for the assessment of the big data veracity. Creating this framework the big data characteristic variety becomes the main challenge. The heterogeneity of big data calls for a framework that is capable of assessing many different types of data. In order to assess semistructured or even unstructured data either (a sample of) this data must be structured before analysis or the source can be examined.

Data quality research has focused on many different aspects and methods of data quality assessment. In this thesis the focus is on assessing data for usefulness and not the organizational processes causing this. The methodologies chosen as starting point are therefore based on assessing the quality of individual data points. These methodologies can be found in Table 3.1.

The methodologies in Table 3.1 are the data-driven strategies identified by Batini (2009). The process-driven strategies are neglected, because they are not relevant to answering the research question. After examining initial search results based on citations, a waterfall literature search has been done to find papers related to quality assessment and papers defining the research area. The resulting definitions of data quality can be found in Table 3.2.

Data and information are used interchangeably unless indicated otherwise. In some contexts a distinction is made between the two, where information means data that has been processed, but most data quality scholars do not make this distinction. In the rest of this thesis data is chosen to avoid confusion, but for the literature review sometimes information is used when the original author preferred it.

Selected Methodology	Description
Total Data Quality Management (TDQM) Wang (1998)	A general-purpose methodology that can be applied in different contexts, the first general methodology published
Data Quality Assessment (DQA) Pipino, Lee, and Wang (2002)	Provide general principles from previous research. Uses subjective and objective metrics. Metrics divided into: simple ratio, min or max value, and weighed average.
Cost-Effect Of Low Data Quality (COLDQ) Loshin (2001)	Provides a data quality scorecard to measure the cost of low data quality for businesses.
Heterogeneous Data Quality Methodology (HDQM) Batini (2011)	Extends Complete Data Quality Batini, Cabitza, Capiello, and Francalanci (2006), which focuses on data quality without context to include semi-structured and unstructured data.

Table 3.1: Data quality assessment methodologies

Metric	Definition
Accuracy	The percentage of data where the descriptions match the real-world objects
Completeness	The degree to which a given data collection includes data describing the corresponding set of real-world objects
Consistency	The percentage of semantic rules violated in the data set. Can be intra-relational constraints and inter-relational constraints
Timeliness	The time passed since the original data was created
Volatility	How quickly the data changes through time

Table 3.2: Data quality metrics to be used for the framework

3.3.1 Overview of current methodologies

Here the methodologies selected in Table 3.1 are described. Starting with their objective and general aim and ending with specific parts of the methodology that may be useful for veracity assessment.

Total Data Quality Management

TDQM by (Wang, 1998) was the first published approach at a formal data quality management methodology. He starts from Total Quality Management (TQM) approaches in other fields, comparing information products to physical products. An information manufacturing system produces these information products and the products have quality dimensions similar to physical products. The following quality dimensions are identified and divided over four categories.

1. Intrinsic IQ
 - Accuracy
 - Objectivity
 - Believability
 - Reputation
2. Accessibility IQ
 - Access
 - Security
3. Contextual IQ
 - Relevancy
 - Value-added
 - Timeliness
 - Completeness
 - Amount of data
4. Representational IQ
 - Interpretability
 - Ease of understanding
 - Concise representation
 - Consistent representation

The information manufacturing systems contains four roles - information suppliers, information manufacturers, information consumers, and information product managers. Suppliers provide the input data, manufacturers maintain the information infrastructure, consumers use the data for their for their work, and product managers are responsible for the system.

As the author notes, the choice of dimensions should be made in view of the context and business goals. Data-centric veracity assessment limits the choice to evaluations made by examining the data as it is. Accuracy, timeliness, completeness, and consistency allow this kind evaluation. These are also identified by the author as the basic IQ dimensions and can be applied to a wide range of data types. The roles, while perhaps useful for a company wide data quality program, are not directly useful to the framework.

Data Quality Assessment

DQA is a methodology created because the authors found that fundamental principles for developing data quality metrics were lacking (Pipino et al., 2002). A distinction is made between objective and subjective assessments. Objective assessments are divided in task-dependent and task-independent, the latter reflects the quality of the data without contextual knowledge. For the purpose of this thesis the objective assessment methods are most useful, since those allow assessing the quality based without involving any potential users and are therefore time and cost effective.

Three functional forms are defined, see Table 3.3. The methodology suggest using these functional forms to grade the data between 0 and 1 on the relevant dimensions. This allows comparison between data.

Simple Ratio	Measures the ratio of desired outcomes to total outcomes
Quality dimensions	free-of-error, completeness, consistency, concise representation, relevancy, and ease of manipulation
Min or Max Operation	Handles dimensions requiring aggregation of multiple data quality indicators
Quality dimensions	Min - believability, appropriate amount of data Max - timeliness, accessibility
Weighted Average	An alternative to min or max when deemed more useful

Table 3.3: Functional Forms of DQA (Pipino et al., 2002)

Cost-effect of Low Data Quality

COLDQ focuses on the economic losses caused by poor data quality (Loshin, 2001). Potential issues related to poor data quality range from service interruptions and scaling difficulty to customer attrition and higher employee turnover. The steps in creating the data quality scorecard are as follows.

1. Map the information chain
2. Interview employees
3. Interview customers
4. Isolate flawed data
5. Identify impact
6. Characterize the economic impact
7. Aggregate the totals
8. Identify opportunities for improvement

It is clear that this methodology focuses on the process mostly, but it does offer the following five data quality of data values measurements. These quality measurements are of the current state assessment and are the only part discussed in more detail. The other parts of the methodology are focused on the process within the organization so are not applicable to new external data sources.

1. Accuracy - The degree with which data values agree with an identified source of correct information. Requires a comparing values with an identified source of correct information and calculating the ratio of correct values.
2. Null Values - Missing values can have different meanings and ratios can be calculated for each of these.
 - Not applicable for this entity
 - Not available for this entity
 - No value in the domain of this attribute that correctly classifies this entity
 - Actual missing value
3. Completeness - The ratio of expected values to total values. What falls within expected must be defined per attribute.
4. Consistency - In this assessment method it refers to consistency between data sets. When linking two data sets consistency can be checked by examining fields that should contain similar entries.
5. Timeliness - The degree to which the information is up-to-date. Requires defining time criteria and constraints. Can be measure using a ratio of up-to-date versus total data or by measuring how much too late the data is delivered to the customer.

Accuracy can be easily implemented by taking a small sample of the data. When acquiring an external dataset claiming to know certain attributes about all inhabitants, a small group of employees can be used to verify the data describing them. Null values can be easily counted and the owner of the data should know their origin. Completeness requires more effort, because an expected set of values needs to be defined for every attribute. When enriching the data, consistency cannot be defined unless parts of the data overlap. When this is the case, for example when a dataset from a data broker contains not only new data, but also data your customers provide you, this can provide insight into the accuracy of the new data as well. The importance of timeliness is highly context dependent. When making predictive models, the most important aspect is that the age of the record is well documented. Also, many types of personal data do not change very often, follow a predictable pattern, or never change at all. Thus, the measurements used and their corresponding definitions depend on both the goal of the data customer as well as the nature of the data source.

Heterogeneous Data Quality Methodology

HDQM is a methodology that focuses on determining the accuracy and currency of all data in an organization, structured, semi-structured, and unstructured. The main idea is "to map the information resources used in an organization to a common conceptual representation and then to assess the quality of data considering such homogeneous conceptual representation" (Batini, 2011).

In this paper the main activities of DQ methodologies are said to be *State reconstruction*, *Assessment/Measurement*, and *Improvement*. State reconstruction is the phase where the databases and data flows are reconstructed so that the quality can be assessed. The outcome of this phase is clear object to investigate using a DQ methodology. During assessment and measurement the data quality is scored according the metrics designed to provide insight into the quality such as the simple ration or weighted average functions. Finally, during improvement changes are made to the process and database with the intention to increase the assessment scores and the cycle

can start again. Two data quality measures are used, accuracy and currency, but the claim is made that this method can easily be extended to include other measures.

Accuracy is defined as how close a value r_i is to another value r in domain D . The author states that the main type of accuracy proposed in DQ methodologies is syntactic accuracy and defines this as the “the distance between r and the elements belonging to D ”. To measure syntactic accuracy it uses the formula 3.1.

$$Acc(t) = \frac{\sum_{i=1}^t 1 - NED(r_i, D(r_i))}{|t|} \quad (3.1)$$

NED is a function that returns a value between 0 and 1 based on the *normalized edit distance*, if there is an exact match between r_i and another element in domain D it will return 0. For values greater than 0 it describes the minimum number of character insertions, deletions, and replacements to convert a value r_i to a value in D . How the total number of edits should be converted to a value between 0 and 1 is unclear from this paper.

To measure currency the normalized currency is used. This is defined as the ratio between actual and optimal currency. The advantage is that normalized currency give a result in percentages which shows the relative importance, instead of a result in the time dimension where direct comparison with related metrics is impossible. Normalized currency is defined as per formula 3.2.

$$Cur = \frac{OptimalCurrency}{ActualCurrency} = \frac{OptimalCurrency}{Age + DeliveryTime - InputTime} \quad (3.2)$$

Where *Age* is defined as the age of the data when it is received from the data provider and *DeliveryTime - InputTime* is the time the data is in the organization before it is used. It is assumed that 1) *Age* > 0, since there will be a delay between a real-world change and a change in the data of the data provider and 2) *OptimalCurrency* = *Age*, since it is impossible to have data more up to date than the data provider.

3.3.2 Conclusions on the data quality methodologies

In conclusion data quality researchers have developed useful metrics and techniques for data quality assessment that can be applied in the veracity assessment framework. The basic metrics of accuracy, completeness, consistency, and timeliness can be used to evaluate any dataset by purely examining the data itself and can therefore be used to evaluate all new data. Metrics relying on surveys among users are only applicable in cases where the data is already part of existing processes and can therefore not be applied to new data sources.

TDQM as the first data quality assessment methodology introduces a wide range of measure that can be used to describe data. Most of these are subjective and context dependent and therefore not applicable to our problem. The roles and information products described are useful in evaluating data analytics processes, but not directly usable in the framework.

DQA introduces mathematical principles to evaluate data quality. For some metrics, depending on the context, different methods of calculations are available. The decision which ones to use will be made in the framework section.

COLDQ reinforces the importance of the 4 data quality metrics selected. It also includes a definition of Null values that can be applied to completeness. Measuring the cost effect, while interesting, is not necessary for veracity assessment though.

HDQM is the first methodology that includes the idea of unstructured and semi-structured data in its evaluation of data quality. However, the conclusion is that data will need to be

structured before quality measurement is possible. Most importantly it offers the idea of syntactic accuracy, where accuracy can be measured without acquiring a dataset containing a ground truth.

It should be noted that this literature review only includes basic background and parts of the methodologies that were deemed useful for this thesis. The actual depth of the methodologies is much greater. While DQA is described in a less than 10 page paper, COLDQ is part of a book describing an enterprise knowledge management methodology. The summaries of their useful parts should thus be viewed as just that, a summary containing only information distilled for this specific use.

3.4 Ethical Big Data Analytics in Insurance

Big data is also a topic that can inspire visions of an Orwellian future, some may argue that we are already there (Schneier, 2015). This thesis would be incomplete without assessing the potential consequences of increased personal data analysis in the insurance industry, since this research is moving it forward. This part of the literature review will support answering the fifth research question.

Several applications of big data in the insurance industry also have clear societal benefits. The detection of fraud and damage prevention can lead to lower premiums and less hassle for customers. Creating systems where customers are directly rewarded for safe behavior can reduce the moral hazard effect of insurance. The moral hazard effect refers to the effect where being insured promotes riskier behavior. A driver will generally be more careful when he knows that his premiums depend directly on his driving style.

New data can also allow customers to be divided into finer grained risk pools, splitting them more precisely according to their risk profiles. It can be perceived as fairer to be in the same risk pool as people exhibiting similar behavior and who therefore have similar risk profile. Why would one have to pay for the recklessness of others? This is a popular argument by insurance marketers to justify more fine grained risk pooling, but there is a fine line between differentiation and discrimination as we will see from the literature.

3.4.1 Critical questions for Big Data

Boyd and Crawford (2012) has written the single by far most cited article when looking for 'Big Data' and 'ethics' on scopus with 543 citations (#2 has 152, #3 has 85). This article discusses many potential issues with Big Data and here the ones that have an impact on the data analysis in the insurance industry are summarized.

Big Data opens new possibilities for the social sciences to quantify and analyze human behavior. Combined with the belief that quantitative research produces facts this can create the impression that social science can produce more objective results thanks to Big Data. There are however many layers of subjective interpretations before the final analysis is made. The models and experiments may be sound, but the data used is not neutral. Which variables needed to be included, how they were collected, and what happened during the ever more important data cleaning step all influence the results in a subjective way.

This is relevant to the data science in insurance, because the new risk factors discovered in their growing data analytics departments are not objective as well. Knowing and accounting for biases in data becomes harder to do when the dataset become combinations of data from several different sources with ever higher numbers of variables. In the case of an insurance company, there may be a bias in the current customer base, that makes conclusions from data analysis inherently biased. If new products are created based on these analyses the bias might lead to a different bottom line than expected.

Another issue raised is the ethics of data collection. In the case of Big Data it becomes infeasible to ask every person for explicit consent. Data being available does not automatically make it ethical to use this data. Users of data should ask themselves where the data comes from, under what circumstances was this data created, and whether the people it describes agree or at least benefit from this use of their data.

A common justification for using data is that the people it describes were in public, so their actions could be recorded by anyone sharing the same public place. In this paper a distinction is made between being public and being in public. Where public people actively try to get attention, people in public just happen to be in a public space and do not necessarily want anyone to record their actions (Boyd & Marwick, 2011).

Although insurance companies can solve this by incorporating some form of consent into the agreements they make with their clients, whether this form of consent will remain sufficient in the future with the new EU data protection rules remains to be seen (EU, 1995).

3.4.2 The data supply chain

Kirsten Martin (2015) published an article evaluating ethical issues within the big data industry using a supply chain perspective. The chain starts with tracking companies, the original data manufacturers. The second step is data aggregators, companies that combine data from several sources and sometimes add additional through profiling. Lastly, there are the consumers of this data, companies that make business decisions based on the content of the data. In this supply chain the insurance industry can be a source of data, but in most cases is a customer of the data aggregators. An important notion is that downstream companies are responsible for checking the ethical practices of their business partners upstream. This is true for electronics manufacturers, clothing designers, and also insurance companies. The end user is also responsible for checking whether the quality of the data is sufficient.

Another consequence of viewing the big data industry on a system level is the question of negative externalities. This refers to transactions where a third party not at the table is negatively impacted. In manufacturing this a common example is the public paying for cleaning up pollution from factories. Similarly Martin looks for negative externalities of the big data industry. She identifies unnecessary surveillance as the pollution created by this industry. Surveillance changes how individuals behave and may hinder personal and social development (Rachels, 1975).

The harmful effect can be further divided into *Value destruction*, *Diminished rights*, and *Disrespect* (Martin, 2015). Incorrect data often causes these effects, but correct data can as well. Correct data can be disrespectful to share or sell when its reason for existence was limited to certain purpose. Some data are only given in a transaction with authorities or companies for a specific event. When data brokers are selling lists of crime victims, people with addictive behavior, and mailing lists for domestic violence shelters (Hicken, 2013) then this is disrespectful to the persons it describes. It seems unlikely that any people on these lists have agreed to this information being sharing or receive benefits from marketeers using this to create personalized ad campaigns.

Big data can cause breaching of privacy, facilitate discrimination, and distort the power relationship between those that have it and those that do not. Every beneficial application of big data comes with the risk that the data is used for unethical practices. *Black box* car insurance, basing premiums on driver performance, may promote safer driving and fairer distribution of premiums. On the other hand, the location data can be used to create extensive profiles of the customers. Some companies even publicly hail the *Internet of Things* for being able to read the customer's minds using the data collected (Moran, 2016).

The idea that users disclose information with a certain purpose and that there exists an

implicit confidentiality agreement in this transaction is proposed by Martin. This idea is similar to the *spheres of justice* from Michael Walzer (Walzer, 1982), also used by Christen et al. (2015) discussed next. The general idea is that different social spheres have different types of goods and that these spheres have to be kept separated. If the spheres mix this could lead to health care waiting list rankings being based on family relationships or parliament seats being allocated based on financial assets. Mixing of spheres can lead to unfair decisions (Romei & Ruggieri, 2013).

3.4.3 Beyond informed consent

Christen et al. (2015) proposed extending the idea of informed consent when using personal data. They claim that "recent developments make it questionable that the consent route is a sufficient and meaningful expression of autonomy in the context of big data". Many people implicitly assume that the data share does not leave the setting they share it in and this assumption is violated increasingly. Furthermore, the extend to which data is collected and stored also exceeds expectations. For example, people are often not aware of all meta-data that is created. Therefore three values with their corresponding definitions are proposed that should guide creation of a new moral landscape.

1. Autonomy - Users ought to be aware of how their data records are used in order to promote their values and gain control over privacy related choices
2. Responsibility - Users should be held responsible and accountable for the ways in which they use their personal information and the information about other people. If some subjects are wronged, it must be possible to attribute personal responsibility for the wrongs in question.
3. Fairness - The benefits of knowledge and information ought to be fairly apportioned to all participants in interactions, so as to rule out inequality of opportunity and exploitation by some at the expense of others.

3.4.4 The institute of business ethics

As the Institute of Business Ethics notes in their recent briefing, in complex systems, the best intentions can still lead to discrimination (ibe, 2016). When acquiring new data the method by which it is obtained may create a bias. For example when collecting data through apps, people without a phone are not represented. A famous example of this is Boston's pot hole app discriminating against the old and poor, the app used sensors in the phone to identify potholes. However, areas with a low amount of smartphone users were left out and therefore this method favored improving the neighborhood of wealthy and tech savvy citizens.

Owning extensive amounts of personal data also distorts the power relationship between company and customer. In extreme cases marketing departments can develop methods that exploit vulnerabilities of their customers instead of trying to find the best value proposition. In insurance this could take the form of engineering feelings of insecurity specifically targeting recent victims or otherwise risk averse people followed by offering the corresponding insurance policy (Calo, 2013).

3.4.5 The insurance industry

The insurance industry themselves have also published recommendations on how to ethically use the large amount of data that is widely available. The Dutch Verbond van Verzekeraars

(association of insurers) has published a green paper on big data with a proposed list of acceptance criteria when evaluating new data sources. The message they want to send to insurers is that mere compliance is not enough. The use has to match customer expectations and customers need to have control over their personal data. The following aspects are defined in the paper as critical factors in using big data.

1. Customer benefit - Do customer and society benefit from this use of data?
2. Transparency - Is the customer aware of this use of data?
3. Predictability - Is the customer aware how their data could be used?
4. Compliance - Is this use of data legal?
5. Accuracy - Does the data accurately describe personal characteristics of the customers?
6. Understandability - Can the insurer link the data to the results of the model?
7. Rectifiability - Can the customer rectify wrong personal data?

In the UK the Chartered Insurance Institute (CII) has also published papers on big data and ethical data use (CII, 2015; Minty, 2016). Providing less specific guidelines than the Dutch insurance organization they do also note that merely compliance is not sufficient. The main ideas are bringing the public voice in at times of key decisions ('how would this look when discussed with friends?') and challenging justifications for sidestepping public interest. Their concern is that public confidence in the sector can be harmed while this is critical for a sustainable insurance sector.

Interestingly, the UK insurers do not mention furthering public interest in the way that the Dutch insurers do, but focus on maximizing profits while carefully managing public confidence. From the UK perspective 'the public have to experience outcomes, both individually and collectively, that they accept or at least recognize as necessary' (CII, 2015). From the Dutch perspective we have to maximize the value for society, customer, and insurer and be responsible in using big data.

3.5 Summary of the Theoretical Background

In this chapter we have taken several important steps defined in the workflow in Table 2.3. A definition of veracity is created so that we know what the framework should measure. The academic literature that will be used as inspiration for the artifact is discussed as well as the literature that forms the starting for the discussion of the long term consequences of data analytics in the insurance industry. Next to this the Big Data analytics chain is used to clarify where the framework is positioned conceptually, namely in the data collection phase.

By doing this sub-question 4 and 6 are both partly answered. Sub-question 4 concerned the use of data quality methodologies as theoretical underpinning of the framework and from the literature several concepts were identified that are deemed usable. These concepts are accuracy, completeness, consistency, and timeliness. These four were chosen, because they can be applied on the data as is, without taking into account opinions of the people using the data. They are not completely objective, the data analyst still has to make a judgment when it comes to what is an acceptable timeliness value, but it is limited to the judgment of the person applying the framework. The question will be fully answered in the post case evaluation where the metrics are examined one by one.

Sub-question 6 has been partly answered by collecting literature that contains future scenarios and potential issues for data analytics in the insurance industry. Both issues with data collection as well as industry development were considered. Issues range identified concern what constitutes meaningful consent for the use of personal data, how to take the customer's privacy into account, the relevance of other values such as fairness and autonomy have to be considered, and what industry institutions see as potential issues. This question will be further answered after the two cases where the application of the framework is discussed.

The Veracity Assessment Framework

In this chapter the steps toward creating the veracity assessment framework are explained and the current state of the framework is described. The steps defined are inspired by the steps outlined for developing a DQ assessment technique (Woodall, Borek, & Parlikad, 2013). First, the goal and success measurement of the framework are clearly defined. Second, the practical context in which the framework will be developed is described. Third, the knowledge base is described, the methodologies and processes currently in place that the framework has interfaces with or is replacing. Fourth, activities must be defined and these activities should be logically ordered. Finally, a first reflection on the frameworks limitations is done along with describing the cases it will be applied to.

This framework was developed during an internship at the insurance company at the actuarial department. Referring to the Action Design Research (ADR) principles, see Table 2.2, this ensured authentic and concurrent evaluation. Next to this during the development of the framework two feedback sessions were held that included a decision-maker. Together this led to an environment where reciprocal shaping could happen.

4.1 Goal and Success Measurement

The insurance industry is looking outside of company data for new insights. Many potential data sources are available online publicly or are being sold by data brokers. The usefulness of this data is often unknown and there is a need for a clear framework that allows new data to be explored and evaluated.

The goal of the framework is to assess the veracity of new data sources. The veracity, defined as fitness for use, is measured by a combination of metrics from data quality assessment methodologies, entity resolution, and the model selection criteria AIC and BIC. Referring to the full definition of veracity in Section 3.1, the appropriateness of the data should be determined beforehand by the analysts when deciding what new data to acquire. When that is done this framework can be used to determine the quality of data and the potential usefulness of the data.

In order to determine whether this framework actually achieved this purpose two test cases are done in the next chapter. The two test cases are two realistic situations in which this framework

could be applied. The first case is using data obtained from a data broker, a method of accessing new data often used by the company. The second case concerns data collected especially for this thesis from a web source, a house trading platform. This is a method currently not used by the company, but with investments in data analytics skills this is a realistic option for future data collection.

Success of the framework is measure by evaluating whether the results of the application led to results that decision makers can understand. Furthermore it is important the upon reflection of the outcomes the ratings obtained are meaningful. Meaningful means that the different ratings obtained for the different cases reflect differences in veracity. At the end of the analysis a reflection on this success criteria is done. So concluding the following success measurements are defined.

1. Metrics can be understood by decision makers
2. Data analysts should be able to apply to framework
3. Differences in scores should indicate differences in veracity of the data sources

4.2 Context – Organization, People, and Technology

In order to reach these goals it is important to first understand the context in which it is to be applied. To facilitate our understanding the environment is defined using the information systems research framework as defined by A. Hevner, March, Park, and Ram (2004) as guideline. First, an overview of the organization’s strategies, structure, and culture is given. Second, the people in the organization that are stakeholders in this research are described. Third, the technological artifacts and current processes that this framework seeks to replace or has interface with are identified.

4.2.1 Organization: strategies, structure and culture

The organization for which this framework is designed is a financial institution with a broad product line including mortgages, pension funds, insurances, and related products. The department in which the framework is developed is the actuarial department focused on individual home contents insurance, but it should be able to be applied to any new personal data considered for analysis.

The organizational insurance strategy that led to this type of framework being valuable is one of increasing differentiation in insurance policy pricing. Increasing differentiation will allow the organization to offer competitive prices to low risk customers and, if it achieves this before its competition, grow market share. It will also allow the company to make better estimation of the value of its current portfolio and determine which customer are worth keeping. Increasing differentiation requires rich databases and thus created a demand for new data.

During this research the structure of organization was going through a major transformation, but before and after the transformation this organization has a relatively flat structure. The amount of management layers between the majority of employees and the board is around 1 to 2 depending on product line. Employees are expected to work independently and proactively.

The culture is correspondingly open and egalitarian. In general any other person in the organization will make time for meeting (new) colleagues and be open to discussing ideas.

4.2.2 People: roles and capabilities

The people in the organization that are stakeholders of this framework can be divided into decision-makers and data analysts. Decision-makers consist of the managers of the data analysts that must determine which new data is worth the investment. Data analysts identify new data and must evaluate its merit to report to the decision maker.

The manager of the data analysts has to set the focus for the analytics team. He determines which analyses are relevant for the business and has can start a process for the acquisition of new data. The capabilities of the decision maker include understanding where to create business value, interpreting the results of analytics, and facilitating the growth in data analytics capabilities.

The data analysts have to find and exploit opportunities for data to create insight. The capabilities of the data analysts include working with relevant analytics tools (i.e. R, Excel, Emblem), understanding potential issues with data, and knowing the right modeling strategy for the insights needed from the data.

4.2.3 Technology: existing artifacts and processes

The process currently used for the assessment of new data is an ad hoc process where the appropriate method of evaluation depends on a brainstorming session among data analysts.

The process before the framework is the process that results in the identification of new data. This similarly to the current method of evaluating the data where brainstorming among the analysts creates leads to be investigated. After obtaining (a sample of) the new data the framework should be able to judge it. Note that even if the whole dataset is already available, the velocity aspect of Big Data leads to the need for periodic updates of this data and the decision to invest long-term can still be made based on the framework conclusions.

The process that follows after the framework is where decision-makers use the results to obtain funds from the company enter enter into a long standing relation with a data supplier or set up recurring data collection efforts within the company.

4.3 Knowledge Base – Foundations and Methodologies

A key part of designing a technological artifact using action design research is the scientific grounding of the artifact. Again the information systems research framework as defined by A. Hevner et al. (2004) is used as a guideline. The foundation has already been discussed in Chapters 2 and 3, but for clarity the relevant parts are referred to here.

The first part of the framework relies on the data quality assessment methodologies discussed in Section 3.3 on page 25. The methods used in the framework are a combination of these methodologies.

The second part of the framework was added during the first attempts at application of the framework. During the construction of the extended datasets connecting the two emerged as a challenge with potential implication for the usefulness of the data. For the development of this part metrics used by Bhattacharya and Getoor (2007) have been implemented.

The third part of the framework is a direct result from the actuarial practice used in the insurance company. The modeling approach is discussed at length in Section 2.4 on page 16. The assumptions underlying this modeling technique and the associated performance measurements have been chosen to match those used in practice and therefore support the interpretation of results by users of the framework.

4.4 Activity Selection

The framework split into three main parts that can be analyzed in parallel. The first group of activities belong to the data quality assessment, the second group of activities come from entity resolution, and the third group of activities comes from modeling techniques.

4.4.1 Activity group 1 - Data Quality Assessment

From the literature review on data quality assessment methodologies the data-centric metrics in Table 4.1 were found that can be applied to all new structured data sources. Measuring each of these metrics defines the activities in this part of the assessment. Figure 4.1 is the workflow corresponding to the tasks outlined in this section.

<i>Metric</i>	<i>Definition</i>
Accuracy	The ratio of data where the descriptions match the real-world objects
Completeness	The ratio and meaning of missing values
Consistency	The ratio of semantic rules violated in the data set. Can be intra-relational constraints and inter-relational constraints
Timeliness	The difference between date of customer data creation and date of new data creation

Table 4.1: Data quality metrics included in the veracity assessment framework

Accuracy

From the data quality methodologies two main methods of measuring accuracy were found. The first is measuring accuracy by having a trustworthy source of data to compare new data to. However, if that source exists and data can be easily collected from there, then the data from source of unknown trustworthiness is irrelevant. The other option is discussed in the HDQM, see 3.3.1, where a domain of accurate possibilities is defined and the accuracy is measured by calculating the shortest edit distance to any value in that domain. This second method can be applied when a domain can be defined and can be used to find whether the values at least belong to right group of values. Method 1 measures *semantic accuracy* and method 2 measures *syntactic accuracy* (Batini & Scannapieca, 2006). The exact definitions are as follows.

1. Determining the ratio of semantically accurate values out of a sample of the total values
 - Semantic accuracy - Measuring the closeness of the value v to the true value v' (Batini & Scannapieca, 2006)
 - Can be used if a sample of trustworthy data is available for comparison
 - A sample can be too small or biased
 - Collecting new data can be cumbersome
 - If a source of trustworthy data is readily available there is not reason to acquire the new data

2. Determining the ratio of syntactically accurate value out of the total values

- Syntactic accuracy - The minimum edit distance from value v to a value in domain D . Where domain D contains all possible true values of v (Batini & Scannapieca, 2006)
- Can always be used
- Does not find syntactically correct, but semantically incorrect values

Syntactic accuracy can always be measured and is therefore recommended to start with in any case. However, for example in cases of acquired data where data brokers have made their own inferences, syntactic accuracy is almost guaranteed. The possible results of their models should only include syntactically accurate values. However, semantic accuracy is highly uncertain. Therefore acquiring a sample of trustworthy data is highly recommended. When the dataset includes personal details, employees could be asked to check their personal details in order to get an indication of semantic accuracy. This will create a bias based on the nature of the company, but if it is large enough and includes people from all kinds of socioeconomic status and background it could provide a reliable estimate.

Completeness

Within the framework, evaluating the completeness of a dataset means evaluating the missing values for each variable. Thus the focus is on column completeness. Completeness most importantly shows what percentage of customers are described by the new data. The missing values can have several different meanings taken from the DQA methodology, see 3.3.1. The definitions are as follows.

Determining the ratio of the different types of missing values to total values

1. Value does not exist in the real world
2. Value exists but is unknown
3. Value exists but is not included in the domain of allowed values
4. Unknown whether value exists

For example, if a potential customer does not own a house, the characteristics of his house would be missing because they do not exist in the real world (and contents insurance would not be very useful to this customer). A value that exists, but is unknown could be any missing that must exist in the real world, such as a missing date of birth. A value outside of the domain of allowed values could be a customer living in a castle. Finally, when the house type of a customer is not known, the type of missing value cannot be classified in any other category than that it is unknown whether the value exists.

Consistency

Consistency measurement is the most subjective of all data quality metrics, because it requires the formulation of rules that the data has to adhere to. After these rules are defined, a ratio of adherence to these rules measures consistency. Examples of rules could be as follows.

Determining the ratio of compliance to different semantic rules, such as

1. Customer age has to be 18 or higher
2. If the *Life phase* variable includes children, *Family composition* should also include children
3. House volume divided by surface area should generally be in the 2-3 range

Timeliness

Timeliness in this case refers to how appropriate it is to link customer records from different datasets. The important characteristic is data creation time. The actuarial model currently uses customer data from 2009 until 2014. Newly acquired datasets could contain outdated values.

When dealing with volatile data, data that changes rapidly over time, timeliness has to be correspondingly low. Houses for example change slowly, so taking house data from 2016 and applying it to 2009 should still be valid in most of the cases. However, if the presence of young children is an important factor in causing insurance claims, then 7 years makes all the difference.

4.4.2 Activity group 2 - Entity Resolution

Entity resolution concerns finding which entities in the different datasets belong to the same real-world entity. In this case the challenge is identifying which observation in the new dataset describes which customer. Two metrics are used to determine the quality of record linking - precision and recall. The definitions of these metrics can be found in Table 4.2.

<i>Metric</i>	<i>Definition</i> (Elmagarmid, Ipeirotis, & Verykios, 2007)
Precision	The percentage of correct matches out of all matches
Recall	The percentage of correct matches out of all possible correct matches

Table 4.2: Entity resolution metric included in the veracity assessment framework

These activities are only required when a deterministic matching procedure with a unique key is not possible. In case of a unique key both metrics should be 100% and linking the records is a trivial effort. However, when a unique key is not available a matching algorithm has to be developed for a probabilistic approach. Entity resolution seems to be main field of science where research into linking datasets is performed, but literature can be found with many other terms as well (e.g. record linkage, deduplication, object identification, etc.). Precision and recall measure the performance of the probabilistic approaches, the major issue in applying these techniques is that testing the algorithms requires a training set to be generated.

Constructing a training set involves manually checking a set of potential matches and even for humans it is possible to make errors doing this. One option for doing this is using crowdsourcing via platforms like Amazon Mechanical Turk, but handling human errors remains a challenge. Another option is using unsupervised or semi-supervised techniques, but the risk of false positives is not well known in that case which is not suitable for use in pricing insurance. The cost of incorrect risk assessments leading to underpricing or overpricing can be high both in terms of attracting unprofitable customers and losing potentially profitable ones.

An example of matching challenges can be found in the evaluation of social media data. If this kind of semi-structured data is considered to be valuable then further research has to be done into developing matching algorithms for each specific source. Precision and recall are then the metrics to be optimized. Fortunately house platform's and broker's data can be matched through primary key.

4.4.3 Activity group 3 - Model improvement

In order to determine the potential usefulness of the data the two information criteria BIC and AIC can be used. Which one to use should depend on considerations discussed in Section 2.4, but both can be used to get different perspectives. The BIC will force stricter feature selection and therefore generally suggest less additional variables to be added to the model than AIC.

The optimal feature selection can be done automatically through exhaustively testing all possible combinations of variables or through an algorithm that attempts to find the optimal combination in a more intelligent and less computationally intensive method. The risk with the latter remains that a local optimum is found instead of the global optimum. For this thesis the `glmulti` package is selected to optimize variable selection (Calcagno, 2015).

This step requires having a dataset merged with new data. This new data has to be evaluated to see whether the linked records represent the entire base dataset well. If for example the new data only contains information about people living in villas then it might improve predictions for this specific group, but have far less overall added value than indicated by this test.

When the data has been examined a model is created using the base variables included in the original model, then a model is created where the new variables are included in the optimization. Potential interaction effects can be included as well. If the optimization leads to new variables being added to the model then the difference in AIC or BIC score gives an indication of the potential value of the new data.

4.5 Application and Evaluation

Application of the framework is done using the test cases in the following chapter. First the framework is applied to a dataset extended with the data broker's data. This case shows how the full framework should operate. Second is the data scraped from the house trading platform, showcasing difficulties with entity resolution and potential solutions as well as the potential value of this data.

Before actually applying the framework an evaluation can already be done on its design so far. It does not fully cover the process when acquiring new data for example. Additional considerations when evaluating new data sources are cost and ethical boundaries.

The framework will also give more accurate assessments for larger sample sizes. A small sample can lead to unreliable results for all three categories of activities. The sample has to be large enough relative to the original dataset to be representative for the data quality assessment results. The sample also has to be able to match a set of customers that is representative for the entire customer database. If, for example, the new data sample is heavily biased toward low income customers the measured model improvement might only be achieved for low income customers. The original data collector could have used collection methods that specifically appeal to a low income population and therefore also score higher on data quality metrics in the sample than on the entire dataset. Although the data still increases insight in customer risk, the cost/benefit ratio might be worse than expected.

A weakness inherently in the framework is that it only applies to already structured data sources. While the issue with big data is the large amount of semistructured and unstructured

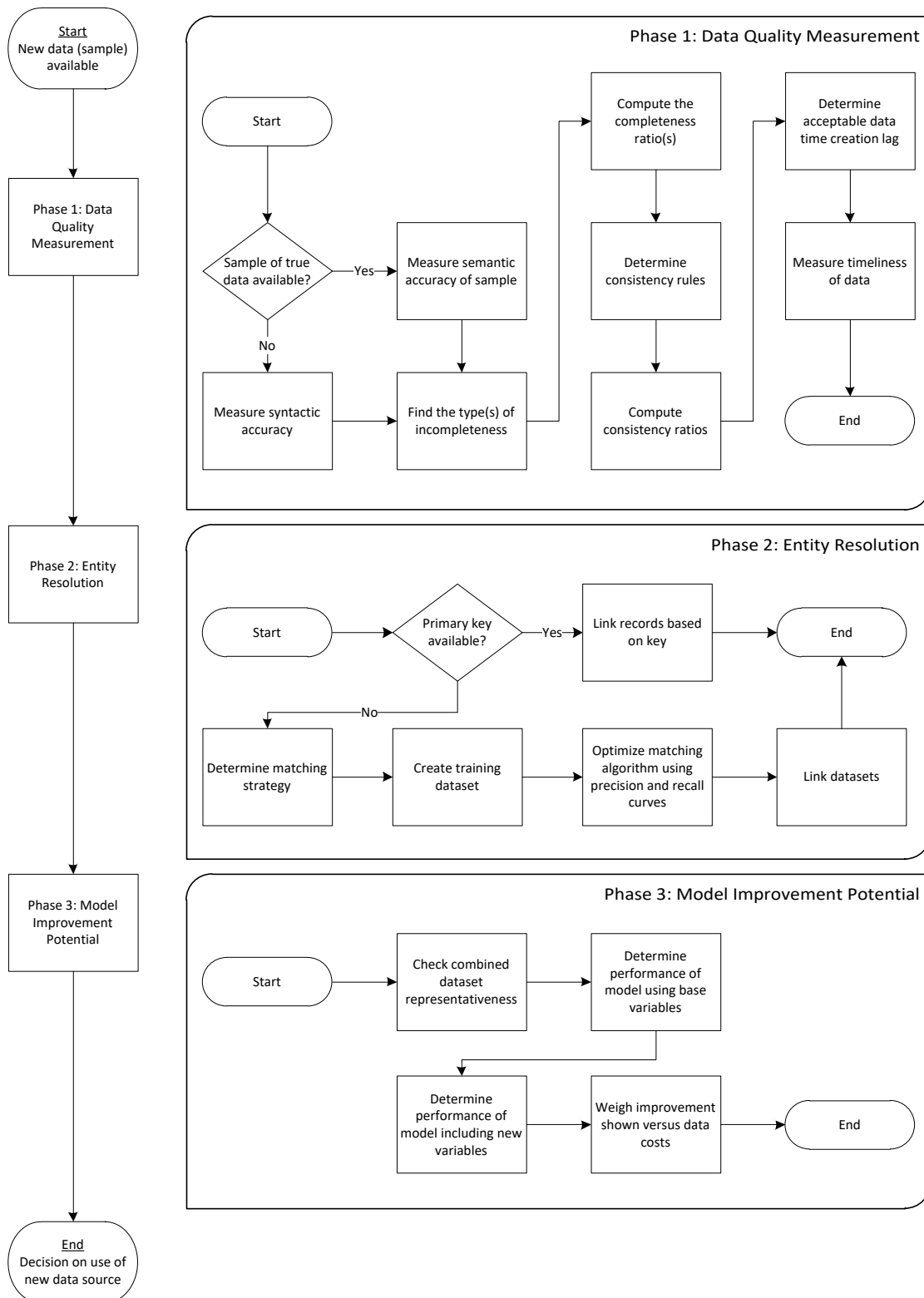


Figure 4.1: Veracity assessment framework activity workflow

data available. Essentially data sources can be viewed as existing on a scale from structured to unstructured and no data is truly without structure. Making the step from unstructured to structured data can be time-consuming and requires solutions that are highly dependent on the type of structure available.

4.6 Summary of Creating the Framework

In this chapter the goals are defined, the context is analyzed, and a framework has been proposed. The goal of assessing veracity has been further specified into three success measures that will be evaluated after its application. These are the following.

1. Metrics can be understood by decision makers
2. Data analysts should be able to apply to framework
3. Differences in scores should indicate differences in veracity of the data sources

In terms of context the following can be said. The organization is a financial institution and is considered to have a flat structure and an open culture. Its strategy with respect to data analytics in insurance is to increase differentiation and thereby being able to offer competitive prices to low risk customers. The people that are important for this research are the data analysts and the decision-makers. The former are required to apply the framework and the latter need to make decisions based on its results. In terms of technology this framework is made to replace an ad hoc process.

The knowledge base for framework creation consisted for a large part of the modeling approach in Chapter 2 and the data quality literature in Chapter 3. Added to this was an entity resolutions step which was not included in the first iteration of the framework. Linking datasets turned out to not be a trivial task in some cases and difficulties can influence the veracity of the data.

The activity selection and their flow make the framework in its current iteration. The framework as such can be found in Figure 4.1. A first evaluation of the framework highlights a few weaknesses as well as considerations that need to be taken into account, but were purposefully excluded. Weaknesses are that its sensitive to sample size, while new data often comes in small samples for analysis as well as that it requires data to be put into a structured form before applying the framework. Excluded considerations are a cost/benefit analysis and ethical considerations.

Although not directly answering any sub-questions, the development of framework is essential to answering the main question. This framework is one of the most important practical and academic contributions of this research. In the next chapter it will be applied on two cases as an example of how to use it and to evaluate its performance.

CHAPTER 5

Analysis & Results

This chapter contains the application of the framework to the two cases, an analysis of its performance, and an analysis of the consequences of expanding these sort of data analytics capabilities in the insurance industry. The analysis starts by describing the data obtained from the company. Followed by the external data sources and the application of the framework to the resulting datasets. In collaboration with actuaries the results are evaluated, the framework is reflected upon and improvements are recommended. Finally, the consequences and risks of continuing to expand data analytics capabilities in the insurance industry are discussed.

This chapter therefore answers sub-questions 3, 4 and 6. Regarding sub-question 3 it becomes clear through the cases what the added value of tapping into the large new pool of Big Data can be. The cases show that several variables are worth considering for improving the risk assessment. The answer to sub-question 4 completed here after evaluation the metrics using the case results. For sub-question 6 the final analysis of the effects of expanding data analytics capabilities for the company and in the insurance industry can be. Next to identifying these risks solutions are proposed. For the company this is done through examining the data collection process, finding risks, and recommend improvements. For the insurance industry long-term effect are evaluated and strategic repositioning options are given.

5.1 Internal Data Overview and Baseline Models

The data set obtained from the company contains types of data that can be found in Table `reftab:companydataset`. The requirements for this research on the data are slightly different then the purpose of the company. Therefore the changes made to the data are described first and then the descriptive statistics of this data are given.

5.1.1 Data cleaning

For the company's analysis a division was made between different parts of an insurance policy. Contents insurance is split into a general part, a electronics part, and a jewelry part. Unless a customer specifically indicates having an unusually large amount of either of the latter two

they are split according to a default key that changes each year. Since for this research we are examining the households, these different parts need to be collapsed.

It started with picking the primary key of the data set and ensuring that this is available for all observations. Keeping in mind that data needs to be linked on a household basis the key was chosen as combination of postal code, house number, beginning of policy, ending of policy, and claim year. The reasons for choosing these are as follows. Postcode and house number uniquely identify any house in the Netherlands, allowing geographically segmented data sets to be linked. Beginning of policy, ending of policy, and claim year uniquely define a specific policy for a specific year. When collapsing the different parts of the policy these will ensure that all parts are actually of the same policy. This means that one customer will in the end have at least one observation per year, since all policy are mapped to a claim year. For the glm analysis a 6 year policy with 1 claim is equivalent to 6 different policies of 1 year where only 1 has that claim in terms of estimating the magnitude of the effect of the predictors so collapsing further is not necessary. Keeping the years separated also allows analysis of trends and dropping of older data when newer data is added.

Another challenge was the recent database migration. This means that some variables have different category codings that need to be harmonized, the same data is contained in different columns, and some data has been lost entirely. The next data issue before collapsing was that some observations had a claim severity of 0 while having a claim frequency up to 6. This affected 20121 observations. In these cases a claim was denied. This could for example be because it did not exceed a persons deductible excess (Dutch: Eigen risico). While these kinds of people might have a justifiably higher risk profile due to being more likely to try to claim something in general, it does not have the same meaning as a legitimate claim. Legitimate claims represent the real risk for the insurer and therefore these rejected claims will not be taken into account. Unfortunately, for all policies having more than 1 claim it cannot be retraced whether any of these are rejected claims. However, since a claim frequency of 1 occurred 121606 times versus 1259 times for more than 1 claim, of which only a small part will have been rejected, this should not influence the predictions noticeably.

When collapsing the amount of observation was reduced by about one-third, which makes sense as each policy is split into three parts (general, electronics, and jewelry as mentioned before). There were also a few remaining orphan electronics or jewelry policies without a general policy. These should not exist since it is impossible to buy only part of policy, but they might contain relevant information about customers so they will be kept in the data set. It should be noted that insured amount for these policies will be low which may cause an unexpected spike in result graphs in this chapter.

After collapsing the policies a check was done on the data linked to each observation. The risk when collapsing is that a general policy has a different house type category assigned than the electronics policy. While this should not occur, it did happen for a few policies. The magnitude of this effect can be found in Table 5.1 and in case of conflicts the category of the general policy part was chosen.

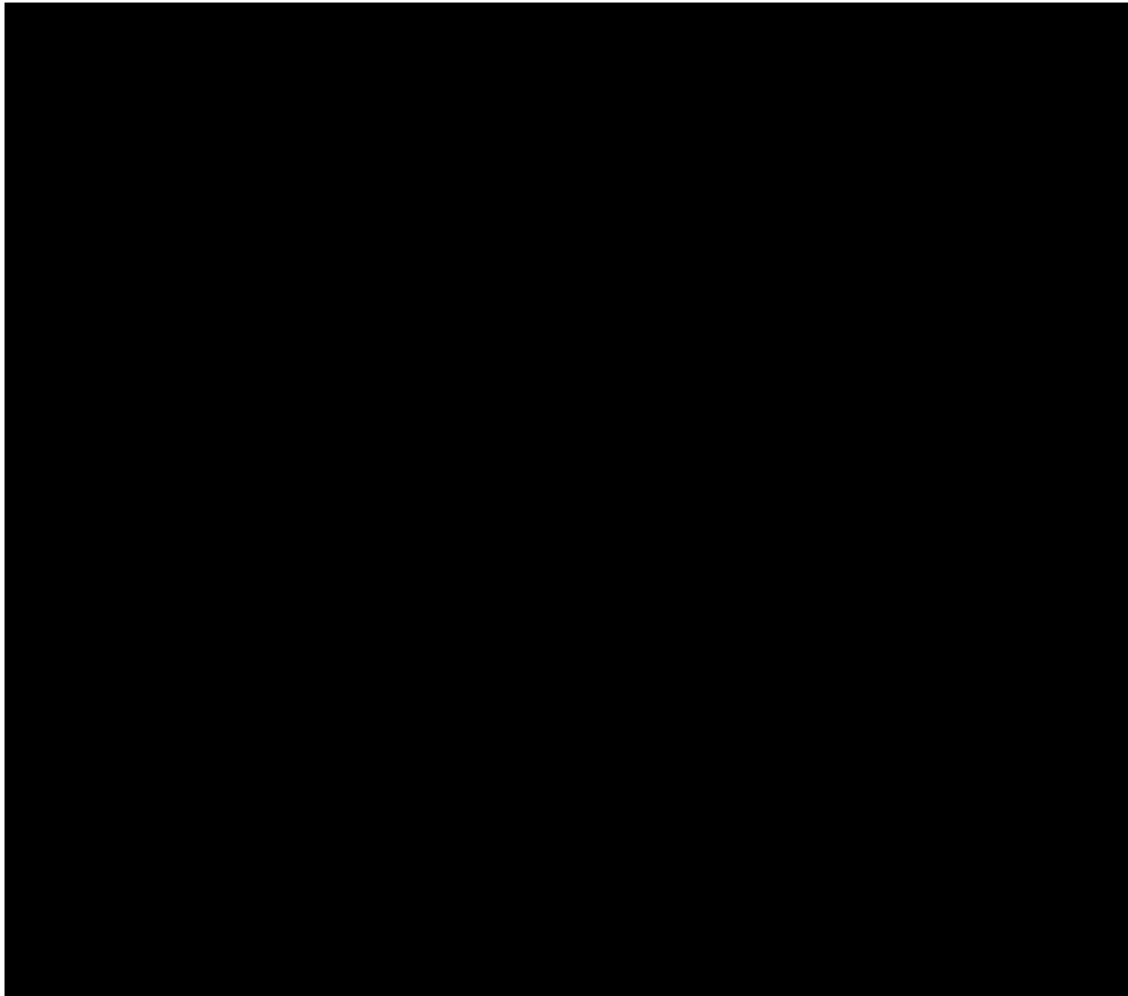


Table 5.1: Original data set cleaning

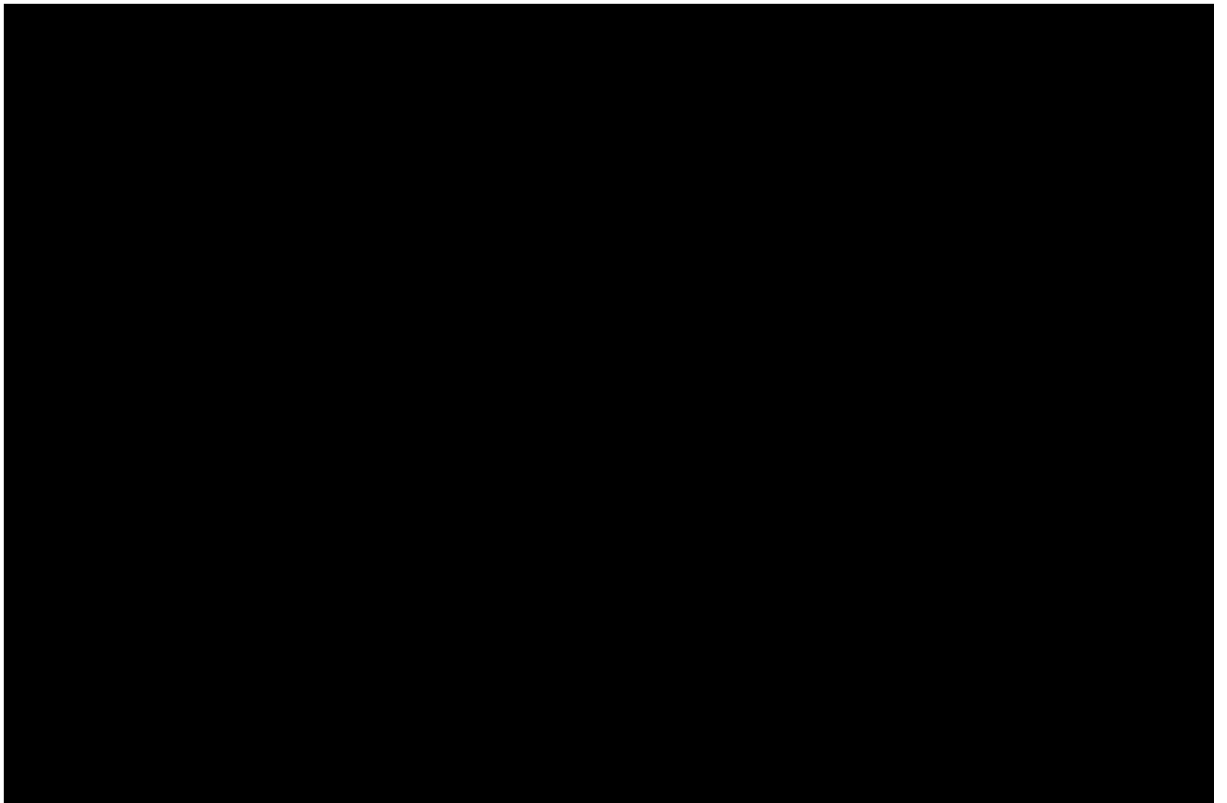


Table 5.2: Original data set overview

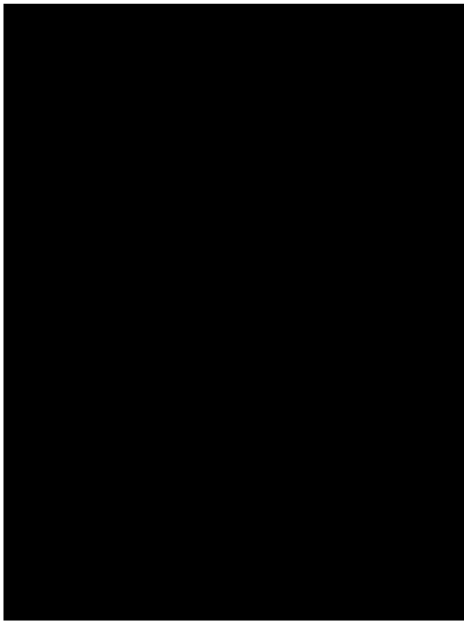


Figure 5.1: Histogram of claims frequency

Table 5.3: Claim frequency table

5.1.2 Internal Data Inspection

In this section the data is examined on an individual basis. Before creating models it is important to take a closer look at the data in order to find potential sources of errors or to gain insights that cannot be derived from the full model. In order to get a first indication of the distribution of categories the total exposure of each category is shown in Figure 5.3. It is immediately clear that several of the categories contain large amounts of missing variables. The origin of these missing variables is data migration issues in the last years. That is the cause of the missing values from 2014.

After working with the company to get the missing data out of the old database the following frequency plots are obtained, see Figure 5.4, which give a much better starting point for analysis. Now the missing values can be considered at an acceptable level and the baseline models can be created. The only category of concern is family composition, but in this case the data is really not available in the company. If the results of the modeling effort show values for this missing category that are far off the known categories or otherwise exceptional, further investigation is required.

The next step is to examine the distributions of the response variables and see whether they match the Poisson and gamma distributions chosen in the research methodology section. The frequency of claims weighted by exposure can be found in Table 5.3 and Figure 5.1. The probability distribution of claim severity can be found in Figure 5.2. It can be seen that both resemble possible probability distributions of the Poisson and gamma family. The severity distribution does show two peaks at 500 and 650. Splitting the data by year shows that a peak at 500 is found in year 2009 to 2013, while 2014 show a strong peak around 650. The cause may be that these figures represent standard claim settlements which are used to facilitate efficient straight through processing.

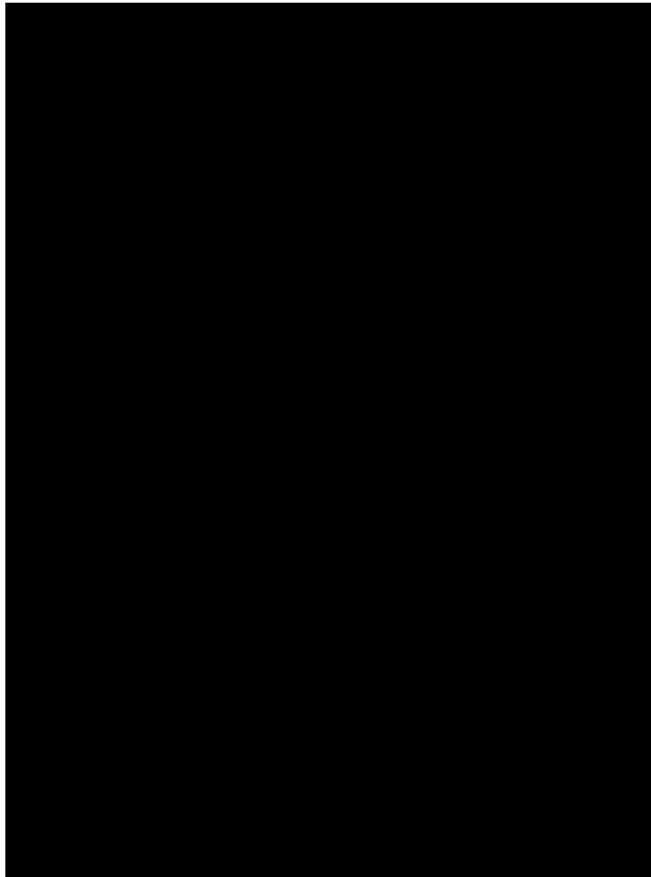
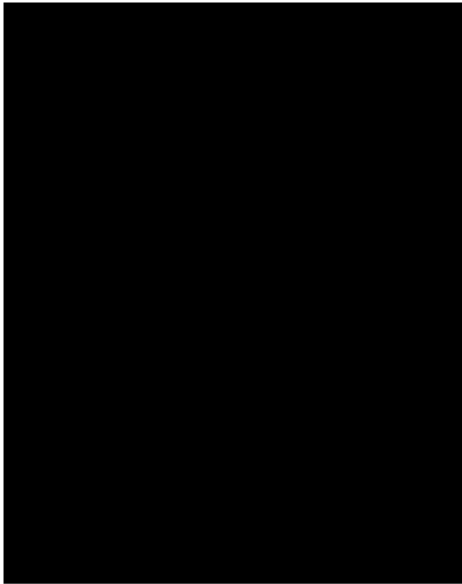
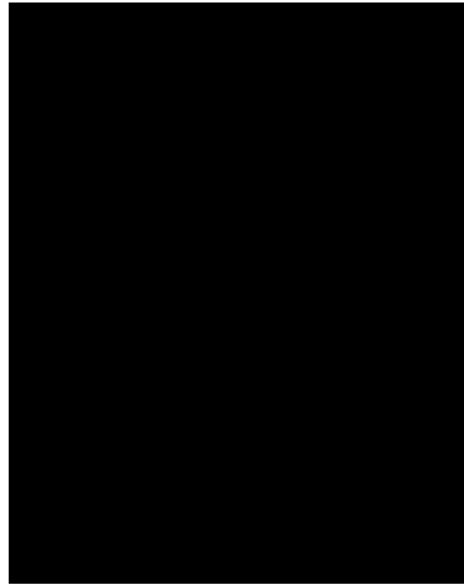


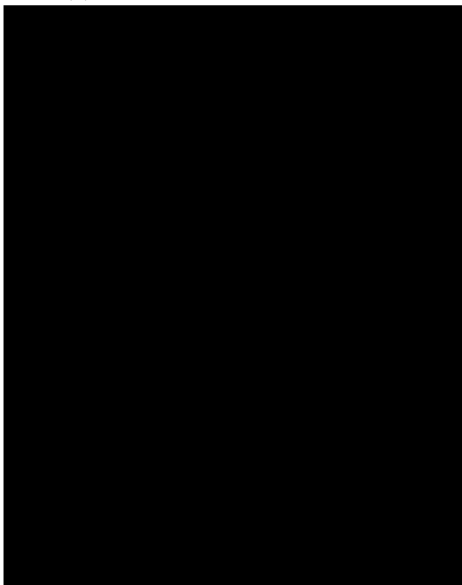
Figure 5.2: Claim severity probability distribution



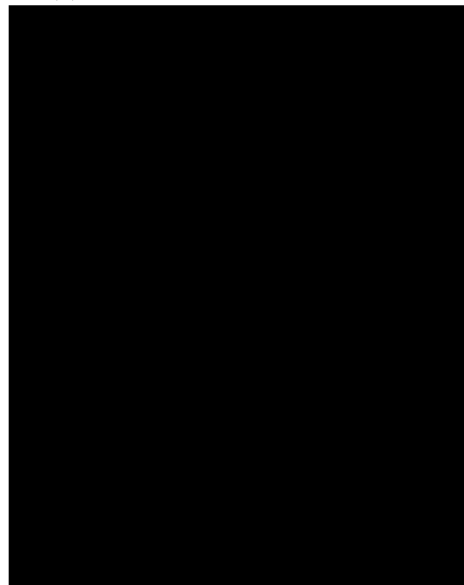
(a) Insured amount frequency plot



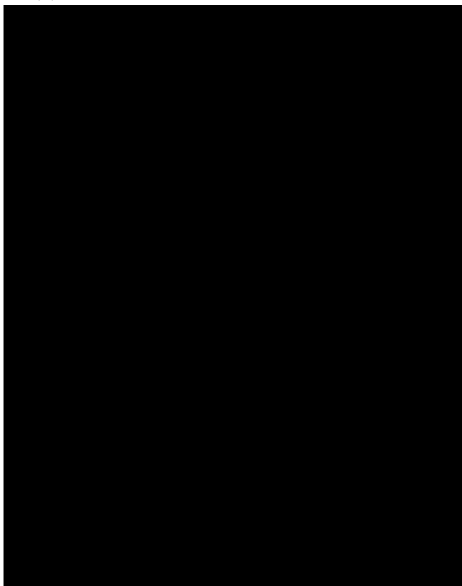
(b) Family composition frequency plot



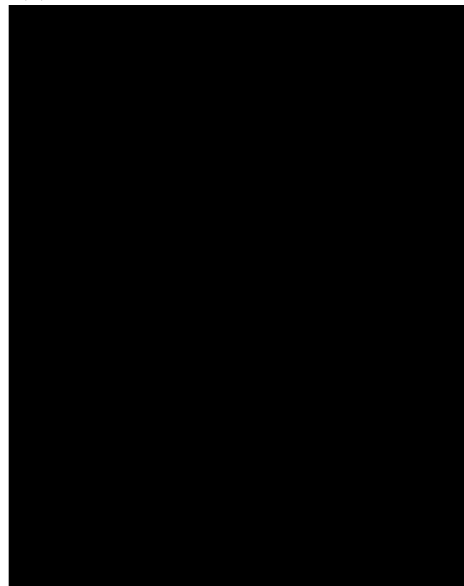
(c) Risk zone composition frequency plot



(d) Credit rating composition frequency plot



(e) Income composition frequency plot

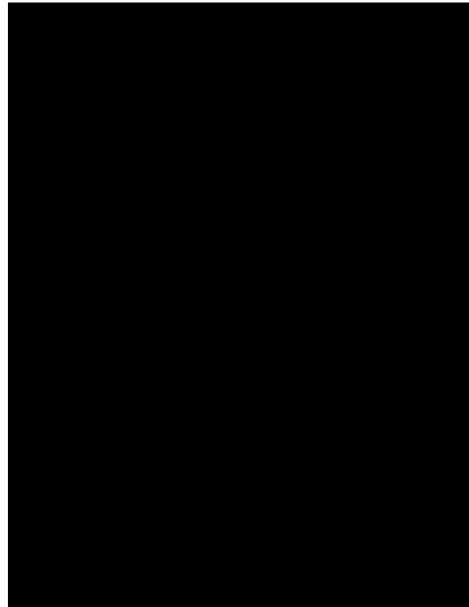


(f) House type composition frequency plot

Figure 5.3: Base variable frequencies weighted for exposure after the data cleaning procedures described



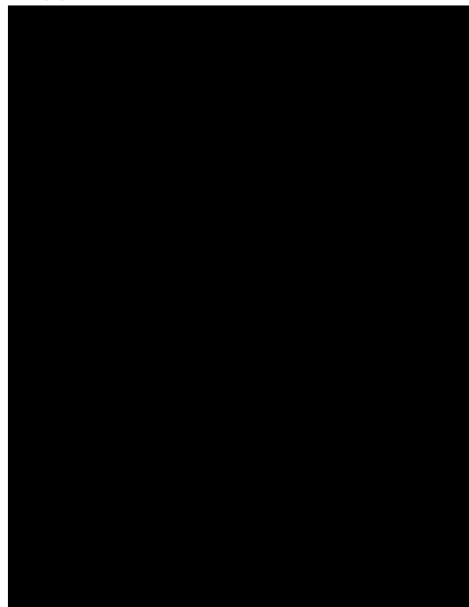
(a) Insured amount frequency plot 2



(b) Family composition frequency plot 2



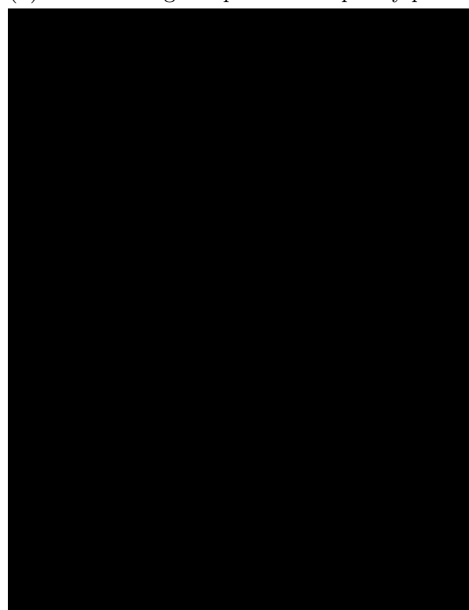
(c) Risk zone composition frequency plot 2



(d) Credit rating composition frequency plot 2



(e) Income composition frequency plot 2



(f) House type composition frequency plot 2

Figure 5.4: Base variable frequencies weighted for exposure after recovering additional in the company

5.1.3 Baseline Models

Here the results of building a generalized linear model in the form described in the research methodology section are given. To recap, for both the severity and frequency models the link function is logarithmic so that the factors are multiplicative. For the frequency model the Poisson distribution is used and for the severity model a gamma distribution is used. The results of the frequency model can be found in Table 5.4, the results of the severity model can be found in Table 5.5. An overview of the coefficients of all the different variables can be found in Figure 5.5

A large black rectangular redaction box covering the content of Table 5.4.

Table 5.4: Frequency model overview using the base data

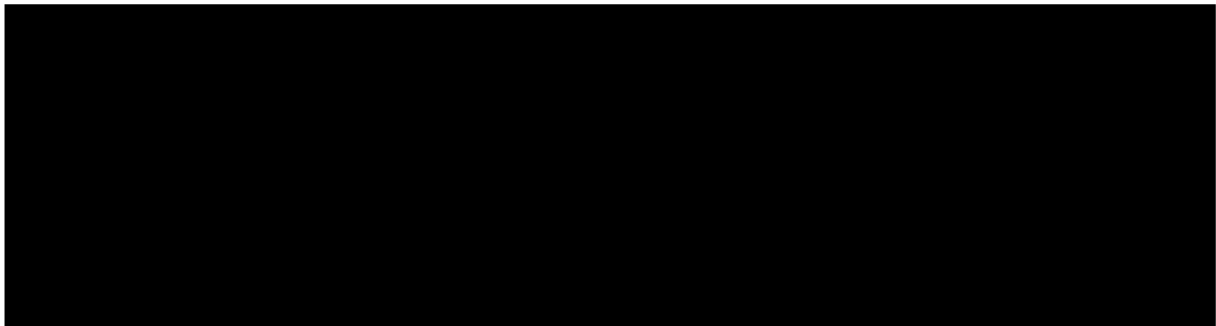
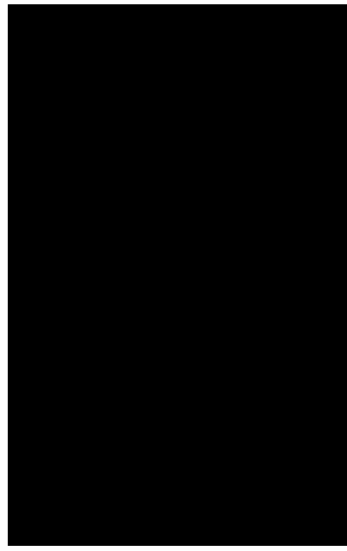
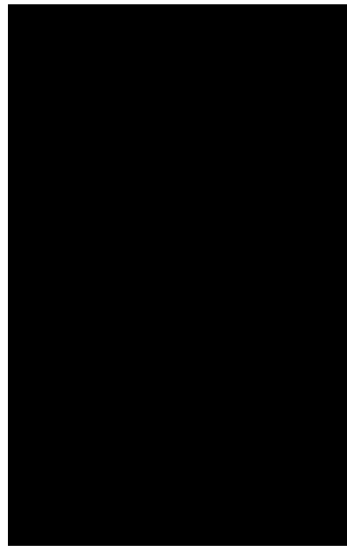
A large black rectangular redaction box covering the content of Table 5.5.

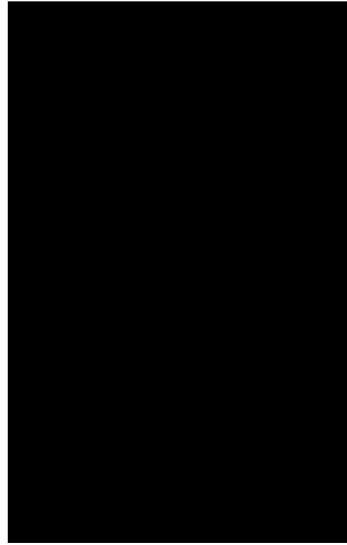
Table 5.5: Severity model overview using the base data



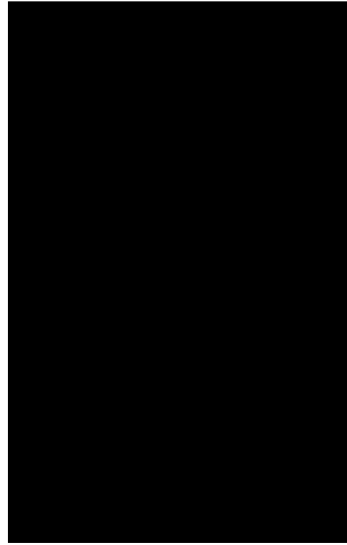
(a) Insured amount coefficients



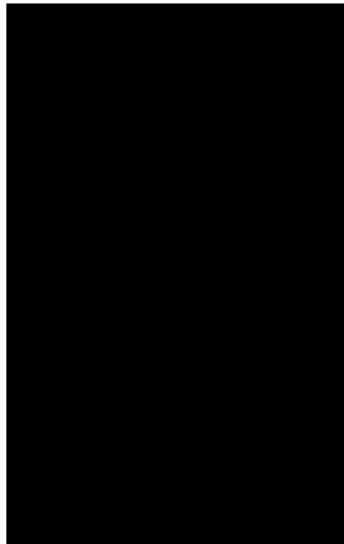
(b) Family composition coefficients



(c) Risk zone coefficients



(d) Credit rating coefficients



(e) House type coefficients

Figure 5.5: Base severity model variable coefficients

5.2 Case 1 – Data Broker

The broker’s data is fully assessed by the framework and the results are shown in this section. Table 5.6 contains short descriptions of the data, a full overview with all coding can be found in Appendix A.

Variable	Coding	Description
<i>Key variables</i>		
Household Key	alphanumeric	Postcode + House number combination, identifies one specific house
<i>Independent variables</i>		
Household income	categorical ordered	Defines household income based on financial and consumer behavior analysis
Life phase	categorical unordered	Describes different age ranges for families with or without children with 8 possibilities
Education 3	categorical ordered	Describes education using 3 classes
Education 4	categorical ordered	Describes education using 4 classes
Social class	categorical ordered	Describes social position of family out of 5 possibilities
Purchase price	categorical ordered	Contains purchases prices of houses divided in 8 ranges
Surface area	categorical ordered	Contains the house surface area divided in 10 ranges

Table 5.6: Data broker’s data set overview

5.2.1 Data Quality Measurement

The metrics to evaluate are accuracy, completeness, consistency, and timeliness. In this section the results of following the step in the flowchart in Figure 4.1 are followed.

Accuracy The first steps relate to accuracy. A set of true data to compare it to is not readily available, so semantic accuracy cannot be measured. Syntactic accuracy can be measured in case of the broker’s data. However, syntactic accuracy is not expected to be meaningful, as the the broker provides a complete dataset where all fields contain a value inside the domain that defines accurate. The data in the set is collected or inferred and only allows values inside this domain. A quick overview of the factor levels available shows that all the data do indeed fit into the domain of coded values. The overview of the data can be found in the bottom part of Figures B.1 and the variable coding domain can be found in Table A.1 in Appendix A.

Completeness This metric examines the missing values of the resulting dataset. Table 5.7 shows how many missing or NA values were contained in the data. Missing means that the data is indicated as missing by the broker. NA means that there was no value for this observation available in the broker’s dataset.

Except for the house price the completeness seems to be at an acceptable level for all variables. A decision to remove house price from the list of variables considered for inclusion would be

understandable. An argument for keeping it in could be that there still is information on about 7 million observations, the model selection approach can indicate whether including it improves prediction.

Variable	Missing	%	NA	%
Life phase	55977	2.5	0	0
Social class	39717	1.8	30428	1.4
Income	46142	2.1	0	0
House price	1561585	69.4	0	0
Surface area	150197	6.7	0	0

Table 5.7: Completeness analysis of the data broker’s extended data set. Total amount of observations is 2249984

Consistency Inconsistencies are data that contradict each other or that does not fit in with rules describing what the data should be.

[REDACTED] Appendix A [REDACTED] Table 5.8 shows how these variables line up.

Having identified the inconsistencies a judgment call is required. Is this amount of inconsistency enough to discard or invest in improving the accuracy of the variables involved? If one is discarded, which one? Attempting to find the cause of these inconsistency can guide the decision. In this the case potential causes are the company data being dated – recall that there is a time lag of up to 6 years between the company data and the data broker’s data, furthermore the data broker has made its own inferences regarding the existence of children whereas the company gets this data directly from the customer. As the existence of children is only part of the information contained in the the life phase variable it is kept for further analysis, but keeping this in mind can assist interpreting the candidate models.

Next to hard inconsistencies as examined in Table 5.8, soft inconsistencies can be explored as well. Soft inconsistencies are soft because they depend on expert judgment. The actuary making the analysis can from experience expect certain correlations to exist, between life phase and house type for example, if this correlation then does not exist up to the level of expectation this can be considered a soft inconsistency. Exploring these inconsistencies assist in getting a feel for the data. An issue is that the amount of soft inconsistencies can be very large and the results too subjective to exclude any data.



Table 5.8: Inconsistencies in the dataset extended with the data broker’s data between [REDACTED]

Timeliness In terms of timeliness the data broker's data is from early 2016. The baseline data contains exposures from 2009 to 2014. This means that there is a 2 to 7 year gap between the original data and the new data. So how to figure out whether this is an issue or not? The volatility of the data has to be examined. The volatility is a metric that shows how quickly the data would be unusable. A maximum age can be defined, but this misses the fact that data will get less accurate with time as well. This accuracy however, should be reflected in a semantic accuracy score.

The data in this case is based on generalizations based on someones postcode and house number combination. So the volatility of the data can be based on the rate of change of houses and neighborhoods. Since these usually do not change rapidly, these are processes taking years if not decades, it can be assumed that the descriptions in 2016 still are an acceptable representation of the situation in 2009.

5.2.2 Entity Resolution

Following the flowchart in Figure 4.1 in this case matching is trivial. A key is provided that allows the data to be linked based on postcode and house number combination. The type of record linking done is a left outer join where the base company data is the left half. This means that the resulting dataset has exactly 2.25 million records as well where all records that had matching data in the broker's dataset have been enriched with this data. To find out how many records this concerns we can look at the completeness analysis.

5.2.3 Model Improvement

In Tables 5.12 and 5.10 the resulting models with performance indicators can be found. The model classifications can be found in Table 5.11 and 5.9. These have been determined by using the `glmulti` package's genetic algorithm in R. Figures B.1, B.2, and B.3 in Appendix B show the coefficients for the variable combination with the lowest AIC score. The AIC, BIC, and their corresponding Akaike weights are calculated using formulas 2.1 to 2.4 on page 18.

First a general note on the coefficients of *Missing*. The missing category often has a large standard error. This is to be expected, because it should include a mix of observations. When the standard error is small it indicated there is a similarity between these observations that should be investigated. Also, when the coefficient estimation of missing is much higher than other estimations, using them as is underestimates prices. A large amount of unexplained risk is in missing. This should be compensated for.

Looking at Table 5.10 and Table 5.9 we can conclude that the broker's data can improve the actuarial models considerably. Comparing the AIC and BIC scores to the baseline scores in Table 5.5 the baseline in both cases is worse than the top 5 models found for the extended dataset. The baseline residual deviance of 64837 gets at most reduced to 64667, a minor improvement. From a NULL deviance of 66009 this means that the explained variance improves from 1.78% to 2.03%. Apparently the value of one's broken possessions does not correlate very strongly with the chosen variables. It could be that the value of the house contents in The Netherlands does not vary greatly in general. Most households have multiple laptops, phones, and perhaps tablets that are among the most likely possessions to break.

Examining the variables, the first thing that stands out is that all base variables are included in all the models. This means that in terms of predictive power none of the new variables are more valuable than the variables already included in the dataset. Of the new variables *Life phase* seems a universally good predictor as it is included in all the models.

Based on the AIC specifically *House price* is a worthwhile additions followed by *House surface area*, *Social class*, *Income*, and *Education 3*. Since models 2-5 are far behind in terms of score to model 1, educations seems to be much worse at predicting claim severity than the other variables. The Akaike weights show from this selection of the models the probability that this model is the best is almost 99.8%. Taking a closer look at the coefficients estimated for the new variables will show us more about the added value of the data, see Figures B.4, B.5, and B.6. This could be as follows.

Income See Figure B.4a and Table B.6. For the severity model SA1, the income coefficients show a general increase of claim severity with increasing income, except for 2x Modaal – twice the mode income. Using this in pricing would go against the principle of having a understandable insurance premium. Combined with the other coefficient estimations being (nearly) insignificant, inclusion of this variable in the model is not recommended.

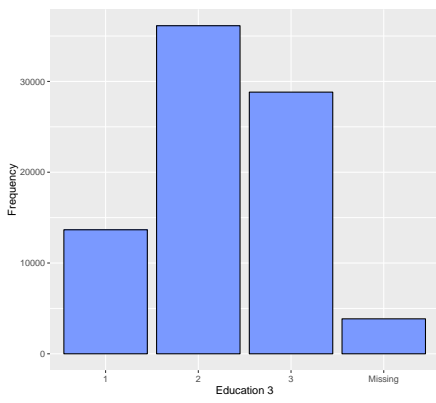
House surface area See Figure B.4b and Table B.6. The house surface area is also problematic. Except for the dummy level which has a low frequency in the data, all coefficient estimations are well within each others standard errors. This means that no reasonable differentiation is possible based on this variable and it should thus be discarded.

House price See Figure B.4c and Table B.6. House price while having relatively little observations, shows a trend that could be useful. Note that category 0 also means the value is unknown, but for a different reason than *Missing*. Missing refers to the data broker not having an entry for this house, while 0 means the house did exist in the broker’s dataset, but the house price is unknown to the broker. If increasing the completeness for this data is possible it may be worthwhile.

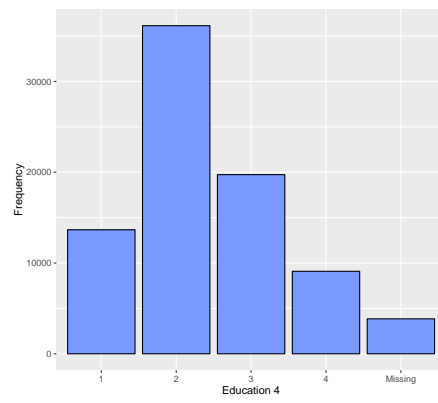
Social class See Figure B.4d and Table B.6. Social class is another promising new variable. While coefficient estimation are modest, the low standard errors indicate that there seems to be a difference between the classes. The issue with missing data occurs here, although it concerns a relatively small amount of data.

Life phase See Figure B.5a and Table B.6. Life phase also shows promise as an addition. Small standard errors with modest coefficients adds some explanatory power to the model. The issue with missing also occurs here. Since the observations indicating missing are exactly the same observation that missed social class data, further research into these observation is required to figure out why these customers had such high claim severity.

Based on the BIC this list is led by *Education 3* and *Education 4*. As it turns out the latter is an exact copy of the former with one category split in two, see Figure 5.6. After education *Social class* and *Income* are also included in top 5 models.



(a) Education 3 frequency barplot



(b) Education 4 frequency barplot

Figure 5.6: Education 3 vs 4 category comparison

5.2.4 Conclusions on the broker's data

For the conclusion each part of the evaluation will be ranked on a scale of 1 to 5. The meaning of scores 1 to 5 are as follows.

1. Unusable
2. Major changes required to be usable
3. Moderately usable, improvements before use recommended
4. Usable with minor issues
5. Usable with no known issues

Going through the categories one by one leads to the overall assessment in Figure 5.7. The justification for the scores is as follows.

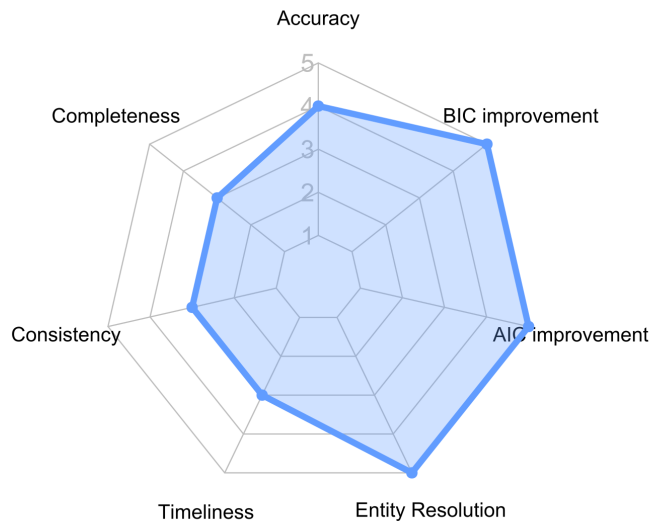


Figure 5.7: Assessment of broker's data

On Data Quality the dataset scores is as follows. For *accuracy* it scores a 4 – while having a 100% syntactic accuracy the fact that semantic accuracy is unknown prevents this from being a five. For *completeness* it scores a 3 – the amount of missing house prices prevents this from being higher, especially since house price turned out to be one of more useful variables. For *consistency* it scores a 3 – the inconsistencies found between family composition and life phase are not necessarily problematic, but the fact that life phase is variable resulting from modeling efforts of the data broker should be kept into mind when making conclusions about this variable. Ideally an attempt would be made to check whether the inconsistency result from time lag in data collection or that the model behind this variable is inaccurate. For *timeliness* it scores a 3 – the time gap between the two datasets is considerable, so investigating volatility of the data is recommended.

On Entity Resolution the dataset scores a 5. The low number of missing values indicated that almost all of the customers are represented in the new data and linking the datasets was

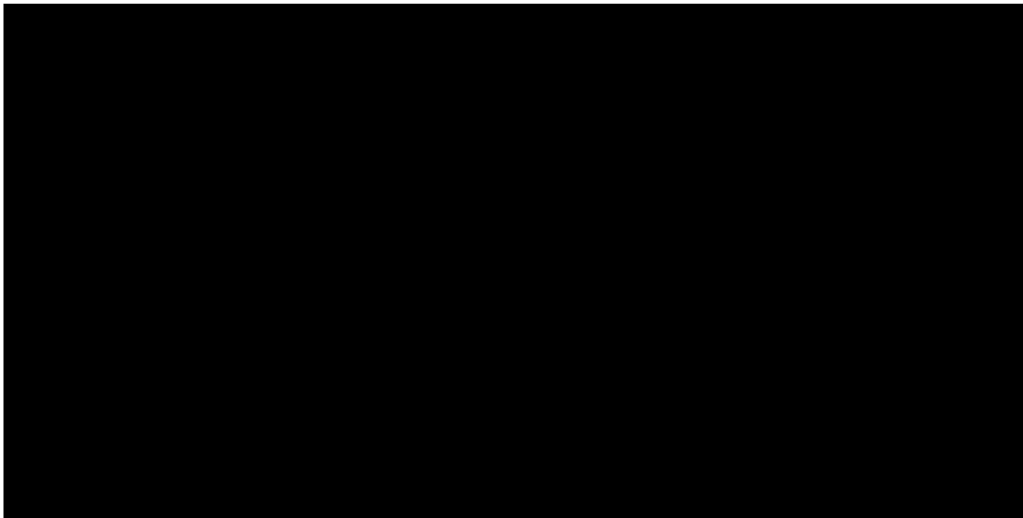


Table 5.9: The top severity models for the data broker's data from the glmulti analysis

done based on the unique postal code/house number combination. Being able to use the full key guarantees a high precision and high recall.

On Model Improvement the dataset scores a 5 for AIC and a 5 for BIC. Both information criteria undoubtedly judge the new variable as improving the model. Upon further examination of the coefficients some variables showed doubtful patterns, but life phase, social class, house price, and education have shown to be useful differentiators.

Model	# var.	$\log(L)$	R.Dev.	df	AIC	Δ AIC	$w(\text{AIC})$	BIC	Δ BIC	$w(\text{BIC})$
SA1	10	-731952	64667	71	1464047	0	0.99772	1464719	115	0.00000
SA2	9	-731964	64680	66	1464060	13	0.00150	1464685	81	0.00000
SA3	10	-731962	64678	69	1464062	15	0.00055	1464715	111	0.00000
SA4	10	-731968	64684	64	1464064	17	0.00020	1464670	66	0.00000
SA5	9	-731972	64689	62	1464068	21	0.00003	1464655	51	0.00000
SB1	7	-732027	64749	48	1464150	103	0.00000	1464604	0	0.81151
SB2	8	-732023	64745	49	1464145	98	0.00000	1464609	5	0.06661
SB3	7	-732023	64745	49	1464145	98	0.00000	1464609	5	0.06661
SB4	8	-732006	64727	52	1464117	70	0.00000	1464610	6	0.04040
SB5	7	-732013	64734	51	1464129	82	0.00000	1464612	8	0.01486

Table 5.10: Performance of the competing severity models. # var refers to the number of variables in the model, $\log(L)$ is the log-likelihood calculated using the R's `logLik()` function, R.Dev. is the residual deviance (null deviance = 66009), df is the degrees of freedom which equals the total number of factor levels included in the model (note: if one level is a linear combination of other included levels it does not count toward df), AIC (see Equation 2.1 on page 18), Δ AIC is the difference between lowest and this AIC, $w(\text{AIC})$ the probability of this model having the best fit according to the AIC, BIC (see Equation 2.2 on page 18, Δ BIC is the difference between lowest and this BIC, $w(\text{BIC})$ is the probability of this model having the best fit according to the BIC

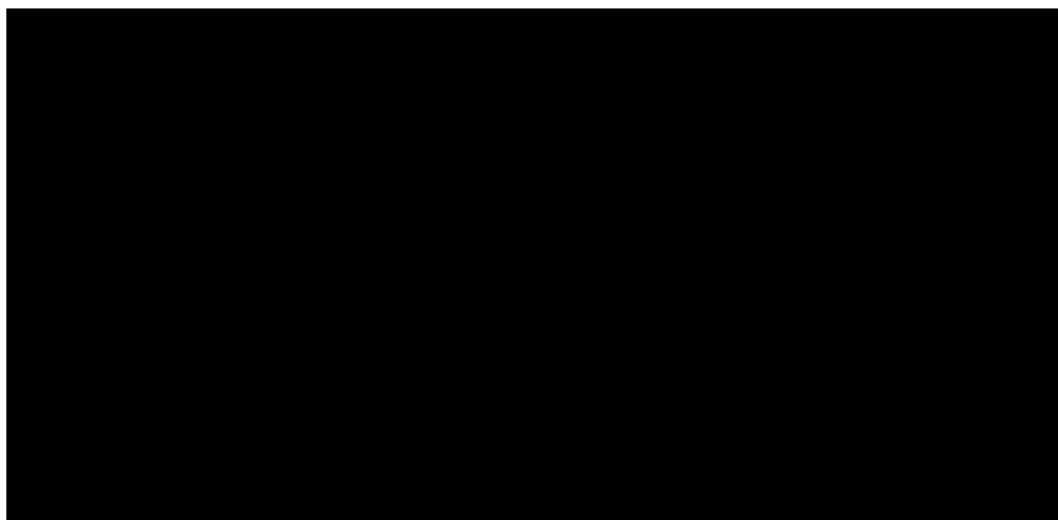


Table 5.11: The top frequency models for the data broker's data from the glmulti analysis

Model	# var.	$\log(L)$	R.Dev.	df	AIC	Δ AIC	$w(\text{AIC})$	BIC	Δ BIC	$w(\text{BIC})$
FA1	10	-357595	544324	70	715331	0	0.9994	716214	71	0.0000
FA2	10	-357605	544344	68	715347	16	0.0003	716205	62	0.0000
FA3	10	-357607	544347	67	715348	17	0.0002	716194	51	0.0000
FA4	9	-357611	544355	65	715351	20	0.0000	716172	29	0.0000
FA5	10	-357608	544349	69	715353	22	0.0000	716225	82	0.0000
FB1	7	-357684	544502	53	715474	143	0.0000	716143	0	0.5114
FB2	8	-357648	544429	58	715411	80	0.0000	716143	0.1	0.4864
FB3	8	-357653	544440	58	715422	91	0.0000	716154	11	0.0021
FB4	8	-357671	544475	56	715453	122	0.0000	716160	17	0.0001
FB5	8	-357647	544427	60	715413	82	0.0000	716171	28	0.0000

Table 5.12: Performance of the competing frequency models. # var refers to the number of variables in the model, $\log(L)$ is the log-likelihood calculated using the R's `logLik()` function, R.Dev. is the residual deviance (null deviance = 567700), df is the degrees of freedom which equals the total number of factor levels included in the model (note: if one level is a linear combination of other included levels it does not count toward df), AIC (see Equation 2.1 on page 18), Δ AIC is the difference between lowest and this AIC, $w(\text{AIC})$ the probability of this model having the best fit according to the AIC, BIC (see Equation 2.2 on page 18, Δ BIC is the difference between lowest and this BIC, $w(\text{BIC})$ is the probability of this model having the best fit according to the BIC

5.3 Case 2 – House Trading Platform

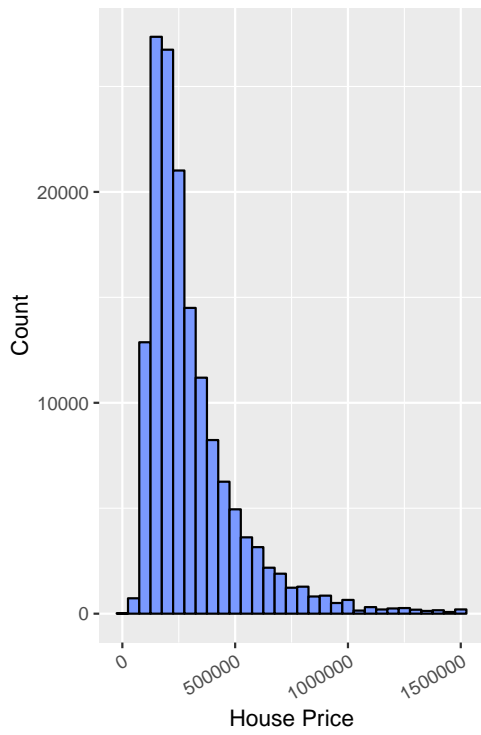
The house trading platform’s data contains a range of house characteristics. The ones in Table 5.13 have been scraped. The reason for excluding data beforehand is because makes computation faster both when extracting and analyzing data.

Variable	Coding	Description
<i>Key variables</i>		
Postcodes	4 digits 2 letters	Defines part of a street, e.g. 1234 AB
House number	alphanumerical	Combined with postcode to identify a specific house
<i>Independent variables</i>		
House price	numerical	The seller’s target price in euros. Will be converted to an ordered categorical variable for analysis
House type	categorical unordered	Apartment, Villa, Corner house, etc.
House surface area	numerical	The total floor surface area in m^2 . Will be converted to an ordered categorical variable for analysis
Number of rooms	categorical ordered	Number of rooms total and number of sleeping rooms in the house
Garden	dichotomous	Whether the house has a garden

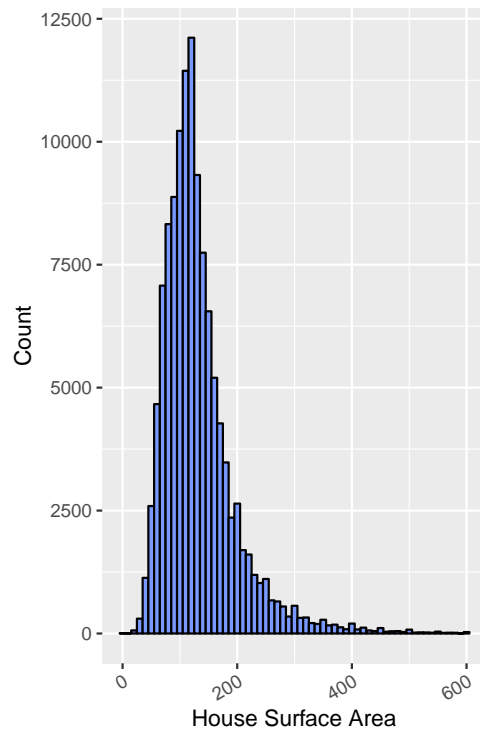
Table 5.13: House trading platform data set overview

5.3.1 Data Quality Measurement

Here the data quality assessment of the house trading platform data is given. An overview of the scraped dataset is given in Figures 5.8 and 5.9. Note that the graphs for surface area and price were cut off in order to have the bulk of the houses clearly represented. There were houses with higher prices and surface areas. An overview of the dataset statistics after linking the company data according to the procedure described in the next subsection can be found in Figure 5.10. In the linked dataset some factor levels are dropped due to either the company data not including any customers that had a house with these characteristics or multiple houses being available in the house trading platform dataset with the same postal code, which resulted in the ones containing that variable level being removed.



(a) House price



(b) House surface area

Figure 5.8: House trading platform scraped data counts

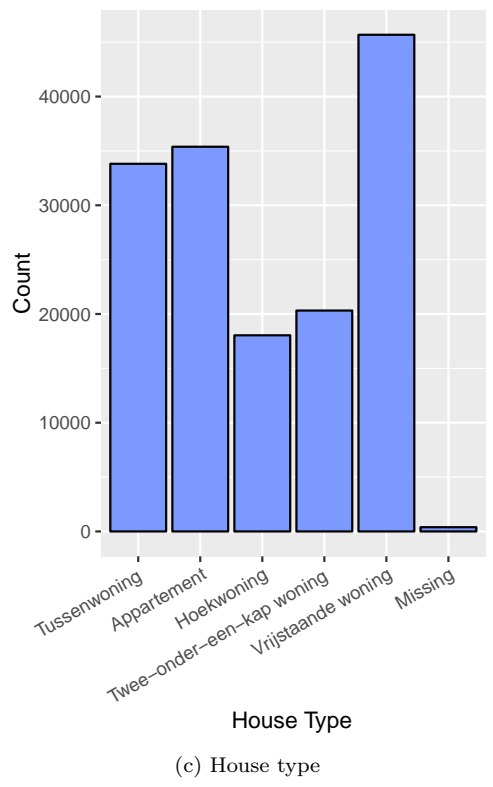
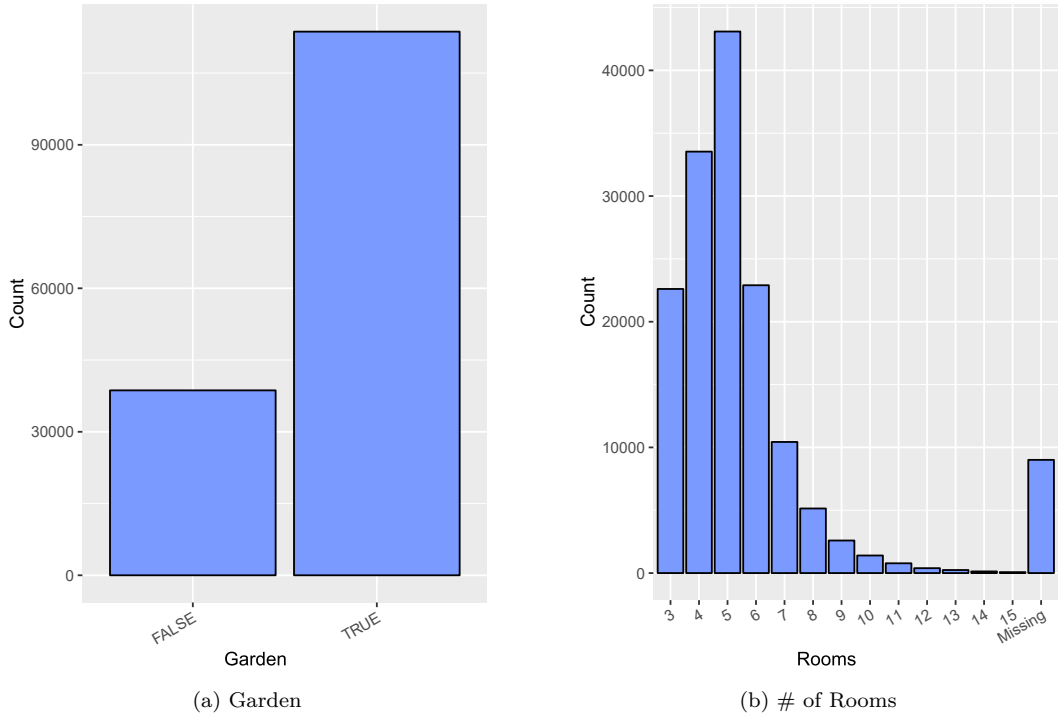


Figure 5.9: House trading platform scraped data counts

Accuracy Only certified house brokers have permission to put up offers on the platform. It is expected that they are well aware of the correct data and therefore the data should be accurate. Any deviations from real world values can be explained by their incentive to make the house attractive, but these deviations are expected to be minor if any. Of the variables chosen, only two have potential issues. The house surface area might be rounded up and the house price is the listed house price so it may differ from the actual market value.

Completeness As the graphs already show, the completeness of the dataset is very high. The room relatively large amount of missings in the Room variable is attributed to errors in data collection combined with the cut off point at 14 rooms during data cleaning forcing houses with more rooms into the missing category.

Variable	Missing	%	NA	%
House price	0	0	0	0
House type	340	0.2	0	0
House surface area	0	0	0	0
Number of rooms	9008	5.9	0	0
Garden	0	0	0	0

Table 5.14: Completeness analysis of the trading platform extended data set. Total amount of observations is 152321.

Consistency In the scraped dataset no potential for inconsistencies is identified. In the linked dataset however, a comparison can be made between the company house types and the trading platform house types. The linked set used in this case is the dataset matched with high precision. While this does not help interpret model result, it does show whether inconsistencies the base data is inconsistent with the company data, see Table 5.15.

	A	R	C	S	D	M	I	I%
Apartment (c)	<i>17054</i>	614	230	83	482	116	1525	8.2
Row house (c)	1037	<i>9078</i>	623	344	1268	18	3290	26.6
Corner house (c)	87	891	<i>4103</i>	197	277	3	1455	26.1
Semi-detached house (c)	86	389	1115	<i>5674</i>	665	6	2261	28.5
Detached house (c)	67	189	234	169	<i>12103</i>	13	672	5.3
Missing (c)	47	19	15	2	134	0	217	100
						Total	9420	16.4

Table 5.15: Inconsistencies between house platform and company data in house types with a dataset matched on postal code and housenumber. Company houses (c) vs Trading platform houses – A = Apartment, R = Row house, C = Corner house, S = Semi-detached house, D = Detached house, M = Missing, I = Total Inconsistent. Total observations is 57432

Timeliness This quality characteristic can distort the analysis. The house data is from up to 1 year ago, while the customer data is from 2009 to 2014. For the analysis the assumption is that people change house infrequently enough for this data to be valid.

5.3.2 Entity Resolution

In terms of linking the data to company dataset a few options were available. A full key for linking is available, but contrary to the data broker's data, this dataset does not include the whole of the Netherlands. The amount of houses traded at one time is limited, so either periodic scraping attempts are required or the company providing platform will need to be contacted to access historical data.

One of the weaknesses of scraping was also exposed, as on secondary scraping attempts the structure of the site was changed so the scripts did not capture all of the same values as the first time. Therefore the dataset analyzed is limited to the first scraping run.

1. Linking based on full postal code + house number combination
2. Linking based on 6 digit postal code
3. Linking based on 4 digit postal code

Option number 1 leaves us with a small dataset of approximately 50.000 observations. This is the option that ensures a high precision since every link can be assumed to be correct. The recall on the other hand will be quite low.

Option number 2 balances the precision and recall based on the assumption that houses with the same 6 digit postal code have similar characteristics. A 6 digit postal code identifies part of a street and while some streets have different configurations of for example apartments and row houses, generally the differences in house types should be similar. Making this assumption will reduce precision, but up recall. A higher amount of data that correctly describes customers will be available at the loss of certainty in the accuracy. This approach results in a dataset of approximately 360.000 observations.

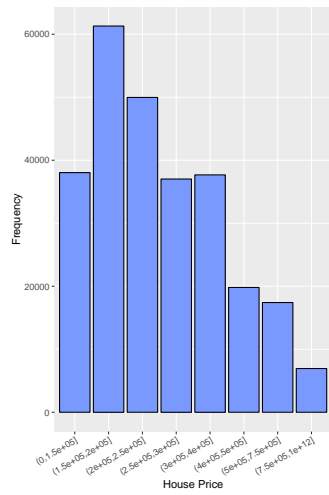
Option number 3 attempts to maximize the amount of information extracted from the new data at an even high loss of accuracy. The 4 digit postal codes identifies a neighborhood. The assumption that houses are similar in the entire neighborhood is extremely lenient. The reduction in precision is considered too high to use this linking strategy.

So in conclusion a balance is chosen in option 2. This creates a dataset of sufficient size to evaluate while maintaining a reasonable level of precision. The exposures of the different variables in this linked dataset can be found in Figure 5.10

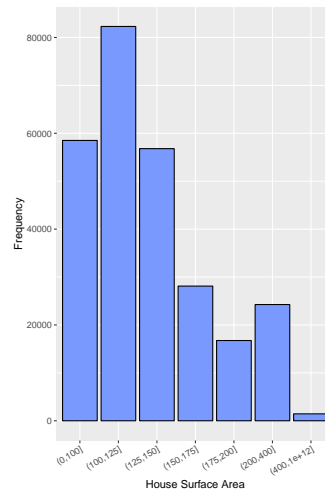
5.3.3 Model Improvement

In order to determine the potential model improvement the improvement in AIC and BIC scores are evaluated again. The best models can be found in Table 5.16, the performance of these models is compared in Table 5.17. The variable exposures and coefficient estimations for the best AIC (H1) and best BIC (H6) models can be found in Appendix B.6 starting on page 119.

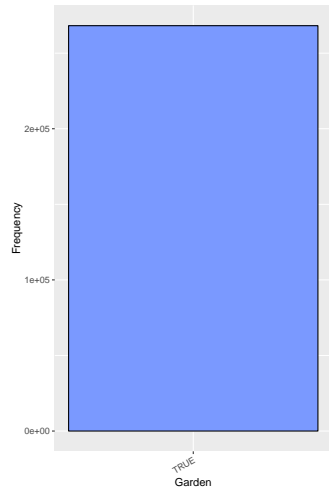
Looking at Table 5.16 the variables that are seemingly valuable are identified. In this case the AIC based analysis show all variables as improving the model, except for the *garden* variable. This is due the linked dataset include solely houses with gardens, so no differentiation is possible. The BIC shows no new variables in the preferred model, but includes the house type of the trading platform in two of the top 5 models, marking this variable as potentially useful.



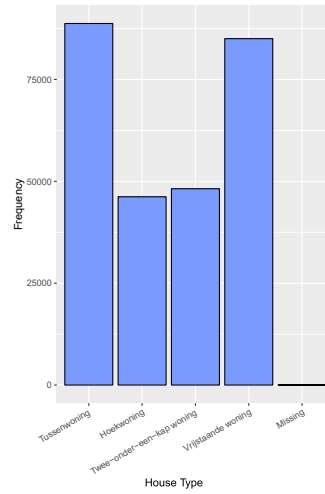
(a) House price



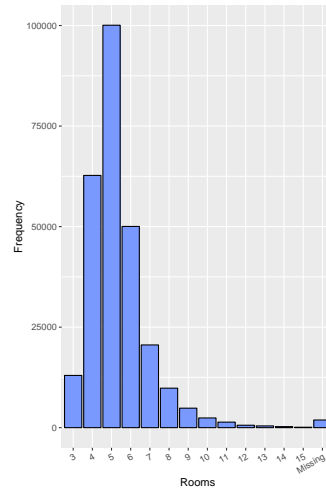
(b) House surface area



(c) Garden



(d) House type



(e) # of Rooms

Figure 5.10: House trading platform linked data exposures

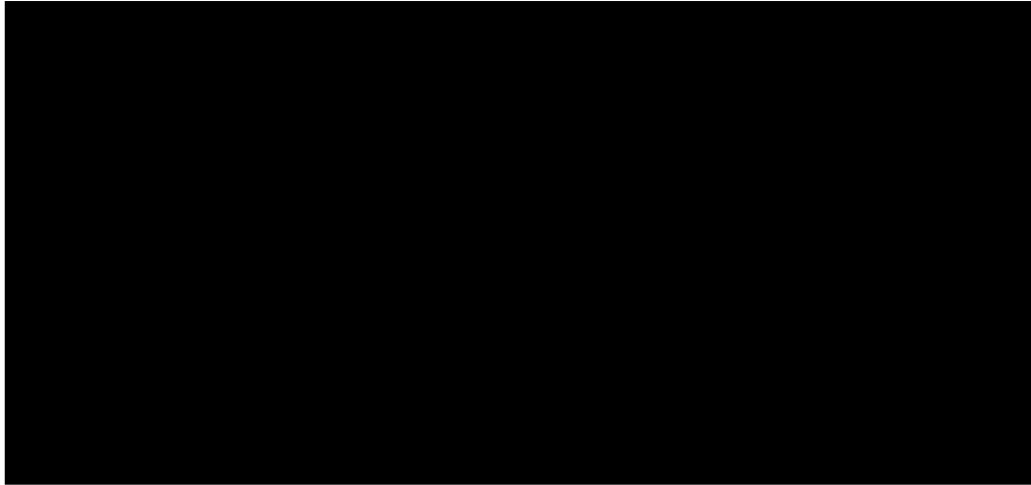


Table 5.16: The top frequency models for the house trading platform data from the glmulti analysis

Model	# var.	$\log(L)$	R.Dev.	df	AIC	Δ AIC	$w(\text{AIC})$	BIC	Δ BIC	$w(\text{BIC})$
H1	8	-64265.41	96845.77	61	128653	0	0.8420	129311	341	0.0000
H2	7	-64271.54	96858.03	57	128657	4	0.1140	129272	302	0.0000
H3	9	-64264.50	96843.96	65	128659	6	0.0419	129360	390	0.0000
H4	7	-64277.51	96869.96	55	128665	12	0.0021	129258	288	0.0000
H5	7	-64277.78	96870.51	57	128670	27	0.0000	129284	314	0.0000
H6	3	-64357.35	97029.65	20	128755	102	0.0000	128970	0	1.0000
H7	3	-64367.60	97050.15	20	128775	122	0.0000	128991	21	0.0000
H8	3	-64376.02	97066.99	20	128792	139	0.0000	129008	38	0.0000
H9	4	-64353.64	97022.23	24	128755	102	0.0000	129014	44	0.0000
H10	4	-64354.56	97024.07	24	128757	104	0.0000	129016	46	0.0000

Table 5.17: Performance of the competing house trading platform frequency models. # var refers to the number of variables in the model, $\log(L)$ is the log-likelihood calculated using the R's `logLik()` function, R.Dev. is the residual deviance (null deviance = 66009), df is the degrees of freedom which equals the total number of factor levels included in the model (note: if one level is a linear combination of other included levels it does not count toward df), AIC (see Equation 2.1 on page 18), Δ AIC is the difference between lowest and this AIC, $w(\text{AIC})$ the probability of this model having the best fit according to the AIC, BIC (see Equation 2.2 on page 18, Δ BIC is the difference between lowest and this BIC, $w(\text{BIC})$ is the probability of this model having the best fit according to the BIC

5.3.4 Conclusions on the House trading platform data

Using the same rating scale as for the evaluation of the data broker's models, which is as follows.

1. Unusable
2. Major changes required to be usable
3. Moderately usable, improvements before use recommended
4. Usable with minor issues
5. Usable with no known issues

Going through the categories one by one leads to the overall assessment in Figure 5.11. is as follows.

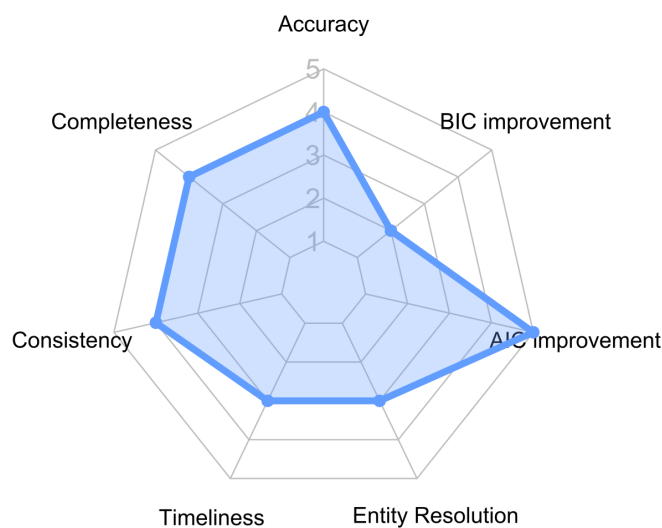


Figure 5.11: Assessment of house trading platform data

The justification for the scores on Data Quality the dataset scores is as follows. For *accuracy* it scores a 4 – for syntactic accuracy a high score is obtained, arguably the missing house types and room amounts count towards lower the syntactic accuracy as these values have been forced into missing for not fitting into any domain category, but they are also counted in completeness. Semantic accuracy was impossible to test, since no sample of ground truth data was available, although the origin of the data being professionals with a goal that supports accuracy and the fact that house buyers can easily check the data inspires confidence. For *completeness* it scores a 4 – the missing room counts make this imperfect, in general no issues are expected here. For *consistency* it scores a 4 – with a 83.6% of the house matching the company data and it may even be the case that the company data is inaccurate. Some further research can be done, but this is unlikely to be an issue. For *timeliness* it scores a 3 – the time gap between the two datasets again is considerable, so investigating volatility of the data is recommended.

On Entity Resolution the dataset scores a 3. Firstly, the dataset so far is too small to be linked to a sizable amount of company customers by using the primary key. In order to enlarge the amount of customer that could be included, assumptions were made that compromised precision.

The result is a dataset that could have issues with accuracy. Gathering more data from the house trading website is recommended so that linking by primary key will provide a sizable dataset where precision and recall are both high.

On Model Improvement the dataset scores a 5 for AIC and a 2 for BIC. The AIC as a criterion that is known to lead to the inclusion of more parameters does suggest models that include all the new data. The BIC on the other hand practically rejects all the new data, suggesting models that even reject some of the base variables included in the base models.

5.4 Evaluation of Case Results

Now that the two cases have been discussed, we can evaluate the performance of the framework. Recall that the requirements defined at the start were the following.

1. Metrics can be understood by decision makers
2. Data analysts should be able to apply to framework
3. Differences in scores should indicate differences in veracity of the data sources

Requirements #1 and #2 are partly insured through the research methodology applied and at the end a presentation is given to large group of analysts and decision-makers to obtain final feedback on the framework. The fulfillment of both these criteria has been ensured in the development process of the framework as during the development multiple moments of interaction were organized where both these stakeholders had the opportunity to give feedback on the current design and thus influence its final form. Especially one of the key data analysts in the company was involved in the design on a daily basis. The final test of these requirement is a presentation given inside the company that includes actors from these groups not involved in the design process the results of which unfortunately could not be included yet. Requirement #3 can be evaluated using the results of the cases.

5.4.1 Decision-maker and data analyst understanding

As per the principles of ADR, see Table 2.2, the framework developed at the company. Its development was supervised by a data analyst and the researcher was located at the departments where stakeholders of this research worked. This allowed the research to be well connected with the daily practice of the company.

Next to this there were the formal feedback moments already mentioned, which allowed both a decision-maker and data analysts that were not involved on a daily basis to provide feedback and steer the design in a direction that would be satisfy these requirements.

Through these processes it seems likely that these two requirement are sufficiently satisfied. The final test is a presentation within the company for a large group of analysts and a few decision-makers. The results of this presentation could unfortunately not be included in the report, but will be included in the defense of the thesis.

5.4.2 Evaluation of metrics

In this section we will go through the metrics one by one to evaluate how they rated the data sources, what that the conclusions therefore were, and make notes on whether the results are meaningful. However, first some general notes are made on the performance of the framework. The results of the two cases can be found in Figure 5.12

General notes Measuring data quality before or after the entity resolution step. Just the model improvement is not enough to judge the inclusion of a variable. Data quality scores are really only meaningful on per variable basis, interpreting these results on an entire dataset clouds underlying issues to decision-makers. The skill and judgment of the analysts is a significant influence on the final scores.

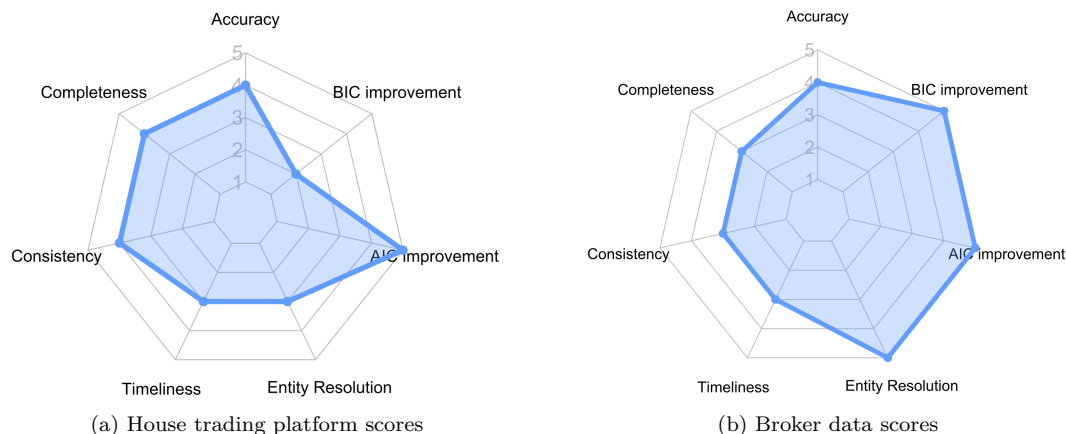


Figure 5.12: Comparison of case results

Accuracy Measuring accuracy has proven to be difficult in both cases. The syntactic accuracy was always high due to the way the datasets were constructed using values from the set of categories. Issues that can lead to low syntactic accuracy such as manual entry errors or errors in data retrieval due to a programming error were not present. Usually these issues will be solved during data cleaning, making syntactic accuracy perhaps not such a valuable metric. Furthermore, semantic accuracy, which would be a far more meaningful form of accuracy to measure, has its own challenges. If obtaining ground truth data on a large scale would be possible, that data should be used instead of data of which the accuracy is measured. So the best solution might be collecting a sample of ground truth data, but this process could be time consuming and the small sample size can make the outcome unreliable.

Completeness Completeness is a straightforward metric that can easily be computed for any dataset. The meaning however is not easily interpreted. As could be seen from the high incompleteness level of the house surface area in the data broker's data, the variable still was deemed a valuable inclusion by the model performance measurement metrics. Even if low completeness scores are obtained, there may be value in the data that is available. So on the positive side this metric indicates which variables to scrutinize if they seem to improve the model and that those variables are not immediately to be discarded if they do not improve the model. On the negative side, scoring this for the entire dataset sums the results for all variable, making it hard for decision-makers to interpret. Scores should really be viewed on a per variable basis.

Consistency Consistency is one of the more subjective measures, as it depends on the creativity of the analyst to define inconsistencies. It is also a variable where the before or after linking the data becomes a highly relevant question. In both cases no rules for inconsistency within the dataset were found, so it was evaluated after linking the data, as the new data and the company data included variable describing the same real world object.

When approaching consistency this way, it becomes one of the more interesting metrics. It can serve as a proxy for accuracy depending on the trust the company has in its own data. Furthermore, it can then serve a proxy for accuracy in the entire new dataset, because it can show the amount care taken in gathering and inferring the inconsistent variable, which can be assumed to be similar for all variables in the new dataset. However, care should be taken to make

sure its really the new data that is inaccurate and not the company data. Since the number of inconsistencies generally is much smaller than the entire dataset, sampling to check which value is semantically accurate should guide the final judgment.

Timeliness Both datasets suffered from large gaps in data collection times for the two dataset. Customer data in the company has been collected for many years, while new dataset described the characteristics of customer as determined in the recent past. The influence of this gap is hard to estimate, but another data quality measurement found in the literature could alleviate this uncertainty. Volatility either as a hard expiration data for certain data or as a measure if a variables increasing inaccuracy over time could be evaluated in conjunction with timeliness in order to create a more complete judgment of time related issues in the dataset. The original reason for not including volatility was that it is cumbersome to define for each variable, but if the new data is valuable enough, the investment can be considered.

Entity Resolution This metric was added to evaluate the quality of the linking process between the two datasets. The measurements are hard to accurately measure though. In the first case it was completely irrelevant due to the availability of a primary key. In the second case it could have been similar, but the small amount of data available in the set made making a more lenient linking assumption an attractive to option to attempt to obtain a larger dataset. A qualitative analysis of the effect on precision and recall guided this decision and the metrics provided a useful perspective for weighing the options.

AIC/BIC Improvement Measuring the potential usefulness of the variable through this metric worked well in the sense that is easy to implement and leads to clear results. However, and this is already hinted at in the first case, this improvement alone is not enough to make a final judgment. The case of *income* in the first case makes this clear. This variable comes out as one the recommended additions, but upon examination of the modeling results, the coefficients and their standard errors, it is shown that this variable follow an understandable pattern. The variable that breaks the pattern may even the only reason this variable was included, because the other estimation do not vary significantly from the dummy level.

As for the actual scores and the reason for using both measures the score difference between case 1 and case 2 clearly demonstrate the need for that. For case 1 is clear that the data provides new insights as demonstrated by full scores on both criteria. For case 2 it is clear that the data provides less new insights than in case 1, as demonstrated by the low BIC score. The difference in variable selection between the two measures, with BIC leading to more parsimonious models, creates more depth in the judgment of the new data than using either single would have done. However, the disadvantage is that it becomes more difficult to say whether the data has a high enough veracity if one score is low and the other high.

Conclusions Overall the conclusion is that while these metric can successfully be used to gain insight into new data, the interpretation needs to be guided by expert judgment. These metrics by itself without knowing the scores for individual variables in the case of data quality or knowing the variable coefficients estimated for the recommended models are not enough to reach a conclusion on the veracity of the data. Although this framework does provide a structured way to evaluate data and with minor extensions can be valuable asset to companies struggling with the need to evaluate new data.

Recommendations The recommendations for improving and extending are as follows. Note that some improvements would also increase the time required for going through the framework so the expected additional insight would need to be weighed on a case by case basis.

1. Report data quality values per variable to the decision-makers
2. Reporting on data quality before and after entity resolution could indicate issues in the entity resolution step and help explain modeling results
3. Analyze the modeling output for the best models and examine the coefficient estimations together with the Chi squared significance of each variable
4. Develop a measure of volatility for a comprehensive time related score

5.5 Consequences of Advancements in Data Analytics

It is clear that big data analysis can have a positive impact on our quality of life. Specifically in insurance benefits include improved fraud detection, ability to offer products closely matching customer needs, and damage prevention through availability of sensors in buildings and cars. Next to this pricing can be done on a more personal level, offering every customer a premium more closely matching the risk profile.

Engaging in higher levels of personal data analysis comes with risks of crossing legal and ethical boundaries. For the ethical evaluation of big data analytics in the insurance industry the focus is on two topics. The first point is ethical risks taken by the company in data collection and how to mitigate these risks by acting more ethically. The second topic is the consequence of using personal data for pricing on an industry scale. The latter is slightly off-topic to the main points in this thesis, but while investigating ethics several ideas came forward that have practical value as well as can lead to directions for future research.

In order to get an informed opinion in these two perspectives the literature from Chapter 3 is combined with interviews with key people in functions related to data analytics. These interviews and the company processes surrounding data collection are described in the next subsection. The second subsection includes the analysis of the company's process versus the ethical boundaries found in the literature. The analysis of the industry consequences is included in the third subsection.

5.5.1 The data collection process

In order to be able to advise the company on policy regarding ethical data collection, first the current process needs to be known. For this purpose interviews were conducted with the people in Table 5.18. The departments collaborate according to the scheme in Figure 5.13. The interviewees were asked to explain their responsibilities in the case of new data being acquired and the ethical issues identified in the literature were discussed to see how these were dealt with in the current process.

It starts with data analysts having decided on a new data source. If the data is personal, the privacy board and data protection officer have to be involved. The privacy board is an internal unit responsible for checking whether the data complies with internal guidelines surrounding the use of personal data. The data protection officer is a position required by European law whose responsibility it is to ensure that the company fulfills its legal obligations concerning the protection of personal data. When the data is acquired from another company, the procurement department can assist in the acquisition process. When personal data is involved, assistance from the procurement department is mandatory. The procurement department works with the screening department to do a background check on ethical behavior of the board of their partner.

The final say and responsibility in this process is with the manager of the data analysts. The decision whether or not to go through with the acquisition is made there, the other departments are supporting that manager. In practice however, when one of the supporting departments indicates that the risks are too high, at the very least the decision is discussed in more detail to see if these risks are manageable or the acquisition is canceled.

Role	Responsibility
Customer intelligence manager	The decision-maker in the process. Determines which data is relevant, where to get it, and has authority to make the acquisition.
Data Protection Officer	Independently ensuring the internal application of the data protection regulation provisions. Keeping an overview of all internal uses of personal data.
Privacy Process Manager	Creating processes that ensure data use in the company complies with privacy regulation
PR Spokesperson	Monitor and manage public opinion. Support decision-makers with this information.
Procurement–Marcom and Privacy	Manage the process of contracting partners for the Marketing & Communications departments. Drafting a contract together with legal and contacting screening for background checks. Checking the origin of the data if it is on a personal level. Checking compliance to privacy regulation and security maturity levels. Negotiating accountability.
Screening	Integrity checks on board of partner company. Analysis of financial situation.
Legal	Supporting decision makers on legal issues. Making impact assessment on new regulation and checking new plans for legal issues.

Table 5.18: Key employees in acquisition of new data

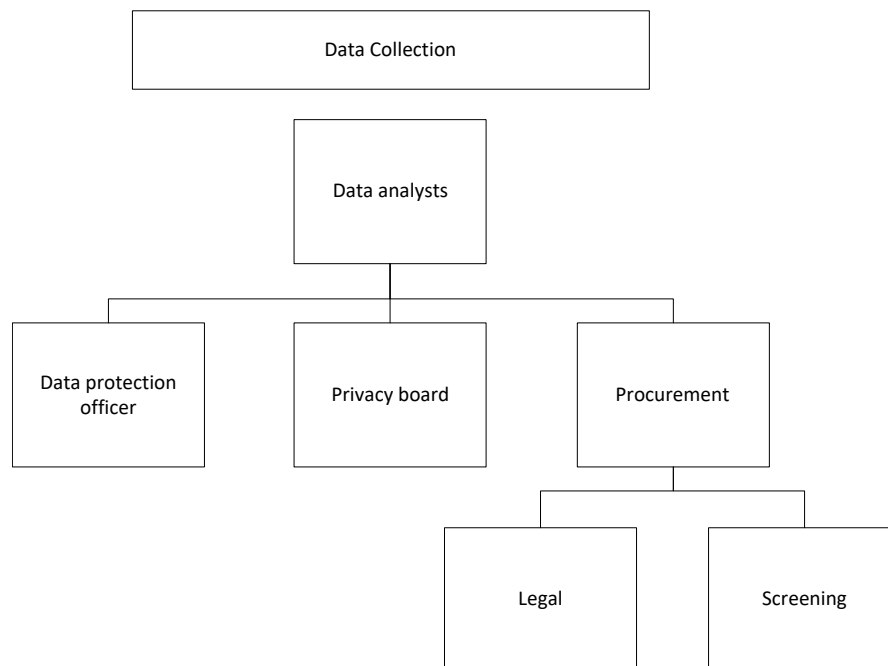


Figure 5.13: When the data analysts require new data the Privacy board and Data protection officer support the evaluation of the impact on Privacy. If the data is acquired from a company, procurement can support negotiating the contract together with legal and checking the company background together with screening.

5.5.2 The current process versus ethical boundaries

In this section the process described above is compared to potential ethical issues identified in the literature and an assessment is made whether these issues are sufficiently dealt with. First lets recap the issues identified in order of appearance in the literature section.

1. Biases in the data may lead to unfair conclusions (Boyd & Crawford, 2012; ibe, 2016)
2. Data being shared in public does not automatically mean consent is given for its use (Boyd & Crawford, 2012)
3. Data users have the responsibility to check the ethical practices of their suppliers (Martin, 2015)
4. Data users should compensate for the negative externalities they cause (Martin, 2015)
5. Data shared in one context does not automatically mean consent is given for its use in another context (Christen, Domingo-Ferrer, Herrmann, & van den Hoven, 2015; Martin, 2015)
6. Next to Privacy values such as Autonomy, Responsibility, and Fairness should be used by decision-makers (Christen et al., 2015)
7. Mere compliance to current laws is not sufficient, the public should be involved in the process of deciding what is acceptable use of personal data (CII, 2015; Verbond van Verzekeraars, 2016)

In the next paragraphs these issues are compared to current practices in the company in order to find potential issues in the current processes.

Biases: 1 Checking for biases is a responsibility of the data analysts. Since data cleaning is currently the responsibility of the data analysts themselves without a formal process these biases may not always be identified or checked for at all. Despite this it considered good practice and managers may ask for biases in the data used, this can create a culture where biases are correctly identified without the need for a formal process. Whether this is the case here is not clear.

Consent: 2, 5 Meaningful consent one of the most discussed issues. What counts as meaningful is up for debate, but the traditional small writing in the middle of an agreement that nobody reads is not sufficient from an ethical viewpoint. Using data originating from a different context without explicitly informing the customer of this use is also more and more viewed as unethical. While often the customer has given permission legally to share his data with third parties, this consent is often given in the meaningless way just described.

Currently the use of data where this would be an issue is avoided. The data used for this thesis however is exactly of this type. If the insurance company wants to start using this kind of data a system has to developed that clearly communicates this with their customers. Such a system is currently not in place, but a solution like a personal account containing all data that the company has and uses can be imagined.

Supply chain: 3 In the acquisition of new data the focus is on the company directly involved in the transaction. The procurement and screening departments work together to analyze this company. To ensure the privacy of their customers the data security is thoroughly discussed and formal requirements in the form of a certified security maturity level is required. The board of the company is screened on a personal level for ethical conduct, but the data itself and companies further down the supply chain are not automatically screened.

The origin of the data has to be disclosed though, so it is known where the data came from. Employees involved in the process will include a judgment on the origin if it is deemed to be unethical, but this is not a formal requirement. Formally evaluating data provenance will increase the chance that risks are identified.

Negative externalities: 4 While the notion of negative externalities caused by the Big Data industry could be valid, further research is required to assess what the impact is and what the extend of this impact is. For now I would advice the company to not take this into account, but keep an eye on any developments of this idea.

Autonomy, Responsibility, and Fairness: 6 Adhering to these values mean being one step ahead of the companies that go for compliance. This can be an asset when attempting to commoditize insurance, especially in a world where concern over personal data use is growing. For people to start valuing their data enough to build a business model depending on these values a large scale scandal involving personal data is probably required to happen first. However, with businesses competing to extract value from Big Data this does not seem unlikely.

Autonomy, giving customers control over how their data records are used, has the added side benefit of allowing customers to correct any mistakes in your data. Assuming customers are motivated in some manner to be honest this would be a highly efficient to ensure high data quality. Crowdsourcing data quality management like this could give the company a competitive advantage.

Responsibility, in this case creating a personal accountability for wronging customers due to mistakes in the use personal information, will cause analysts to be very prudent. Standards will have to be created that when analysts follow these standards they can be confident that no action will be taken toward them or that the company will take responsibility.

Fairness, the equal distribution of the advantages gained through participating in the sharing of data, should be explained to the customer. The benefits of sharing their personal data should be clear.

Compliance and public involvement: 7 From the industry association themselves comes the recommendation to involve the public in deciding what is acceptable use of personal data. This recommendation was echoed by the spokesperson interviewed, including institutions representing the public on the topic of personal data allows a wide range of opinions into the process. When properly managed this should prevent public backlash against decisions made.

Concerns identified during the interviews

In this section concerns identified are discussed. First, the risks that all interviewees agreed on and their current place in the process are described. These are direct consequences that the company might suffer due to to mismanaging data acquisition. Second, the tension between data analytics and risk controlling parties is described.

During the interviews it became clear that two real negative results are considered. The first is the risk of being fined for privacy, data security, or discrimination related missteps. The second

is reputation risk, public backlash against what it deems discriminatory or privacy infringing decisions.

The first risk is managed in the process by checking data security and privacy policies of companies worked, demanding a certain certified security maturity level and carefully controlling who gets access to the data. Also, one of the most negotiated contract details is how the accountability for fines is shared between the insurance company and the data aggregator. This ensures that if a fine results from the behavior of the partner company, they are responsible for settling it.

The second risk of public backlash is currently not formally included in the process. All employees said to report it whenever they noticed potentially unethical actions, but responsibility to identify and report this was not clearly assigned as far as these interview have shown. The PR spokesperson and data protection officer seem to be the ideal persons to involve since their work is closest to the public, but they are as of now least involved in the acquisition of new data. Even if the partner company will pay the fines, the insurance company risks suffering reputation damage.

A tension between data analysts and risk managing parties exists, because the analysts want to be able to investigate their ideas as fast as possible while the risk managing parties need time to move every request through their processes. According to the interviewees this is amplified by the difference in project structure. Data analysts are working increasingly *Agile*, meaning that the time between idea and execution is inherently short. Combining this with risk managers needing to go through a process that can exceed the time of one sprint means that the analysts are sometimes unable to obtain approval in time. Knowing that the risk management departments are formally supporting and decision power lies with the analysts this could lead to a situation where data is acquired without all supporting departments being involved.

Recommendations for process improvements

To deal with the issues identified, three changes are proposed. These changes have practical implications that may be too costly for the risk involved at this time, but they can at least provide a starting point for a discussion on how to improve the process.

The first suggestion is tracking data provenance. For every variable included in a dataset it should be clear how it was originally collected, at what time, where it moved in between, and what kind of operations were done on the data. This would help address all 7 issues from the literature. Biases and negative externalities from data collection can be found, it should be clear what type of consent was given, and communicating about the data to your customers can be done more transparently fostering mutual understanding.

The second suggestion is to formalize an ethical judgment about new data. This can be focused on the issues that are viewed most important and could thus potentially alleviate all of them. For example the privacy board could be extended to ethics board and report on more issues than just privacy. Also, public data protection organizations (e.g. Bits of Freedom) could be invited for a periodical dialogue with the spokesperson, analysts, and/or the data protection officer.

The last suggestion is to involve the risk management parties directly into the *Agile* project teams. This can provide clarity to the analysts about what is possible within the next sprint and allow risk assessments to be made in a timely manner if the value of the data is clear. If the risk manager already sees the value from the team meetings the analysts will not have to separately convince him. It can also help risk management departments to anticipate on the needs of the analysts and make sure approvals are ready in time.

5.5.3 The industry perspective

In this section the insurance industry is evaluated as a whole. What happens to the industry when all insurance companies are expanding their data analytics capabilities? The focus is on the profitability of the industry and what an insurance company could do to ensure a strong future market position.

The goal of the data analysts is to find customers that have a lower risk profile than other customers. New ways to dissect the group of potential customers that allow the company to target low risk customers before other insurance companies have figured out that this variable is a good predictor of future claims. This allows an insurance company to build a portfolio where it enjoys higher margins than its competitors or gain a larger market share by offering insurances to these customers at competitive prices.

Assuming that all insurance companies invest in increasing their data analytics capabilities, that the same data is accessible by all companies, and that all reach a similar skill level in analytics, all companies will eventually identify low risk customer groups, potentially different groups. It should be noted that models are not without uncertainty as has been described in the modeling done for this thesis. If all companies have different groups these groups will overlap and every company will attract customers whose risk is underestimated. This eventually makes every portfolio have negative margins.

Then there are the high risk customers. These customers will find insurance to become very expensive, potentially unaffordable. By pushing these customers out of the market the insurance industry is essentially reducing its own market. The companies would be better off in an industry where large risk pools exist so they can cover the largest amount of risk, but industry pressures push them to differentiate in pricing. This would also allow insurance companies to focus their analytics on damage prevention instead of differentiation.

So how can an insurance company position itself for success in this environment? A few suggestions are given here. The first suggestion is that if the road to increased differentiation is chosen, the company must be a leader in analytics. The first companies will be able to enjoy increased margins for as long as they keep their advantage.

The second suggestion is to decommo-ditize the insurance market. Improving the value offering, differentiating the company from other insurance companies and thereby having the ability to be successful at perhaps a higher price point than the competition. Improving the value could for example be done through improved claim handling and offering damage prevention analytics (e.g. in combination with smart homes). Other solutions and estimating the customer's willingness to pay for these services will have to be researched.

The third suggestion is, if the industry is going down this road, to start an insurance of last resort. With high risk customers being denied by regular insurers the market for insurers of last resort will grow. This allows the insurance company to recapture the risk that it has pushed out of the normal insurance market.

The fourth suggestion is to not go down this road at all. Solutions can consist of the government regulating insurance price differentiation or the industry self-regulating their price differentiation. In terms of societal consequences this seems to be most efficient use of resources. The race to the bottom will be avoided, insurers will be allowed to focus on damage prevention analytics, and the solidarity of insurance will be kept intact. Practically this is also by far the hardest results to achieve. Insurers will still be incentivized to differentiate even if it is just for targeted marketing campaigns or knowing which customers to offer discounts to if they threaten to leave. How to design regulation to avoid most of the differentiation should be carefully researched.

CHAPTER 6

Conclusions & Discussion

In this chapter the conclusions of this thesis can be found. The coming paragraphs discuss the answers found to each of the sub-questions. After this recommendations both to the company and for future research are given. We will start with the overall conclusions.

6.1 General Conclusions & Discussion

Starting by going to back to the main research question and the principles of Action Design Research (ADR) we can conclude the following. The question concerns assessing Big Data veracity for the use in insurance risk assessment models. The answer to this question is the veracity framework developed in Chapter 4. The framework is an example of exaptation, using existing knowledge to solve new problems, as is common in design research.

Using the definition of veracity as proposed in Chapter 3 – 'The ability of the data to support a decision making process by being appropriate, useful, and of sufficient quality in the context in which it is analyzed', we can say that this framework provides insight in the usefulness and quality aspects of veracity. By examining research from the domains of data quality, entity resolution, and actuarial science, a combination of metrics was found that can be used to assess new data sources as they become available and make a judgment about their veracity.

The principles of ADR say that the research must be inspired by practical challenges and the resulting artifact ingrained with academic theory. By using academic theory from the aforementioned disciplines to solve the practical problem of assessing veracity we can say that these two requirements have been satisfied. Furthermore, the principles are that the artifact must be a result of guided emergence in the organizational setting where it should be used and shaped reciprocally through concurrent evaluation by researchers and practitioners. As the research took place at the company in where the researcher was supervised by practitioners these principles have also been adhered to. Lastly, the outcomes should be generalized to an abstract problem and this part will be treated here in more detail.

The general problem that has been discussed in Big Data research concerns assessing the veracity of data. This research has focused on defining and measuring veracity. The definition of veracity can be used in other Big Data research projects to unify them. The measures for

veracity can also be used outside of the insurance context. The data quality metrics are relevant for any structured dataset, entity resolution is relevant in any case where dataset need to be linked, and model improvement is relevant in any context where improving predictions is the goal. While the first two have measures that are applicable to any structured data, the model improvement metrics depend on the modeling approach used for decision-making.

An extensive discussion of the weaknesses of the framework and the research can be found in the limitations section in this chapter, but first the practical and academic contribution are explicitly stated, followed by a discussion of the sub-questions.

6.1.1 Practical Contributions

The first practical contribution is the framework itself. This framework can be applied within the company to formalize the analysis of new data sources replacing the current ad hoc process. This will make it easier for decision-makers to evaluate and compare different data sources. Given their vision to increase price differentiation in insurance policy pricing, this framework can assist them evaluate any future data sources.

The cases that demonstrated the application of the framework are also a practical contribution. Both applications concerned data the company actually considered using and therefore the results of these cases give the company insights that can support decisions about using these sources. Next to being data sources that the company would have evaluated themselves, they are also examples on how to apply the framework.

Furthermore the analysis of the company's current data acquisition process and the associated risks give the company an opportunity to improve it. Suggestion on how to improve have also been provided. These can be found in the recommendations sections at the end of this chapter.

Finally, the insurance industry analysis can facilitate a discussion about strategic direction for the company in the near future. Knowing potential future scenarios allows the company to prepare and adapt ensuring continuation of its business.

6.1.2 Academic Contributions

The first academic contribution of this thesis is the definition of veracity. The definition of Veracity proposed is as follows.

Veracity

The ability of the data to support a decision making process by being appropriate, useful, and of sufficient quality in the context in which it is analyzed.

The main idea is that the Veracity assessment of data should serve decision making and thus it depends on the context in which the data is used. The appropriateness of the data in question should be determined by analysts before collecting the data. The usefulness and quality are treated in the framework. The definition resulted from reviewing definitions in general and Big Data academic literature, but focusing on characteristics that are common in Big Data science. So far many different scholars have defined veracity differently in their research so a common definition is necessary to unify Big Data veracity research.

Reflecting on the definition, it did become more broad than the original use of veracity in other academic literature. It also risks overlapping with another characteristic proposed in academic literature, namely value. Value however, can focus more on analytics techniques and processes to make the data part of a business model. Veracity, according this definition, focuses on the potential that the data has to become a part of the business model. Therefore this definition

complements the other characteristic of Big Data in a way that supports the creation of tools that allow us to benefit from Big Data.

Another contribution is demonstrating the application of a relatively novel research methodology. Looking back at how it influenced the process and the results, it is concluded that this methodology can be successful in creating useful artifacts, but that the researchers have to pay special attention to ensure making an academic contribution. When developing an artifact within a company that must be useful to the company the focus easily drifts to practical contributions over academic ones. This part of the research, where a reflection is made on the theory used, is possibly more important than for other research designs. In other research answering the main question is a contribution by itself. While it can be argued that this counts for design research as well, care has to be taken that at least the problem treated is not solved before or that the techniques used to solve it are new. We will see that this is the case in this research.

As with any design research, the artifact developed is one of the academic contributions as long it satisfies the requirements of solving a new problem or using a new solution. This artifact is an example of exaptation, using existing knowledge to solve a new problem. The problem of assessing Big Data veracity has only recently started to get academic attention and methods to judge it have been called for in Big Data research. While the framework bears similarity to data quality assessment methodologies due to being partly inspired by it, there are also key differences. This framework takes another perspective than data quality, focusing on how to evaluate Big Data to support decision-making. Thus striving for veracity assessment is fundamentally different than striving for data quality assessment. That makes this a first step in a new direction. Potential extensions of this framework can be found in the recommendations section at the end of this chapter.

Next to developing a framework that goes in a new direction, this research also included two cases as environments for evaluation. This allowed the framework to be tested and receive immediate feedback from the context for which it was designed.

The development of the framework also led to general insights about the fields of science that were used as inspiration. The following paragraphs reflect on each field specifically. Note that the model improvement part was based on the practices of the company. This part of the framework should be adapted based on the model type that the data is acquired for and is not considered part of the academic background.

Data quality assessment The data quality methodologies reviewed contained definitions that were especially useful to assessing structured data. One of the challenges in assessing veracity is the prevalence of unstructured and semi-structured data. The last reviewed methodology, Heterogeneous Data Quality Methodology, started with the purpose to be able to assess less structured data, but the researchers concluded that the metrics provided still required the data to be put into a structured format first. So the first reflection is that development of metrics for semi-structured data could be valuable to Big Data veracity assessment.

The second reflection on data quality assessment is that the measures for scoring data quality all are subjective and context dependent when it comes to the final judgment. Several quantitative methods for determining quality scores are available, but in subjectively choosing one the outcome of the evaluation can be influenced. Thus, creating new scoring methods tailored to the challenges in Big Data can be a future contribution. For example, measures that score veracity of a statement based on meta data or syntax, metrics that score trustworthiness of a source by evaluating their processes, or other metrics that circumvent the need for a structured dataset to be available.

Entity Resolution While this step was added to the framework it did not arise as a major issue in both cases. In the first the datasets could be linked through primary key and in the second case it was evaluated qualitatively. So far the conclusion is that if no key is available using the new data would become incredibly cumbersome for the company. Developing reliable algorithms for entity resolution is a difficult challenge. Therefore, reflecting on this, the creation of tools to ease this process would be a valuable addition.

6.2 Discussion of Sub-questions

In this section all the sub-questions are answered in order. Sub-questions 1 and 2 are treated in chapters 2 and 3. Sub-question 3 is treated in chapter 5. Sub-question 4 is treated in chapters 3, 4, and 5. Sub-question 5 is treated throughout the research, but specifically answered in the previous section. Sub-question 6 is treated in chapters 3 and 5. The sub-questions are as follows.

1. What criteria are currently used for assessing claim risk?
2. Which data sources can potentially improve the current risk assessment??
3. What are the key characteristics of data sources for insurance risk assessment?
4. How can data and information quality assessment methodologies support the assessment of the veracity of these data sources?
5. What can be contributed to Big Data research from the development of the veracity assessment framework?
6. What are the consequences of increased data analytics capabilities in the insurance industry?

6.2.1 1 – Criteria for assessing claim risk

The goal of adding new data is to be able to further differentiate between customer groups. This allows competitive pricing for low risk customers and thus expand the insurance portfolio in a healthy way. Customers that the company identifies as high risk will likely be able to obtain lower prices at other insurance companies and, if these customer actually were high risk, negatively impact the expected value of the competitors portfolios.

This data describes customers and is thus by definition personal data. This means that it is subject to personal data protection regulation that is becoming increasingly strict. Obtaining consent from customers and in the near future allowing a customer to access and correct this data is mandatory. The distinction between differentiation and discrimination is also important here. In the case of contents insurance the company has no obligation to enter into an insurance contract with anyone, so whatever information is available about the person could be used for risk evaluations. Some variables such as ethnicity or religion are considered discriminatory and are therefore off-limits.

Pricing decisions are based on the risk assessments. It is important that the pricing structure follows a for customers understandable pattern (i.e. even if the analysis suggest that people making twice the modal income have a much lower risk than everyone else, see Figure 5.4a, a pricing structure that increases prices by increasing income is required).

Finally, the DNB (De Nederlandse Bank) supervises the company and thus modeling and pricing decision should be able to withstand its scrutiny.

6.2.2 2 – Data sources that can potentially improve the risk assessment

Improving the risk assessment models allows more detailed differentiation between high and low risk customers. If low risk customers can be identified these customers can be offered a competitive premium. This allows the portfolio to grow in a profitable way while offering fairer premiums to low risk customers.

So data sources that contain information describing characteristics of customers that indicate a higher or lower risk should be considered. These must be appropriate concerning the type of insurance and must be variables that the company is legally and by their policy allowed to differentiate on. This excludes discriminatory data such as for example ethnicity.

The type of variables as such could either be characteristics of the customer or of objects related to the type of insurance. The former can be found in the data broker's dataset, where a variable like life phase differentiates customers of different age groups. The latter can be found in house trading platform data. Since this research was done for contents insurance, it made sense to include house characteristic from a house trading platform as these can be related to the contents of a house.

6.2.3 3 – Key characteristics for data sources

Key characteristics of new data sources can be divided in two categories. The first is veracity related characteristics. The data has to be of sufficient quality, had to be able to be linked to the current dataset, and describe customers in a way that can improve risk predictions. Secondly, there are security, legal, and privacy concerns. Upcoming regulation has to be taken into account and the source of the data has to be checked for compliance on ethical standards.

New data sources have to be of sufficient quality to be able to differentiate successfully. Low quality can lead to mispricing customers. Pricing them too high will cause the organization to lose potentially profitable customers, pricing them too low will cause the organization to have to pay out more claims than expected. The new data also has to be able to linked to the current data in a reliable way to avoid low dataset quality that can cause the same mispricing.

Next to the risk of mispricing, there are data security and privacy issues that have to be taken into account. This data is always personal in nature. This means that security has to be carefully managed to ensure that none of this data ends up with unauthorized people.

It is also important that ethical boundaries concerning the use of personal data are respected. Especially with upcoming reform of the EU data protection rules. Increased accountability and responsibility requires organizations to design processes that ensure compliance. Key changes are that consent is required by means of a clear affirmative action, free and easy access to data for customers must be possible, and the right to be forgotten is reinforced.

6.2.4 4 – The value of data quality research to veracity assessment

There is some overlap to criteria important in traditional data quality measurement and the assessment of veracity. Therefore data quality research has served as an inspiration for the veracity assessment framework. In the current version the focus is on the data-centric quality measures accuracy, completeness, consistency, and timeliness. These metrics can provide an indication of the quality of any data source as long as the data is presented in a structured format.

As it turns out the metrics can be difficult to measure and in some cases are not very meaningful. Semantic accuracy requires having ground truth data, but this is often difficult to gather on a large enough scale and if it was possible then this ground truth data should be collected instead of acquiring the new data. Syntactic is meaningless in the most common

situation of a data broker's data, where the data broker created the dataset from the domain with to which the values in dataset would be tested. Completeness does provide an exact statistic into the data, although even high degrees of incompleteness might still make the data valuable. Consistency is the least exact as the conclusions depends highly on the effort put into finding potential inconsistencies. Timeliness finally is highly dependent on the estimated volatility of the data. Although a qualitative pass/fail judgment can usually be made, a better approach would be to determine a volatility function from which an acceptable age can be found.

Still, despite these limitations, the metrics can still allow an organization to make a reasonable judgment about the usefulness of the data. It allows comparing different data sources and it can structure the discussion when contemplating the acquisition of new data.

6.2.5 5 – Contributions to Big Data research

The main contributions to the Big Data research are the framework itself and the definition of veracity. As stated in the introduction of this thesis, veracity assessment is one of the big challenges in Big Data. This framework provides new ways of measuring veracity by drawing upon other scientific disciplines. Both the framework and the definition have already been discussed in the academic contributions section.

6.2.6 6 – Consequences of increased data analytics capabilities in the insurance industry

More data combined with increases in computational power enable the construction of increasingly detailed models. Ultimately every customer could have a unique risk profile allowing the insurer to reward low risk customers with lower premiums. The side effect is that high risk customers would end up with increasingly higher premiums. While this may seem fair, it also undermines the risk sharing principle of insurance. Consequences have been viewed from two perspectives, the single organization and the industry as a whole.

For an individual insurance company increased dealing with personal data comes with certain risks. The two direct risks identified are the risk of being fined for violating EU data protection regulations and the risk of damage to the company image. The literature review in Section 3.4 on page 31 shows where the way the company handles its data can trigger these events. Most importantly consent for data use being violated and basic values being compromised.

In order to deal with these issues three process changes are proposed. First, keeping data provenance records that include at least the data collection method, type of consent given, and operations done on the data. Second, formalizing ethical reporting on new data. Third, involving risk management parties in the data analyst project teams.

In the insurance industry, as more and more insurers are advancing their data analytics capabilities, they too will be able to offer similar competitive prices and the ability to enjoy larger margins will have diminished again. A potential end game might be a commoditized insurance industry where customers are mostly selecting insurance based on the premium they are offered. Uncertainty is in the nature of data analytics and therefore the customers an insurer manages to sell insurance to will be those where the risks are underestimated. This leads to a situation with low or even negative margins after large industry wide investments in data analytics.

A few solutions are proposed to position yourself as an insurance company in this environment. If the investments in data and data analytics capabilities are deemed worthwhile, only the industry leaders will be able to enjoy increased margins. Another path to take perhaps simultaneously is to search for new ways to add value in order to decommo-

market. For example it could be that a niche exists with customers with a higher willingness to pay for improved customer service. A third option is to get into the potentially growing market for insurers of last resort, offering insurance solutions for customers that suddenly fall on the high risk side of the spectrum. Finally, new legislation limiting the possibilities for price differentiation could be made that prevent this race to the bottom, but still leave the question of how insurers can improve their business models.

6.3 Limitations

The limitations can be split in limitations specifically following from the research design and general limitations to the potential impact of the study. During the cases there also were limitation to the availability and quality of data, but since these would also exist in real life applications they are not so much limitations to this study as limitations to the scores of those dataset.

For example, in the house trading platform case only a small amount of data was available resulting in a too small sample for proper assessment of the data. This resulted in a relatively low veracity while in reality the house trading platform could provide high veracity data.

6.3.1 Study design limitations

Action design research is a relatively free creative process. The sources of inspiration used for creating the framework depend on the knowledge of the researcher and the stakeholders involved. This means that other researchers tasked with solving the same problem could reach completely different solutions.

This being essentially a solo research project with only sparse access to the time of other people limits the possibilities for realizing the action in action design research. The company was involved enough to provide the required feedback, but it was not comparable to examples described by Sein et al. (2011) where teams of multiple researchers and practitioners were devoted to the developing, testing, and implementing the artifact in an organizational setting.

The strength of having a high chance that a practically useful artifact is developed comes at the cost of having generalizable outcomes.

6.3.2 Impact limitations

While the framework provides a measure for veracity, it does not deal with challenges posed by the other V's of Big Data (Volume, Variety, Velocity, Veracity). It can be used on high volumes of data given sufficient computational power, but has not been designed with this as a requirement. For application to high velocity data the framework would need to be automated, while the metrics chosen can be adapted to provide automated scores, in its current form the judgment of the data analyst still has a critical role. The most problematic however is high variety data. Currently the framework requires that data is available in a structured format, while a large part of Big Data consists of unstructured data. Adapting or extending the framework to cope with these challenges can be an important contribution of future research.

Next to this the information provided by this framework is only part of the information needed by a decision-maker when evaluating new data. As found in this research there are ethical boundaries that need to be respected when considering personal data, there are industry wide implication of strategies that drive the need to acquire new data, and there is of course a cost/benefit ratio that the framework makes no statements about. However, these are not part of veracity and while extending the framework to includes these evaluations would cause it to lose focus.

6.4 Recommendations

The recommendations can be divided in practical recommendations for the insurance company and recommendation for further academic research.

6.4.1 Recommendations to the Organization

Practical recommendations can again be split, related to opportunities in data analytics, related to the ethical use of data, and related to the future of the insurance industry. The recommendations are given as a list and discussed in the text following it. First general opportunities in data analytics and framework related recommendations are given.

1. Evaluate different methods of modeling
2. Get more computational resources
3. Multi peril analysis
4. Crowd source data quality
5. Avoid unstructured and semistructured data sources such as social media
6. Look for datasets with variables of which a ground truth exists within the company for better accuracy estimation
7. Analyze more than the top 5 models to see which variables are included in most models
8. Include interaction terms

During this research it has become clear that next to the Generalized Linear Model (GLM) used in this thesis, recent advancements in data science combined with increased computational power allow for other forms of modeling as well. In a search for improved risk assessments different types of models from the area of machine learning or a GLM using Bayesian statistics could give the company an edge in analytics. Even if some of these techniques have a hard time complying to the understandable pricing rule, marketing and customer retention decisions can in any case be made based on these techniques.

Tying in with the previous point is the amount of computational resources available. When using automated model selection techniques as provided by the `glmulti` R package, analysis is computationally cumbersome. With the size of dataset used in insurance, a dedicated server should boost productivity.

On another note, models can be split according to peril. Splitting the models by cause of damage can provide interesting new opportunities for external data. Data specific to each peril can be found and this could result in models with a much higher explanatory power.

Data quality can be improved by having customer correct the data describing themselves. Implementing a system where every customer's data is available through an online portal where they are allowed to correct it has multiple advantages. Next to improving data quality it ensures compliance with part of the 2018 EU data protection regulation which demands that customers can access their data without too much difficulty.

Making sense of unstructured data is difficult and its analysis produces uncertain results. There is so much potential in structured data sources that focusing on acquiring new structured data and improving analysis methods is expected to yield more results with more certainty in less time. Techniques for interpreting unstructured data and linking it to an existing dataset are still immature and the value derived from doing so unclear.

Then finally there are some recommendations related to the use of the framework. For estimating accuracy it could be interesting if new datasets include a variable of which a variant is already available within the company. This would for an easy comparison and gives an indication of the accuracy of the rest of the data.

One easy improvement to the model given more computational power is to determine the 100 best models instead of 5 best and include interaction terms as well. The interaction terms give more depth to the model, improving understanding of the predictors. Looking for the 100 best models allows an analysis of term importance which determines the importance of variables based on their inclusion in models weighed by the IC scores of those models.

If the new data is found to be too uncertain to base pricing decisions on, it can still be used as an alternative method to calculate Customer Lifetime Value (CLV). CLV is the prediction of the profit obtained from the entire future relationship with the customer. This CLV can be used to target new customers with a high CLV, even though there is too much uncertainty in the value to base pricing decisions on. Another use for this CLV can be to assist decision making for customer retention. Customers with a high CLV can be offered incentives to stay, whereas customers with lower CLVs can be let go.

The following recommendations result from the analysis of the data collection process.

1. Formalizing ethical considerations into data collection process
2. Involve digital rights organizations (e.g. bits of freedom) in the analytics process

The main recommendation concerning company data ethics to formalize an ethical judgment that extends beyond privacy and compliance when new data is considered. With increasing regulation aiming to give customers more power over their data, it becomes important to consider whether current practices are future proof. Involving digital rights organizations and similar groups when having new ideas concerning data analytics can be a way of outsourcing the ethical judgment.

The following recommendations follow from the consequences of the increased data analytics in the insurance industry as a whole.

1. Be the leader in data analytics
2. Find ways to decommoitize the insurance market
3. Start an insurer of last resort

Only the leader manages to improve its market position. The laggards will be left with a portfolio containing underestimated risk. Decmoitizing, improving the value proposition by offering other services, can be a way out . If it turns the scenario of high risk customers being pushed out of the regular insurance market becomes true, spinning off an insurer of last resort could be profitable.

6.4.2 Recommendations for Future Research

The following topics are recommendations for future research that follow from this thesis.

1. Apply the framework in different contexts
2. Create metrics that can be applied to unstructured data
3. Automate the framework in order to efficiently score high velocity data

4. Develop methodologies to efficiently solve the entity resolution problem
5. Extending and evaluating other veracity measures
6. Future scenarios for the insurance industry

The first recommendation concerns the context dependency of this research. This framework was created with and for an insurance company. It is expected to also be able function in other environments, but this should be tested in further research.

The second recommendation concerns the high variety in Big Data. The level of structure in data differs considerably. For example, the house trading platform data was available in semi-structured form on a website. For the framework to be applied on it however, it needed to be available in a relational table format. This means that the data first had to be scraped from the website before it could be analyzed. In collecting the data from this website errors may have entered the data that reduced the veracity score. In the originally proposed case of social media, the practical challenge of gathering data and linking it to the base dataset was too complicated to deal with for this thesis (on top of the ethical issues). Being able to directly analyze this would allow a wider range of sources to be assessed before going through the effort of structuring the data.

The third recommendation concerns the high velocity of Big Data. The data sources used are updated more and more frequently. The veracity of this data may vary with time. Automating the framework would allow periodic or even continuous veracity assessment to ensure that data used by the company is always scoring sufficiently high.

The fourth recommendation goes into the entity resolution phase. Improvements in the field of entity resolution will positively impact Big Data analytics. If it becomes more practical to link existing customers to datasets that cannot be linked through primary key then the amount of data sources increases considerably. Social media data could become a useful source of insight.

The fifth recommendation concerns the metrics used in the framework. A property of design research is that the sources that inspire the researcher have a large influence on the resulting artifact. Another researcher given the same assignment would probably have produced a different results. Therefore through finding and comparing a wide range of measures veracity assessment could be improved.

Finally, the analysis of how the insurance industry could change through industry wide increases in data analytics raises several questions. The likelihood and societal impact of differentiating insurance policy premium pricing down to the individual level should be researched. If this is found to either negatively impact the insurance industry itself or society as a whole, strategies for regulating the insurance industry in such a way that it can remain a functional asset to society become interesting topics for research.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on automatic control*, 19(6), 716-723.
- Anagnostopoulos, I., Zeadally, S., & Exposito, E. (2016). Handling big data: research challenges and future directions. *The Journal of Supercomputing*, 72(4), 1494–1516.
- Ashwin, K., Kammarpally, P., & George, K. (2016). Veracity of information in twitter data: A case study. In *2016 international conference on big data and smart computing, bigcomp 2016* (p. 129-136).
- Batini, C. (2011). A data quality methodology for heterogeneous data. *International Journal of Database Management Systems*, 3(1), 60-79.
- Batini, C., Cabitza, F., Capiello, C., & Francalanci, C. (2006). A comprehensive data quality methodology for web and structured data. In *1st international conference on digital information management* (p. 448-456).
- Batini, C., Capiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), Art.16 p 1-52.
- Batini, C., & Scannapieca, M. (2006). *Data quality concepts, methodologies and techniques*. Springer Computer Science eBooks.
- Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 1-36.
- Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication and Society*, 15(5), 662-679.
- Boyd, D., & Marwick, A. (2011). Social privacy in networked publics: teens' attitudes, practices, and strategies. In *Privacy law scholars conference*.
- Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: a practical information-theoretic approach*. Springer-Verlag New York Inc.
- Calcagno, V. (2015). *Package glmulti, model selection and multimodel inference made easy*.
- Calo, M. R. (2013). Digital marketing manipulation. *The George Washington Law Review*, 82(4), 995-1051.
- Camacho, J. (2014). Tackling the big data 4 vs for anomaly detection. In *Proceedings - ieeef infocom* (p. 500-505).
- Ceci, S., Ross, D., & Toglia, M. (1987). Suggestibility of children's memory: Psycholegal implications. *Journal of Experimental Psychology: General*, 116(1), 38-49.

- Chessell, M. (2014). *Ethics for big data and analytics*. IBM whitepaper.
- Christen, M., Domingo-Ferrer, J., Herrmann, D., & van den Hoven, J. (2015). Beyond informed consent – investigating ethical justifications for disclosing, donating or sharing personal data in research. In *Joint conference of the international society for ethics and information technology and the international association for computing and philosophy*.
- CII. (2015). Big data and insurance: a conversation. *CII research reports*.
- Conger, K. (2016). LinkedIn sues anonymous data scrapers. *TechCrunch.com*.
- Del Boca, F., & Noll, J. (2000). Truth or consequences: The validity of self-report data in health services research on addictions. *Addiction, 95*(SUPPL. 3), S347-S360.
- Diebold, F. X. (2012). A personal perspective on the origin(s) and development of "big data": The phenomenon, the term, and the discipline. *Unpublished manuscript, University of Pennsylvania*.
- Earl, J., Martin, A., McCarthy, J., & Soule, S. (2004). The use of newspaper data in the study of collective action. *Annual Review of Sociology, 30*, 65-80.
- Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication, 11*(2), 27-62.
- Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1-16.
- EU. (1995). *Eu directive 92/46/ec*.
- Freiling, A. (2015). 2015 european insurance outlook. *EY Insurance*.
- Frizzo-Barker, J., Chow-White, P. A., Mozafari, M., & Ha, D. (2016). An empirical study of the rise of big data in business scholarship. *International Journal of Information Management, 36*(3), 403-413.
- Funda. (2016). *Gebruikersvoorwaarden – funda.nl/gebruikersvoorwaarden*.
- Gregor, S., & Hevner, A. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly: Management Information Systems, 37*(2), 337-355. (cited By 288)
- Hall, K. (2013). Data quality more important than fixating over big data, says shell vp. *ComputerWeekly.com*.
- Hardy, J. (1997). Amyloid, the presenilins and alzheimer's disease. *Trends in Neurosciences, 20*(4), 154-159.
- Harrel, F. (2001). *Regression modeling strategies*. Springer-Verlag New York Inc.
- Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning - data mining, inference, and prediction*. Springer.
- Henfridsson, O. (2011). *Action design research introduction presentation*. (University of Oslo, Viktoria institute)
- Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly: Management Information Systems, 28*(1), 75-105. (cited By 4003)
- Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems, 19*(2), Art. 4.
- Hicken, M. (2013). *Data brokers selling lists of rape victims, aids patients*.
- Hocking, J., Wood, A., Dally, N., Pan, K., Lin, B., Ban, H., ... Lee, S. (2014). Insurance and technology, evolution and revolution in a digital world. *Morgan Stanley Research Global*.
- ibe. (2016). Business ethics and big data. *Business Ethics Briefing*.
- Janssen, M., van der Voort, H., & Wahyudi, A. (2016). Factors influencing big data decision-making quality. *Journal of Business Research, 70*, 338-345.
- Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). *Modern actuarial risk theory*. Springer.

- Kamel Boulos, M., Maramba, I., & Wheeler, S. (2006). Wikis, blogs and podcasts: A new generation of web-based tools for virtual collaborative clinical practice and education. *BMC Medical Education*, 6(41), 1-8.
- Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 110(15), 5802-5805.
- Landrum, J., & Bone, R. (2001). Lutein, zeaxanthin, and the macular pigment. *Archives of Biochemistry and Biophysics*, 385(1), 28-40.
- Laney, D. (2001). 3d data management: Controlling data volume, velocity, and variety. *Application Delivery Strategies – META Group*.
- Langlois, J., Kalakanis, L., Rubenstein, A., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? a meta-analytic and theoretical review. *Psychological Bulletin*, 126(3), 390-414.
- Liu, J., Wu, S., & Zidek, J. (1997). On segmented multivariate regression. *Statistica Sinica*, 7(2), 497-525.
- Lopez, V., del Rio, S., Benitez, J. M., & Herrera, F. (2015). Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data. *Fuzzy Sets and Systems*, 258, 5 - 38. (Special issue: Uncertainty in Learning from Big Data)
- Loshin, D. (2001). *Enterprise knowledge management - the data quality approach*. Academic Press.
- Lu, R., Zhu, H., Liu, X., Liu, J., & Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, 28(4), 46-50.
- Lukoianova, T., & Rubin, V. (2013). Veracity roadmap: Is big data objective, truthful and credible? In *Advances in classification research online* (Vol. 24).
- Manyika, J. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*.
- Marr, B. (2015). *A brief history of big data everyone should read*.
- Martin, K. E. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive*, 14(2), 67-85.
- McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*.
- McGranahan, D. (2015). Global insurance industry insights. *McKinsey Global Insurance Pools*.
- Minty, D. (2016). Price optimisation for insurance. optimising price; destroying value? *Think-piece CII*.
- Moran, J. (2016). *Iot: The next best thing to reading your customer's mind*.
- Nelder, J., & Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370-384.
- Pipino, L., Lee, Y., & Wang, R. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211-218.
- Power, D. J. (2014). Using 'big data' for analytics and decision support. *Journal of Decision Systems*, 23(2), 222-228.
- Press, G. (2013). A very short history of big data. *Forbes*.
- Rachels, J. (1975). Why privacy is important. *Philosophy & Public Affairs*, 4(4), 323-333.
- Romei, A., & Ruggieri, S. (2013). A multidisciplinary survey on discrimination analysis. *The knowledge engineering review*, 29(5), 582-638.
- Saha, B., & Srivastava, D. (2014). Data quality: The other face of big data. In *Proceedings - international conference on data engineering* (p. 1294-1297).
- Schneier, B. (2015). *Data and goliath*. W. W. Norton & Company.
- Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics:

- The real-world use of big data. *IBM Institute for Business Value*.
- Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6(2), 461-464.
- Sein, M. K., Henfridsson, O., Purao, S., Rossi, M., & Lindgren, R. (2011). Action design research. *MIS Quarterly*, 35(1), 37-56.
- Shyr, J., & Spisic, D. (2014). Automated data analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(5), 359-366.
- Snow, D. (2012). Adding a 4th v to big data - veracity. *Dwaine Snow's Thoughts on Databases and Data Management (IBM)*.
- Stvilia, B., Gasser, L., Twidale, M., & Smith, L. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), 1720-1733. (cited by 149)
- Verbond van Verzekeraars. (2016). Green paper big data: grip op data. *Verbond van Verzekeraars*.
- Vlassov, V., Lozano, M. G., Rosell, M., & Franke, U. (2015). Toward automatic veracity assessment of open source information. In *2015 IEEE International Congress on Big Data*.
- Wagenmakers, E.-J., & Farrell, S. (2004). Aic model selection using akaike weights. *Psychonomic Bulletin & Review*, 11(1), 192-196.
- Walzer, M. (1982). *Spheres of justice: A defense of pluralism and equality*. New York City: Basic Books.
- Wang, R. (1998). A product perspective on total data quality management. *Communications of the ACM*, 41(2), 58-65.
- White, M. (2012). Digital workplaces vision and reality. *Business Information Review*.
- Woodall, P., Borek, A., & Parlikad, A. K. (2013). Data quality assessment: The hybrid approach. *Information and Management*, 50, 369-382.
- Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(2), 1-6.

List of Figures

2.1	Response variable distributions	17
(a)	Frequency counts	17
(b)	Poisson distribution	17
(c)	Probability distribution of damages	17
(d)	Gamma distribution	17
3.1	Steps and transfer points in the big data chain	24
4.1	Veracity assessment framework activity workflow	43
5.1	Histogram of claims frequency	49
5.2	Claim severity probability distribution	50
5.3	Base variable frequencies weighted for exposure	51
(a)	Insured amount frequency plot	51
(b)	Family composition frequency plot	51
(c)	Risk zone composition frequency plot	51
(d)	Credit rating composition frequency plot	51
(e)	Income composition frequency plot	51
(f)	House type composition frequency plot	51
5.4	Base variable frequencies weighted for exposure 2	52
(a)	Insured amount frequency plot 2	52
(b)	Family composition frequency plot 2	52
(c)	Risk zone composition frequency plot 2	52
(d)	Credit rating composition frequency plot 2	52
(e)	Income composition frequency plot 2	52
(f)	House type composition frequency plot 2	52
5.5	Base severity model variable coefficients	54
(a)	Insured amount coefficients	54
(b)	Family composition coefficients	54
(c)	Risk zone coefficients	54
(d)	Credit rating coefficients	54
(e)	House type coefficients	54
5.6	Education 3 vs 4 category comparison	59

(a)	Education 3 frequency barplot	59
(b)	Education 4 frequency barplot	59
5.7	Assessment of broker's data	60
5.8	House trading platform scraped data counts	65
(a)	House price	65
(b)	House surface area	65
5.9	House trading platform scraped data counts	66
(a)	Garden	66
(b)	# of Rooms	66
(c)	House type	66
5.10	House trading platform linked data exposures	69
(a)	House price	69
(b)	House surface area	69
(c)	Garden	69
(d)	House type	69
(e)	# of Rooms	69
5.11	Assessment of house trading platform data	71
5.12	Comparison of case results	74
(a)	House trading platform scores	74
(b)	Broker data scores	74
5.13	Departments involved in the acquisition of new data	78
B.1	Broker's data extended model coefficients AIC 1	105
(a)	Income coefficients	105
(b)	House surface area coefficients	105
(c)	House price coefficients	105
(d)	Social class coefficients	105
B.2	Broker's data extended model coefficients AIC 2	106
(a)	Life phase coefficients	106
(b)	House type coefficients	106
(c)	Risk zone composition coefficients	106
(d)	Credit rating composition coefficients	106
B.3	Broker's data extended model coefficients AIC 3	107
(a)	Insured amount coefficients	107
(b)	Family composition coefficients	107
B.4	Broker's data extended model coefficients SA1 1	114
(a)	Income coefficients SA1	114
(b)	House surface area coefficients SA1	114
(c)	House price coefficients SA1	114
(d)	Social class coefficients SA1	114
B.5	Broker's data extended model coefficients SA1 2	115
(a)	Life phase coefficients SA1	115
(b)	House type coefficients SA1	115
(c)	Risk zone composition coefficients SA1	115
(d)	Credit rating composition coefficients SA1	115
B.6	Broker's data extended model coefficients SA1 3	116
(a)	Insured amount coefficients SA1	116
(b)	Family composition coefficients SA1	116
B.7	House platform extended model coefficients H1 1	121

(a)	House surface area coefficients H1	121
(b)	House price coefficients H1	121
(c)	Insured amount coefficients H1	121
(d)	Family composition coefficients H1	121
B.8	House platform data extended model coefficients H1 2	122
(a)	House type (trading platform) coefficients H1	122
(b)	House type (company) coefficients H1	122
(c)	Risk zone composition coefficients H1	122
(d)	# of Rooms composition coefficients H1	122
B.9	House platform extended model coefficients H6 1	124
(a)	House type (company) coefficients H6	124
(b)	Insured amount coefficients H6	124
(c)	Family composition coefficients H6	124

List of Tables

1.1	Definitions of veracity	3
1.2	Challenges versus issues to deal within Big Data	8
2.1	Comparing DR, AR, and ADR	10
2.2	Stages and principles of Action Design Research	12
2.3	Workflow	13
3.1	Data quality assessment methodologies	26
3.2	Data quality metrics to be used for the framework	26
3.3	Functional Forms of DQA (Pipino et al., 2002)	28
4.1	Data quality metrics included in the veracity assessment framework	39
4.2	Entity resolution metric included in the veracity assessment framework	41
5.1	Original data set cleaning	47
5.2	Original data set overview	48
5.3	Claim frequency table	49
5.4	Frequency model overview using the base data	53
5.5	Severity model overview using the base data	53
5.6	Data broker’s data set overview	55
5.7	Completeness analysis of the data broker’s extended data set	56
5.8	Inconsistencies in the dataset extended with the data broker’s data	56
5.9	The top severity models for the data broker’s data	61
5.10	Performance of the competing severity models	62
5.11	The top frequency models for the data broker’s data	62
5.12	Performance of the competing frequency models	63
5.13	House trading platform data set overview	64
5.14	Completeness analysis of the trading platform extended data set	67
5.15	Inconsistencies between house platform and company data	67
5.16	The top frequency models for the house trading platform data from the glmulti analysis	70
5.17	Performance of the competing house trading platform models	70
5.18	Key employees in acquisition of new data	78

A.1	Data coding	102
B.1	Baseline model coefficients of the severity model	104
B.2	Data broker's model coefficients of the frequency model AIC	108
B.3	Anova FA1	109
B.4	Data broker's model coefficients of the frequency model BIC	110
B.5	Anova FB1	111
B.6	Data broker's model coefficients of the severity model AIC	112
B.7	Data broker's model coefficients of the severity model BIC	117