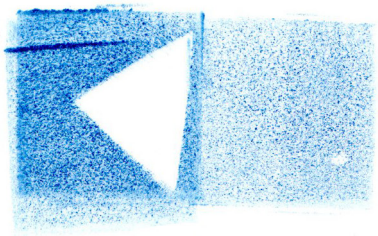
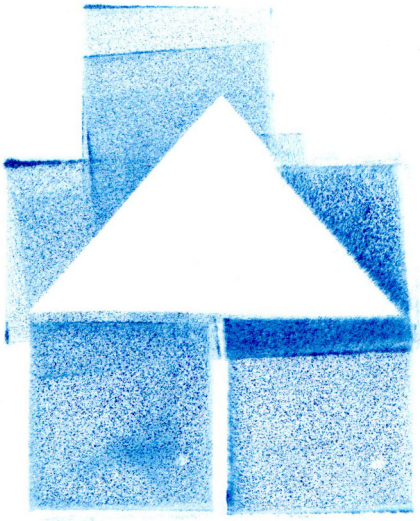
on rainstorm damage to building structure and content





PROPOSITIONS

accompanying the thesis

On rainstorm damage to building structure and content

Matthieu Spekkers
Delft, 7 January 2015

1. The amount of repair costs of a house that has been damaged by rainfall cannot be explained by rainfall data, but depends on other factors related to the house and its owner.
2. In the Netherlands, damage due to intrusion of sewage into houses already occurs for rainstorms that are less intense than the design storms that are being used to design urban drainage systems.
3. The rainfall clause that has been introduced in the year 2000 in most Dutch private property and content insurance policies, does not account for the short, intense rainstorms that can overload sewer systems. This should therefore be adjusted.
4. In twenty years, the design of urban drainage systems in the Netherlands will be based on risk management models where the use of damage data from insurance companies will play an important role.
5. Depression has evolutionary roots: it is a mental adaptation that enables people to focus on solving analytical problems over a long time (Andrews and Anderson, 2009, doi:10.1037/a0016242). For this reason, a depression helps in the process of writing a good Ph.D. thesis.
6. The existence of 'Big Data' tells us more about the human obsession for collecting data than it tells us about the societal problems that could possibly be solved with it.
7. The assertion that science only takes place in a laboratory is a myopic point of view; the scientific method nowadays also comprises the testing of hypotheses on data that, strictly speaking, have not been collected under controlled conditions, but nevertheless provide useful results.
8. The assumption that flood depth is the most important predictor for building structure damage, which is underlying most damage models for river flooding, is incorrect when applied to flooding from urban drainage systems in flat areas.
9. Anyone with a normal brain can solve the Rubik's Cube within one minute with only a few days of training; but those that are smart, spend their time better.
10. The location of a paint stain on a cycling path contributes to the understanding of possible failure mechanisms related to the transport of a paint container by bike and is also a good predictor for a do-it-yourself store to be present in the vicinity of the paint stain.

These propositions are regarded as opposable and defendable, and have been approved as such by the supervisor prof. dr. ir. F.H.L.R. Clemens.

STELLINGEN

behorende bij het proefschrift

On rainstorm damage to building structure and content

Matthieu Spekkers
Delft, 7 januari 2015

1. De hoogte van de herstelkosten van een door regen getroffen woning kan niet worden verklaard op basis van neerslaggegevens, maar is afhankelijk van andere factoren die met de woning en de woningeigenaar te maken hebben.
2. In Nederland doet schade door binnendringend rioolwater in woningen zich al voor bij buien die minder intensief zijn dan de standaardbuien die de basis vormen voor het ontwerp van rioolstelsels.
3. De neerslagclausule, die sinds het jaar 2000 in de meeste Nederlandse particuliere inboedel- en opstalverzekeringspolissen is opgenomen, is niet ingesteld op de korte intensieve buien die leiden tot overbelasting van rioolstelsels en dient daarom aangepast te worden.
4. Over twintig jaar is het ontwerp van rioolstelsels in Nederland gebaseerd op risicomanagement waarbij het gebruik van schadegegevens van verzekeraars een belangrijke rol gaat spelen.
5. Depressiviteit heeft evolutionaire wortels: het is een mentale aanpassing die de mens in staat stelt zich voor langere tijd extreem goed te kunnen concentreren op het oplossen van analytische problemen (Andrews and Anderson, 2009, doi:10.1037/a0016242). Om die reden helpt een depressie bij het schrijven van een goed proefschrift.
6. Het bestaan van 'Big Data' zegt meer over de menselijke obsessie om maar van alles te willen registeren dan dat het wat zegt over de maatschappelijke problemen die er mogelijk mee opgelost kunnen worden.
7. De bewering dat wetenschap zich slechts in een laboratorium afspeelt, is een kortzichtige opvatting; de wetenschappelijke methode omvat tegenwoordig ook het testen van hypotheses op gegevens die strikt gesproken niet onder gecontroleerde omstandigheden verzameld zijn, maar die desondanks bruikbare resultaten opleveren.
8. De veronderstelling die als uitgangspunt dient voor de meeste schademodellen voor rivieroverstromingen en die stelt dat overstromingsdiepte de belangrijkste voorspeller is voor woonhuisschade, is onjuist als ze toegepast wordt op overstromingen van rioolstelsels in vlakke gebieden.
9. Iedereen met een normaal stel hersenen kan met slechts een paar dagen trainen de Rubiks kubus binnen één minuut oplossen; echter zij die slim zijn besteden hun tijd beter.
10. De locatie van een verfvlek op een fietspad draagt bij aan het begrip van de mogelijke faalmechanismen van het vervoeren van een verfbus op een fiets en is bovendien een goede voorspeller voor de aanwezigheid van een bouwmarkt in de directe omgeving van de verfvlek.

Deze stellingen worden opponeerbaar en verdedigbaar geacht en zijn als zodanig goedgekeurd door de promotor prof. dr. ir. F.H.L.R. Clemens.

ON RAINSTORM DAMAGE TO BUILDING STRUCTURE AND CONTENT

About the flipbook animation in the top-left corner

By quickly flipping the pages, an animation is obtained of a front of rain and thunderstorms that crossed the Netherlands on 26 May 2009, leaving a trail of damage in its wake. The animation is based on weather radar images and runs from the beginning to the end of the book. The black dots are areas with significant rainstorm damage, based on a nationwide home insurance database. Shades of grey indicate the rainfall intensity, with darker shades corresponding to higher rainfall intensities (up to 30 $\mathrm{mm\,h^{-1}}$). The real time between the first frame (this page) and last frame is around five hours. Data sources are discussed in this thesis.

ON RAINSTORM DAMAGE TO BUILDING STRUCTURE AND CONTENT

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen
op woensdag 7 januari 2015 om 15:00 uur

door

Matthieu Hendrik SPEKKERS
civiel ingenieur
geboren te Purmerend.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. F. H. L. R. Clemens

Copromotor:

Dr. ir. J. A. E. ten Veldhuis

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. F. H. L. R. Clemens,	TU Delft, promotor
Dr. ir. J. A. E. ten Veldhuis,	TU Delft, copromotor
Prof. dr. K. Arnbjerg-Nielsen,	Danmarks Tekniske Universitet
Prof. dr. ir. P. Willems,	KU Leuven
Prof. dr. ir. M. Kok,	TU Delft
Prof. dr. ir. P. H. A. J. M. van Gelder,	TU Delft
Dr. H. Kreibich,	Helmholtz Centre Potsdam
Prof. dr. ir. J. B. van Lier,	TU Delft, reservelid

ISBN 978-94-6108-852-9

Printed by Gildeprint – Enschede
Cover designed by Elsbeth Ciesluk
An electronic version of this dissertation is available at http://repository.tudelft.nl/

Contents

CHAPTER 1

Introduction

1.1 Impacts of heavy rainfall in cities

The topic of this thesis is the analysis of damage to building structure and content caused by rainfall. In a broader context, different pathways can be considered that describe how rainfall leads to damage. For instance, damage can be caused by rainfall inducing river flooding (e.g. Jonkman et al., 2008; Merz et al., 2010) or landslides (e.g. Brunetti et al., 2010; Segoni et al., 2014). At the scale of cities, two other damage pathways can be studied. Firstly, that of pluvial flooding, where flooding is caused by stormwater being unable to enter urban drainage systems or flowing out of urban drainage systems when capacity is exceeded (e.g. Ten Veldhuis, 2010). Secondly, that of direct rainwater intrusion due to defects in the building envelope. The damage that results from these two pathways are central in this thesis and is, in the remainder of the thesis, referred to as "rainstorm damage".

A number of severe damage events have demonstrated the serious consequences of rainstorms in cities. On July 2011, for example, Copenhagen was hit by 150 mm of rainfall in three hours, which resulted in surcharging of sewer systems, leakages of roofs, flooding of basements, shops, roads and railways. Danish home insurers received more than 90 000 claims and paid out more than 800 million euros (2011 value) in compensation (Garne et al., 2013). Another example is the heavy rainfall event of autumn of 1998 in the Netherlands that was associated with a return period of about 125 years and caused around 410 million euros (1998 value) of direct damages to households, agriculture and industries in the Netherlands. Damage assessment experts from the Dutch insurance sector identified a total number of 10 660 agricultural companies, 2470 buildings, 1220 other companies and 350 governmental agencies as being damaged by pluvial flooding (Jak and Kok, 2000). Other rainstorm damage events that are well-documented are the summer floods of 2007 across the UK that are believed to be for a great deal related to pluvial flooding (Pitt, 2008; Coulthard

and Frostick, 2010), and the 2004 and 2006 floods in Heywood, Greater Manchester (Douglas et al., 2010).

There is also evidence that minor rainstorms can produce considerable damage in the long run due to their high frequency of occurrence. The Association of British Insurers report for the year 2012 that U.K. insurers paid out 1.5 billion euros to flood claims. Half of it was estimated to be related to pluvial flooding or flooding from small urban streams. Although damage of individual flooding events were small, the annual losses ranked among the highest in the U.K. (Risk Management Solutions, 2013). Similarly, Einfalt et al. (2009) state that many small-scale flood events remain unnoticed, but together constitute for several millions of euros of flood damage per year for Germany. For the case of lowland areas, Ten Veldhuis (2011) estimated that the cumulative damage of 10 years of successive pluvial flood events to residential buildings is of the same order of magnitude as a single event with a return period of 125 years.

Rainstorm damage will likely increase in the future. Over the past 60 years, the frequency and intensity of heavy rainfall events has increased in many parts of the world (Hartmann et al., 2013). It is likely that the frequency and intensity of heavy rainfall events will continue to increase in the next decades as a consequence of climate change (Kirtman et al., 2013). The impacts of climate change on sewer flooding and combined sewer overflow in terms of frequency and volume are uncertain, not only due to uncertainties in climate projections, but also because of uncertainties in hydrological and hydraulic modelling (Willems et al., 2012; Arnbjerg-Nielsen et al., 2013). Another driver for a likely future increase in rainstorm damage is ongoing urbanisation and urban densification. It has led, and probably will continue to lead to an increase in the percentage of impervious areas, which in turn accelerates run-off of rainwater and thus add to the probability of pluvial floods (Ashley et al., 2005). Furthermore, an increase in economic wealth and population can make urban societies more vulnerable to rainstorms.

1.2 The need for damage data and damage models

Many authors, active in research areas related to different kinds of weather-related risks (e.g. hailstorms, landslides, river flooding, coastal flooding), recognize that damage data on natural hazards are generally lacking or incomplete, which is limiting the development of reliable models for damage estimation (e.g. Pielke and Downton, 2000; Hohl et al., 2002; Elmer et al., 2010a; Merz et al., 2013; André et al., 2013). A definition of damage data is data reporting statistics about the adverse consequences of a damage event, collected during or in the aftermath of an event.

On the topic of rainstorms, little research has focused so far on the collection of rainstorm damage data, the understanding of mechanisms causing damage and the deepening of statistical methods to analyse data. Among exceptions are studies by Busch (2008); Smith and Lawson (2012); Einfalt et al. (2012); Cheng (2012); Zhou et al. (2013); Climate Service Center (2013), who analysed damage data sources (i.e. insurance databases, newspaper archives, emergency call data) and their relationships to rainfall data, and Ten Veldhuis (2011), who quantified the cumulative damage of successive pluvial flooding events based on municipal call data related to

urban drainage problems. In most of these studies, however, the spatial and temporal resolutions of rainfall data were insufficient to capture the characteristics of short, high-intensity rainfall events. Moreover, the studies scarcely considered other explanatory variables besides rainfall variables. Because the availability of damage data is generally lacking, there is no strong foundation for the development of prediction models for rainstorm damage.

There are a number of possible explanations for the lack of rainstorm damage data availability. To begin with, the damage of individual rainstorms is usually too small and localised to trigger water authorities, media or homeowners to report damage. Rainstorm damage is generally lower, on an event basis, than damage from other hazard events such as river flooding, and therefore less disruptive for society. Moreover, damage databases, such as those from insurers or national health services, are hard to access because of strict privacy regulations and because they contain company-specific confidential information which may not be shared in public (Lawson and Carter, 2009; Garne et al., 2013; André et al., 2013). Furthermore, damage data may be available, but unpublished, because there is too little contact between researchers and potential data providers.

Damage data and damage models have a high potential of providing valuable information to homeowners, water authorities, insurers and meteorologists to support damage prevention and reduction. Homeowners who consider waterproofing their houses can benefit from information on the efficiency of precautionary measures and the potential damage reduction (Thieken et al., 2005; Gersonius et al., 2008; Poussin et al., 2014). Water authorities responsible for the prevention of pluvial flooding have to decide on flood control measures such as constructing stormwater detention ponds and increasing storage in sewer systems (Hauger et al., 2006). They may benefit from information on locations that historically received much damage to prioritise investments and ensure their effectiveness. Meteorologists and flood forecasting centres can use damage data to develop or validate weather alarms (Hurford et al., 2011, 2012; Falconer et al., 2009) and flash flood guidance (Norbiato et al., 2009). Damage models can help insurers to estimate how much they will spent on compensations over a certain period of time and for a specific hazard portfolio (Bortoluzzo et al., 2011) and, thus, to raise the right amount of capital in the case of severe damage events. Furthermore, damage data can potentially be used to validate flood simulation models by comparing observed and predicted flood depths and locations.

So far, models related to water damage have been mainly developed for river flooding. These damage models, or stage-damage functions, usually consider flood depth and building class as the primary damage-influencing factors (Grigg and Helweg, 1975; Smith, 1994; Merz et al., 2010; Jongman et al., 2012). This approach is likely to be unsatisfactory for pluvial flood damage estimation. In recent years, an increasing number of studies has shown that flood depth alone cannot sufficiently explain damage variability (Merz et al., 2004; Thieken et al., 2005; Pistrika and Jonkman, 2009; Merz et al., 2010; Freni et al., 2010; André et al., 2013) and that many other factors play an important role, such as the level of precaution and socioeconomic status of households (Changnon et al., 2000; Kreibich et al., 2005; Thieken et al., 2005; Merz et al., 2013; Poussin et al., 2014). In particular for pluvial flooding, uncertainties in urban drainage models are not yet well understood to make reliable flood depth calculations

in areas where interactions between streets and sewers are dominant (Deletic et al., 2012). A source of uncertainty relates to incomplete knowledge of failure mechanisms that lead to flooding. For example, blockages of sewer pipes and inlets contribute considerably to pluvial flooding (Ten Veldhuis et al., 2011), but this process is usually ignored in urban drainage models. Moreover, existing damage models are calibrated based on damage data from flood events that involve a range of flood depths up to several meters, but insufficiently describe damage associated with flood depths of several decimetres. For pluvial flooding, some authors attempted to assess damage for case studies using a simple threshold method, where a unit cost price is allocated to an object when flood depth has exceeded a critical, object-specific threshold (Zhou et al., 2012; Sušnik et al., 2014). This method has not been compared with real damage data from pluvial flooding, so its reliability is unknown.

Alternatively, damage models can be developed based on statistical relationships between damage and explanatory data. For instance, Merz et al. (2013) applied a decision-tree model to a damage database related to building structure damage after major river floods in Germany. Through this approach, they were able to identify variables, beyond flood depth, with strong explanatory value. The same technique was used by Lozano et al. (2008) to explore relationships between fire occurrence and environmental factors. Castañeda Vera et al. (2014) applied logistic regression to model the occurrence of rainstorm damage to tomato crops as a function of meteorological, topography and management variables. Such an approach is in fact being used to support weather index-based insurance in agriculture, where insurance payouts are based on measurements from weather stations that strongly correlate with crop damage, rather than actual damage experienced by the policyholders (Barnett and Mahul, 2007; Dick et al., 2011). Other examples of statistical models derived from damage data include models for hailstorm damage (Hohl et al., 2002; Botzen et al., 2010) and storm damage (Dorland et al., 1999). While some research has been carried out on statistically modelling in other natural hazard sciences, there have been only a few investigations into the modelling of rainstorm damage to building structure and content. This justifies the collection of damage data and to development of damage models for rainstorms.

1.3 The potential of mining insurance damage data

There are a number of sources for damage data that can potentially be used for the analysis of rainstorm damage. A non-exhaustive list of damage data sources and their key features is given in Table 1.1. Damage data sources have clearly different natures; they are collected by different stakeholders, in different ways and for multiple purposes (Elmer et al., 2010b). Dedicated data processing and analysis techniques are therefore needed to enable combined use of these data sources.

In this thesis, insurance databases are analysed. Insurance databases often contain many claim records that have been collected continuously in time. Disadvantages are the restricted access and the limited recordings of process information, such as flood depth and extent measurements, details on damage causes, and building and socioeconomic information (Elmer et al., 2010b; Thieken, 2011; Zhou et al., 2013). Moreover, insurance damage data may be subject to a number of biases that can lead

Table 1.1: Damage data sources and their key features.

Source	Key features	References
Interview surveys taken in the aftermath of a damage event	– Collection of process information (e.g. flood depth, duration, damage cause) – Standardized data collection method – Object-scale information (e.g. level of precaution, building-related and socioeconomic variables) – Specific to case studies – Time-consuming, costly	Thieken et al. (2005); Elmer et al. (2010b)
Newspapers archives	– Archives can go far back in time – Contain damage information about objects and infrastructures usually not reported in call or claim data, such as closure of shops, blocked roads and tunnels – Biased by interpretation of reporter – Only newsworthy events are reported – Mostly qualitative information – Sensitive to temporal biases (e.g. changes in reporter team, changes in identity of newspaper)	Smith and Lawson (2012); Lawson and Carter (2009); Septer and Schwab (1995)
Emergency call data from local and regional authorities (e.g. police and fire brigade records, municipal call databases)	– Many records – Calls are usually recorded during or shortly after an events, which limits data distortion – Covering primarily localised, small damage events – Subject to interpretation and classification biases – Information on damage causes and flood characteristics may be incomplete or missing	Ten Veldhuis et al. (2011); Visser (2014); Rodríguez et al. (2012); Caradot et al. (2011); Busch (2008); Lawson and Carter (2009)
Insurance databases	– Many years of continuously collected records – Large number of policyholders – Quantitative data, restricted to tangible damages – Lack of process information – Lack of object-scale information – Biased because of differences between insurers (i.e. data format, data quality, insurance policy) – Privacy restrictions – Often only aggregated data available for research purposes – Quality standards set by insurer – Possibly biased because of changes in insurance policies over time	Busch (2008); Freni et al. (2010); Cheng (2012); Zhou et al. (2013); André et al. (2013)
National disaster databases, based on assessment reports from damage experts commissioned by a state government	– Only cover rare, catastrophic events – Detailed reports on financial losses per building or district – Process information to some extent available	Wind et al. (1999); Jak and Kok (2000)

to misinterpretations of damage information (Gall et al., 2009): there may be differences in data formats and quality between insurers, differences in insurance policies or the way data are recorded and stored. There is a risk of censoring small claim sizes, because of insured not taking the trouble of making a claim. Furthermore, insurance data only account for tangible damage, such as rainstorm damage to buildings, businesses, vehicles and crops (Changnon et al., 1996; Botzen et al., 2009; Castañeda Vera et al., 2014); intangible damages such as car accidents, traffic delays and health risks (Ten Veldhuis et al., 2010; De Man, 2014) are not included.

A few number of studies have been using insurance data of rainstorm damage to building structure and content. In a study by Zhou et al. (2013), 1000 insurance damage claims related to sewer surcharging for the case of Aarhus, Denmark, showed that claim size was not explained by rainfall-related variables. They did find a significant relationship between daily rainfall volume and hourly rainfall intensity and total damage per day. Based on home insurance data for two heavy rainfall events in Germany, Climate Service Center (2013) analysed the feasibility of using weather radar data to derive relationships between rainfall intensities and high rainstorm damages. They were able to identify a rainfall threshold above which damage starts to occur; however, no strong linear relationships between rainfall intensity and claim frequency could be established. To improve relationships, they recommend to include spatial data, such as information on topography, land use and level of imperviousness. Freni et al. (2010) conducted a damage assessment based on the outcomes of two urban drainage models, a distributed reservoir model and a 1D/1D dual drainage model, in combination with stage-damage functions derived from around 600 insurance damage claims and water depth measurements for a case study in Palermo, Italy. They concluded that the uncertainty in stage-damage functions was higher than the accuracy gained by adopting a detailed hydrodynamic model, which emphasizes the need to develop and validate damage models. For sewer flooding events in four cities in Ontario, Canada, Cheng (2012) studied relationships between a rainfall index and monthly-aggregated insurance damage data related to residential buildings and businesses. They determined critical thresholds of the rainfall index for triggering high numbers of claims. However, the validity of the identified thresholds has not been tested on an independent data set, thus the predictive power of the thresholds remains uncertain. As also stated by the authors, the strength of the relationships was strongly limited by monthly resolution of the rainfall and damage data. In a recent publication, The Center for Neighborhood Technology (2014) analysed pluvial flood and sewer-backup damage data from private insurance companies, disaster assistance programs and an online survey, for the case study of Cook County, Illinois. They found that highest damage amounts were observed in districts with low household incomes. Moreover, results of surveys among affected homeowners suggest that besides the economic costs of flooding, stress and health issues may be important too.

In conclusion, rainstorms can have considerable impacts to urban societies. The lack of rainstorm damage data has hampered the development and validation of damage models. Insurance databases can be considered as a promising means to analyse rainstorm damage data as shown by aforementioned studies. These studies, however, concentrate on only small numbers of rainfall events and case study sites, and are limited by the availability, resolution and quality of weather and insurance data.

Moreover, previous research mainly focuses on rainfall variables as predictor for damage, while many other variables are possibly important. As a result, there is still a poor understanding of the factors contributing to rainstorm damage variability, which is the motivation of this thesis.

1.4 Objective, research questions and outline

The general objective of this thesis is to explain variability in rainstorm damage based on multiparameter statistical analyses of home insurance data and a wide range of explanatory data, including weather, building-related, topographic and socioeconomic data. The following research questions are addressed, with the corresponding chapters denoted in brackets:

1. To what extent can information from insurance damage databases be used for the analysis of rainstorm damage? (Chapter 2, 3, 4 and 5)
2. What are relative contributions of different damage mechanisms to the occurrence of rainstorm damage? (Chapter 5)
3. Can rainfall thresholds be identified that trigger the occurrence of insurance damage claims? (Chapter 2 and 5)
4. To what extent can rainstorm damage be predicted based on weather variables? (Chapter 2, 3, 4 and 5)
5. To what extent can rainstorm damage be predicted based on other contextual variables besides weather variables? (Chapter 4)
6. What are appropriate statistical approaches to model variability in rainstorm damage data? (Chapter 2, 4 and 5)

The research data in this thesis are drawn from two home insurance databases from Dutch insurance industry:

- A nationwide insurance database covering water-related damage claims for the period 1998–2011, based on data from a number of large insurance companies (used in Chapter 2, 3 and 4).
- A detailed, property level insurance database of water-related damage claims, for a case study in Rotterdam, the Netherlands, for the period 2007–2013 (used in Chapter 5).

The overall structure of the thesis takes the form of six chapters, including this introductory chapter and a concluding chapter. Chapters 2–5 of the thesis are based on papers that have been published in peer-reviewed journals or are under review, and a peer-reviewed conference paper. Chapter 2 starts with a description of the nationwide insurance database. A logistic regression model is applied to the damage data with the aim to explain claim probability as a function of rainfall characteristics derived from a national rain gauge network. In Chapter 3, an attempt is made to use weather radars as an alternative source of rainfall data to investigate correlations with damage locations and characteristics. The use of decision-tree models is explored in Chapter 4

to study the effects of weather and other contextual variables on claim probability and size. Chapter 5 describes the property level insurance database in more detail. This chapter is about the failure mechanisms causing rainstorm damage to building structure and content, and the extent to which the occurrence of these damage causes relate to weather variables.

CHAPTER 2

Predicting claim probability based on rain gauge measurements

Summary. In this chapter, a nationwide insurance database of water-related damage claims related to building structure and content damage was analysed, for the Netherlands. The aim was to investigate whether the probability of occurrence of rainstorm damage is associated with the intensity of rainfall. Rainfall data were used for the period 2003–2009 based on a network of 33 automatic rain gauges operated by the Royal Netherlands Meteorological Institute. Insurance data were selected within a range of 10 km from rain gauges. Through a logistic regression model, the claim probability was linked to maximum rainfall intensity, with rainfall intensity based on 10-min to 8-h time windows. Rainfall intensity proved to be a significant damage predictor; however, the explained variance, approximated by a pseudo-R^2 statistic, was at most 34 % for building structure damage and at most 30 % for building content damage. When directly comparing predicted and observed values, the model was able to predict 5–17 % more cases correctly compared to a random prediction. No important differences were found between relationships with building structure and building content damage data.

2.1 Introduction

In the autumn of 1998 extreme rainfall caused around 410 million euros (1998 value) of direct damages to households, agriculture and industries in the Netherlands. Damage experts from the Dutch insurance sector identified a total number of 10 660 agricultural companies, 2470 buildings, 1220 other companies and 350 governmental agencies as being damaged by rainwater (Jak and Kok, 2000). The rainfall event with an associated return period of about 125 years resulted in flooding of areas before rainwater was able to enter natural or engineered drainage systems. Other severe events that are well documented are the summer floods of 2007 across the UK, for example in

This chapter is based on: Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2013b). A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences*, 17(3):913–922, doi:10.5194/hess-17-913-2013.

the City of Hull, that are believed to be for a great deal related to pluvial flooding (Pitt, 2008; Coulthard and Frostick, 2010), and the 2004 and 2006 floods in Heywood, Greater Manchester (Douglas et al., 2010). These events are just a few of the many examples that illustrate the serious consequences of high-intensity rainfall. But also minor events with relatively small flood volumes and extensions can produce considerable damage in the long run due to their high frequency of occurrence (Freni et al., 2010; Ten Veldhuis, 2011). The aforementioned events have demonstrated that pluvial floods often occur at much smaller ranges of spatial and temporal scales than fluvial and coastal floods.

An increasing number of authors have acknowledged that a lack of data availability and quality have been important limitations in quantitative flood damage estimations (e.g. Freni et al., 2010; Merz et al., 2004; Hurford et al., 2011). In the absence of damage data, a common approach in flood damage estimation is to combine simulated flood depths and/or flow velocities and stage-damage curves (e.g. Ernst et al., 2008; Jonkman et al., 2008; Pistrika and Jonkman, 2009; De Moel and Aerts, 2010; Middelmann-Fernandes, 2010). The stage-damage curves are usually related to direct damages occurring in large catchments and are derived through synthetic and/or empirical approaches. Only few studies have focused on modelling damages of pluvial floods related to the malfunctioning of urban drainage systems (e.g. Zhou et al., 2012).

Insurance databases are a promising source for flood damage data. These databases often contain many claim records that have been collected continuously in time. Disadvantages are the restricted access and the limited recordings of process information, such as flood depth and extent measurements, details on damage causes, and building information (Elmer et al., 2010a; Thieken, 2011; Zhou et al., 2013).

A few recent studies have analysed insurance data related to pluvial floods. Freni et al. (2010) conducted a damage assessment based on the outcomes of a simple and a detailed hydrodynamic model in combination with stage-damage functions derived from around 600 insurance damage claims and water depth measurements for a case study in Palermo, Italy. They concluded that uncertainty in stage-damage function (40–50 % of average value) was higher than the accuracy gained by adopting a detailed hydrodynamic model. In another study, 1000 insurance damage claims related to sewer surcharging for the case of Aarhus, Denmark, showed that costs per claim were not explained by rainfall (Zhou et al., 2013). They did find a significant relationship between rainfall and total costs per day. These studies confirmed the need to obtain accurate damage data to further investigate costs of pluvial floods.

In this chapter, data from an insurance database containing 20 years of water-related claims for private properties and contents in the Netherlands, provided by the Dutch Association of Insurers, were analysed. The analysis built on earlier work by the Dutch Association of Insurers, where relationships between rainfall and claim data were studied at a regional scale (Ririassa and Hoen, 2010). Using simple linear regression, they found significant relationships between the total amount of damage in a province (roughly 2500–3500 km^2 in size) and hourly rainfall data (one or two rain gauges per province), but the explained variance was low (4 % for building content and 12 % for building structure). It can be argued that, given the size of a province and the limited number of rain gauges used, the model does not account for variations

in damage caused by local rainfall, whilst local convective rainfall is probably an important contributor to damage. The aim of this chapter was to investigate whether high numbers of damage claims are associated with high rainfall intensities, considering rainfall at scales most closely related to functioning of urban drainage systems. In an exploratory study, various damage statistics were correlated with rainfall intensity and the strongest correlation was found between rainfall intensity and the number of damage claims. Rainfall intensity was selected to characterise rainfall events as it was hypothesized to be the most critical rainfall characteristic in relation to damage generating mechanisms such as overloading of sewer systems. Separate relationships were analysed between rainfall data and building structure (i.e. property) damage data as well as building content damage data, through statistical analysis. A better understanding of relationships between rainfall extremes and floods is useful in the development of, for example, warning systems for pluvial floods (Hurford et al., 2012; Parker et al., 2011; Priest et al., 2011).

The chapter is structured as follows. In Sect. 2.2 data sources as well as the statistical model to link rainfall and insurance damage data are described. Results of the statistical analysis are discussed in Sect. 2.3, as well as the significance of predictor variables and the model performance, followed by a discussion in Sect. 2.4. Conclusions and recommendations are summarised in Sect. 2.5.

2.2 Methods

2.2.1 Rainfall data

Rainfall data are based on two networks of rain gauges held by the Royal Netherlands Meteorological Institute (KNMI): a network of 300+ manual rain gauges (see Fig. 2.1, triangular markers) and a network of 33 automatic rain gauges (solid circles). The temporal resolution of the automatic network is 10 min, and the spatial density is about 1 station every 1000 km^2 (see also Table 2.1), with most of the rain gauges located in rural areas or close to city boundaries. The manual network measures daily volumes based on 08:00 UTC–08:00 UTC intervals. The spatial density of the manual network is about 1 station every 100 km^2. All gauge data have been extensively validated by KNMI using well-documented methods (KNMI, 2000).

2.2.2 Insurance data

The insurance databases cover water-related damages to private properties and building content in the Netherlands and are summarised in Table 2.1. Data related to

Table 2.1: Summary of rainfall and insurance data sources.

Data source	Temporal resolution	Spatial resolution	Availability	Records
Manual rain gauge network	daily volumes	$\approx 1/100\,km^2$	1950–today	
Automatic rain gauge network	10-min volumes	$\approx 1/1000\,km^2$	2003–today	
Building structure damage database	by day	district level	1986–2009	$\approx 300\,000$
Building content damage database	by day	district level	1992–2009	$\approx 270\,000$

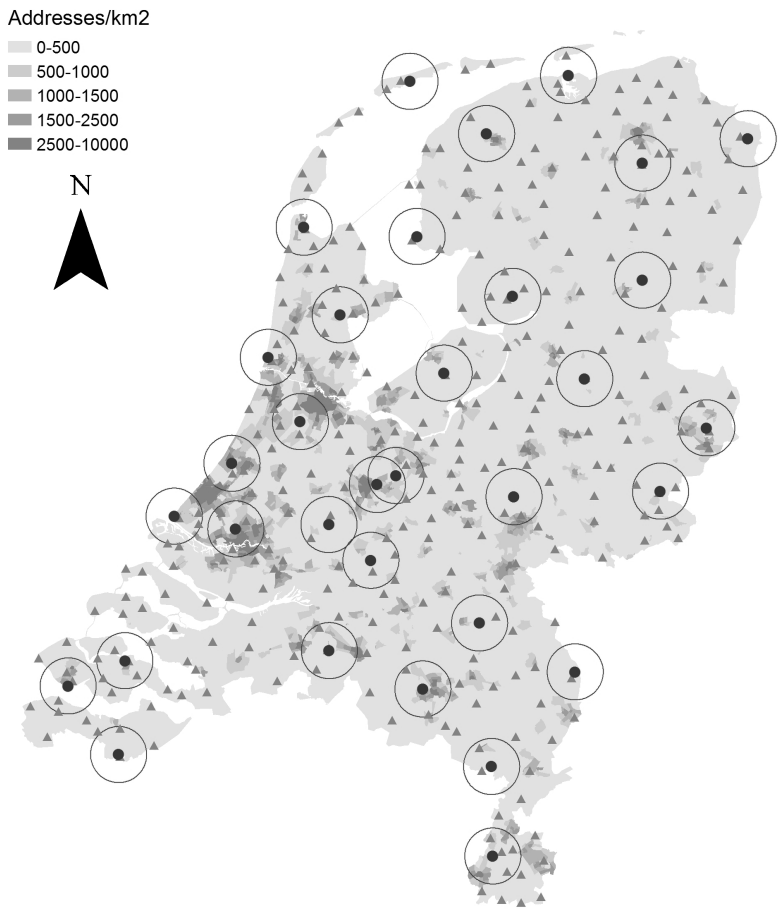


Figure 2.1: Locations of 33 automatic rain gauges (solid circles) and 300+ manual rain gauges (triangular markers) and the area within a 10-km radius of automatic rain gauges (open circles). Urban density (addresses/km^2) is presented in grey scales.

building structure and content damage are available from 1986 until 2009 and from 1992 until 2009 respectively. The database consists of data from a number of large insurance companies in the Netherlands, covering about 20–30 % of the Dutch market. The average number of insurance policies in the database is approximately 1 million per year for building structure and 2 million per year for building content. Homeowners can insure both building structure and content; tenants can only insure building content, while the rented building is considered a commercial building. Commercial buildings are covered in a separate database that is not used in this study.

Water-related damages can be divided into two groups: (1) non-rainfall-related damages and (2) rainfall-related damages. Examples in the first group are bursts of water supply pipes and leakages of washing machines. Examples in the second group are leakages of roofs and flooding from urban drainage systems or local watercourses. This distinction is not explicitly made in the data provided by insurance companies. Insurance companies use different systems to classify claims, and the quality with which claims are assigned to groups varies between companies.

Damage due to pluvial flooding is included in most of the insurance policies after 2000 following advice issued by the Dutch Association of Insurers (Ministry of Transport Public Works and Water Management, 2003). Damage due to pluvial floods should be directly and solely related to local extreme rainfall for a claim to be accepted. Flooding from rivers, sea or groundwater is not commonly insured in the Netherlands, and therefore if pluvial flooding coincides with other flood types, the damage is not insured. Rainfall is considered "extreme" when "rainfall intensity is higher than 40 mm in 24 h, 53 mm in 48 h or 67 mm in 72 h at or near the location of the damaged property", without "near" being precisely defined. The intensities are associated with occurrence frequencies of once every 3 to 7 years in the Netherlands. It is unclear how and to what extent fulfilment of this requirement is examined by the insurance companies. Upon further inquiry, companies have indicated that detailed rainfall data to examine individual cases of local rainfall are usually lacking.

The insurance database consists of four sub-databases: (1) a damage claim database with records related to building structure; (2) a damage claim database with records related to building content; (3) a database with policy holder information related to building structure (i.e. property) insurances; and (4) a database with policy holder information related to building content insurances. The databases with policy holder information related to building content and building structure are separate databases, and it is impossible to link them. Therefore, building structure and content claims cannot be related to a single household. The variables that are included in the database are listed in Table 2.2. The address of the insured household is available at 4-digits postal district (i.e. neighbourhood) level. Typical surface areas of districts are 1–5 km^2 for urban areas and 10–50 km^2 for rural areas. Recorded damages include the costs of cleaning, drying and replacing materials and objects and the costs of temporarily rehousing of people. For the analysis in this chapter, it is assumed that the number of insurance policies is constant during one year. In case an insurance policy is only active for a part of the year, the insurance policy is counted proportionally for that year. Duplicate records were removed, as well as records with missing or incorrect date, location or damage value (around 6 % of the original database). Records with damage value equal to zero were also removed (around 1 % of the records), as

Table 2.2: A brief overview of variables recorded in insurance databases held by the Dutch Association of Insurers. The damage claim records can be linked to the policy holder information through the policy ID key.

Damage claim records	Policy holder information
Damage value claimed	Type of building
Damage value paid out	Policy coverage
Date damage occurred	Start date of policy
Damage cause	End date of policy
Policy ID key	Insured sum of property
	Insured sum of content
	4-digits postal district code
	Policy ID key

these are damage claims that did not meet the policy conditions. First and last day of the month were excluded as they, in a few cases, showed unrealistically high claim numbers compared to other days. These days are probably due to software defaults when exact damage date was unknown or not entered by the insurer's employee.

2.2.3 Aggregating rainfall and insurance data

In this chapter data from April 2003 to 2009 is considered. Insurance damage data were selected within a 10-km radius from the automatic rain gauges based on the distance between the district's centroid and its nearest automatic rain gauge (version shapefile of districts: March 2011). It is assumed that rainfall measured at the rain gauges is uniformly distributed in the rain gauge area. Rain gauge data are generally assumed to be representative within a range of several kilometers. Several ranges were tested and a 10-km range proved to be the best compromise between distance from rain gauges and number of data covered. In Overeem et al. (2011) it is expected that the decorrelation distance for Dutch rainfall events is larger than 15 km. They refer to a study by Berne et al. (2004) where a decorrelation distance of 15 km was found for typical intense Mediterranean rain events, which are on average more intense and more convective compared to rainfall events in the Netherlands. This justifies selecting the claims within 10 km from a rain gauge. Figure 2.2 shows two rain gauges and their neighbouring districts. Insurance data were converted to count data: the number of water-related claims k_i and number of insured households K_i were aggregated by day and by rain gauge area. The subscript i denotes the index of the observation. The number of insured households per rain gauge area ranges from around 300 to 55 000 for property insurance and from around 300 to 120 000 for building content insurance. The higher number of building content insurances is explained by the fact that property insurance only concerns homeowners, whereas building content insurance concerns both homeowners and tenants. Observations with less than 5000 households were filtered out as they were found to be very sensitive to errors in data. The maximum rainfall intensity $I_{i,z}$ is determined for each day and rain gauge area, where subscript z denotes the length in minutes of the moving time window, for z values 10 (original data), 20, 30, 40, 50, 60, 70, 80, 90, 120, 180, 240 or 480 min.

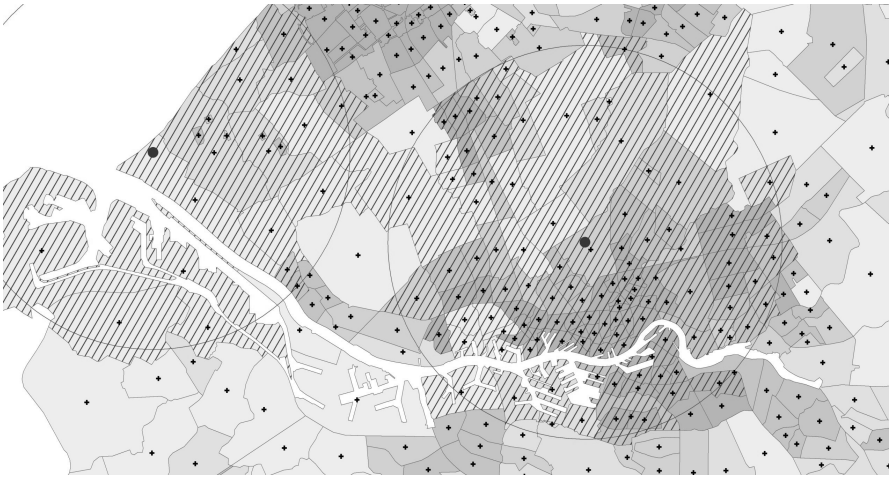


Figure 2.2: Example to illustrate the subsetting of insurance data. The two solid dots are rain gauges and the open circles the rain gauge areas. The crosses are the centroids of the districts. The shaded areas are the districts that have been subsetted.

2.2.4 Distinguishing rainfall-related and non-rainfall-related events

The distinction between non-rainfall-related and rainfall-related claims is not explicitly made in the data provided by insurance companies. Non-rainfall-related claims occur throughout the year, whereas rainfall-related claims are clustered on wet days. Consequently, a high number of claims in a rain gauge region on a particular day is more likely to be associated with rainfall. In the remainder of this chapter, these observations are labelled as "damage events".

The number of claims that can be expected on dry days was estimated based on claims recorded on dry days in 10-km ranges from the network of 300+ manual rain gauges, in order to obtain an independent estimate of the data associated with gauges in the automatic network. Observations were only selected in case of two subsequent dry days, because the daily volumes recorded by manual gauges are based on 08:00 UTC–08:00 UTC intervals. It was found that the number of non-rainfall-related claims is well described as a binomially distributed random variable:

$$k_i \sim B(K_i, \zeta), \tag{2.1}$$

where K_i is the number of insured households and ζ the probability that an individual, insured household will have a non-rainfall-related claim on a day. It is assumed that ζ is constant in both time and space. Best fits with data were found for $\zeta = 3.2 \times 10^{-5}$ (building structure data) and $\zeta = 1.3 \times 10^{-5}$ (building content data). The probability of obtaining y claims at least as extreme as k_i, the one observed, given the number of insured households K_i (i.e. p value) is therefore

$$\Pr(y \geq k_i \mid K_i) = 1 - \sum_{y=0}^{k_i - 1} \binom{K_i}{y} \zeta^y (1-\zeta)^{K_i - y}. \tag{2.2}$$

Any p value below a significance level α indicates occurrence of a damage event, as it is unlikely to be associated with non-rainfall-related claims. Different levels of significance ($\alpha = 1 \times 10^{-2}$, 1×10^{-3}, 1×10^{-4} and 1×10^{-5}) are used to study its effect on the results. A binary variable Y_i is introduced to classify the observations that are considered a damage event $Y_i = 1$ and those that are not $Y_i = 0$:

$$Y_i = \begin{cases} 1 & \text{if } p \text{ value} < \alpha \\ 0 & \text{if } p \text{ value} \geq \alpha. \end{cases} \tag{2.3}$$

2.2.5 Linking binary outcome to maximum rainfall intensity

The outcome, damage event or not, can be linked to the maximum rainfall intensity (maximum within one day for the chosen time window z) using various types of models for binary data (McCullagh and Nelder, 1989). In this study a logistic function was used, which yields

$$\text{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 I_{z,i}, \tag{2.4}$$

where θ_i is the probability of a damage event ($Y_i = 1$) and β_0 and β_1 are regression coefficients. The regression coefficients are estimated using maximum likelihood estimation. The likelihood ratio (LR) test is used to test if β_1 is significantly different from zero, i.e. if maximum rainfall intensity is a parameter that contributes to high numbers of damage claims. There is no universally accepted goodness-of-fit measure in logistic regression that represents the proportion of variance explained by the predictors, such as R^2 in ordinary least squares regression. Several pseudo-R^2 statistics have been developed that mimic the R^2 in evaluating the variability explained, which is one of the approaches used in this chapter. In this chapter McFadden's R^2 is used, which compares the log-likelihood of the model without predictor and log-likelihood of the model with predictor (Long, 1997, p. 104). The other approach directly compares observed and predicted values from the fitted model using contingency tables, using a cutoff point of $\theta = 0.5$.

2.3 Results

2.3.1 Logistic regression results

In Table 2.3 the results of the logistic regression are summarised. Results are based on the 60-min rainfall intensity. The significance levels α, used for the dichotomization of damage data, range from 1×10^{-2} to 1×10^{-5}. Table 2.3 lists estimates for slope coefficient β_1, since this is the most important parameter for interpretation of logistic regression results. The standard error in β_1 is denoted as SE. The slope coefficient is expressed in exponential form, $\exp(\beta_1)$, which is the odds ratio. The odd ratio should be interpreted as the factor with which the odds (probability of a damage event divided by probability of no damage) change as an effect one unit change in the maximum rainfall intensity. For a large number of observations, LR $\sim \chi^2$ with degrees of freedom equal to the number of parameters being estimated.

Table 2.3: Logistic regression results for model fits on building structure and content data. The results are based on $z = 60$ min and a range of α levels. The regression coefficient β_1 has units in $\mathrm{h\,mm^{-1}}$.

data	α	β_1	SE	LR	d.f.	p	$\exp(\beta_1)$	95 % C.I. $\exp(\beta_1)$ Lower	Upper
building structure	0.01	0.265	0.0093	766	1	<0.001	1.30	1.28	1.33
	0.001	0.309	0.0113	723	1	<0.001	1.36	1.33	1.39
	0.0001	0.319	0.0126	626	1	<0.001	1.38	1.34	1.41
	0.00001	0.325	0.0141	528	1	<0.001	1.38	1.35	1.42
building content	0.01	0.248	0.0081	882	1	<0.001	1.28	1.26	1.30
	0.001	0.281	0.0097	782	1	<0.001	1.32	1.30	1.35
	0.0001	0.276	0.0107	597	1	<0.001	1.32	1.29	1.35
	0.00001	0.282	0.0118	516	1	<0.001	1.33	1.30	1.36

The slope coefficient is significantly different from zero in all cases (at $p < 0.05$ level), which means the maximum rainfall intensity is a significant predictor for the probability of occurrence of rainstorm damage. The odd ratios ($\exp(\beta_1)$) vary between 1.28–1.35 for building structure damage and 1.26–1.30 for building content damage, indicating a 28–35 % (building structure) and 26–30 % (building content) increase in odds of a damage event for each $\mathrm{mm\,h^{-1}}$ change in rainfall intensity. Different time windows ranging from 10 min to 8 h have been investigated and produce similar results.

In Fig. 2.3 four examples of logistic functions are plotted as well as the data on which models were fitted. The plots are related to cases of building structure damage (with the dichotomization based on $\alpha = 1 \times 10^{-3}$) and 10-, 20-, 30- and 90-min rainfall intensities. The function links the probability of a damage event θ on the y-axis to maximum rainfall intensity I_z on the x-axis. The steepness of the slope of the logistic function is determined by β_1 (see also Table 2.3); a large slope coefficient makes the transition between "low damage" and "damage event" more abrupt. The grey dots are the observations, either $Y = 0$ in case of "low damage" or $Y = 1$ in case of a "damage event". A jitter function was applied to better visualize the density of the data points. The open circles are the calculated empirical proportions (number of observed $Y = 1$ in a bin divided by total number of observations in a bin n) for eight non-overlapping equally sized bins. The error bars represent one standard deviation σ of uncertainty on the empirical proportion estimate, where $\sigma = \sqrt{\theta(1-\theta)/n}$.

Most observations without damage ($Y = 0$) are associated with low-intensity rainfall; e.g. 99 % of the observations without damage are below 6.9 mm in 10 min. Few observations of low damage are associated with high-intensity rainfall. The $Y = 1$ observations are distributed over a larger range of rainfall intensities. The differences in the distributions of $Y = 0$ and $Y = 1$ are also reflected in the empirical proportions (open circles), with increasing values for higher rainfall intensities. Due to the low number of observations for high rainfall intensities, large uncertainty ranges occur for values of $\theta > 0.5$.

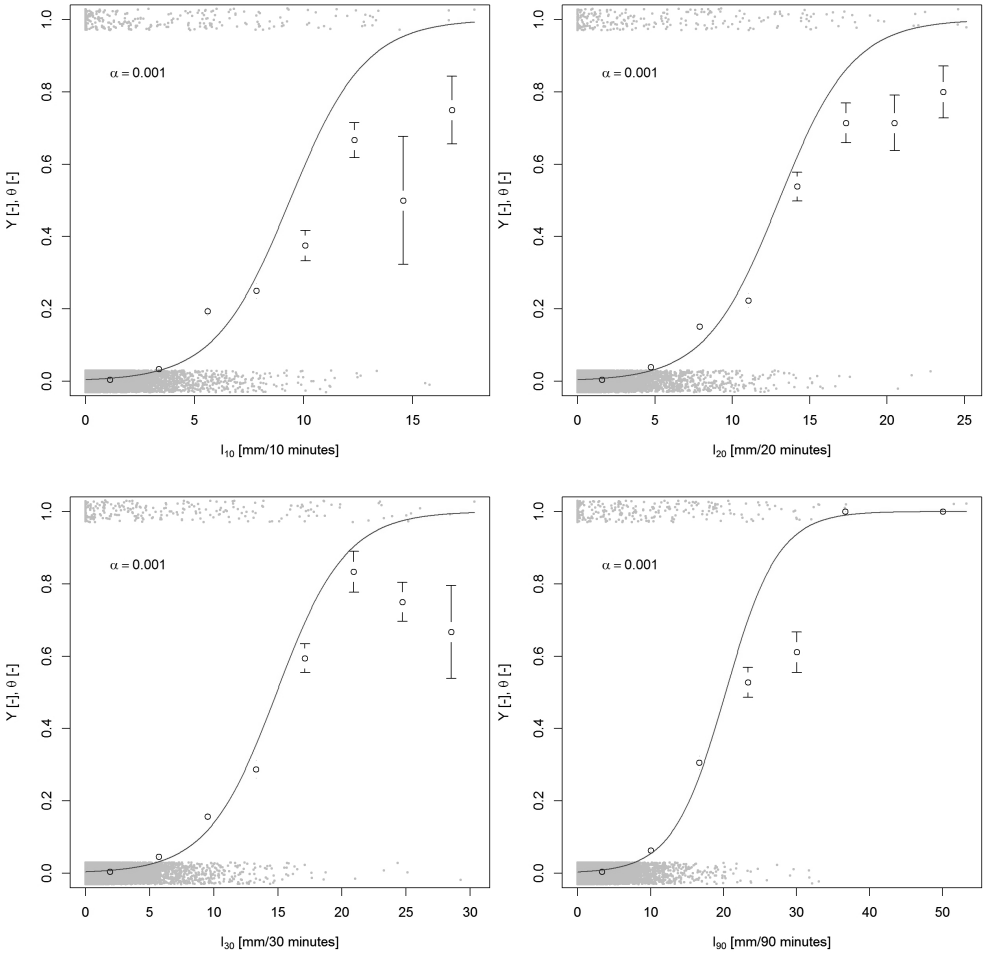


Figure 2.3: Logistic functions (solid lines) fitted on building structure damage data. Plots are related to the cases of $z = 10$, 20, 30 and 90, using $\alpha = 1 \times 10^{-3}$. The small solid dots are the binary observations, either $Y = 0$ or $Y = 1$. A jitter function was applied on the binary observations to better visualize the density of the data points. The open circles are the calculated empirical proportions for eight non-overlapping, equally spaced bins. The error bars represent one standard deviation of uncertainty on the empirical proportion estimate.

2.3.2 Goodness-of-fit using pseudo-R^2

McFadden's R^2 statistic was calculated using different time windows (z) and thresholding criteria (α). Results are listed in Table 2.4. The maximum rainfall intensity accounts for at most 34 % (for building structure damage) and at most 30 % (for building content damage) of the variance explained, taking into account that these values are approximations and depend on the selected pseudo-R^2. There is a slight improvement in the model predictability if rainfall intensity is based on longer time windows, with an "optimum" between 2 and 4 h. The differences are, however, rather small to be conclusive about what time window best predicts damage. An optimum, if true, may reflect the temporal scale at which failure mechanisms (e.g. floodings, leakages of roofs) have caused damage. It would be interesting to have more detailed information on the cause of a damage claim, which would possibly allow characterising temporal scales of different damage generating mechanisms. The results suggest that for this kind of analysis there is no need to collect rainfall data with temporal resolutions smaller than 10 min. Lowering the significance level α, and hence selecting observations that are related to a larger number of claims, improves the predictability by high rainfall intensities. In other words, the results indicate that observations related to a larger number of claims are more likely to be associated with rainfall data than observations related to a smaller number of claims. Building structure damage is better explained by rainfall than building content damage, although the differences are marginal (1–4 % point).

2.3.3 Goodness-of-fit using contingency tables

Another way to look at model performance is to directly compare observed and predicted values using contingency tables. The model is said to have predicted a significant damage event if the estimated θ is greater than or equal to 0.5 and no damage if θ is smaller than 0.5. The rainfall intensity for which the probability of success equals 0.5 is here defined as the rainfall threshold, although it does not necessarily imply a sudden transition from "no damage" to "damage". The rainfall thresholds are listed in Table 2.5 for different α and z. The thresholds are slightly higher for lower significance levels and higher for building content damage compared to building structure damage; however, these differences are small compared to uncertainty introduced by assuming gauge measurement to be representative for the area in a 10-km range of the rain gauge.

Table 2.4: Evaluation of model performance using McFadden's R^2. Outcomes are given for ranges of z and α.

		$z=10$	$z=20$	$z=30$	$z=40$	$z=50$	$z=60$	$z=90$	$z=120$	$z=180$	$z=240$	$z=480$
building structure	$\alpha=0.01$	0.102	0.111	0.114	0.117	0.118	0.120	0.123	0.124	0.126	0.127	0.126
	$\alpha=0.001$	0.186	0.205	0.212	0.215	0.218	0.220	0.224	0.228	0.230	0.227	0.222
	$\alpha=0.0001$	0.234	0.255	0.263	0.268	0.273	0.275	0.277	0.278	0.280	0.275	0.269
	$\alpha=0.00001$	0.280	0.305	0.314	0.323	0.329	0.331	0.335	0.339	0.344	0.340	0.333
building content	$\alpha=0.01$	0.092	0.099	0.103	0.107	0.109	0.110	0.114	0.116	0.118	0.116	0.110
	$\alpha=0.001$	0.167	0.177	0.183	0.189	0.192	0.195	0.202	0.207	0.212	0.210	0.196
	$\alpha=0.0001$	0.190	0.201	0.209	0.217	0.223	0.227	0.237	0.244	0.250	0.248	0.239
	$\alpha=0.00001$	0.232	0.244	0.256	0.266	0.272	0.277	0.285	0.292	0.298	0.294	0.284

Table 2.5: Rainfall thresholds: rainfall intensity in mm h^{-1} for time window z at which probability of a damage event $\theta = 0.5$.

		$z=10$	$z=20$	$z=30$	$z=40$	$z=50$	$z=60$	$z=90$	$z=120$	$z=180$	$z=240$	$z=480$
building structure	$\alpha=0.01$	52.2	36.3	27.8	22.7	19.3	17.0	12.6	10.3	7.8	6.4	4.0
	$\alpha=0.001$	56.2	39.1	29.8	24.4	20.8	18.2	13.5	10.9	8.2	6.8	4.3
	$\alpha=0.0001$	60.1	42.0	32.1	26.2	22.2	19.4	14.5	11.8	8.8	7.3	4.6
	$\alpha=0.00001$	64.5	45.2	34.6	28.2	23.9	20.9	15.6	12.5	9.3	7.7	4.8
building content	$\alpha=0.01$	56.3	39.4	30.1	24.5	20.8	18.2	13.5	10.9	8.2	6.8	4.4
	$\alpha=0.001$	60.8	43.1	33.2	27.0	22.8	20.0	14.7	11.9	8.8	7.2	4.6
	$\alpha=0.0001$	67.8	48.4	37.3	30.3	25.7	22.4	16.5	13.2	9.8	8.0	5.0
	$\alpha=0.00001$	71.6	51.2	39.6	32.2	27.2	23.8	17.6	14.1	10.4	8.6	5.3

Table 2.6: Contingency table, cutoff point $\theta = 0.5$ ($\alpha = 1 \times 10^{-5}$, $z = 60$, building structure data).

	Damage predicted $I_z \geq 20.9$	No damage predicted $I_z < 20.9$	Total
Damage observed	$a = 19$	$b = 101$	120
No damage observed	$c = 13$	$d = 34\,056$	34 069
Total	32	34 157	n = 34 189

In a 2×2 contingency table the observed Y (0 – no damage observed or 1 – damage observed) is compared with the predicted Y (0 – no damage predicted or 1 – damage predicted). Table 2.6 presents the contingency table for $\alpha = 1 \times 10^{-5}$ and $z = 60$ based on building structure damage data. The percentage of correct predictions ($= \frac{a+d}{n} = 0.997$) is heavily skewed in this case due the high number of days without damage. An alternative performance index, less sensitive to skewness of observations, is the sum of fractions of correctly predicted observations ($= \frac{a}{a+b} + \frac{d}{c+d}$) (Kennedy, 2003). Using this approach, scores are presented in Table 2.7 for a range of z and α. The models score around 5–17 % better compared to random predictions. In most cases, building structure damage is better predicted by rainfall than building content damage, although the differences are small and for a few cases scores are equal. The scores do not improve when lowering the significance level from 1×10^{-4} to 1×10^{-5}. The highest scores are obtained for time windows between 30 and 50 min, which are smaller than the 2 to 4 h found using McFadden's R^2.

Table 2.7: Scores using alternative performance index ($= \frac{a}{a+b} + \frac{d}{c+d}$).

		$z=10$	$z=20$	$z=30$	$z=40$	$z=50$	$z=60$	$z=90$	$z=120$	$z=180$	$z=240$	$z=480$
building structure	$\alpha=0.01$	1.05	1.07	1.07	1.07	1.07	1.08	1.07	1.07	1.07	1.07	1.06
	$\alpha=0.001$	1.08	1.13	1.14	1.14	1.14	1.12	1.12	1.11	1.10	1.10	1.10
	$\alpha=0.0001$	1.11	1.16	1.17	1.16	1.16	1.15	1.15	1.14	1.13	1.11	1.12
	$\alpha=0.00001$	1.11	1.15	1.17	1.16	1.16	1.16	1.16	1.16	1.13	1.14	1.12
building content	$\alpha=0.01$	1.04	1.05	1.06	1.06	1.07	1.07	1.06	1.06	1.07	1.06	1.05
	$\alpha=0.001$	1.07	1.09	1.11	1.10	1.10	1.10	1.11	1.11	1.11	1.10	1.08
	$\alpha=0.0001$	1.06	1.08	1.10	1.12	1.12	1.12	1.14	1.12	1.13	1.12	1.10
	$\alpha=0.00001$	1.07	1.07	1.09	1.11	1.13	1.12	1.12	1.14	1.14	1.12	1.12

2.4 Discussion

The contingency tables can be used to address the fractions of type 1 errors and type 2 errors. Type 1 errors (b in Table 2.6) can be indicative of local rainfall that caused damage, while it was not recorded by the local rain gauge due to insufficient spatial density of the rain gauge network. They can also indicate that rainfall intensity does not sufficiently represent the damage generating mechanism and that other exploratory variables such as total rainfall volume, wind speeds or building characteristics need to be added to the model. Type 2 errors (c in Table 2.6) can be related to local rainfall that hit the rain gauge, but not the surrounding urban area. They can also be related to cases of overnight rainfall where people claim the day after. The time window approach used in this study allowed rainfall intensity to be based on rainfall prior to midnight; still rainfall that fell before the start of the time window was not analysed. Both types of errors could be reduced with a higher spatial resolution of rainfall data. Weather radar data are able to provide a better representation of spatial variability, although it is less accurate in determining the intensity than gauge measurements.

The need to reduce type 1 and type 2 errors can be different for different stakeholders. As an example from the water manager's perspective, a decision to open or not to open a water storage facility may lead to unpreparedness in case of a type 1 error or unnecessary costs in case of a type 2 error. A more risk-seeking attitude (accepting some damage) of a potential decision-maker allows a larger cutoff point ($\theta > 0.5$), and a more risk-averse attitude (accepting no damage) allows a smaller cutoff point ($\theta < 0.5$).

A considerable fraction of the variance is left unexplained, which emphasizes the need to study other explanatory variables. There are a few aspects that need to be considered when taking other explanatory factors into account: (1) the explanatory variable should be available and parameterized at the level of 4-digits postal districts, as this is the scale at which insurance data are available; (2) data should be available nationwide if the analysis is performed on the whole insurance database; and (3) since additional data come from different sources, different levels of data quality need to be taken into account. Explanatory factors that are worthwhile to investigate in a future study are topographical properties, urban drainage system properties (e.g. drainage capacity, age of infrastructure, percentage of surface water), level of urbanization, socio-economic indices (e.g. income of households, property value), and district properties (e.g. percentages of low-rise and high-rise buildings, percentage impervious surface).

The results of this study are of practical relevance for insurers, water managers and meteorologists. Some insurers have indicated that the staffing of their call centres (that receive the claims) during extreme events is an issue, and that a better knowledge of what events are likely to cause considerable calls (tens of times more than on a regular day) can be helpful to adjust the capacity of their call centres. It can also be relevant for insurers when reconsidering their policy conditions. The current "rainfall clause" that is being used (see Sect. 2.2.2) has some flaws. For example, the rainfall intensity criteria that are mentioned in this clause are not related to capacities of urban drainage systems. Dutch urban drainage systems are designed to cope with

approximately 20 mm h^{-1}*; the "40 mm in 24 h" criterion, for example, normally should not cause sewer flooding. The results of this study show that short-duration intense rainfall already results in a significant number of claims. Another interesting application is the development or validation of weather alarms, which are usually based on some meteorological thresholds. Climate researchers may use the model to extrapolate probabilities of rainfall damage given some projected change in rainfall extremes.

The extent to which the available insurance data can be used for pluvial flood damage models is limited for two main reasons. Firstly, it is hard to distinguish those claims that are related to pluvial floods from those claims related to other failure mechanisms (e.g. leakages of roofs). Insurers use different definitions for pluvial flooding and different systems to categorize claims. A better and more systematic documentation of claim data could overcome this problem. Secondly, the building addresses are available at the level of 4-digits postal districts (i.e. neighbourhoods), and therefore it is impossible to relate claims to attributes of individual households, such as the level of precaution, basement use and door threshold level. Simplified damage assessment may be possible at the level of neighbourhoods, taking into account district-specific properties.

2.5 Conclusions and recommendations

In this chapter relationships were investigated between water-related damage data provided by insurance companies and rainfall extremes for the period 2003–2009 in the Netherlands. The results show that high claim numbers related to building structure and content damages were significantly related to maximum rainfall intensity, based on a logistic regression, with rainfall intensity for 10-min to 8-h time windows. The variance explained by rainfall intensity, approximated by a pseudo-R^2 statistic, was 34 % for building structure damage and 30 % for building content damage, based on a time window of 3 h. When directly comparing predicted and observed values, the model was able to predict 5–17 % more cases correctly compared to a random prediction. No important differences were found between building structure and content damage data. A considerable fraction of the variance is left unexplained, which emphasizes the need to study damage generating mechanisms and other explanatory variables, such as wind speed or building characteristics. A better documentation of exact damage causes in insurance databases is essential to detail relationships with damages caused by failure mechanisms of urban drainage systems. A limitation of the present study was that rainfall data were insufficiently representative of local rainfall conditions in the vicinity of the claim. Since most claims are located in urban areas, this indicates the need for rainfall data of high spatial resolution at the urban scale.

*In the 70s, sewers were designed to cope with 60 or 90 L s^{-1} ha^{-1} for flat and hilly areas respectively (Koot, 1977). These values correspond to rainfall intensities of 21.6 and 32.4 mm h^{-1}. In the 80s, hydrodynamic calculations in urban drainage became common practice in the Netherlands, which principles where standardized in the 90s (Van Mameren and Clemens, 1997). Hydrodynamic models are being used to test the hydraulic design of sewers based on design storms with usually a return period of 2 years (Van Luijtelaar and Rebergen, 1997; Stichting RIONED, 2004), which is approximately 20 mm h^{-1}.

CHAPTER 3

Spatial analysis of rainstorm damage using weather radar

Summary. The aim of this chapter was to explore the extent to which weather radars can be helpful to predict damage locations and characteristics, as they provide better spatial resolution compared to rain gauge networks. Rainstorm damage data were analysed based on a nationwide home insurance database for the Netherlands. A 14-year (1998–2011) database of corrected C-band radar images by the Royal Netherlands Meteorological Institute was used to extract characteristics of rainfall events. These characteristics were linked to various damage variables at district level. Results are based on a selected data set representing the top 150 days of largest damage amounts nationwide. Rainfall and damage locations show similar spatial patterns when visualized on maps, which was particularly the case for maximum hourly rainfall intensity and rainfall volume. In a quantitative analysis, highest correlation coefficient was found between claim frequency and maximum hourly rainfall intensity, although the relationship is moderate ($r = 0.38$). The average claim size does not show any significant correlation with the rainfall variables, except a weak relationship with maximum hourly rainfall intensity ($r = 0.12$). This implies that more intense rainfall mainly affects the number of households claiming and not so much the amount of damage per individual household.

3.1 Introduction

Intense rainfall may locally cause considerable damage in urban areas, for instance, as a result of flooding from urban drainage systems or rainwater intrusion through defects in the building envelope. It is of interest for many stakeholders to understand the process of how rainfall results in damage. In the case of building structure and content

This chapter is based on: Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2013a). A spatial analysis of rainfall damage data using C-band weather radar images. In Butler, D., Chen, A. S., Djordjevic, S., and Hammond, M. J., editors, *Proceedings of the International Conference on Flood Resilience: Experiences in Asia and Europe*, Exeter, UK. Centre for Water Systems, University of Exeter.

damage, for instance, insurers are interested to know how characteristics of rainfall explain damage claim frequency and size and to what extent these relationships can be extrapolated to estimate damage under climate scenarios. Authorities responsible for the management of sewer flooding may prioritize their investments by knowing the amount and location of historic flood damage. Meteorologists can improve the effectiveness of weather alarms when there is empirical evidence for rainfall thresholds that trigger high damage (Hurford et al., 2012).

In this context, it is useful to have regression models that link rainfall characteristics to damage variables. Such regression models may reveal predominant rainfall characteristics that cause high damage. Damage data used for such analysis may potentially come from insurance companies, as insurance databases usually cover many records that are continuously collected in time; however, strict privacy regulations often limit the amount of data that is available for research. Few studies have examined relationships between rainfall and insurance damage data, mainly rainfall-related or water-related building structure and content damage data (Zhou et al., 2013; Einfalt et al., 2012; Cheng, 2012). General conclusions, however, cannot be drawn from this limited number of studies as the studies varied greatly in terms of temporal and spatial resolution, length and quality of the available damage and rainfall data.

The aim of this chapter is to explore the extent to which weather radar data can be helpful to predict damage locations and characteristics. For this purpose, a nationwide insurance database was provided by the Dutch Associations of Insurers and covers claimed building structure and content damages in the Netherlands for the period 1998–2011. In Chapter 2 it was shown that insufficient representativeness of rainfall data for local conditions, especially in cities, was a possible explanation why only weak relationships between rainfall and damage were found. For this end, a database of corrected C-band radar images by the Royal Netherlands Meteorological Institute (Overeem et al., 2009) was used in this chapter to extract various rainfall characteristics.

Section 3.2 describes the C-band weather radar data and insurance damage data and the variables used for regression analysis. Results of regression analysis are discussed in Section 3.3. Conclusions and recommendation are summarised in Section 3.4.

3.2 Methods

3.2.1 Insurance and weather radar data

The insurance database, provided by the Dutch Association of Insurers, covers water-related damages to building structure and content in the Netherlands and are summarised in Table 3.1. The insurance claims are partly related to rainstorm damages, such as rainwater intrusion through defects in the building envelope and flooding from sewers or watercourses. They are also related to other, non-rainstorm causes, such as bursts of water supply pipes or leakages of washing machines. Records include the costs of cleaning, drying and replacing materials and objects and the costs of temporarily rehousing of people. Daily data are available at the level of 4-digits postal districts, i.e. neighbourhood level. These districts have typical surface areas of 1–5

Table 3.1: Summary of rainfall and insurance data. The availability of radar data is based on the fraction of available 5-minute composites, see Overeem et al. (2009, 2011).

Data source	Temporal resolution	Spatial resolution	Period	Availability
C-band weather radar data	1 scan per 5 min	2.5 km x 2.5 km pixels	1998–2008	83.5 %
	1 scan per 5 min	1 km x 1 km pixels	2009–2011	≈100 %
Building structure damage database	by day	district level	1986–2011	order 10^5
Building content damage database	by day	district level	1992–2011	order 10^5

km^2 for urban areas and 10–50 km^2 for rural areas. The databases have been extensively checked on missing or incorrect values (e.g. blanks, zeros and incorrect dates) and inconsistencies as described in Chapter 2.

A database of adjusted C-band weather radar images were provided by the Royal Netherlands Meteorological Institute (Overeem et al., 2009). Data are available for the entire land surface of the Netherlands with a 5-minute temporal resolution and a 2.5-km (1998–2008) and 1-km (2009–2011) spatial resolution (Table 3.1). The images are composites based on two C-band Doppler radars, which have been corrected for various biases using data from manual and automatic rain gauges (Overeem et al., 2009). The availability of radar data is 83.5 % in the period 1998–2008 and almost 100 % in the period 2009–2011, based on the fraction of available 5-minute composites (Overeem et al., 2009, 2011).

3.2.2 Data selection

Insurance and rainfall data are used for the period 1998–2011. This work discusses results based on a selection of days. The top 120 days with largest damage nationwide were selected and ranked according to their total number of claims per insured household for both insurance databases. Table 3.2 lists the first 10 days, with separate lists for building structure and content damage claims. The dates of both lists together made a list of 150 unique days. Due to missing radar images, in particular for the 2.5-km radar images (see Table 3.1), 16 out of 150 days (11 %) were discarded from the analyses. Furthermore, first days of the month were excluded, because it is sometimes used by insurers as a default date when claim date was unknown or not entered correctly by the insurers' employee. Another eight days were removed because on these days (almost) no rainfall was observed, but nonetheless showing considerable claim numbers. Although not confirmed with precipitation data, claims on these days may be related to snowfall as most of the days happen to be in December or January. The 150 unique days cover 16 % of the total number of claims in the databases.

3.2.3 Damage variables

For each day and district (4019 districts for the year 2011), the following damage statistics are available: number of claims, number of insured households and total amount of damage. From these, claim frequency, normalized total damage and average claim size are calculated; see Table 3.3 for definitions. Damage values before 2002 were converted from guilder to euro using the conversion ratio 1 guilder = 0.454 euro. All values are in 2011 euros. Every value associated with a year before 2011 was

Table 3.2: Top 10 days with largest water-related damage nationwide in the period 1998–2011. Days are sorted by the normalized number of claims per household. Days with missing radar data are listed in the table, but are excluded from the analyses.

Insurance type	Rank	Date	Normalized number of claims per household [-]	Normalized amount of damage per household [-]	Radar images available	Plotted in Fig. 3.1
Building structure	1	06-06-1998	1.00	1.00	no	no
	2	05-07-1999	0.69	0.90	yes	no
	3	14-09-1998	0.61	0.67	yes	no
	4	18-01-2007	0.54	0.51	yes	no
	5	26-08-2010	0.49	0.73	yes	yes
	6	30-06-2003	0.49	0.87	yes	no
	7	25-11-2005	0.46	0.61	yes	no
	8	14-07-2011	0.45	0.46	yes	yes
	9	26-05-2009	0.37	0.50	yes	no
	10	22-06-2008	0.34	0.50	yes	no
Building content	1	06-06-1998	1.00	0.89	no	no
	2	14-09-1998	0.95	0.95	yes	no
	3	05-07-1999	0.83	1.00	yes	no
	4	10-07-2010	0.76	0.64	yes	no
	5	28-06-2011	0.73	0.63	yes	no
	6	14-07-2010	0.70	0.46	yes	no
	7	26-08-2010	0.67	0.97	yes	yes
	8	26-05-2009	0.65	0.67	yes	no
	9	12-07-2010	0.64	0.48	yes	no
	10	22-06-2008	0.49	0.26	yes	no

adjusted for inflation according to the correction indices in Table 3.4.

The distinction between rainfall-related and non-rainfall-related claims is not explicitly made in the data. Non-rainfall-related claims occur throughout the year, whereas rainfall-related claims are clustered on wet days. Consequently, a high number of claims in a district is more likely to be associated with rainfall. In Chapter 2 a method is proposed to label districts that show high numbers of claims on a particular day compared to what is expected on dry days. This method is based on the distribution of the number of claims observed in districts on dry days. Given the statistical properties of this distribution, the probability of any observation to be drawn from the distribution was calculated. Observations with probabilities smaller than a significance level, which was set to 0.001, are likely to be related to rainfall. In the remainder of this chapter, these observations are labelled as "damage events".

Districts with only a small number of insured households are more likely to show extreme values of damage variables and may distort results. Based on visual inspection of calculated confidence intervals of claim frequency, districts with less than 300 insured households were therefore left out. For the selected 150 days, a total number of 2514 damage events were found that met the aforementioned criteria and were included in the analysis.

Table 3.3: Definitions of rainfall and damage variables. Damage-related variables are aggregated by day and district and separately for building content and building structure damage.

Variable abbr.	Description	Unit
Damage-related		
cf	Claim frequency = number of claims in a district per day divided by number of insured households in a district per day	-
dtot	Normalized total damage = total damage in a district per day divided by number of insured households in a district per day	euro d^{-1}
acs	Average claim size = total damage in a district per day divided by number of claims in a district per day	euro d^{-1}
Rainfall-related		
rmax	Maximum hourly rainfall intensity = maximum intensity of a rainfall event based on an 1-hour moving time window	mm h^{-1}
rvol	Rainfall volume (of event)	mm
rdur	Rainfall duration (of event)	h
rmean	Mean rainfall intensity = rainfall volume of event divided by rainfall duration of event	mm h^{-1}
rtime	Time of rainfall peak, ranging from -1 to 1, giving the relative time of the rainfall peak between 00:00 the day before and 24:00 the same day	-
rvolbp	Rainfall volume before rainfall peak	mm

Table 3.4: Inflation adjustment according to the online database of Statistics Netherlands (2012). The average inflation per year for the Netherlands is used (second column), based on the consumer price index. Every damage value associated with a year before 2011 was multiplied with a correction index (third column).

Year	Inflation [%]	Correction
1998	2.0	1.31
1999	2.2	1.28
2000	2.6	1.25
2001	4.5	1.19
2002	3.4	1.16
2003	2.1	1.13
2004	1.2	1.12
2005	1.7	1.10
2006	1.1	1.09
2007	1.6	1.07
2008	2.5	1.04
2009	1.2	1.03
2010	1.3	1.02
2011	2.3	1.00

3.2.4 Rainfall variables

Damage events were linked to various rainfall characteristics listed in Table 3.3. The procedure to link rainfall characteristics from radar images to damage events is as follows. Firstly, rainfall time series are processed on individual pixel level. Rainfall data were abstracted for all damage days and for previous days. Then independent rainfall events were selected based on intermediate dry period of at least 12 hours, with "dry" being defined as < 0.083 mm for a 5-minute time step. Only rainfall events that coincide at least for one time step with the damage day are kept. This results in either zero, one or two independent rainfall events that can be associated with a damage day. In the case of zero events, all rainfall characteristics are assigned zero values, except the time of rainfall peak, which is marked as not available. In the case of two events, the maximum value out of the two events is taken. This way, maps can be plotted with the spatial distribution of rainfall characteristics as is done in Fig. 3.1. Secondly, the radar pixel value at the district's centroid is selected to be representative for the district.

3.2.5 Log-linear model

A linear regression model was applied using log-transformed values of damage variables. Distributions of damage variables encountered in insurance data are typically strongly non-normal (De Jong and Heller, 2008), which is also the case here. In case the distribution is log normal, the values of damage variables are log-transformed to approximate normality and linearity assumptions of a linear model. In this study, a log transformation works out well for average claim size, but in a lesser extent for claim frequency and normalized total damage (not shown here). Nevertheless, small deviations of the distribution from log normal were assumed acceptable.

Policyholders are not subject to a deductible, which, if it was the case, puts a lower limit to the amount of damage policyholders may claim. It is therefore assumed that the distribution of the damage per claim is not left truncated. Some left truncation of the data can be expected as people may choose not to take the trouble of claiming small damages; however, this factor is ignored here. The distribution of the average claim size is assumed not to be censored by the insured sum, as water-related damages are typically much smaller than the insured sum.

3.3 Results and discussion

3.3.1 Spatial patterns of rainfall and damage

To compare rainfall patterns and damage locations spatially, two days were selected for which the spatial variability of maximum hourly rainfall intensity and rainfall volume were plotted on a colour map (Fig. 3.1). The two days were selected from the top 10 days, indicated with "yes" in the last column of Table 3.2. The damage events are marked on the map with red dots (related to building content) and black crosses (related to building structure).

By comparing the rainfall and damage data visually, it can be concluded that rainfall and damage show similar patterns. For example, on 26 August 2010, both

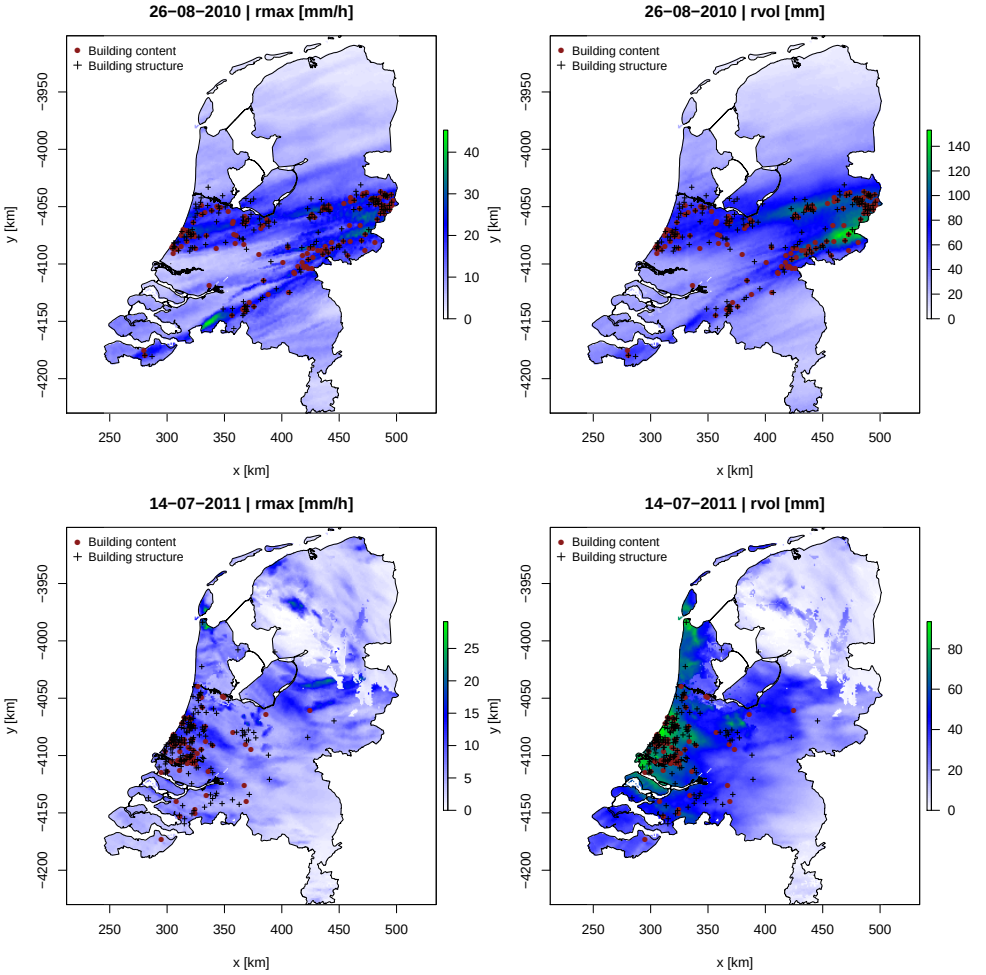


Figure 3.1: Maps with event maximum hourly rainfall intensity (left) and event total rainfall volume (right) for 26 August 2010 and 14 July 2011. These days are selected from top 10 list in Table 3.2. The red dots (related to building content) and black crosses (related to building structure) mark "damage events", i.e. districts with significantly high numbers of claims compared to what is expected on dry days. Significance level is set to 0.001. Note that colour bar legends have different value ranges.

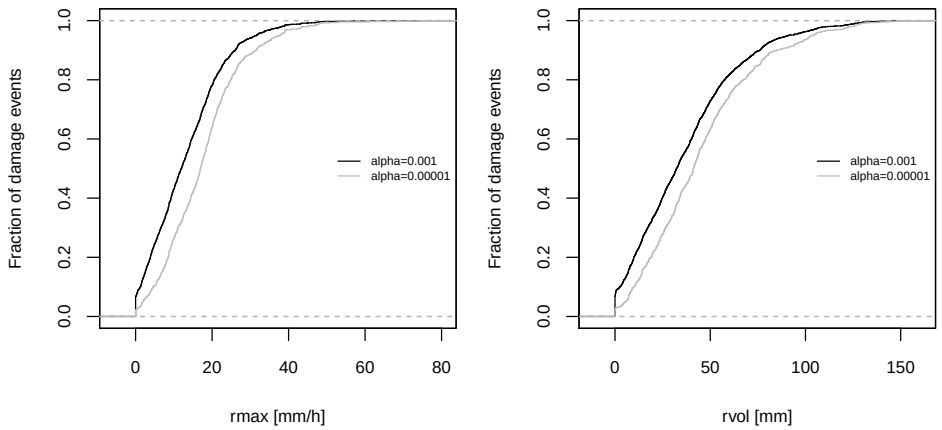


Figure 3.2: Cumulative density functions of maximum hourly rainfall intensity (left) and rainfall volume (right) associated with the occurrence of damage events. The curves represent the fraction of damage events that is below a particular value of a rainfall characteristic. The black line is related to significance level of 0.001 and the grey line shows the effect of setting a stricter significance level ($\alpha = 1 \times 10^{-5}$).

rainfall extremes and damage locations are concentrated in a horizontal band across the centre of the Netherlands, with rainfall intensities of 20 mm h^{-1} or more, whereas in the rest of the Netherlands, with rainfall intensities less than 20 mm h^{-1}, no significant damage was reported. On 14 July 2011, rainfall volumes were highest along the west coast, with rainfall volumes of 70 mm or more, while most of the damage events are clustered in the same region.

In Fig. 3.2, the empirical cumulative density functions are given for the maximum hourly rainfall intensity (left) and rainfall volume (right) associated with the occurrence of damage events. The curves represent the fraction of damage events that is below a particular value of a rainfall characteristic; 6.8 % of the damage events ($\alpha = 0.001$) is associated with no rainfall, which may be caused by errors in the data. Another reason is that the significance level, used to label damage events, was set too loose. If significance level is set to 1×10^{-5}, than 2.1 % of the damage events is unrelated to rainfall. Half of the damage events are observed when rainfall intensity is 12 mm h^{-1} or less and rainfall volume is 32 mm or less ($\alpha = 0.001$). The shape of curve for $\alpha = 0.001$, having a steep slope near the left of the figure, indicates that no clear rainfall threshold exists for occurrence of damage.

3.3.2 Regression analysis

Figure 3.3 shows a correlogram of rainfall-related and damage-related variables. The direction and size of the triangle depicts the sign and magnitude respectively of the Pearson correlation coefficient between two variables. An upward pointing triangle indicates a positive correlation and a downward pointing triangle a negative correl-

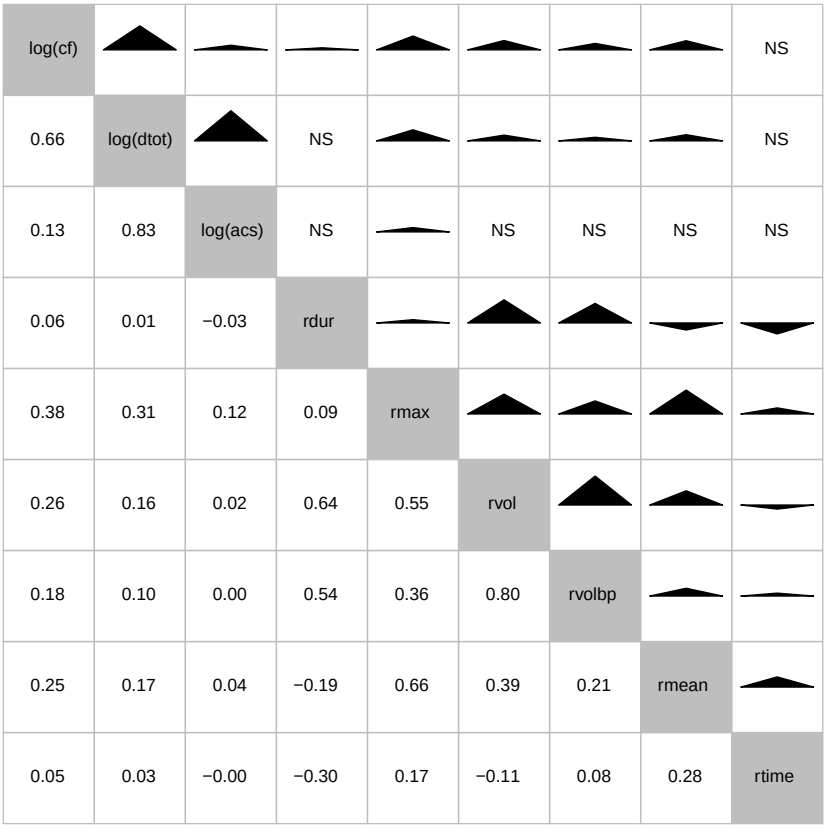


Figure 3.3: Correlogram of correlations among variables. The direction and size of the triangle depicts the sign and magnitude respectively of the Pearson correlation coefficient between two variables. An upward pointing triangle indicates a positive correlation and a downward pointing triangle a negative correlation. Not statistically significant relationships (1 % significance level) are denoted with "NS".

ation. Not statistically significant relationships (1 % significance level) are denoted with "NS".

Highest correlation score is found between maximum hourly rainfall intensity and claim frequency ($r = 0.38$); rainfall volume and mean rainfall intensity are the second and third best predictors for claim frequency ($r = 0.26$ and $r = 0.25$). Slightly lower correlation coefficients were found when normalized total damage was taken as dependent variable. Although these relationships are significant, the strength of correlations is moderate. The average claim size is only significant with respect to maximum hourly rainfall intensity, but the relationship is weak ($r = 0.12$). Time of rainfall peak is insignificant with respect to any of the damage variables.

Scatter plots in Fig. 3.4 of normalized total damage as a function of maximum hourly rainfall intensity (left) and rainfall volumes (right) confirm the moderate relationships, showing large spread of data around the linear fit. Nevertheless, the linear model with log-transformed dependent variable is an appropriate model choice, as the residuals are randomly dispersed around the horizontal axis (lower figures). Similar plots can be made using the log-transformed claim frequency as dependent variable.

To summarise, more intense rainfall mainly results in more households claiming and not so much the amount of damage per individual household. This suggests that variations in the average claim size are probably caused by a large extent to local characteristics, such as properties related to building and household. The results have implications, for instance, for damage modelling. It is suggested to focus on rainfall thresholds based on rainfall intensity and to a lesser extent on rainfall volume or mean rainfall intensity. However, rainfall as single predictor lacks predictive power. Districts may respond differently to similar rainfall events and efforts should be made to collect other contextual variables that describes these district-specific thresholds.

3.4 Conclusions

The aim of this chapter was to investigate the extent to which rainfall characteristics, extracted from C-band radar images, can explain rainstorm claim statistics related to building structure and content damage. In this chapter results were discussed based on data from the 150 days with largest damage amounts in the Netherlands in the period 1998–2011. By comparing damage locations and spatial variability of rainfall visually, it can be concluded that rainfall and locations of reported damages show similar spatial patterns. No clear rainfall thresholds could be identified below which no damage occurs. Using linear regression with log-transformed damage variables, highest correlation coefficient was found between claim frequency and maximum hourly rainfall intensity ($r = 0.38$). Rainfall volume is a slightly less important predictor for damage compared to maximum hourly rainfall intensity. The average claim size does not show any significant correlation with the rainfall variables, except a weak relationship with maximum hourly rainfall intensity ($r = 0.12$). This implies that more intense rainfall mainly effects the number of households claiming and not so much the amount of damage per individual household. A large part of the variance in damage variables is left unexplained. Therefore, in Chapter 4 the inclusion of a larger number of other contextual variables, defined at the district scale, are investigated, such as socio-economic characteristics of households and building properties.

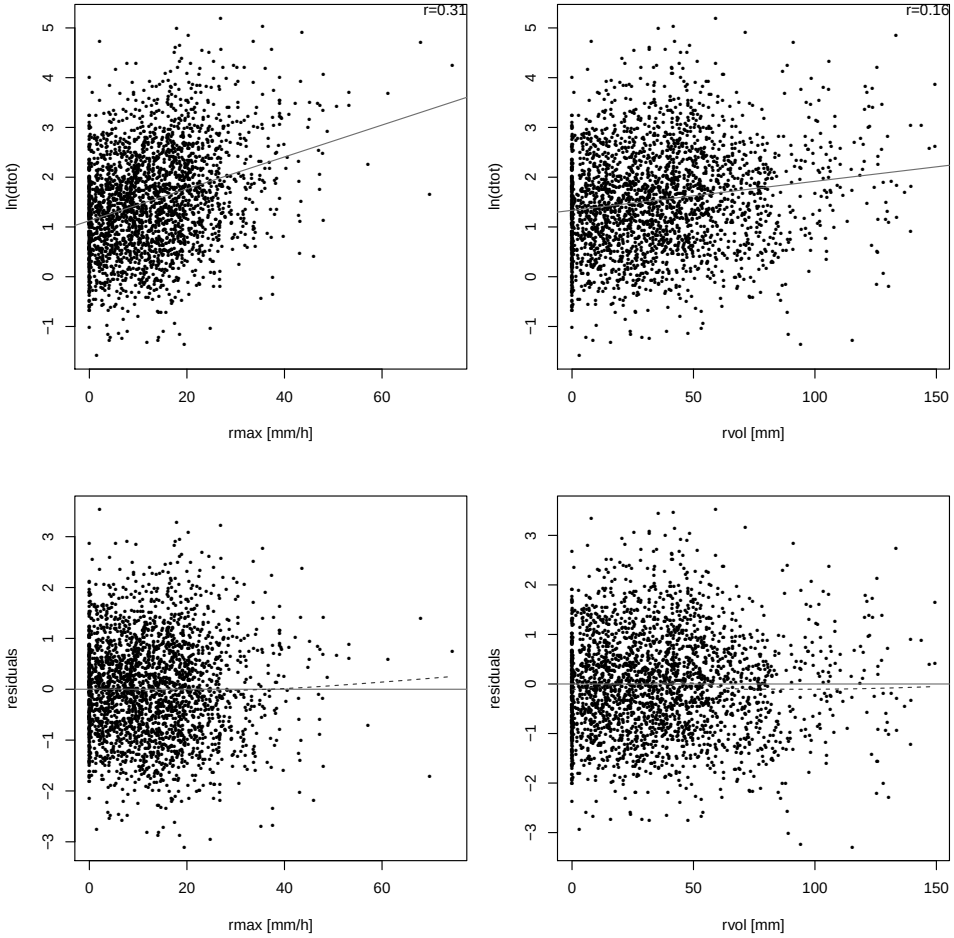


Figure 3.4: Scatter plots of log-transformed normalized total damage (dtot) against maximum hourly rainfall intensity (top left) and rainfall volume (top right). The solid line is the linear regression model. The lower figures show the model residuals. The dashed line is the locally weighted regression curve.

CHAPTER 4

Tree analysis of contextual factors influencing rainstorm damage

Summary. In this chapter, a wide range of damage-influencing factors and their relationships with rainstorm damage was investigated, using decision-tree analysis. For this, district-aggregated claim data from home insurance companies in the Netherlands were analysed, for the period 1998–2011. Response variables being modelled are average claim size and claim frequency, per district, per day. The set of predictors include rainfall-related variables derived from weather radar images, topographic variables from a digital terrain model, building-related variables and socioeconomic indicators of households. Analyses were made separately for property (or building structure) and content damage claim data. Results of decision-tree analysis show that claim frequency is most strongly associated with maximum hourly rainfall intensity, followed by real estate value, ground floor area, household income, season (property data only), buildings age (property data only), fraction of homeowners (content data only), and fraction of low-rise buildings (content data only). It was not possible to develop statistically acceptable trees for average claim size. Cross-validation results show that decision trees were able to predict 22–26 % of variance in claim frequency, which is considerably better compared to the 11–18 % of variance explained by the global multiple-regression models.

4.1 Introduction

A key aspect of flood risk management is the analysis of flood-damage data and the development of flood-damage prediction models. A considerable amount of literature on this topic is associated with catastrophic river floods that involve large catchments

This chapter is based on: Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2014b). Decision-tree analysis of factors influencing rainfall-related building structure and content damage. *Natural Hazards and Earth System Science*, 14(9):2531–2547, doi:10.5194/nhess-14-2531-2014.

(Merz et al., 2010; Jongman et al., 2012). Comparatively little research focused on damage of small-scale floods in urban areas that are a result of localised heavy rainfall (e.g. Ten Veldhuis, 2011; Hurford et al., 2012; Blanc et al., 2012; Zhou et al., 2012). One possible explanation for this is that the adverse consequences at the scale of river catchments are possibly larger than at the urban scales Moreover, information and data on impacts from urban flooding are rare, as well as appropriate methods to analyse these. Meanwhile, reliable damage models for this type of flood can help insurers and water authorities to respond more adequately to rainfall extremes.

Severe pluvial floods in the UK in 2004, 2006 and 2007 (Pitt, 2008; Coulthard and Frostick, 2010; Douglas et al., 2010) have demonstrated that local high-intensity rainfall can have large impacts on society. Another example is the heavy rainfall event of 1998 in the Netherlands that caused around 410 million euros (1998 values) to private buildings and agriculture (Jak and Kok, 2000). Recent figures related to building damage due to heavy rainfallm show that Danish insurance industry has compensated around 300 million euros per year between the years 2009 and 2011 (Garne et al., 2013).

The objective of a damage model is to predict damage that is related to single objects (e.g. buildings) or spatially aggregated units (e.g. postal districts, neighbourhoods), based on a set of explanatory variables. In particular, building damage and the factors contributing to damage has been object of research in many natural hazard sciences, such as building damage due to landslides (e.g. Chiocchio et al., 1997), hailstorms (e.g. Hohl et al., 2002) and coastal flooding (e.g. André et al., 2013). For river flooding, traditional building damage models usually consider flood depth and building class as the primary damage-influencing factors (Merz et al., 2010). In recent years, an increasing number of studies have shown that flood depth alone cannot sufficiently explain damage variability (Merz et al., 2004; Thieken et al., 2005; Pistrika and Jonkman, 2009; Merz et al., 2010; Freni et al., 2010) and that many other factors play an important role, such as the level of precaution and socioeconomic status of households (Kreibich et al., 2005; Thieken et al., 2005; Merz et al., 2013). In particular for pluvial flooding, uncertainties in urban drainage models are not yet understood well enough (Deletic et al., 2012) to make reliable flood depth calculations. A source of uncertainty relates to incomplete knowledge of failure mechanisms that lead to flooding. For example, blockages of sewer inlets contribute largely to pluvial flooding (Ten Veldhuis et al., 2011), but this process is usually ignored in urban drainage models.

Instead, Merz et al. (2013) argue that "there is a need for multi-variate statistical analyses of comprehensive flood damage data to quantify the interaction and influence of various factors and to further develop reliable damage models". They successfully applied tree-based data-mining techniques on a comprehensive damage data set related to building damage after major river floods in Germany. Through this approach, they were able to investigate a large variety of potential damage-influencing characteristics, beyond the ones that are used in traditional flood-damage models, and identify parameters with strong explanatory value, such as floor area, building value, flood return period, contamination, flood duration and level of precaution.

The use of tree-based models, or decision trees, is also explored in this chapter in the context of modelling damages related to heavy rainfall. Decision trees have

proved to be useful to explore the structure of complex data sets. Decision trees have been applied in a large variety of fields, such as ecology (e.g. Rejwan et al., 1999; De'ath and Fabricius, 2000) and medicine (e.g. Hess et al., 1999), but the study by Merz et al. (2013) was the first to explore the concepts for flood-damage modelling.

In this chapter, results of decision-tree analysis are presented, based on a large insurance database of district-aggregated damage data. The data represent water-related damages to residential buildings, for the period of 1998–2011, covering the whole of the Netherlands. In a previous study based on the same database (Ririassa and Hoen, 2010) and in Chapter 3, relationships between various characteristics of rainfall events and various damage variables were investigated. These studies found that rainfall characteristics explain only part of the variance in water-related damage data. Similar conclusions were drawn by Cheng (2012); Einfalt et al. (2012); Zhou et al. (2013) and Climate Service Center (2013), who also analysed water-related insurance claim data in relationship to rainfall data. There may be two reasons for the variance that is left unexplained. Firstly, global regression models were used in the aforementioned studies, but, given the complexity of the problem, they may not be the most appropriate model choice. Secondly, the analyses were limited to rainfall-related factors only, while, in reality, many more factors are relevant for damage.

Building upon the research by Merz et al. (2013), this chapter aims to investigate a wide range of damage-influencing factors, defined by the scale of districts and their relationships with average size and frequency of insurance damage claims, using decision-tree analysis. The set of explanatory variables includes rainfall-related variables derived from weather radar data sets, topographic variables from a digital terrain model, building-related variables, and variables related to the socioeconomic status of households. Variables related to functioning of urban drainage systems (e.g. storage capacity, sewer type) were not included because these were not available on a nationwide basis. Separate analyses were made for property (or building structure), and content damage data. The chapter is structured as follows. First of all, an overview of the data sources and a description of how response and explanatory variables were derived from the data is given (Sect. 4.2). In Sect. 4.3, more background is given on the various choices that were made to construct decision trees. Results of the decision-tree analysis and a comparison with results from a global multiple-regression model are presented in Sect. 4.4, followed by a discussion in Sect. 4.5. Finally, Sect. 4.6 summarises conclusions and recommendations.

4.2 Data

4.2.1 Damage variables

Insurance damage data were provided by the Dutch Association of Insurers, an organization that represents the interests of private insurance companies operating in the Netherlands (Table 4.1). The data include daily records of water-related damage claims related to residential buildings and building contents in the Netherlands from a number of large private insurance companies. The database covers policy data of on average 22 % of all households in the Netherlands, in the period 1998–2011 (Fig. 4.1). In the Netherlands, almost all privately owned buildings are insured for property

Table 4.1: Overview of data sources used in this chapter.

#	Data source	Temporal resolution	Spatial resolution	Period	Related references
1	Databases from Dutch Association of Insurers				Ririassa and Hoen (2010)
	Property damage claims	By day	District level	1998–2011	
	Content damage claims	By day	District level	1998–2011	
2	C-band weather radar data set from the Royal Netherlands Meteorological Institute	1 scan/5 min	2.5 km × 2.5 km pixels	1998–2008	Overeem et al. (2009)
		1 scan/5 min	1 km × 1 km pixels	2009–2011	See Sect. 2.3 in Overeem et al. (2011).
3	Databases from Statistics Netherlands				
	Real estate values	By year	Per object	1998–2011	
	Housing stock register	By year	Per object	2006–2011	
	Integrated household income data	By year	Per household	2003–2011	
	Highest level of education achieved data	By year	Per person	1999–2010	
	Demographic background of persons data	By year	Per person	1995–2011	
4	National Building Register	By day	Per object	Dynamic	Online viewer: http://bagviewer.pdok.nl/.
5	Digital terrain model of the Netherlands	1 scan	5 m × 5 m pixels	Obtained in the period 2007–2012.	Online viewer: http://ahn.geodan.nl/ahn/. More background: Van der Sande et al. (2010); Van der Zon (2013).

damage that may result from a wide range of risks, such as fire, hail, rainfall and storms. Such insurance is commonly obliged in the case of a mortgage. The data are aggregated at the level of 4-digits postal districts, i.e. neighbourhood level. The Netherlands has around 4000 districts, with surface areas varying between 1 km^2 and 50 km^2.

Water-related damages can have a wide range of causes, such as rainwater intrusion through roofs and pluvial flood water that enters buildings through doors and wall openings. Cases of fluvial flooding are not included in the data, as these are not commonly covered by property and content insurance policies in the Netherlands (Seifert et al., 2013). Insurers typically compensate for the costs of cleaning, drying and replacing materials and objects, and the costs of temporarily rehousing people.

Damage values before 2002 were converted from guilder to euros using the conversion ratio 1 guilder = 0.454 euros. All values are in 2011 euros. Every value associated with a year before 2011 was adjusted for inflation according to the correction indices in Table 3.4 (Chapter 3). Extensive checks on missing and incorrect values (e.g. blanks, zeros and incorrect dates) and inconsistencies in the data are discussed in Chapter 2. Figure 4.2 shows that property insurance is well represented in the database in most regions of the Netherlands (insurance density of $> 10\,\%$), but are poorly represented in parts of the northern provinces (insurance density of $\leq 10\,\%$). This is mainly the case for property insurance, as almost all districts have content insurance density of $> 10\,\%$.

The response data being modelled are average claim size and claim frequency, per district, per day (see Table 4.2 for definitions).

4.2.2 Subsetting data

A case (i.e. a row in the data table) is a unique combination of a day and a district. Cases were filtered out for a number of reasons. Cases with few recorded claims are often not related to rainfall, but to other causes of water-related damage, such as bursts of water supply pipes and leakages of washing machines. These non-rainfall-

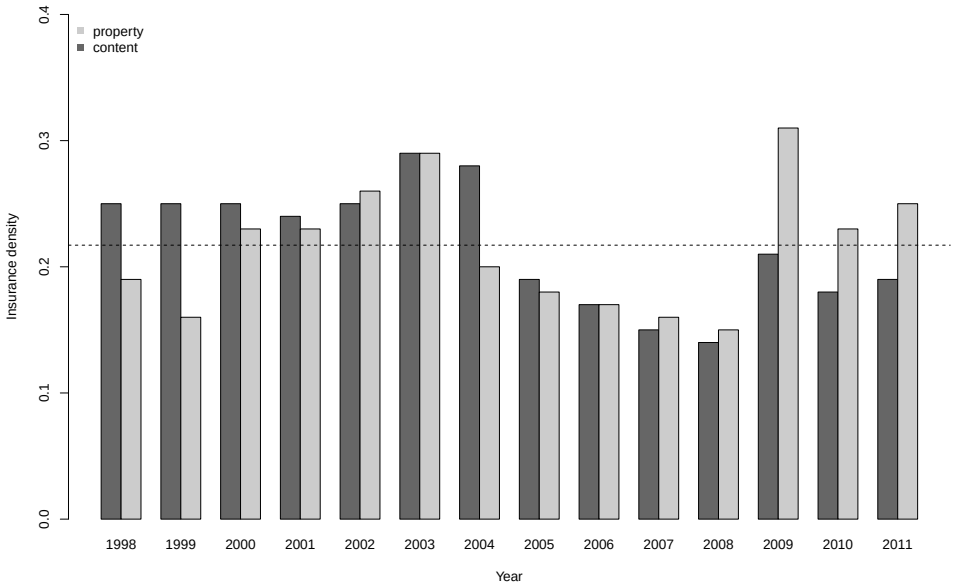


Figure 4.1: Insurance density per year: the number of insured households in the database from the Dutch Association of Insurers per year, divided by the total number of households in the Netherlands per year. Light bars represent property insurance, and dark bars represents content insurance. The dashed horizontal line (= 22 %) represents the average insurance density for the period 1998–2011 (the same percentage for content and property insurance).

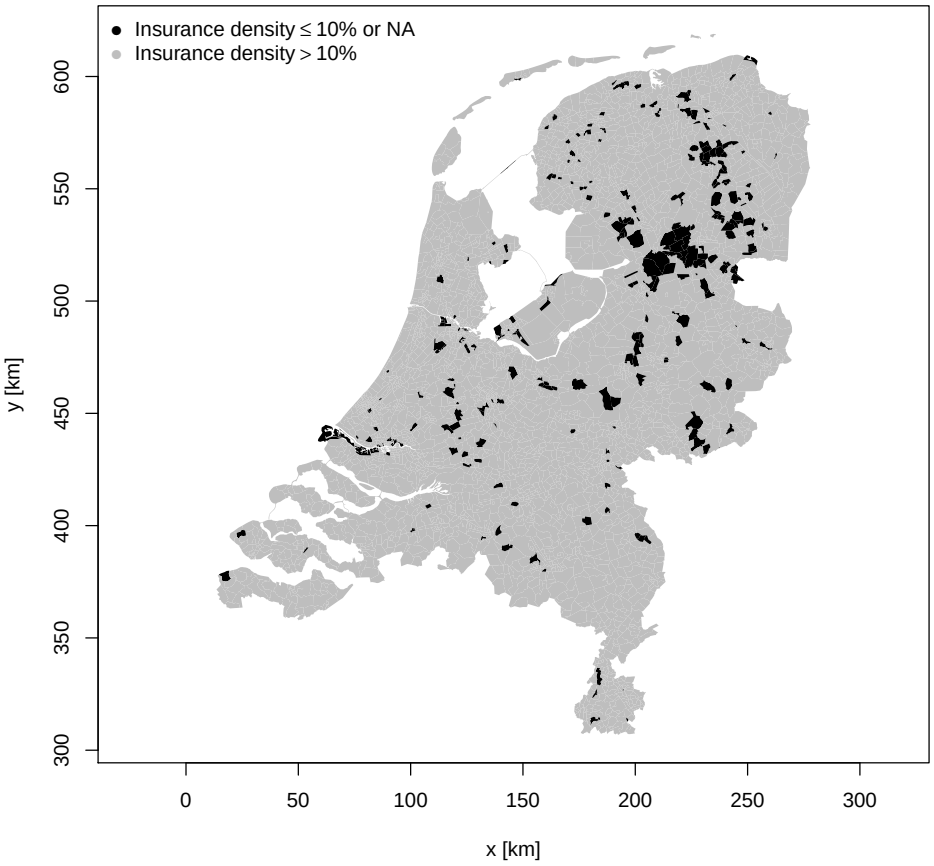


Figure 4.2: Property insurance density: the percentage of homeowners included in the database from Dutch Association of Insurers, averaged over the years 1998–2011. Dark areas denote districts that have an insurance density of less than 10 % or where values are not available. Note that this figure is slightly different for individual years.

Table 4.2: Model variables, variable definitions and value ranges.

Variable name	Definition	Min – Max (Median) Property data	Min – Max (Median) Content data	Source (Table 4.1)
Response variables				
Claim frequency (cf)	Number of claims per day per district divided by number of policyholders per district	0.0007–0.0933 (0.0039)	0.0006–0.0812 (0.0026)	1
Average claim size (acs)	Total damage per day per district divided by number of claims per day per district (euros)	43–80 520 (1024)	12–28 282 (674)	1
Rainfall-related variables				
Maximum rainfall intensity (rmax)	Maximum intensity of rainfall event at the building-weighted centroid of a district, using an 1 h moving time window ($mm\,h^{-1}$)	0–97 (4)	0–97 (8)	2
Mean rainfall intensity (rmean)	Mean intensity of rainfall event at the building-weighted centroid of a district ($mm\,h^{-1}$)	0–38 (1)	0–46 (1)	2
Rainfall volume (rvol)	Volume of rainfall event at the building-weighted centroid of a district (mm)	0–149 (12)	0–154 (17)	2
Rainfall duration (rdur)	Duration of rainfall event at the building-weighted centroid of a district (h)	0–48 (10)	0–48 (11)	2
Socio-economic variables				
Household income (inc)	Median disposable household income per district, adjusted for inflation according to Table 3.4 and classified in 10-percentile groups: 1= lowest 10 % of data, 10= highest 10 % of data	1–10 (5)	1–10 (3)	3
Education of breadwinner (edu)	Mean level of highest education obtained by main breadwinner per district, according to Dutch education index: 1 = lowest: e.g. kindergarten, 7 = highest: e.g. degree in medicine	2.6–5.3 (3.9)	2.6–5.2 (3.7)	
Age of breadwinner (age1)	Median age of main breadwinner per district (years)	24–68 (51)	27–72 (50)	3
Fraction of homeowners (own)	Number of owner-occupied buildings per district divided by the total number of residential buildings per district	0.08–0.95 (0.62)	0–0.98 (0.52)	3
Building-related variables				
Real estate value (rev)	Median real estate value of residential buildings per district, adjusted for inflation according to Table 3.4 (euros)	39 371–1 068 136 (184 508)	34 132–773 468 (145 774)	3
Fraction of low-rise buildings (low)	Number of residential addresses that have their entrance at ground level divided by the total number of residential addresses per district	0–1 (0.91)	0–1 (0.85)	4
Building age (age2)	Median age of residential buildings per district (years)	2–251 (41)	1–253 (42)	4
Ground floor area (floor)	Mean area of the ground floor of a building per district (m^2)	7–385 (63)	17–263 (62)	4
Topographic variables				
Slope (slope)	Median slope at building pixels (°) per district, according to Horn (1981)	0.29–7.29 (0.62)	0.29–6.48 (0.65)	5
Position index, 25 m (tpi1)	Median topographic position index at building pixels (m) per district, according to Weiss (2001) using 25 m × 25 m window	−0.02–0.16 (0.04)	−0.01–0.16 (0.04)	5
Position index, 255 m (tpi2)	Median topographic position index at building pixels (m) per district, according to Weiss (2001) using 255 m × 255 m window	−1.55–0.95 (0.11)	−0.73–1.24 (0.11)	5
Position index, 1005 m (tpi3)	Median topographic position index at building pixels (m) pre district, according to Weiss (2001) using 1005 m × 1005 m window	−16.76–7.20 (0.14)	−9.85–7.2 (0.12)	5
Others				
Season (seas)	Season of the year: winter = Dec–Feb, spring = Mar–May, summer = Jun–Aug, autumn = Sep–Nov	NA	NA	NA

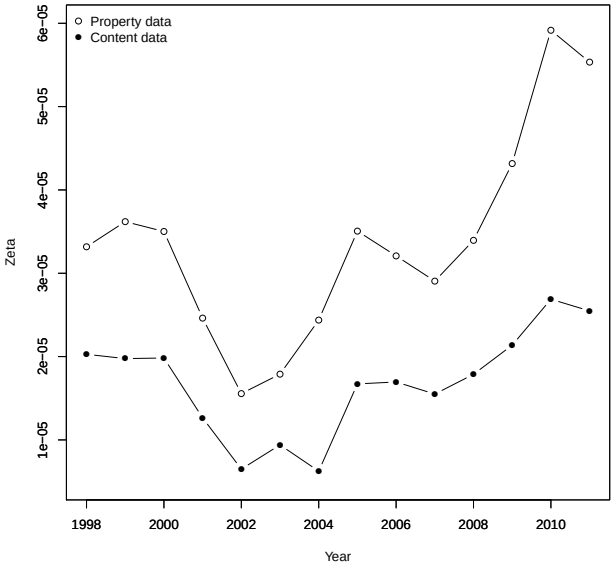


Figure 4.3: Average probability of a non-rainfall-related claim per day per policyholder for the years 1998–2011. The white dots are related to property claim data, the black dots to content claim data.

related claims occur throughout the year, whereas rainfall-related claims are clustered on wet days. Cases were therefore selected based on a statistically higher number of claims than expected on dry days. For this, a filter approach proposed in Chapter 2 was applied. A binomial probability law was applied to dry days in the data set to derive the probability of y claims at least as extreme as k_i, the number of claims observed for case i, given K_i, the number of insured households for case i (i.e. p value):

$$\Pr(y \geq k_i \mid K_i) = 1 - \sum_{y=0}^{k_i-1} \binom{K_i}{y} \zeta^y (1-\zeta)^{K_i - y}, \tag{4.1}$$

where ζ is the probability of a non-rainfall-related claim on a day for an individual, insured household. Figure 4.3 shows the estimated ζ per year for content and property claims, based on cases for which no rainfall was recorded. The variations of ζ between years may be related to annual changes in the participating insurers; among insurers, there may be different policies towards claim compensation. Additionally, there can be changes in people's claiming behaviour. Cases were selected if the p value (according to Eq. 4.1) was below a significance level of 0.01 (1 %), with a minimum of two claims per case. This implies that relationships between variables are investigated given a likelihood of 99 % of rainfall-related damage.

Furthermore, cases were discarded if insurance density was less than 10 %, the value of claim frequency was unrealistically large (> 0.1), or the number of policy-

holders was less than 100. The last rule was applied to reduce the risk of cases with few policyholders to show high claim frequencies just by chance. The final subsets related to property data and content data contain around 6000 cases ($\approx$ 15 500 claims) and around 6300 cases ($\approx$ 19 000 claims) respectively. Figure 4.4 shows the distributions of the response variables for the subsets; the distribution are skewed to the right.

4.2.3 Contextual variables

Rainfall-related variables

For each case in the subset, rainfall volume, rainfall duration, maximum and mean rainfall intensity were extracted from weather radar data (Table 4.2). Definitions of these variables can be found in Table 4.2. A database of C-band weather radar images was used, provided by the Royal Netherlands Meteorological Institute (Table 4.1). The images are composites based on two C-band Doppler radars, which have been adjusted for various biases using data from manual and automatic rain gauges (Overeem et al., 2009). The rainfall-related variables were obtained using the following steps, as is also described in Chapter 3.

Firstly, rainfall time series are processed at individual pixel level. Rainfall data were extracted for claim days (i.e. the days related to the cases) and for one previous day. Then, independent rainfall events were selected based on an intermediate dry period of at least 12 h, with "dry" being defined as < 0.083 mm for a 5 min time step. The dry period of 12 h interval relates to the time of a sewer systems to restore to equilibrium state (i.e. a state with only dry weather flow) after a rainfall event. Dutch sewers are designed to restore to an equilibrium state in around 10 h to 24 h (Stichting RIONED, 2008). Only rainfall events that coincide with a claim day for at least one time step are kept. This results in either zero, one or two independent rainfall events that can be associated with a claim day. In the case of zero events, all rainfall characteristics are assigned zero values. In the case of two events, the maximum value out of the two events is taken.

Secondly, the radar pixel value at the building-weighted centroid of a district is selected. The weighting was based on the locations of residential buildings in the district according to the National Building Register (see "Building-related variables" in Sect. 4.2.3). The building-weighted centroid better links radar data to urbanised areas compared to the geometric centroid, particularly for larger districts with spatial variation of urban density (Fig. 4.5).

Topographic variables

A digital terrain model (DTM) of the Netherlands was used to characterise districts in terms of their steepness (Table 4.1). Steep catchments are prone to depression filling, where rainwater runs down a slope and fills up depressions at the bottom if no drainage facilities are available (Ten Veldhuis et al., 2011). The DTM used is a representation of the natural terrain, excluding semi-permanent objects like vegetations and buildings. The spatial resolution of the DTM was aggregated to 5 m $\times$ 5 m tiles (Van der Zon, 2013). Data gaps in the DTM were filled using linear interpolation. More background

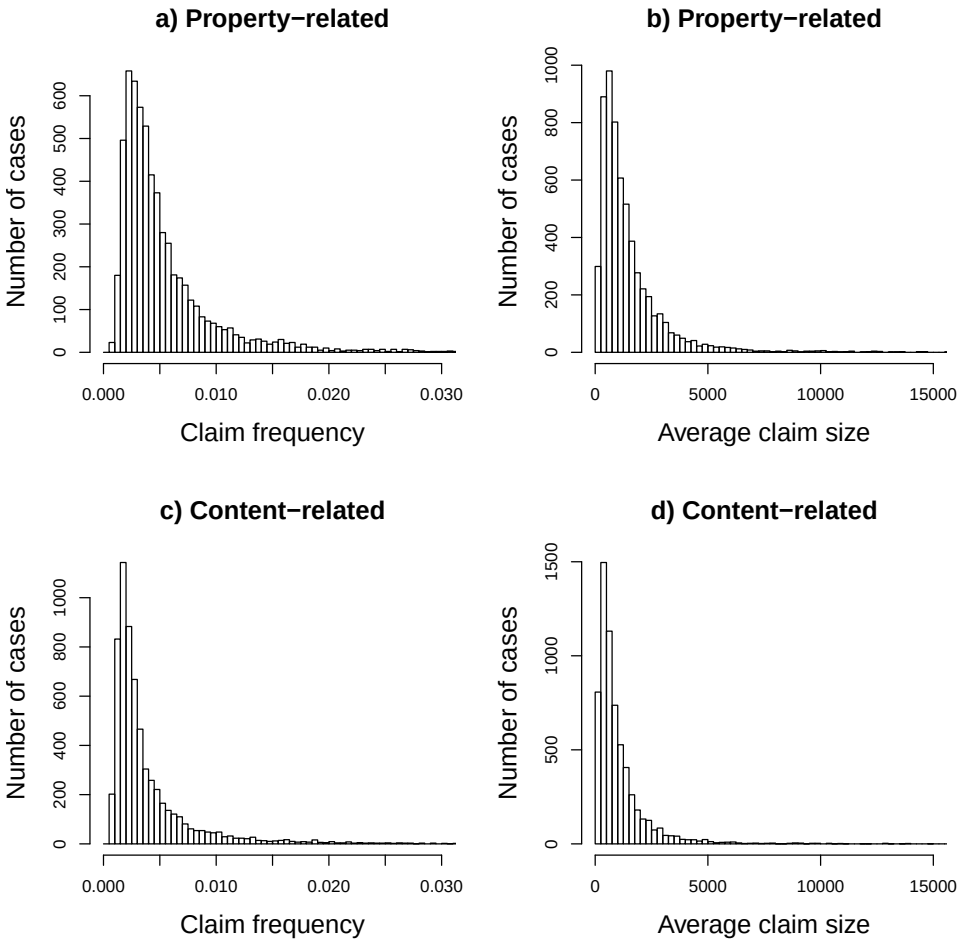


Figure 4.4: Histograms of response variables in subset data: **(a)** claim frequency of property-related cases, **(b)** average claim size of property-related cases, **(c)** claim frequency of content-related cases and **(d)** average claim size of content-related cases. Histograms of claim frequency and average claim size have a bin size of 0.0005 and 250 euros respectively.

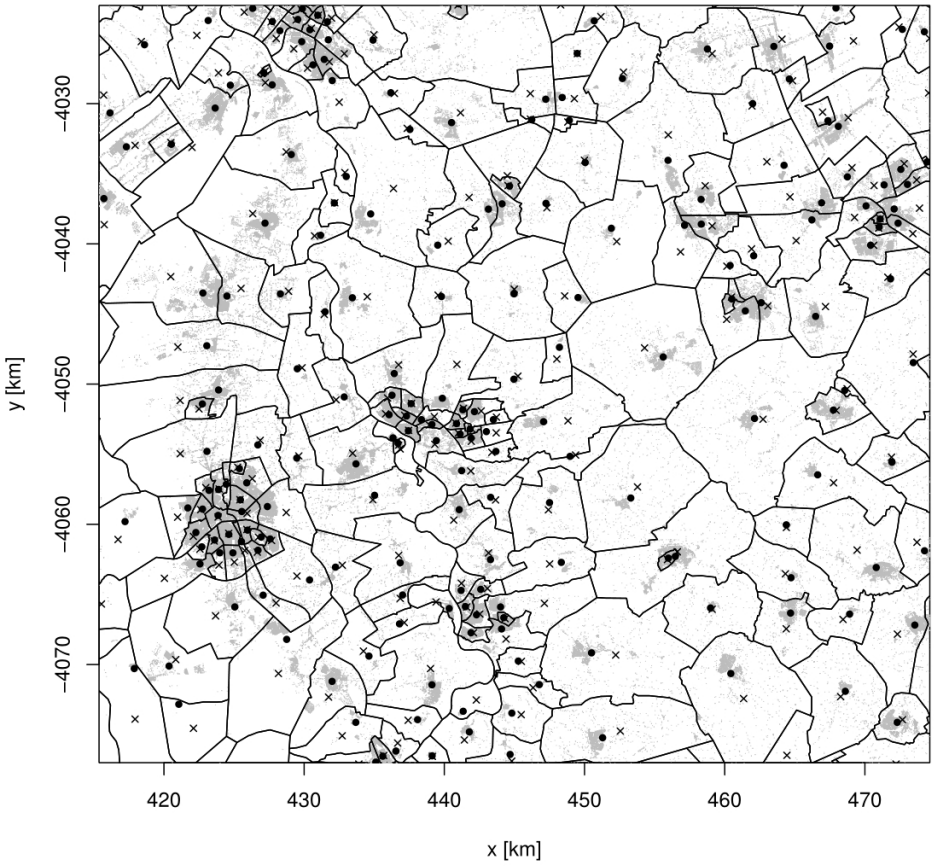


Figure 4.5: An example map showing postal districts (polygons), their geometric centroid (crosses) and their building-weighted centroids (dots). The grey dots are residential areas used in the weighting.

on the laser scanning campaign and data quality can be found in Van der Sande et al. (2010) and Van der Zon (2013).

There is a wide range of techniques to calculate topographic variables from raster data. For example, see Wilson et al. (2007) for an extensive review. This study focused on two variables: topographic position index (TPI) and slope (Table 4.2). TPI compares the elevation of a cell to the mean elevation of a specified neighbourhood around that cell (Weiss, 2001). A positive TPI value means that the cell is a locally high point within the analysis window, whereas a negative TPI value corresponds with a locally low point. TPI was calculated using three sizes of analysis windows, i.e. a 25 m × 25 m, 255 m × 255 m and 1005 m × 1005 m window. Slope was assessed according to the procedure discussed in Horn (1981), where the maximum rate of change in value from the cell to its eight neighbours was calculated.

Values of the topographic variables were assigned to residential buildings, based on the pixel in which the geometric centroid of the building was located. Building locations were derived from the National Building Register (Table 4.1) using the reference data of 31 December 2011. The derived values were then spatially aggregated to obtain median variable values per district. Median values, rather than mean values, were used to reduce the effect of outliers. Although there may be changes in the housing stock between years, it was assumed that the district-aggregated topographic variables are constant for the entire study period.

Socioeconomic variables

Previous studies have shown socioeconomic data of households, such as ownership structure, to be significantly correlated to property and content damage (e.g. Thieken et al., 2005). The relationships between socioeconomic variables and the damage may be weaker when studied at the level of districts (compared to the level of individual households), in particular when districts are heterogeneous. For example, when there is a large variance in household incomes.

Databases of Statistics Netherlands were used to derive a number of basic socioeconomic variables (Table 4.1 and 4.2). The variables are district-aggregated statistics. Median values were used instead of mean values for variables that showed strong variance within districts (i.e. age of breadwinner and household income) to reduce the influence of outliers. Because only homeowners can take property insurance, the variable "fraction of homeowners" is only relevant for content-related response variables.

Building-related variables

Building-related variables were based on the National Building Register (NBR), a geodatabase of all buildings and addresses in the Netherlands (Table 4.1), except for real estate values, which are based on databases of Statistics Netherlands. The NBR contains many building attributes, such as construction year, type of use and ground floor area. The database effectively tracks changes in the housing stock: i.e. new building are added, old buildings are marked "not in use". For any historic point in time, subsets of the housing stock can be made. Subsets of the data were made for each year (reference data: 31 December) of objects with a residential function, possibly combined with shopping or business function, and for which the building

status was marked "in use". From each case, three variables were derived: fraction of low-rise buildings, building age, and ground floor area (Table 4.2). Fraction of low-rise buildings was indirectly determined from the data; overlapping points (i.e. points representing addresses at different storeys of a flat) were removed and residual points were then counted and compared to original point data. In the cases that multiple addresses were sharing the same building polygon, the ground floor area was adjusted by dividing the total polygon area by the number of addresses.

Others

For each case, the season of the year was included to account for seasonal effects, such as occurrence of snow and hail and blockages of rain gutters or sewer inlets due to leaf fall.

4.3 Methods

4.3.1 Decision trees and splitting criteria

The two response variables, claim frequency and average claim size, are separately modelled as a function of the candidate explanatory variables (Table 4.2), using decision trees. The advantages of tree models are that they "can deal with non-linear relationships, high-order interactions and missing data" (De'ath and Fabricius, 2000).

The philosophy of this approach is to learn a tree by finding an explanatory variable that splits the data into two groups, or nodes, such that variance of the response variable is minimized. A data set is split into two groups by a chosen reference value of an explanatory variable: a group for which values are lower than the chosen reference value and a group for which values are higher than or equal to the chosen reference value. From all possible splits of all explanatory variables, the one that minimises the variance of the response variable in the resulting groups, is selected. This process is recursively repeated on each subgroup until a large tree is learned. Trees are trained based on the complete data set.

An important aspect in learning trees is the choice of the splitting criterion. A general expression of a goodness-of-split measure is the difference between the within-node deviance of the response data in the parent group, D_{P}, and the sums of within-node deviance of the response data in the left and right child group, D_{L} and D_{R} (Therneau and Atkinson, 2014):

$$\phi = D_{\mathrm{P}} - D_{\mathrm{L}} - D_{\mathrm{R}} \tag{4.2}$$

A split that maximizes Eq. 4.2 is sought out. The expression of the within-node deviance is specified depending on the type of response data. For continuous data, as is the case of average claim size, the within-node deviance is commonly defined as the sum of squares about the group mean (Table 4.3). The class of trees that are based on this deviance function are referred to as regression trees (Breiman et al., 1984). The summary statistic, or model outcome, that is given at each terminal node is the group mean.

Table 4.3: Within-node deviance functions. Symbols: k_i = number of claims per day per district, K_i = number of policyholders per day per district, w_i = case weight, n = number of cases.

Response variable	Distribution	Within-node deviance	Parameter estimation
$\log(\text{Average claim size}) = y_i$	Normal $(\mu;\sigma)$	$D = \sum \left[w_i (y_i - \hat{\mu})^2 \right]$	$\hat{\mu} = \frac{\sum w_i y_i}{n}$
$\text{Claim frequency} = \frac{k_i}{K_i}$	Poisson (λ)	$D = 2 \sum \left[k_i \log\left(\frac{k_i}{\hat{\lambda} K_i}\right) - k_i + \hat{\lambda} K_i \right]$	$\hat{\lambda} = \frac{\sum k_i}{\sum K_i}$
	Truncated Poisson (λ)	$D = 2 \sum \Big[k_i \log\big(h^{-1}(k_i)\big) - h^{-1}(k_i) - \log\big(1 - \exp(-h^{-1}(k_i))\big) - k_i \log\big(\hat{\lambda} K_i\big) + \hat{\lambda} K_i + \log\big(1 - \exp(-\hat{\lambda} K_i)\big) \Big]$ where $h(x) = \frac{x}{1 - \exp(-x)}$	$\hat{\lambda}$ using maximum likelihood estimation

Note: $h^{-1}(x)$ needs to be calculated numerically, which is inconvenient for decision tree learning where deviance needs to be evaluated for every split.

Similarly to ordinary least-square regression, the variance of the response variable needs to be constant for any group mean, otherwise greater weight is given to groups with higher variations (De'ath and Fabricius, 2000; Moisen, 2008). The average claim size was therefore log-transformed to stabilize variance. Note that there is no need to transform explanatory variables, as regression trees are invariant to monotonic transformations of explanatory variables (Breiman et al., 1984). To make analysis more robust for outliers, the numbers of claims on which average claim size is based were used as case weights.

For event rate data, which is the case of claim frequency, a more appropriate goodness-of-split measure is one that is based on the deviance function of Poisson distributed data (Table 4.3) (Therneau and Atkinson, 2014). Note that claim frequency is calculated by dividing the number of claims by the number of policyholders, where the number of policyholders may vary from district to district. The summary statistic that is given at each terminal node is the Poisson mean. Trees of this class are referred to as Poisson trees, following the naming convention by Lee and Jin (2006). From a theoretical point-of-view, the deviance function of a zero-truncated Poisson distribution gives a better description of the within-node deviance (Table 4.3), because only non-zero counts are considered here. Parameter estimation of this deviance function has the disadvantage of requiring an iterative process that is computationally much more demanding than the Poisson deviance function. For this reason, results are based on the splitting criterion that uses the Poisson deviance function. More on this issue can be read in Sect. 4.5.

The main source of missing data was rainfall data, due to weather radars not being operational. To deal with missing data, a common approach in decision tree learning is to impute missing data using surrogate variables (Breiman et al., 1984). Surrogate variables are variables that would split data into two groups similar to the split by the original, or primary, splitting variable. This method is, however, not appropriate for missing rainfall data, because none of the other explanatory variable considered in the present study can act as a suitable surrogate. Alternatively, the cases without rainfall data were discarded (8–11% of the cases). Still, surrogate variables were recorded at

each node for the purpose of calculating variable importance (see Sect. 4.3.2).

A total number of four trees were generated for the various responses: property claim frequency, content claim frequency, average property claim size, average content claim size. For all trees, explanatory variables listed in Table 4.2 were used as model input, except for "fraction of homeowners" in the case of property claim data.

4.3.2 Determining size of tree and variable importance

The large tree is then trimmed back to a simpler tree that still contains most of the predictive power of the large tree (De'ath and Fabricius, 2000; Therneau and Atkinson, 2014). The right size of tree is determined using 10-fold cross-validation. The following explanation of this procedure is based on the papers by De'ath and Fabricius (2000) and Moisen (2008): the data is randomly divided into ten mutually exclusive subsets of equal size. Then, ten trees are built using nine subsets each time, dropping out one subset in turn. The fitted trees are used to predict the omitted subset, such that the average error of all trees can be estimated. The error of a tree is defined as the amount of variance in the terminal nodes that is left unexplained compared to the variance of the undivided data. This is repeated for each tree size. In contrast to the error of a tree that is fitted on training data, the average error of cross-validation trees will eventually reach a plateau (a tree size where a next split does not add any value to the prediction). Because of the imprecision of determining the exact tree size at which the plateau is reached, the 1-SE rule is applied (Breiman et al., 1984), the smallest tree is taken, such that the average error is within one standard deviation of the minimum error of the cross-validation trees. This tree is referred to as the "pruned tree".

Decision trees can also be used to identify important variables. Variable importance is defined as the sum of the goodness-of-split measure (Eq. 4.2) of each split for which the variable was the primary or the surrogate splitting variable, scaled to sum to one.

Various softwares are available for decision-tree analysis. The Recursive Partitioning and Regression Trees (RPART) library for R 2.15.3 was used for this study, developed by Therneau and Atkinson (2014).

4.3.3 Global multiple-regression models

Results of decision-tree analysis were compared to results of global multiple regression analysis. A Poisson regression model was used to explain claim frequency as a function of various combinations of explanatory variables, which yields:

$$\log(k_i) = \log(K_i) + \beta_0 + \beta_1 \mathrm{x}_{1i} + \cdots + \beta_n \mathrm{x}_{ni}, \tag{4.3}$$

where k_i is the number of claims observed for case i, K_i is the number of insured households for case i, and $\beta_0, \ldots, \beta_n$ the regression coefficients. Regression coefficients are estimated using maximum likelihood estimation. A linear regression model was used to explain claim size, using a log-transformed response variable:

$$\log(y_i) = \beta_0 + \beta_1 \mathrm{x}_{1i} + \cdots + \beta_n \mathrm{x}_{ni} + \varepsilon_i, \tag{4.4}$$

Table 4.4: Spearman's pairwise correlation coefficients. Non-significant relationships ($p < 0.001$) are denoted with a hyphen.

	Property claims		Content claims	
Variable	Frequency	Average size	Frequency	Average size
rmax	0.32	0.07	0.40	0.12
rmean	0.30	0.04	0.35	0.09
rvol	0.29	–	0.31	0.10
rdur	0.18	–	0.14	–
inc	−0.21	–	0.24	–
edu	−0.10	0.07	0.12	0.11
age1	–	–	0.15	–
own	n/a	n/a	0.35	–
rev	−0.20	0.14	0.24	0.13
low	–	–	0.22	−0.06
age2	0.17	–	–	–
floor	0.09	–	0.26	–
slope	0.10	–	–	0.05
tpi1	–	–	–	–
tpi2	–	–	0.10	–
tpi3	0.05	–	0.14	–

where y_i is the average claim size for case i, and ε_i the error term of case i. Tree models and global regression models were compared in terms of variance explained by the models. Since the only interest here is to quantify the performance of an entire set of explanatory variables in predicting claim frequency, and not the individual contributions of the variables, it is safe to ignore any correlation that may exist between the explanatory variables. Note that the categorical variable "season" was not included in the models.

4.4 Results

4.4.1 Explorative analysis

To explore data, pairwise correlations between explanatory and response variables were analysed (Table 4.4). Spearman's correlation coefficients were calculated to account for the non-normal distributions of response data (Fig. 4.4). Note that the categorical variable "season" is not listed in Table 4.4. In general, there is no explanatory variable with strong predictive power. The strongest relationships were found between rainfall-related variables (except for rainfall duration) and claim frequency ($\rho = 0.29$–0.40). Other significant factors associated with claim frequency (with $|\rho| > 0.20$) include household income, real estate value, fraction of homeowners (content data only), fraction of low-rise buildings (content data only) and ground floor area (content data only). Interestingly, household income and real estate value are negatively correlated with claim frequency for property-related data ($\rho = -0.21$ and $\rho = -0.20$ respectively), but positively correlated for content-related data (both have $\rho = 0.24$). This is probably because data sets contain different groups of households: property-related data involves homeowners only, whereas content-related data

include tenants and homeowners. As a consequence, the data sets cover different variable value ranges; content-related data are associated with lower household incomes and real estate values (see Table 4.2). Another explanation could be that more expensive houses are better maintained or have better construction quality, and they are therefore less prone to flooding. Moreover, income is probably related to better maintenance, thereby indirectly affecting the claim frequency.

There are a larger number of significant links between explanatory variables and claim frequency than between explanatory variables and average claim size. In general, relationships between explanatory variables and average claim size were weak or non-existent. Maximum and mean rainfall intensity (and rainfall volume for content-related claims) were significant rainfall-related variables. Moreover, education and fraction of homeowners were significantly correlated with average claim size, for property-related and content-related claims.

Note that correlations reflect relationships based on the entire data set. Variables that turn out not to be important globally may therefore still be important locally.

4.4.2 Decision-tree analysis

In contrast to pairwise correlation analysis, decision-tree analysis allows to investigate relationships that exist locally within subgroups of data. The Poisson tree in Fig. 4.6 explains the property-related claim frequency, by dividing the original data into 14 subgroups (i.e. terminal nodes). The tree uses eight variables for splitting: two variables related to rainfall (maximum rainfall intensity and rainfall volume), three variables related to buildings (real estate value, building age and ground floor area), slope, season and household income. Maximum rainfall intensity is the top splitting variable and also the variable that makes the second split to the right. As a consequence, the data space is effectively split into three rainfall intensity levels: 0–15 $\mathrm{mm\,h^{-1}}$, 15–37 $\mathrm{mm\,h^{-1}}$ and $> 37\,\mathrm{mm\,h^{-1}}$, with most claims (67 %) falling into the lowest rainfall intensity group. Figure 4.7 illustrates the splitting method for the top split; the claim frequency is plotted against maximum rainfall intensity (see top of Fig. 4.7) and a split value for maximum rainfall intensity is sought that maximizes the goodness-of-split measure (see bottom of Fig. 4.7). For cases associated with rainfall intensities larger than 37 $\mathrm{mm\,h^{-1}}$, no further subgroups were found. The next splits down in the tree are related to real estate value. Real estate value correlates negatively with claim frequency; higher claim frequencies are associated with less expensive buildings. Building age only appears to be significant for cases with low rainfall intensities (node 4, rmax $< 15\,\mathrm{mm\,h^{-1}}$). At two nodes (node 5 and 12), season was the best splitting variable, but both splits were not consistent: autumn and winter were found to be either associated with relative low or high claim frequencies. Ground floor area correlates positively with claim frequency at nodes 25: larger buildings receive around 60 % more claims compared to small buildings. The tree explains 32 % of the variance in training data (i.e. $R^2 = 1 - \frac{\text{sum of deviance at terminal nodes}}{\text{deviance of undivided data}}$) and, on average, 26 % of the variance in cross-validation data sets (Fig. 4.8).

The regression tree, explaining content-related claim frequency, has 12 terminal nodes and its splits are based on four splitting variables: maximum rainfall intensity, fraction of homeowners, ground floor area and fraction of low-rise buildings (Fig. 4.9).

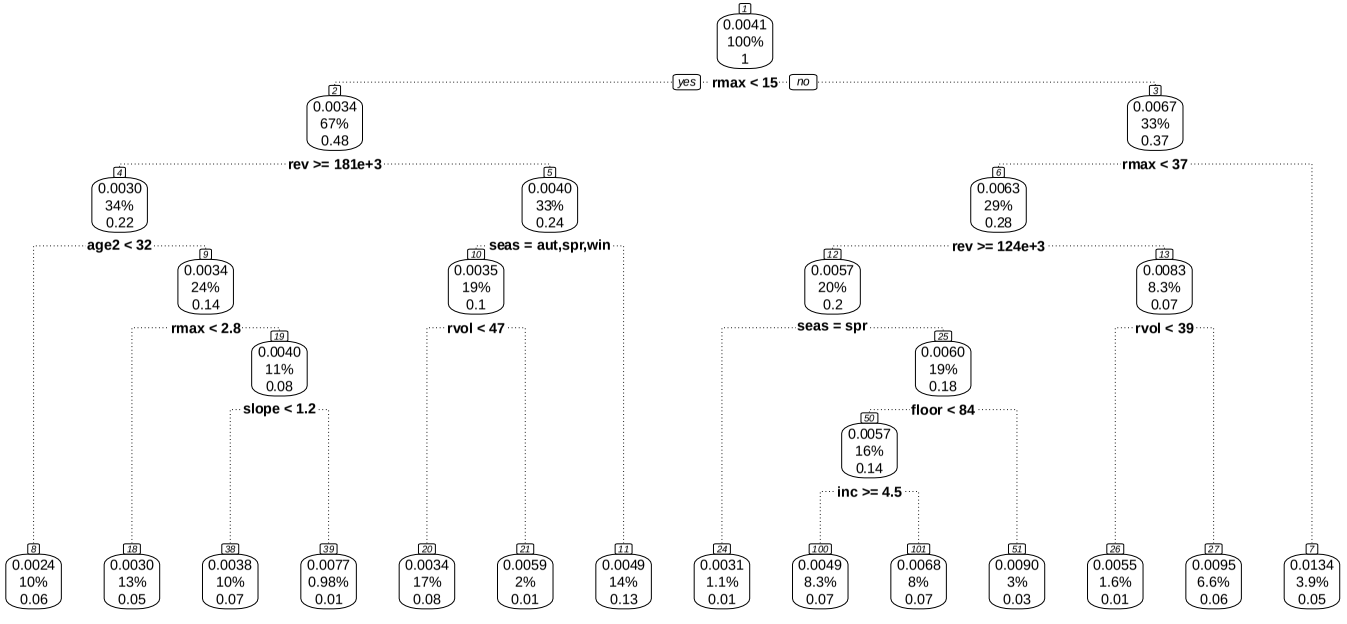


Figure 4.6: Pruned Poisson tree explaining the property claim frequency as a function of rainfall-related, building-related, socioeconomic and topographic variables (tree size = 14). The values at nodes are, from top to bottom: (1) node index, (2) claim frequency (i.e. Poisson group mean), (3) percentage of claims falling into the group and 4) remaining deviance relative to the deviance of the undivided data.

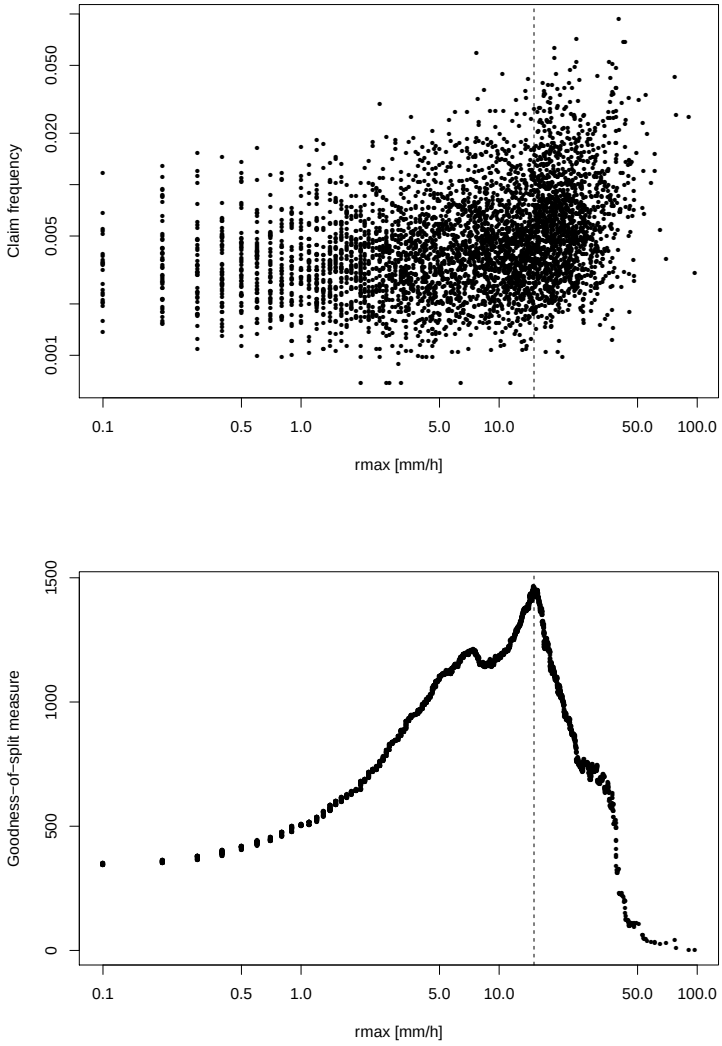
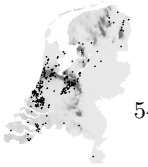


Figure 4.7: Scatter plot of claim frequency against maximum rainfall intensity, for the undivided data (top figure). The dashed vertical line represent splitting value that maximizes the goodness-of-split measure (bottom figure).



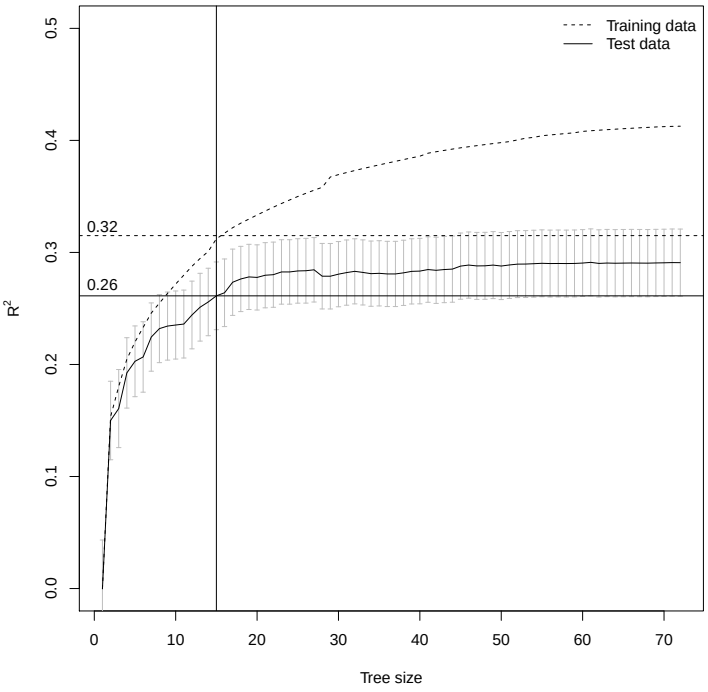


Figure 4.8: Performance of the Poisson tree for property claim frequency: the fraction of variance explained by the tree as a function of tree size, based on training data (curved dashed line) and validation data (curved solid line). The error bars represent one standard deviation of uncertainty. To determine the optimal size of the tree (indicated by the vertical solid line), the smallest tree is taken such that the explained variance is within one standard deviation of the maximum explained variance of the cross-validation trees, i.e. the intersection of the black solid line and horizontal line.

Similar to the previous tree, maximum rainfall intensity is the top splitting variable and also the value of the split ($16\,\mathrm{mm\,h^{-1}}$ vs. $15\,\mathrm{mm\,h^{-1}}$) is consistent between trees. Maximum rainfall intensity appears two more times lower down in the tree (node 4 and 6), which emphasises the importance of this variable in explaining claim frequency. For low-intensity rainfall events ($\mathrm{rmax} < 16\,\mathrm{mm\,h^{-1}}$), fraction of homeowners is a significant variable; districts with relatively many owner-occupied buildings ($\mathrm{own} > 0.52$) receive more claims than districts with relatively many rented buildings ($\mathrm{own} < 0.52$). Highest claim frequencies are observed for cases with high rainfall intensities ($\mathrm{rmax} \geq 16\,\mathrm{mm\,h^{-1}}$), relatively large and mostly low-rise buildings ($\mathrm{floor} \geq 86\,\mathrm{m^2}$, $\mathrm{low} \geq 0.59$, 3.3 % of all claims). The splits at node 15 and 22 (both having "ground floor area" as splitting variable) only reduce the deviance of the undivided data by less than 1 %. Thus, an even smaller tree can be proposed by considering these nodes terminal, without loosing much of the explained variance. The tree explains 30 % of the variance in training data and 22 % of the variance in validation data (not shown here), which means that claim frequency of content-related damage is slightly less predictable than claim frequency of property-related damage.

It was not possible to develop statistically acceptable trees for average claim size. The only meaningful splitting variable that was found for property-related average claim size was the real estate value. Cases with real estate values smaller than 97 000 euros were associated with an average claim size of 820 euros (11 % of the claims), whereas cases with real estate values larger than or equal to 97 000 euros had an average claim size of 1152 euros (89 % of the claims). Thus, rainfall-related variables were not used as a splitting variable. No splits were found for content-related average claim size.

4.4.3 Variable importance

The importance of variables in predicting claim frequency are listed in Table 4.5. Variables that correlate positively with claim frequencies are denoted with a plus sign, negative correlations with a minus. For education of breadwinner, the direction of the correlation is different from node to node (including surrogate nodes). For both content-related and property-related claim frequency, the most important variables are maximum rainfall intensity (importance score: 0.38), mean rainfall intensity (0.14–0.15) and rainfall volume (0.12–0.13). Although mean rainfall intensity did not show up in any of the trees, it was used as surrogate variable for maximum rainfall intensity most of the time. Real estate value is ranked high for property-related claim data (0.08), but is less important for content-related claim data (0.03). For content-related claim data, ground floor area and fraction of homeowners are important (0.08–0.11) after the rainfall-related variables, which is in line with the ordering of splitting variables in the tree of Fig. 4.9.

4.4.4 Comparison with global regression models

Table 4.6 summarises the regression results after fitting various global regression models to the same data that were used to learn the decision trees. Various combinations of explanatory variables were attempted to explain claim frequency and average claim size.

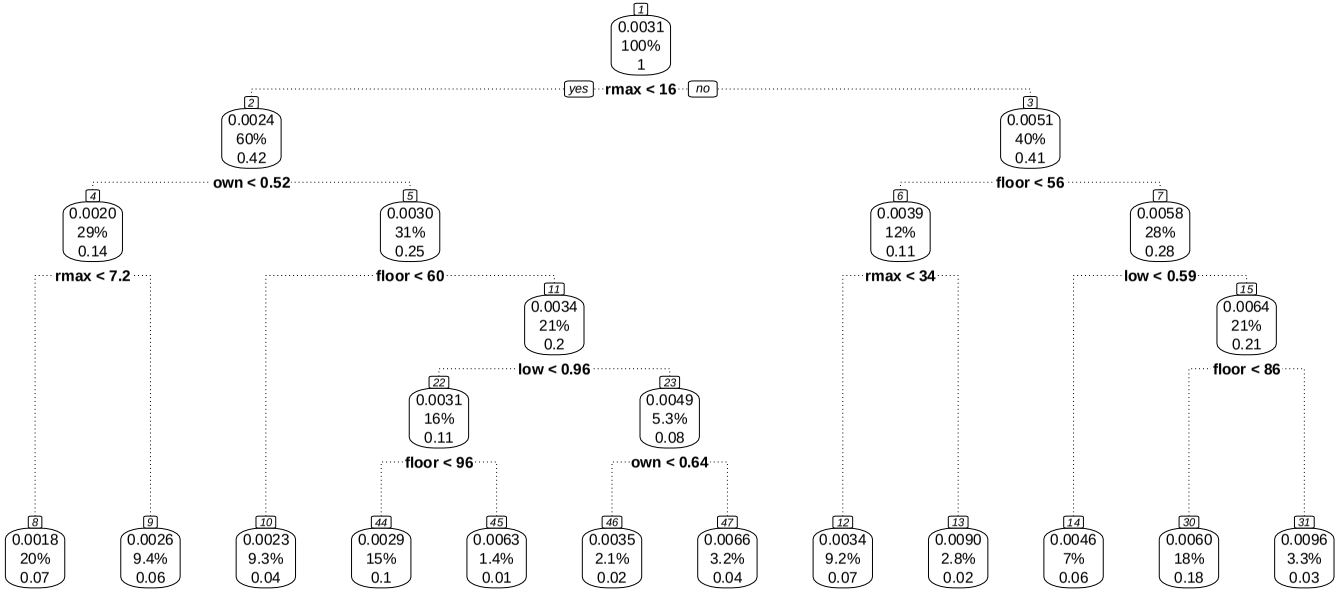


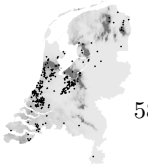
Figure 4.9: Pruned Poisson tree explaining the content claim frequency (tree size = 12). The values at nodes are, from top to bottom: (1) node index, (2) claim frequency (i.e. Poisson group mean), (3) percentage of claims falling into the group and (4) remaining deviance relative to the deviance of the undivided data.

Table 4.5: Variable importance for predicting claim frequency. The variable importance is the sum of the goodness-of-split measure of each split for which the variable was the primary or surrogate variable, scaled to sum to one. Surrogate variables are variables that split data most similar to the primary variable. Values smaller than 0.02 are omitted.

Property claim frequency			Content claim frequency		
Variable	Importance	Type of relationship	Variable	Importance	Type of relationship
rmax	0.38	+	rmax	0.38	+
rmean	0.15	+	rmean	0.14	+
rvol	0.13	+	rvol	0.12	+
rev	0.08	−	floor	0.11	+
seas	0.05	n/a	own	0.08	+
inc	0.05	−	low	0.06	+
age2	0.04	+	inc	0.05	+
slope	0.03	+	rev	0.03	+
edu	0.03	±	edu	0.02	+
floor	0.02	+			
rdur	0.02	±			

Table 4.6: Results of global regression and decision-tree analyses. Response variables are modelled as a function of (1) the maximum rainfall intensity, (2) all rainfall-related variables, (3) the variables actually used in the decision tree and (4) the variables with importance score $>$ 0.02 (for claim frequency) or all variables (for average claim size). For the global regression models, the cross-validated coefficient of determination, r^2_{cv}, is calculated using a similar approach, as discussed in Sect. 4.3.2.

	Global model		Tree model	
Response variable ∼ Explanatory variables	r^2	r^2_{cv}	r^2	r^2_{cv}
Property claim frequency ∼				
1: rmax	0.18	0.09	-	-
2: rmax + rmean + rvol + rdur	0.19	0.10	-	-
3: rmax + rev + age2 + slope + seas + rvol + floor + inc	0.27	0.18	0.32	0.26
4: rmax + rmean + rvol + rev + seas + inc + age2 + slope + edu + rdur	0.28	0.18	-	-
Content claim frequency ∼				
1: rmax	0.19	0.08	-	-
2: rmax + rmean + rvol + rdur	0.20	0.10	-	-
3: rmax + own + floor + low	0.25	0.11	0.30	0.22
4: rmax + rmean + rvol + own + floor + low + inc + rev + edu	0.26	0.12	-	-
Property average claim size ∼				
1: rmax	0.01	0.01	-	-
2: rmax + rmean + rvol + rdur	0.01	0.01	-	-
3: rev	0.02	0.02	0.02	0.00
4: all variables	0.04	0.03	-	-
Content average claim size ∼				
1: rmax	0.02	0.02	-	-
2: rmax + rmean + rvol + rdur	0.02	0.02	-	-
4: all variables	0.05	0.05	-	-



Best fits were found for the Poisson regression models for claim frequency that were based on the combination of variables, which were actually used in the decision trees (variant 3 in Table 4.6): $r^2_{cv} = 0.18$ and $r^2_{cv} = 0.11$ for property-related and content-related data respectively. Adding more variables (variant 4 in Table 4.6) hardly improves the predictive power of the models. The variance explained by the Poisson regression models (11–18 %) is considerably less than the variance explained by the cross-validated Poisson trees (22–26 %). Although linear regression models for average claim size were found to be significant, all models show weak explanatory power.

4.5 Discussion

The results of the tree analyses relate to correlations between variables, which does not necessarily imply causal relationships between variables. The results, therefore, need to be interpreted with caution. For future research, variable importance (i.e. Table 4.5) may give hints on variables that are closely connected to the mechanisms that generate damage. For instance, maximum hourly rainfall intensity was found to be the rainfall characteristic that best explains claim frequencies, which suggest that the process that causes damage is most sensitive to high-intensity rainfall events. For example, roofs may start to leak if rainfall exceeds the capacity of the system that drains rainwater from roofs. Similarly, real estate value, which ranked high on variable importance after rainfall-related variables, may be associated with better, more waterproof, materials and constructions. More research is needed here to understand the actual damaging process.

Topographic variables were not found to be important factors. There may be several explanations for this. One explanation relates to the aggregation of the topographic variables. Within a district, presence of buildings at locally higher, as well as, lower elevations may averaged out topographic variability. Another explanation may be that buildings and/or sewers in hilly areas have been more adapted to floods, i.e. people retrofitting their houses after severe floods.

The findings of this study are relevant for insurers. They contribute to the development of damage assessment tools that can be used to improve customer services. For example, a damage model that is able to spatially map expected damages based on weather forecasts or nowcasts, makes it possible to send out damage experts to customers more quickly and efficiently. Moreover, knowledge on customer groups associated with high claim frequencies may give hints on where damage prevention programmes are most likely to have impact. Insights into damage-influencing factors may also be helpful for meteorologist to improve weather-alert services. Rather than relying solely on meteorological thresholds, weather alerts may be enhanced by also taking into account district-specific thresholds (Parker et al., 2011; Priest et al., 2011).

Using decision trees, 22–26 % of the variance in claim frequency can be explained. Still, a large part of the variance is left unexplained, for which there are several possible explanations. A possible explanation might be that variations in data on a subdistrict scale lead to unexplained variance. The postal districts used here are specially designed for postal services; they are not necessarily statistically homogeneous units in terms of socioeconomics, topography and buildings. For instance, some districts

clearly show two distinct modes of the household income distribution. This makes it difficult to capture characteristics of districts in single variable values. Similarly, the spatial resolution of radar images (1–2.5 km) may be too coarse to capture the spatial variability of rainfall at the subpixel scale (Jaffrain and Berne, 2012; Peleg et al., 2013). Consequently, rainfall peaks of convective cells are underestimated. Another possible explanation is that important explanatory variables are missing. As mentioned in the introduction, variables related to urban drainage systems (e.g. sewer storage capacity, sewer type, soil type and percentage of impervious surface) may be important but were not included, because these were not available on a nationwide basis. Another variable that may be associated with rainfall-related damage, but was not included, is wind speed. Strong winds, in combination with precipitation, may cause damage to roofs, resulting additionally in rainwater intrusion. It is unlikely, however, that additional explanatory variables will have strong predictive power, given that none of the current variables have it. Finally, a source of unexplained variance may be related to data errors, in particular errors in insurance data, such as incorrect claim dates or policyholder counts. The insurance databases used in this chapter lack a consistent classification system, making it hard to subset data that is solely related to flood causes. A better classification of damage causes can give more accurate subsets and likely better model fits. Moreover, it was not possible to link content and property databases to individual policyholders. As a consequence, models could not be developed describing total damage per policyholder.

Although not researched in detail here, the explained variance may be underestimated as a result of the function that was applied to calculate the within-node deviance. The Poisson deviance function that was used allows responses to be zero (i.e. no claim). However, only cases with claims were considered here. A splitting criterion based on a deviance function of a distribution that does not allow the response value to be zero, such as the truncated Poisson distribution, can probably give a better description of the within-node deviance. An attempt was made to learn trees based on an alternative splitting criterion, using the deviance function of a zero-truncated Poisson distribution (Table 4.3). Parameters of this deviance function cannot be estimated explicitly and requires an iterative process. As a consequence, computational times to learn trees increased tremendously ($\sim$ days on a 8-core 2.5 GHz processor), which became even longer when cross-validation runs needed to be performed (time increases proportional to the number of runs). Preliminary results, based on trees only showing the first few splits, show that splits are almost similar to the ones presented in this chapter and are slightly better in reducing the deviance at nodes related to smaller claim frequencies. Given the long computational times, the alternative approach is not favourable unless advanced processors are available.

Claim frequency was calculated by dividing the number of claims per day per district by the number of policyholders per district, thereby assuming that every policyholder in a district is equally likely to generate claims as a result of rainfall. This assumption, however, may not always hold. In the case of a convective rainfall cell hitting a district whose size is smaller than the rainfall cell, it is safe to assume that every policyholder is exposed to rainfall, while in a district much larger than the rainfall cell only part of the policyholders is exposed. Thus, claim frequencies may be underestimated in the case of localised rainfall in large districts.

The structure of a tree is sensitive to a number of aspects. First of all, it is sensitive to the filtering rules that were applied to subset data (Sect. 4.2.2). Moreover, the choice of splitting criterion effects the way data is partitioned. There may be more appropriate splitting criteria for event rate data than the ones tested in this study, for example, splitting criteria based on other distributions for count data, such as the binomial or the negative binomial distribution. Furthermore, trees are sensitive to small changes in the learning data, for instance, when one of the explanatory variables is left out. Although not explored here, bagging and boosting approaches may be considered to overcome this problem, as was done in the study by Merz et al. (2013). With such approaches, results are aggregated over an ensemble of trees, where each tree is based on random but realistic changes in the training data (Elith et al., 2008; Borisov, 2009; Strobl et al., 2009).

It was not possible to develop statistically acceptable trees for average claim size. Attempts were made to build trees for average claim size and log-transformed average claim size. The latter was done to approximate normal distribution as distributions of average claim size are skewed to the right. Median, instead of average claim sizes, were not considered. In many insurance schemes, deductibles may affect claiming behaviour of people and cause censoring of small claim sizes. However, insurance policies related to the present database (i.e. water-related risks) do not have deductibles. There may be other changes in insurance policies (e.g. changes in damage causes that are covered) that may have affected claim sizes through time and caused failures to derive models. These were not accounted for in the present study, because this type of information was not readily available for all insurers in the database. Another possible explanation for failure to derive models for average claim size is that the costs to clean and dry walls and goods may be independent of the amount of rainwater that enters a building, i.e. a wet carpet has to replaced anyway, regardless of flood depth. Moreover, damage assessments are inherently uncertain, because of interpretation errors of insured and damage experts, which are difficult to capture in a model.

Similar to the conclusions by Merz et al. (2013), this chapter shows that decision tree models perform better than global regression models in terms of variance in damage data that is explained. This implies that decision tree models are better able to capture non-linear relationships in the data. For property damage, the decision tree reveals that maximum rainfall intensity effectively splits the data into three branches, each of them describing different relationships between explanatory variables and claim frequency.

In this chapter, tree models for claim frequency and average claim size were investigated given a likelihood of 99 % of rainfall-related damage. Tree models for the probability of occurrence of rainstorm damage were not considered, while it is worthwhile to study this, too, as part of a wider, risk-based approach.

4.6 Conclusions and recommendations

In this chapter, a wide range of factors potentially explaining variability in rainstorm damage were investigated, using decision-tree analysis. To this end, district-aggregated claim data from private-property insurance companies in the Netherlands were analysed, considering claim frequency and average claim size per day. Ana-

lyses were made separately for property and content damage claim data. This study has found that claim frequency is most strongly associated with maximum hourly rainfall intensity, followed by real estate value, ground floor area, household income, season (property data only), buildings age (property data only), fraction of homeowners (content data only) and fraction of low-rise buildings (content data only). It was not possible to develop statistically acceptable trees for average claim size. It is recommended to investigate explanations for the failure to derive models for claim size. These require the inclusion of other explanatory factors that were not considered in this chapter, an investigation of the variability in average claim size at different spatial scales and the collection of more detailed insurance data that allows to distinguish between the effects of various damage mechanisms to claim size. Cross-validation results show that decision trees were able to predict 22–26 % of variance in claim frequency, which is considerably better compared to results from global multiple-regression models (11–18 % of variance explained). Therefore, decisions trees are better able to capture local characteristics of claim data. Still, a large part of the variance in claim frequency is left unexplained, which is likely to be caused by variations in data at subdistrict scale and missing explanatory variables. The findings of this study have an important implication for insurance practice: for damage assessments, more detailed, high-quality damage data are required to sufficiently improve predictive power of damage models. There is, therefore, a definite need to improve insurance databases and to collect explanatory data at scales much closer to that of individual buildings.

CHAPTER 5

Failure mechanisms causing water damage to individual properties

Summary. This chapter is about the relative contribution of different failure mechanisms to the occurrence of rainstorm damage and how these mechanisms relate to weather variables. Relationships were investigated based on a detailed, property level home insurance database of around 3100 water-related damage claims for a case study in Rotterdam, the Netherlands. Records include comprehensive transcripts of communication between insurer, insured and damage assessment experts, which allowed claims to be classified according to their actual damage cause. Results show that roof and wall leakage is the most frequent failure mechanism causing precipitation-related claims, followed by blocked roof gutters, melting snow and sewer flooding. Claims related to sewer flooding were less present in the data, but are associated with significantly larger claim sizes than claims in the majority class, i.e. roof and wall leakages. Rare events logistic regression analysis revealed that maximum rainfall intensity and rainfall volume are significant predictors for the occurrence probability of precipitation-related claims. Moreover, it was found that claims associated with rainfall intensities smaller than 7–8 mm in a 60-min window are mainly related to failures processes in the private domain, such as roof and wall leakages. For rainfall events that exceed the 7–8 $\mathrm{mm\,h^{-1}}$ threshold, failure of systems in the public domain, such as sewer systems, start to contribute considerably to the overall occurrence probability of claims. The communication transcripts lacked information to be conclusive about the extent to which sewer-related claims were caused by overloading of sewer systems or failure of system components.

This chapter is based on: Spekkers, M. H., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2014a). On the occurrence of rainstorm damage based on home insurance and weather data. *Natural Hazards and Earth System Sciences Discussions*, 2(8):5287–5313, doi:10.5194/nhessd-2-5287-2014.

5.1 Introduction

Heavy rainfall causes considerable damage to building structure and content all over the world. Research on this topic has mainly concentrated on the adverse consequences of river flooding (Douglas et al., 2010; Jongman et al., 2012). Little research focused on damage caused by malfunctioning of urban drainage systems and direct water intrusion due to defects in the building envelope. Severe rainstorms have demonstrated that the impact of local high-intensity rainfall to cities can be large. On July 2011, Copenhagen was hit by 150 mm of rainfall in three hours, which resulted in surcharging of sewer systems, flooded houses, shops, roads and railways. Danish insurers received more than 90 000 claims and paid out more than 800 million euros (2011 value) in compensation (Garne et al., 2013). Another example is the heavy rainfall event of autumn 1998 in the Netherlands, which was associated with a return period of about 125 year and caused around 410 million euros (1998 value) to private buildings and agriculture (Jak and Kok, 2000). But also the cumulative damage of minor rainfall events can be considerable in the long run due to their high frequency of occurrence (Ten Veldhuis, 2011).

Many authors, from fields related to different kinds of weather-related risks (e.g. hailstorm, landslides, river flooding, coastal flooding), have recognized that damage data is lacking or biased and that this is limiting the development of reliable damage models (e.g. Pielke and Downton, 2000; Hohl et al., 2002; Elmer et al., 2010a; Gall et al., 2009; André et al., 2013). The same is true for rainstorms; little research focused on the collection of rainstorm damage data, the understanding of mechanisms causing damage and the deepening of statistical methods to analyse damage data. Among exceptions are studies by Busch (2008), Smith and Lawson (2012), Einfalt et al. (2012), Cheng (2012), Zhou et al. (2013) and Climate Service Center (2013), who analysed damage data sources (i.e. from insurance industry, local media, rescue service reports) and their relationships to rainfall data. As a result, there is no strong foundation for the development and validation of prediction models for rainstorm damage. Such models could help homeowners and water authorities to make better decisions on measures to prevent or reduce damage (e.g. retrofitting of buildings and early warning systems).

A potential source of damage data are insurance damage databases. They contain claims often collected over many years and from a large number of insured. A difficulty of insurance databases is that information on the mechanisms that cause damage and building-related, weather and socioeconomic variables are not or only limitedly available in claim data or cannot easily be retrieved from insurers' data archives (André et al., 2013).

This study aims to quantify the relative contribution of different failure mechanisms to the occurrence of building structure and content damage induced by rainstorms and to investigate to what extent the probability of occurrence of these processes is related to weather variables. For this purpose, a property level database of around 3100 water-related damage claims was analysed, for a case study in Rotterdam, the Netherlands. An interesting feature of this database is that it includes comprehensive transcripts of communication between insurer, insured and damage assessment experts, which allowed classification of claims based on the failure mech-

anisms causing damage. This information is, however, stored in an unstructured way that required substantial data classification efforts before data could be used for the analysis in this chapter.

The outline of this chapter is as follows. In Sect. 5.2 insurance damage data and classification of claims are described, as well as, the statistical method used to model probability of claim occurrence as a function of weather variables. Results of data analyses and regressions are presented in Sect. 5.3, followed by discussion in Sect. 5.4. In Sect. 5.5, conclusions and recommendations are summarised.

5.2 Methods

5.2.1 Case study description

This work focuses on Rotterdam, which is, with a population of around 620 000 ($\approx 3000/\mathrm{km}^2$), the second largest city of the Netherlands (Statistics Netherlands, 2014). Because the city is relatively flat (maximum ground level variations of 10–15 m), floods from heavy rainfall are typically characterized by flood depths up to a few decimetres and limited surface run-off. Rotterdam's sewers are mainly combined systems ($\approx$ 1800 km), some parts of the city have separate systems for wastewater and stormwater ($\approx$ 500 km) (City of Rotterdam, 2011). The average density of sewer pipes in the city centre is 15.6 km km^{-2} and 13 % of the area is surface water (i.e. city canals and ponds, not rivers) (Statistics Netherlands, 2013; City of Rotterdam, 2014). The majority of the buildings in Rotterdam was constructed in the 20th century. Rotterdam's urban fabric is characterised by a combination of terraced houses and high-rise residential and commercial buildings (Kadaster, 2013). It is assumed that within the study period no changes have been made to the sewer infrastructure and the building portfolio of Rotterdam that have significantly affected results of this study.

5.2.2 Insurance data

Insurance damage data were provided by a Dutch insurance company that is part of the Achmea insurance group*. Data are available at property level for the period of January 2007–October 2013 (data collected on: February 2014), containing around 3100 water-related claims. A claim relates to building structure or content damage or a combination of the two, depending on the available insurance policies at the risk address.

From each claim, the following information is available: risk address, type of insurance coverage, damage date, amount of compensation and detailed transcripts of communication between insurer, insured and damage assessment experts (e.g. calls, abstracts from reports). The database has been checked on missing and incorrect values, such as duplicated records, inconsistencies in date formats and claim coding. Every value associated with a year before 2013 was adjusted for inflation according to the correction indices in Table 5.1. On average, the data set contains information of around 16 000 risk addresses, which is 6 % of the total number of households in

*Website of Achmea insurance group: http://www.achmea.nl.

Table 5.1: Inflation adjustment according to the online database of Statistics Netherlands (2014). The average inflation per year for the Netherlands is used, based on the consumer price index. Every damage value associated with a year before 2013 was multiplied with a correction index.

Year	Inflation [%]	Correction
2007	1.6	1.12
2008	2.5	1.10
2009	1.2	1.08
2010	1.3	1.07
2011	2.3	1.05
2012	2.5	1.02
2013	2.5	1.00

Rotterdam. These risk addresses constitute a total number of around 21 000 insurance policies, of which around 6000 insurance policies relate to building structure insurance and around 15 000 to building content insurance. These numbers relate to data from one insurance company of the Achmea insurance group and do not reflect the market share in Rotterdam of the Achmea insurance group as a whole. Table 5.2 summarises the key features of the home insurance policy related to the present database.

The general rule for a claim to be accepted is that damage should be unforeseen and have occurred suddenly. Damage due to river flooding is not covered. Damage due to pluvial flooding is covered, provided that damage is directly and solely related to localised heavy rainfall (Ministry of Transport Public Works and Water Management, 2003).

5.2.3 Classification of claims

For the purpose of this study, claims were manually classified according to the actual cause of damage using the information in the communication transcripts. Transcripts contain telegram style summaries of calls and abstracts from reports and typically vary in length between a few lines to a few thousand words. When a claim is first reported at the insurer's call centre, the client is asked a few basic questions to verify if the client was indeed insured at the time of the damaging event and to make a quick assessment on the severity of the damage (e.g. "Is the risk address still habitable?"). Follow-up calls typically describe the actual cause of damage, an inventory of damaged goods and materials and the costs related to cleaning, drying, repairing or replacing goods and materials.

An easy-to-use web interface and SQL database was built based on the classification scheme listed in Table 5.3. Failure mechanisms described in Table 5.3 are also shown graphically in Fig. 5.1. Per claim only one cause class could be selected. Labels were given to each cause class to indicate whether the class relates to precipitation or not and whether the class relates to failures of systems in the public domain (i.e. responsibility of water authorities) or private domain (i.e. responsibility of homeowner, landlord or housing cooperative). Next to the classification scheme, a number of checkboxes was available to specify if (1) building or content was underinsured, (2)

Table 5.2: Key features of the home insurance policy related to the damage database used in this chapter.

	Content insurance	Property insurance
For whom?	homeowners and tenants	homeowners, landlords, housing cooperatives, homeowners associations
Covers physical damage to	– Portable goods – Semi-permanent objects (e.g. curtains, laminate, carpet, window blinds, shutters) – Additions or refurbishments to the property which enhance the property value that have been made by a tenant ("tenants improvements")	– Building – Building foundation – Garden, garden sheds – Permanent floors (e.g. floor tiles, glued wooden floors) – Kitchens
Damage assessment is based on	Replacement value or current value if replacement value is less than 40 % of current value	Costs to repair or rebuild (part of) the building, deprecation costs
Other compensations	Temporary housing, costs of damage experts, costs to clean and dry goods and materials and costs to detect and repair leakages	
Damage assessment by means of	"small" claims → proofs of payment "large" claims → independent damage assessment expert	
Grounds for rejection	– Negligence by insured (e.g. windows or doors that were left open during rainfall events, valves of the central heating system that were not closed properly after refilling the system, no leaf basket installed in rain gutter) – Lack of maintenance (e.g. poor quality sealant joints between walls and floors, rain gutter clogged with leafs) – Damage caused by "slow" processes (e.g. rotting, moisture intrusion through walls) – Construction errors (i.e. liability of building contractor or water company) – Costs not covered (e.g. costs to repair leakage are in some cases not compensated) – Floods from rivers or sea – Groundwater flooding	
Others	– In case of underinsurance (i.e. insured sum is less than asset value), compensation is proportional to the level of underinsurance – The insurance policies do not have deductibles	

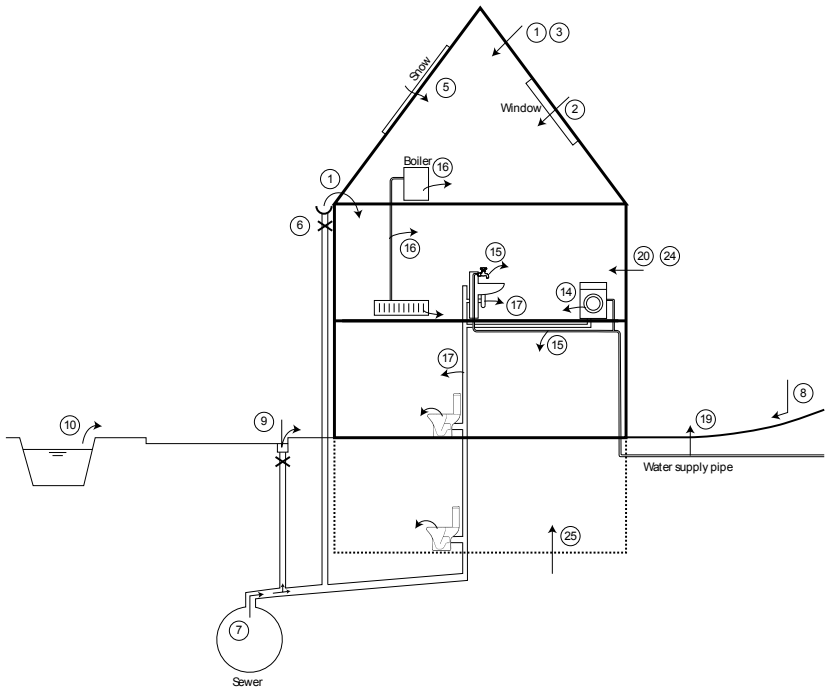


Figure 5.1: Water-related failure mechanisms applicable to residential buildings. Situation with a combined sewer system is displayed. Descriptions of the numbers are listed in Table 5.3.

insured has not responded to a request for a long time, (3) claiming process is still ongoing (typical processing time is a few months), (4) claim was rejected because of a lack of building maintenance and (5) damage was (partly) not insured.

Classification was done by three persons, by dividing the data set into three independent subsets containing 60, 36 and 4 % of the claims. On average, classification took four minutes per claim. The entire text was read first while making preliminary classification choices. A second reading was used to verify and finalise selections. If the available information was unclear or multi-interpretable, claims were flagged for investigation by one of the other two persons. This happen to be the case for 7 % of the claims.

5.2.4 Weather variables

A set of weather variables was derived for each combination of risk address and day (i.e. a case) to investigate explanations for claim occurrence (Table 5.4). Rainfall volume and maximum rainfall intensity were extracted from weather radar data, provided by the Royal Netherlands Meteorological Institute (KNMI), according to a method described in Chapter 3 and 4. Maximum rainfall intensity was calcu-

Table 5.3: Classification scheme of water-related failure mechanisms applicable to residential buildings. The column "Precipitation?" indicated if the claim is related to precipitation. The column "Domain" indicates whether damage prevention mainly concerns homeowners (private) or water authorities (public).

Id	Short name	Description	Precipitation?	Domain
1	Roof and wall leakages	Rainwater intrusion through roofs, facades, walls, wall-window interfaces and closed doors, which includes rainwater intrusion as a result of overloaded rain gutters	Yes	Private
2	Rainwater through open window	Rainwater intrusion through open windows, open doors	Yes	Private
3	Hail impacting roofs	Hail impacting roofs or windows	Yes	Private
4	Precipitation-related in private domain[1]	Precipitation-related in private domain, but other or unknown actual cause	Yes	Private
5	Melting snow	Intrusion of melting snow and ice, in particular snow blowing up under roof tiles	Yes	Private
6	Blocked roof gutters	Overflowing of roof gutters due to blockages in gutter or downpipe (e.g. by leafs or ice)	Yes	Private
7	Sewer flooding	Flood water entering buildings through doors or openings as a result of overloaded public sewer systems, including sewer backups	Yes	Public
8	Depression filling	Flood water entering buildings through doors or openings as a result of depression filling, i.e. rainwater filling up depressions if no drainage facilities are available	Yes	Public
9	Blocked sewer inlets	Flood water entering buildings through doors or openings as a result of blocked sewer inlets	Yes	Public
10	Flooding from local watercourses	Flood water entering buildings through doors or openings as a result of flooding from local watercourses (e.g. city canal, pond)	Yes	Public
11	River flooding	Flood water entering buildings though doors or openings as a result of flooding from river systems	Yes	Public
12	Precipitation-related in public domain[1]	Precipitation-related in public domain, but other or unknown actual cause	Yes	Public
13	Precipitation-related[1]	Precipitation-related, but other or unknown actual cause	Yes	Unknown
14	Leakages of household appliances	Leakages of household appliances (e.g. washing machines, dishwashers, aquaria, waterbeds)	No	Private
15	Bursts of household water supply pipes	Bursts of household water supply pipes, including attached facilities	No	Private
16	Leakages of central heating systems	Leakages of central heating systems, which includes boilers, radiators and pipes	No	Private
17	Blocked or leaking household wastewater systems	Flooding of wastewater due to blockage in or leakage of wastewater system located inside the building	No	Private
18	Non-precipitation-related in private domain[1]	Non-precipitation-related in private domain, but other or unknown actual cause	No	Private
19	Bursts of public water supply pipes	Bursts of water supply pipes owned by water supply company	No	Public
20	External water discharges	External water discharges (e.g. extracted groundwater from a construction site, fire extinguishing water)	No	Public
21	Blocked public wastewater system	Flooding of wastewater due to blockage in sewer lateral or sewer main, not related to rainfall events	No	Public
22	Non-precipitation-related in public domain[1]	Non-precipitation-related in public domain, but other or unknown actual cause	No	Public
23	Non-precipitation-related[1]	Non-precipitation-related, but other or unknown actual cause	No	Unknown
24	Water discharge from neighbours[1]	Water discharge from neighbours, but other or unknow actual cause	Unknown	Private
25	Groundwater flooding	Groundwater flooding due to persistent rainfall or sudden wall failure	Unknown	Unknown
26	Water-related[1]	Water-related, but other or unknown actual cause	Unknown	Unknown

[1] Residual group; a group of claims for which exact failure mechanisms could not be derived from communication transcripts.

Table 5.4: Definitions of explanatory variables and variable value ranges.

Variable name	Definition	Min – Median – Max
Rainfall volume (vol)	Volume of rainfall event at the radar pixel intersecting the building's centroid (mm)	0 – 4.9[1] – 86.2
Maximum rainfall intensity (max_{15})	Maximum intensity of rainfall event at the radar pixel intersecting the building's centroid, using a 15 min moving time window ($mm\,h^{-1}$)	0 – 3.8[1] – 102.3
Maximum rainfall intensity (max_{30})	Maximum intensity of rainfall event at the radar pixel intersecting the building's centroid, using a 30 min moving time window ($mm\,h^{-1}$)	0 – 2.8[1] – 62.3
Maximum rainfall intensity (max_{60})	Maximum intensity of rainfall event at the radar pixel intersecting the building's centroid, using a 60-min moving time window ($mm\,h^{-1}$)	0 – 2.0[1] – 34.2
Maximum temperature (temp)	Maximum temperature measured at the KNMI Rotterdam weather station (°)	−6 – 14.8 – 35
Daily-averaged wind speed ($wind_d$)	Daily-averaged wind speed measured at the KNMI Rotterdam weather station ($mm\,s^{-1}$)	0.7 – 4 – 14.3
Maximum hourly wind speed ($wind_h$)	Maximum hourly-averaged wind speed measured at the KNMI Rotterdam weather station ($mm\,s^{-1}$)	2 – 6 – 16
Wind gust ($wind_g$)	Wind gust measured at the KNMI Rotterdam weather station ($mm\,s^{-1}$)	3 – 11 – 28
Season (seas)	Season of the year: winter = Dec–Feb, spring = Mar–May, summer = Jun–Aug, autumn[2] = Sep–Nov	NA

[1] Median based on non-zero values only, [2] The level "autumn" was dropped to avoid multicollinearity.

lated using a 15-min, 30-min and 60-min moving time window to study typical time scales of failure processes. Rainfall duration was not considered, because results from Chapter 3 and 4, which are based on similar type of insurance databases, have shown that rainfall duration has no significant or weak effect to rainfall-related damage. Maximum temperature, daily-averaged wind speed, maximum hourly wind speed and wind gust were obtained from an automatic weather station operated by the KNMI, located in the north of the city, around 10 km from the city centre. The season of the year was included to account, for instance, for the occurrence of snow and hail and blockages of rain gutters due to autumn leaf fall.

5.2.5 Modelling the probability of claim occurrence

The modelling objective was to test the significance of weather variables in explaining the occurrence of precipitation-related claims. For each case, a unique combination of risk address and day, the outcome (Y_i) can be a reported claim (1) or not (0). The binary outcome can be linked to a set of weather variables ($x_1, \ldots, x_n$) using various types of models for binary data (McCullagh and Nelder, 1989). In this study a logistic regression model was used:

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni}, \tag{5.1}$$

where θ_i is the probability of claim occurrence ($Y_i = 1$) and $\beta_0, \ldots, \beta_n$ are regression coefficients. The regression coefficients were estimated using maximum likelihood estimation. The significance of the regression coefficients is tested using the Wald test. Logistic regression is known to generate biased estimates for rare events data, i.e. data series in which only a low percentage of events occur, resulting in an underestimation of the probability of rare events (King and Zeng, 2001). In present database, only

1031 precipitation-related claims were recorded in the period of 2007–October 2013, which is on average 2.67×10^{-5} claims per day per risk address. King and Zeng (2001) proposed a method, called rare events logistic regression, to deal with rare events data. This method encompasses a case-control design where ten times more non-events (i.e. no claim from an insured) are selected than events (i.e. a claim from an insured). The method first estimates regression coefficient using an ordinary logistic regression model (Eq. 5.1), then correcting regression coefficients for finite sample and rare events bias. For this purpose, the Rare Events Logistic Regression (relogit) routine from the Zelig package (Imai et al., 2007) for R was used. Collinearity among explanatory variables was tested by calculating the Pearson's correlation coefficient between each pair of explanatory variables. None of the correlation coefficients yielded values > 0.7, which means that collinearity effects can be neglected (Dormann et al., 2013).

The likelihood ratio and a pseudo-R^2 statistic were used to evaluate goodness-of-fit of a model. The likelihood ratio compares the likelihood of a model with predictors to the likelihood of a model without predictors (i.e. intercept-only model), which tests if adding explanatory variables to a model significantly improves model fit. For logistic regression, there is no universally accepted measure that represents the proportion of variance explained by the predictors, such as R^2 for ordinary least-squares regression. Several pseudo-R^2 statistics exist; however, these statistics generally score much lower than their equivalent in ordinary least-square regression and are therefore found less informative. They can be used, nevertheless, to compare predictability of nested models. In this study McFadden's R^2 is used (e.g. Long, 1997).

5.2.6 Discarded data

During the validation process of the insurance data, it was found that on three extremely stormy days (i.e. storm Kyrill on 18 January 2007 and storms on 27 July 2013 and 28 October 2013), despite occurrence of rainfall, no or hardly any precipitation-related claims were recorded. Upon further inquiry, the insurer has indicated that on extremely stormy days, precipitation-related claims are often inaccurately recorded as storm-related claims. These three days are therefore excluded from the logistic regression analysis.

5.3 Results

5.3.1 Relative occurrence frequencies and costs of claims

Analyses of the relative occurrence frequencies of damage causes show that leakage of roofs and walls is the most frequent failure mechanism generating water-related claims, followed by burst of household water supply pipes, blockage or leakage of household wastewater systems and leakage of household appliances (Fig. 5.2). Besides roof and wall leakages, other common precipitation-related failure mechanisms are blocked roof gutters, snow melting under roof tiles and sewer flooding.

In general, 34 % of the claims were related to precipitation and 43 % to non-precipitation causes. For the remaining 23 %, it was unknown if the claim was related

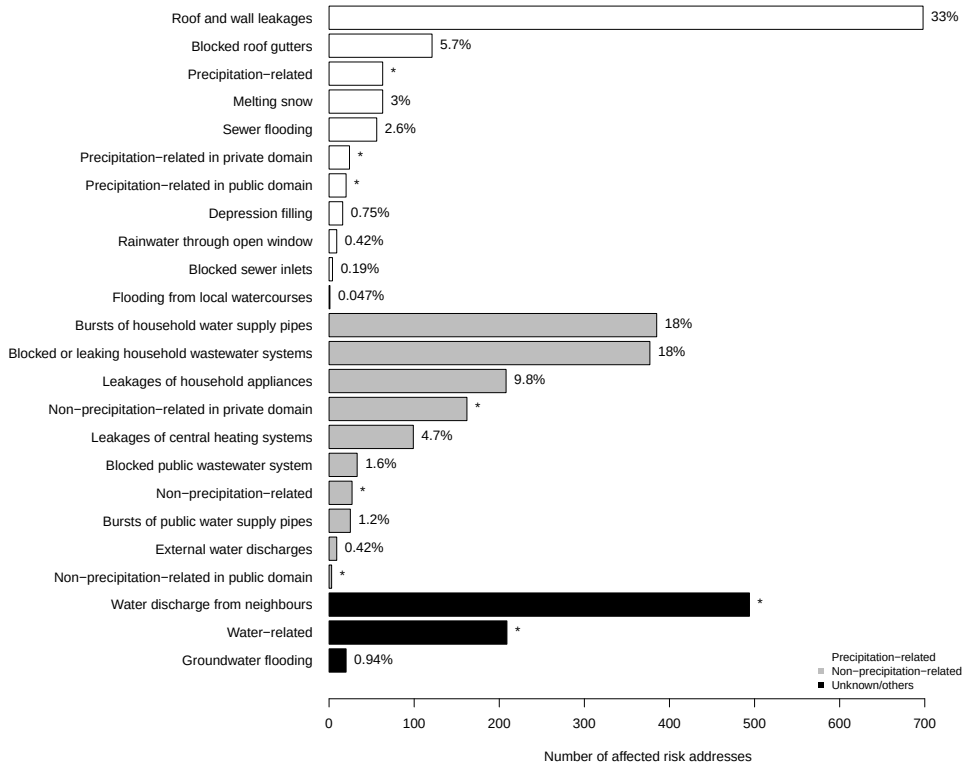


Figure 5.2: Occurrence rates and relative occurrence frequencies of failure mechanisms causing water-related claims ($n = 3126$). An asterisk next to a bar indicates a residual group: a group of claims for which exact failure mechanisms could not be derived from communication transcripts. Percentages are based on the number of claims in the non-residual groups.

to precipitation or not. These unknowns include claims caused by water discharges from neighbouring properties and groundwater flooding. In particular for groundwater flooding, insufficient information was available to distinguish between floods as a result of persistent rainfall or because of sudden wall failures not related to rainfall. For insurers, there is no strong need to collect information on the actual cause of groundwater flooding as they do not compensate for this type of flood (see also Table 5.2).

Top of Fig. 5.3 shows the yearly distribution of precipitation-related claims (white), non-precipitation-related claims (grey) and claims for which the cause was unknown (black), for the years 2007–2013. There is an increase in the number of claims through the years. This increase is most apparent between 2008 and 2010 for precipitation-related claims and between 2009 and 2012 for non-precipitation-related claims. Possible explanations of these trends are discussed in Sect. 5.4. Most precipitation-related claims are recorded in July–August and December–January (bottom of Fig. 5.3); the December–January claims can partly be explained by the damage due to melting

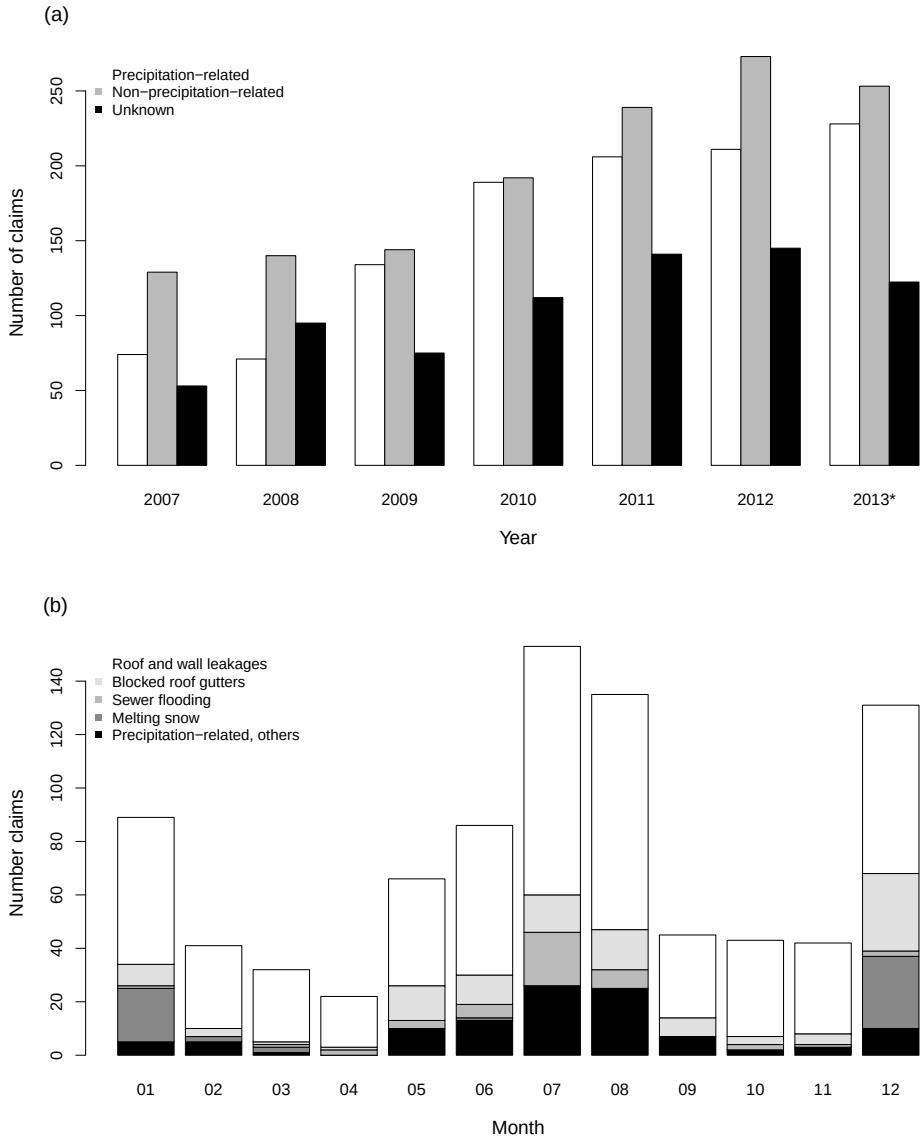


Figure 5.3: Yearly and monthly distribution of the number of claims: **(a)** The yearly distribution of precipitation-related claims (white), non-precipitation-related claims (grey) and claims for which the cause was unknown (black), for the years 2007–2013. The number of claims related to claims with unknown damage cause are depicted by the black bars. The values of the year 2013 (denoted with an asterisk) are estimated, because no data was available for the months November and December; **(b)** The monthly distribution of precipitation-related claims for the years 2007–2012, per cause class. The year 2013 is excluded because data was not available for the entire year.

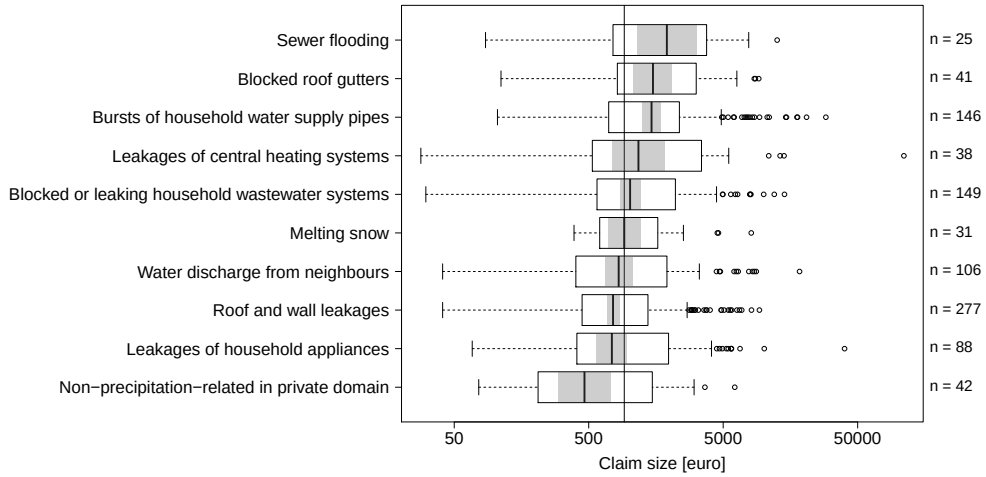


Figure 5.4: Distribution of claim sizes associated with various failure mechanisms. Claim size is the sum of property and content damage. Only risk addresses are included for which both property and content insurance were available. Results are only shown for failure mechanisms with at least 20 claim records. The grey rectangles display the 95 % confidence interval around the median. If the grey rectangles of two boxplots do not overlap, there is a strong indication that the median are statistically different. The vertical solid line represent the median claim size. The number of claims (n) within each class are given next to the boxplots.

snow. Claims related to sewer flooding mainly occur in June–August.

Although claims related to sewer flooding were less present in the data, they are associated with significantly larger claim sizes (1150–3160 euros, based on the 95 % confidence interval around the median in Fig. 5.4) than claims generated by roof and wall leakages (680–840 euros), the majority class. Sewer floods are costly because of the required (chemical) cleaning of sewage spills and replacement of goods that cannot be cleaned properly. In contrast, costs related to roof and wall leakages, which usually do not involve large water volumes, are relatively low and limited to the repair of the leak and the painting of walls and ceiling.

Based on a qualitative analysis of outliers, it was found that exceptionally large claim sizes are related to cases where water leakage could not be stopped easily (e.g. burst of water supply pipe just outside property), flooding occurred while no one was at home or temporary housing was required.

5.3.2 Effects of rainfall intensity on claim occurrence probability

In Fig. 5.5, the empirical probability of precipitation-related claim occurrence per day per risk address is shown, as a function of the rainfall intensity (black dots, based on 1031 claims). Within the subset of precipitation-related claims, a further distinction was made between the occurrence probability of claims caused by failure of systems in the private domain (grey dots, 876 claims) and the public domain (light grey dots, 89 claims), according to column 5 ("Domain") of Table 5.3. The empirical probability is calculated as follows: within a bin, with a size of 5 % of the range of x values,

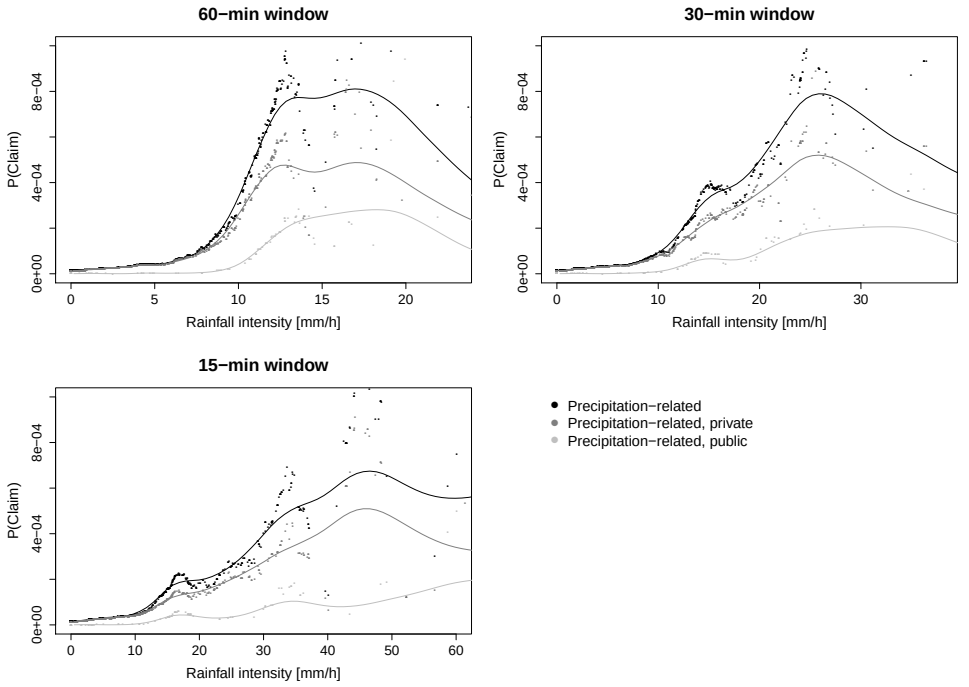


Figure 5.5: The empirical probability of precipitation-related claim occurrence per day per risk address as a function of rainfall intensity, using a 60-min (top left panel), 30 min (top right panel) and 15 min moving time window (bottom left panel). Results are related to precipitation-related claims (black dots), broken down to those classified as "private" (grey dots) and "public" (light grey dots). The range of x values is from 0 mm h^{-1} to the rainfall intensity associated with a return period of 5 year. Locally-weighted regression lines are based on penalized B-splines.

the number of successes (i.e. claims) are divided by the sample size (i.e. number of combinations of days and risk addresses that can generate claims). Empirical probabilities are evaluated at each x value that corresponds with a claim.

The top-left plot of Fig. 5.5, based on a 60-min window, shows that the occurrence probability of precipitation-related claims increases with increasing rainfall intensity and that it increases considerably when rainfall intensity exceeds 7–8 mm h^{-1}. For events with rainfall intensities smaller than 7–8 mm h^{-1}, the occurrence probability of precipitation-related claims is mainly determined by failure processes in the private domain, which are primarily roof and wall leakages. Thus, damage due to roof and wall leakage already occur at small rainfall intensities, which suggest that leaks may be latent before first observed during a rainfall event. For rainfall events that exceed the 7–8 mm h^{-1} threshold, failure processes in the public domain start to contribute substantially to the overall occurrence probability. Similar conclusions can be drawn from the other two plots related to a 30-min (top-right) and a 15-min window (bottom-left), with the difference that rainfall threshold shift to 9–10 and 12 mm h^{-1} respectively. The locally-weighted regression lines reveal that relationships using the 30-min and 15-min window have a less linear nature than the ones based on a 60-min

Table 5.5: Goodness-of-fit measures of logistic regression models.

		Claims related to failure of private systems			Claims related to failure of public systems		
	Model	Likelihood ratio	d.f.	Pseudo-R^2	Likelihood ratio	d.f.	Pseudo-R^2
1	$\max_{60}$	134.97***	1	0.033	57.72***	1	0.094
2	$\max_{60} + \text{vol}$	253.66***	2	0.063	112.68***	2	0.183
3	$\max_{60} + \text{vol} + \text{temp}$	261.95***	3	0.065	114.24***	3	0.186
4	$\max_{60} + \text{vol} + \text{temp} + \text{seas} + \text{wind}_\text{d}$	276.11***	7	0.069	120.4***	7	0.196
5	$\max_{60} + \text{vol} + \text{temp} + \text{seas} + \text{wind}_\text{h}$	284.49***	7	0.071	121.01***	7	0.197
6	$\max_{60} + \text{vol} + \text{temp} + \text{seas} + \text{wind}_\text{g}$	302.71***	7	0.075	122.16***	7	0.199
	comparison models 2–1	118.69***	1		54.96***	1	
	comparison models 3–2	8.29**	1		1.56	1	
	comparison models 4–3	14.16**	4		6.16	4	
	comparison models 5–3	22.54***	4		6.77	4	
	comparison models 6–3	40.76***	4		7.92	4	

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

Note: because the *relogit* routine does not report goodness-of-fit statistics, statistics are based on the ordinary logistic regressions.

window.

5.3.3 Logistic regression results

Logistic regression analyses were performed to test the significance of various combinations of explanatory variables in explaining the occurrence probability of precipitation-related claims. Separate analyses were made for the occurrence of claims caused by failures of systems in the public and private domain (according to column 5 in Table 5.3). From the three variants of maximum rainfall intensity, the one based on a 60-min window was used for modelling. Regression coefficients were estimated based on the data in the rainfall intensity range of 5 to 12 mm h^{-1} (60-min window). Data associated with 12 mm h^{-1} or larger are scarce and are, therefore, likely biased towards single rainfall events. In a first attempt to fit a logistic regression model to data in the range of 0–12 mm h^{-1}, it was found that much weight was given to the data in the range of 0–5 mm h^{-1}, resulting in a poor fit to data in the higher rainfall intensity range. Possibly, claims associated with rainfall intensities of 0–5 mm h^{-1} are generated by a different process than the claims associated with rainfall intensities larger than 5 mm h^{-1}. More on this can be read in Sect. 5.4.

The goodness-of-fit measures of the various models, including a comparison of the likelihood ratio statistics between models are summarised in Table 5.5. The models that combine maximum rainfall intensity and rainfall volume result in better fits compared to the intercept-only models. Maximum temperature, wind parameters and season significantly improve the model fit for claims caused by failures of private systems, but not for claims caused by failures or public systems. Most of the explanatory power derives from maximum rainfall intensity and rainfall volume. Of all wind parameters, wind gust has best explanatory power.

Table 5.6 lists the estimates of the regression coefficients for the two models that include all explanatory variables (using wind gust as wind parameter), further referred to as the "private model" and the "public model". The categorical variable "season" was modelled as four separate binary variables, where one level was dropped to avoid multicollinearity. The summer season was found to positively correlate with the occurrence of claims related to failure of private systems. Moreover, regression

Table 5.6: Estimates of regression coefficients of the rare event logistic regression models.

	Claims related to failure of private systems		Claims related to failure of public systems	
	β (SE)	exp(β) (95 % C.I.)	β (SE)	exp(β) (95 % C.I.)
(Intercept)	-14.841^{***} (0.605)		-22.263^{***} (3.504)	
Maximum rainfall intensity (60-min)	0.192^{***} (0.046)	1.21 (1.16–1.27)	0.479^{*} (0.191)	1.61 (1.33–1.95)
Rainfall volume	0.048^{***} (0.006)	1.05 (1.04–1.06)	0.053^{**} (0.017)	1.05 (1.04–1.07)
Maximum wind gust	0.082^{***} (0.019)	1.09 (1.06–1.11)	0.154 (0.084)	1.17 (1.07–1.27)
Maximum temperature	0.083^{***} (0.019)	1.09 (1.07–1.11)	0.156 (0.080)	1.17 (1.08–1.27)
Season: spring	0.095 (0.265)	1.10 (0.84–1.43)	0.211 (0.890)	1.23 (0.51–3.00)
Season: summer	0.521^{*} (0.242)	1.68 (1.32–2.15)	0.319 (0.825)	1.38 (0.60–3.14)
Season: winter	0.324 (0.251)	1.38 (1.08–1.78)	0.054 (0.976)	1.05 (0.40–2.80)
Likelihood ratio χ^2	302.71		122.16	
d.f.	7		7	
p value	0.000		0.000	
Pseudo-R^2	0.075		0.199	

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

Note: standard error (SE) of estimate is given between brackets; the upper and lower bound of the 95 % confidence interval (C.I.) are exp($\beta \pm 1.96$ SE), assuming normality on the log odds scale.

analysis revealed that the regression coefficient of the maximum rainfall intensity is larger for the public model than for the private model, which means that rainfall intensity more strongly affects the claim occurrence by failures of public systems than private systems. The odds ratio (exp(β)) related to maximum rainfall intensity varies between 1.16–1.27 for the private model and 1.33–1.95 for the public model, which means a 16–27 % and a 33–95 % increase in odds for each mm h^{-1} change in rainfall intensity, for private and public model respectively.

5.4 Discussion

Based on the insurance data for the case study in Rotterdam, a distinct rainfall intensity threshold could be defined above which failures of public systems start to contribute considerably to the occurrence of damage claims (Fig. 5.5). Interestingly, this threshold of 7–8 mm h^{-1} (based on a 60-min window) is not in line with the design standards of sewers in the Netherlands. Dutch sewers are designed to cope with rainfall intensities of 20 mm h^{-1}, which is associated with an event return period of approximately 2 years (see also Sect. 2.4). This suggest that the threshold relates to some other damaging process than simply overloading of sewer systems, for example, blockages in sewer pipes or malfunctioning of non-return valves in sewer laterals. On closer inspection of the communication transcripts of claims labelled as "sewer flooding", it was found that most cases relate to sewer backups from toilets or floor drains and to a lesser extent to run-off entering buildings at ground level. Still, communication transcripts were inconclusive about the extent to which these claims were related to overloaded sewer systems or failure of system components.

Findings of present work have implications for pluvial flood risk management. The return period of design storms as currently being used to design sewer systems in the Netherlands is largely based on political consensus. Potentially the results presented here can be used to obtain an objective design criterion based on risk assessment. Furthermore, this chapter provides insights into contributions of urban drainage systems to flood damage at city level. Results will support urban water managers in the evaluation of urban drainage system capacity and decisions about

the need for and prioritisation of investment to increase drainage capacity. Further research is needed to explain why damage related drainage capacity occurs below the level of design capacity; this will help water managers to focus efforts on ensuring that their systems reach design capacity.

Results of this chapter have practical relevance for insurers. From present case study, it became evident that the majority of the water-related claims are caused by roof and wall leakages. Thus, damage prevention programmes focussing on these causes may be helpful. When it is raining heavily ($>$ 7–8 mm in a 60-min window) insurers can expect more claims related to sewer flooding that require special services for the cleaning of sewer spills.

In the higher range of rainfall intensities in Fig. 5.5, relationships between rainfall intensity and claim occurrence probability become less distinct, which can partly be explained by the limited amount of claim data associated with extreme rainfall events. The present insurance database covers almost seven years of claim data (2007–October 2013), where around 80 % of the precipitation-related claims relate to rainfall events with return periods smaller than 2 year. Around 10 % of the claims can be attributed to two exceptional rainfall event with a return period of 14–18 year. As a consequence, empirical probabilities in the higher range of rainfall intensities are unreliable and biased towards single rainfall events.

Claims associated with rainfall intensities of 0–5 mm h^{-1} in Fig. 5.5 (60-min window) are possibly generated by a different process than the claims related to rainfall intensities larger than 5 mm h^{-1}. It maybe the case that more specific damage processes can be distinguished within the existing cause classes. For example, the class "roof leakages" may contain two processes; one related to the presence of latent leaks that are first observed when it is raining and another one related to the exceedance of the "hydraulic capacity" of roofs. The hypothesis could not be tested based on present database, because it lacked information to distinguish between the sub-processes.

In the top panel of Fig. 5.3, an increasing trend is observed in the number of water-related claims in the period 2007–2013. There are a number of possible explanations for this trend. To start with, the number of policyholders may have increased in time. This could not be verified, because in present study only policyholder data were available for a single snapshot in time. Another explanation may be related to bursts of household water supply pipes, which is the most frequent cause of non-precipitation-related claims. Based on an unpublished report, the insurance company has observed a substantial increase in defects in water supply systems in the recent years, mainly because of incorrectly installed compression fittings. Other explanations that may be worthwhile to investigate are differences in climate variables between years and the effect of 2007–2008 financial crisis on the claiming behaviour of people.

There are a number of aspects with regard to uncertainty in insurance data. The occurrence of claims that relate to causes that are not covered by insurers (e.g. groundwater flooding) are probably underestimated by the data, simply because people may be aware of the fact that damage is not covered and, thus, not make a claim. Moreover, the reported claim date may not always be the date on which the damage occurred, for example, because the exact damage data is unknown, which may be the case when people are on holidays. Furthermore, addresses of insured are based on static policyholder information, i.e. situation on a snapshot in time (reference date: 31

July 2013). Errors in addresses may occur if policyholder information has changed in time (e.g. policyholders moving to another address).

Failure of public systems (e.g. sewer system) will probably mostly affect buildings that occupy ground floor. In this study, no distinction was made between terraced or detached houses and high-rise buildings (i.e. houses that occupy first floor or higher). As a consequence, claim occurrence probabilities related to failure of public systems is likely higher than the probabilities estimated in present study, which is based on all building types.

5.5 Conclusions and recommendations

The main goal of this chapter was to investigate the relative contributions of different failure mechanisms to the occurrence of rainstorm damage to building structure and content, as well as the extent to which the probability of occurrence of these failure mechanisms relate to weather variables. For this purpose, a property level home insurance database of around 3100 water-related damage claims was analysed, for a case study in Rotterdam, the Netherlands.

The results of this investigation show that leakage of roofs and walls is the most frequent failure mechanism causing precipitation-related claims, followed by blocked roof gutters, snow melting under roof tiles and sewer flooding. Although claims related to sewer flooding were less present in the data, they are associated with significantly larger claim sizes (1150–3160 euros, 95 % confidence interval around the median) than claims generated by roof and wall leakages (680–840 euros), the majority class. Rare events logistic regression analysis revealed that maximum rainfall intensity and rainfall volume are significant predictors for the occurrence probability of precipitation-related claims. Moreover, it was found that claims associated with rainfall intensities smaller than 7–8 mm in a 60-min window are mainly caused by failures of systems in the private domain, such as roof leakages and blocked roof gutters. For rainfall events that exceeds the 7–8 $\mathrm{mm\,h^{-1}}$ threshold, failure of systems in the public domain, such as sewer systems, start to contribute considerably to the overall occurrence probability of claims. The communication transcripts, however, lacked information to be conclusive about to extent to which sewer-related claims were caused by overloading of sewer systems or failure of system components.

It is worthwhile to investigate spatial distributions of water-related claim data in a future study, considering local conditions such as building age and type and percentage of impervious area. An important limitation of this study is that the number of claims associated with extreme rainfall events was relatively small. Given the fact that manual classification took considerable amount of work, it is recommended to explore methods to automate and standardize classification of claim data, with the aim to process more data in future analyses.

CHAPTER 6

Conclusions and recommendations

The general objective of this thesis was to explain variability in rainstorm damage based on statistical analyses of home insurance data and a wide range of explanatory data, including weather, building-related, topographic and socioeconomic data. In this thesis, rainstorm damage is defined as damage that results from pluvial flooding or rainwater intrusion through defects in the building envelope. In previous chapters results of the statistical analyses have been presented, based on two home insurance databases provided by the Dutch insurance industry: a district-aggregated, nationwide database and a detailed, property level database for a case study in Rotterdam. In this chapter general conclusions are drawn and recommendations for insurance practice and further research are given.

6.1 Conclusions

1. Scientific studies that have analysed rainstorm damage data are scarce, which hampers the development of prediction models for rainstorm damage. Based on literature review (Chapter 1), the following is concluded. Many authors, active in research areas related to different kinds of weather-related hazards, recognize that damage data are generally lacking or incomplete. This is limiting the understanding of damage mechanisms and, therefore, the development of damage models. This is especially true for rainstorms damage: little research focused on the collection of rainstorm damage data, possibly because rainstorm damage is relatively small and too localised to trigger authorities and homeowners to report damage. Moreover, rainstorm damage is generally lower, on an event basis, than damage from other hazard events such as river flooding, and therefore less disruptive for society. Furthermore, damage databases, such as those from insurers or national health services, are hard to access because of strict privacy regulations. As a con-

sequence, there is currently no strong foundation for the development of prediction models for rainstorm damage.

2. A promising source of rainstorm damage data are insurance databases; however, information from insurance databases is prone to human errors and usually incomplete for scientific purposes and, therefore, substantial efforts are required to validate and classify data and to collect complementary data. Findings of related studies and the analyses of two home insurance databases show that information about hazard characteristics, damage causes and building-related and socioeconomic variables is not, or only limitedly, available in insurance databases or cannot easily be retrieved from insurers' data archives. When available, information is sometimes stored in an unstructured way that requires substantial data classification efforts before data can be used for scientific analysis. Moreover, data validation is required to check for incorrect and missing claim information, such as the amount of compensation and damage date. Furthermore, complementary data need to be collected and appropriately converted to formats that allow analysis of variables influencing damage. This process calls for a considerable amount of work and, moreover, depends on the availability and quality of other data sources.

3. A relatively small number of rainstorm damage claims relate to public system failure, such as sewer flooding and depression filling. The majority of claims are caused by water intrusion due to defects in the building envelope. Data analysis of home insurance data for a case study in Rotterdam for the period 2007–2013 revealed that leakage of roofs and walls is, by far, the most frequent failure mechanism generating precipitation-related claims, followed by blocked roof gutters, snow melting under roof tiles and sewer flooding (Chapter 5). Although claims related to sewer flooding are rare, they are associated with significantly larger claims sizes (1150–3160 euro, based on 95%-confidence interval around the median) than claims generated by roof and wall leakages (680–840 euro), which can partly be explained by the required cleaning of sewer spills. With a small sample size, outcomes must be interpreted with caution, as the findings are mainly related to minor rainstorms with short return periods.

4. Statistical analysis of home insurance damage data shows that public system failures contribute to the occurrence of damage claims, starting from a minimum rainfall intensity threshold. Based on a case study in Rotterdam (Chapter 5), it was found that claims associated with rainfall intensities smaller than 7–8 mm in a 60-minute window are mainly related to failure processes in the private domain, such as roof and wall leakages. For rainfall events that exceed the 7–8 mm h^{-1} threshold, failure of systems in the public domain, such as sewer systems, start to contribute considerably to the overall occurrence probability of claims. Interestingly, this threshold is not in line with the design standard of sewers in the Netherlands. Dutch sewers are designed to cope with rainfall intensities of approximately 20 mm h^{-1} (see also Sect. 2.4). The communication transcripts lacked information to be conclusive about the extent to which sewer-related claims were caused by overloading

of sewer systems or failure of system components.

5. Of all weather variables studied, the maximum hourly rainfall intensity is the most important predictor for the occurrence of rainstorm damage claims. The risk of a claim is made up of two components: claim probability and claim size. Using logistic regression and decision-tree analysis (Chapter 2, 4 and 5), the most important predictor for claim probability proved to be the maximum hourly rainfall intensity during an event, followed by storm event total rainfall volume. Still, a considerable fraction of the variance was left unexplained. No meaningful relationships were found between weather variables and claim size. The advantage of weather radars over rain gauge networks as a source of rainfall data is that they provide better spatial coverage even if the accuracy of rainfall estimates is lower than for rain gauges. Dense rain gauge networks are lacking in cities, while the majority of claims occur in cities (Chapter 2 and 3).

6. A number of relationships between building-related and socioeconomic variables and claim probability are significant, some of them only exist locally within subgroups of damage claim data. Decision-tree analyses revealed that, besides maximum rainfall intensity, other statistically significant variables for claim probability are real estate value, ground floor area, building age, building type (i.e. low-rise or high-rise), household income and ownership structure (Chapter 4). The role of topographic variables remains unclear. The identified variables can be used to focus future data collection efforts.

7. Decision trees are better able to capture local characteristics in damage claim data than global regression models. Many variables influence the occurrence of rainstorm damage claims, most of them are to some extent intercorrelated (e.g. maximum rainfall intensity and rainfall volume, see Chapter 3; real estate value and household income, see Chapter 4). Decision trees outperform global regression models in terms of explained variance due to intercorrelations and threshold behaviour (Chapter 4). Moreover, decision trees were found to support intuitive understanding, because of their hierarchical presentation. The decision-tree approach asks for fewer assumptions about the data; a disadvantage of this method is, however, that it requires many records to sufficiently populate tree subgroups. This may be inconvenient as rainstorm damage events are rare and only small subsets of claim data may be available. Alternatively, a rare events logistic regression model can be applied to insurance damage data when the number of claims is limited and claim probability is low (Chapter 5).

6.2 Recommendations for insurance practice

1. The findings of this study have a number of implications for Dutch insurance practice with regard to the current "rainfall clause" (see also Sect. 2.2.2). Presently, insurers use rain gauge data to verify the validity of rainstorm damage claims and to check compliance with the "rainfall clause" that states that

pluvial flood damage is only covered when "rainfall intensity is higher than 40 mm in 24 h, 53 mm in 48 h or 67 mm in 72 h at or near the location of the damaged property". The density of operational rain gauge networks is too low to capture local rainfall characteristics especially for convective rainstorms, so rainfall data from rain gauges may not be representative for the conditions experienced by the insured. This may result in a case where the insured experiences damage without being compensated or a case where the insured may be compensated without damage. A point that was also made in a Nature article from 1911 titled "Insurance Against Rain" (1911), where an unknown author discusses the limitations of a holiday insurance scheme against rainfall in England. A combination of weather radar and rain gauge measurements will likely give more accurate results. Moreover, the "rainfall clause" is not meaningful in the context of urban drainage systems. Sewers are designed to cope with 40 mm of rainfall as long as the rainfall volume is not concentrated in a short time window (∼ 1–2 hours). As a consequence, insured may experience damage from pluvial flooding even if daily rainfall volumes are below 40 mm or may not if the daily volumes exceed the 40 mm threshold (and rainfall being evenly distributed throughout the day). Results provide evidence that short-duration intense rainfall, with relative small volumes, already results in considerable number of rainstorm damage claims, which may be partially attributes to public systems failures. Taken together, it is recommended to consider rainfall criteria in the "rainfall clause" that more closely match time scales of pluvial flooding in cities.

2. It is recommended to investigate to what extent a more detailed, more systematic classification of damage causes can help insurers to improve their customer services and business efficiency. The aims with which insurers collect information of damage events are, for instance, to quickly and efficiently handle claims and to improve services to customers. Investments to collect more detailed information about the actual causes of claims comes with a price: more efforts need to be put into data collection. For example, insurers' call centre employees have to pose more specific questions to the insured to determine the actual damage cause. Moreover, longer lists of damage causes to select from, may become unclear to the user, resulting in interpretation errors and, thus, a reduced quality of data. Therefore, insurers need to consider both costs and benefits before collecting additional data. A benefit of more detailed information about damage causes is that it provides a better understanding of insurance risks, and thus helps to better assess the potential effects of preventive measures. In recent publications, the Dutch Association of Insurers argue that the current availability and level of detail of damage information is amendable (Hoen and Van Leeuwen, 2012). They have proposed a new cause classification scheme for private and business insurance that can be advantageous for insurance practice (Hoen and Van Leeuwen, 2012, 2013). In this thesis, the Rotterdam case study has shown that classifying claim data based on communication transcripts from insurers' call centres provides more accurate information on the actual damage causes and, thus, on how an insurance portfolio is built up. Claim classification allows the analysis of the most important damage contributors to an insurance portfolio, which can help to promote damage prevention measures among insured more effectively and enhance service to customers. For instance, the large percentage of claims related to roof

leakages suggests that prevention programmes targeted at reducing roof leakages can be helpful. Another reason to classify claim data is that this can help to send out damage experts to customers more quickly and more efficiently. For example, when it is raining intensely, insurers can expect more claims related to public system failures. This type of claims require specialized expertise, such as the cleaning of sewer spills. The aforementioned examples from the Rotterdam case may not be applicable to other cities that are, for instance, situated in sloped areas or are different in terms of urban fabric and urban drainage system; see also Sect. 6.3 on the applicability of results to other regions.

3. It is suggested to adjust staffing capacity of insurers' call centres based on forecasts and nowcasts of weather conditions. Some insurers have indicated that the staffing of their call centres during extreme events is an issue. Findings from this thesis can be used to make a better estimate about the number of expected claims based on storm event maximum rainfall intensity. To this end, a combined use of weather radar and rain gauge measurements can help to quickly assess the severity of damage events and improve estimates of claim frequency.

4. It is worthwhile to study if the insurance industry can benefit from standardizing and automating data formats and data collection methods. From a scientific perspective, a standardized and automated approach can result in less data distortion and allows a better comparison between databases of insurers; a point that has also been emphasized by André et al. (2013). Data distortion may occur, for instance, because of insurers using different terminology, claim coding systems, and softwares (used to store and process data), and having different policies towards claim compensation. A good example is the definition of "extreme rainfall" which can vary greatly from country to country (see e.g. Garne et al., 2013). Another example is how an insurer treats the date being assigned to a claim, which can be the actual damage data or the date on which the claim was made or paid out. Even within an insurance company, comparability of damage databases may be hampered because of different coding systems being used in different units of the organisation, software updates, and changes in insurance policies. From the insurers' point of view, standardisation can limit competitive advantages: as soon as all insurance companies are bound to collect the same set of variables, these variables cannot be used to outperform competitors. Without standards, insurers can make the trade-off between collecting more data at higher costs versus having more knowledge that can give a competitive advantage (A. Hoen, personal communication, 18 August, 2014). A decision to standardise data and methods should therefore always be accompanied with detailed analyses of costs and benefits. The present work indicates that predictive power of damage models can potentially be improved if high-quality contextual and damage data are available at scales close to that of individual properties. Standardisation is a way to possible achieve this goal. In this context, it is recommended to provide metadata of damage databases that describe the insurance policy conditions under which data have been collected, such as the kinds of water damage that are covered, how damage is being assessed and definitions of terminology being used. Moreover, in present study a number of significant variables were found that explain

claim probability, like maximum hourly rainfall intensity and real estate value. It is worthwhile to investigate the extent to which standardising these variables can help to improve the efficiency of damage prevention programmes, and thus reduce portfolio risks. Furthermore, the cause classification scheme proposed in Chapter 5 can, as a start, be used to consider improvements to existing coding systems. Alternatively, conversion tables can be developed to convert between different coding systems, although this may introduce additional interpretation errors. Complementary to data standards, the storage and processing of data can be automated. A possible benefit of an automated processing of claim data is that validated data is more quickly available, allowing the insurer to make quick damage assessment during and in the aftermath of a rainfall event. Related to that, the use of weather nowcasts and forecasts from weather radar data can potentially be automated to verify rainstorm damage claims instantly as they occur.

6.3 Recommendations for further research

1. The results of current work on rainstorm damage data demonstrate the potential for better understanding of damage mechanisms using insurance claim data. Similar studies can be conducted for other damage types, such as storm and hailstorm damage. Future studies can also investigate the co-occurrence of different kinds of weather-related hazards and their relationships with weather variables. This study focused on home insurance data, whereas also data from other industry domains can be considered, such as crop, car and business insurance. Eventually, a full weather-related damage prediction model could be developed incorporating a wide range of weather variables and taking into account interdependencies between weather-related hazards.

2. There is a need to investigate the applicability of present findings to regions that are different than the Netherlands in terms of topography, urban drainage systems, urban fabric, and insurance. This thesis is based on data from Dutch insurers only; results are therefore subject to contextual biases. The Netherlands is exceptionally flat compared to many other European countries. Floods from heavy rainfall are therefore typically characterized by flood depths up to a few decimetres and limited surface run-off. As a result, the insurance data used in this thesis do not cover a large range of flood depths and, thus, do not reflect the damage processes that involve large water volumes. Because of the minor damage per flood event, higher flood frequencies are generally accepted in the Netherlands (Ten Veldhuis, 2010). Dutch urban drainage systems are designed to cope with rainfall events with a return period of around two years. Rainfall thresholds found in this study may therefore be specific to characteristics of Dutch urban drainage systems. Moreover, due to high flood safety standards in the Netherlands, flooding from primary and secondary river systems is rare and, as a consequence, co-occurrence of river and pluvial flooding is rare too. This is not always the case in other European countries, where interactions between river and pluvial flooding is more common. Present results also depend to some extent on the characteristics of the buildings

under study. They may not hold if different building types and construction methods are considered than the buildings in present database. Findings can also be subject to systematic biases that arise from differences between insurers and insurance policies across regions and countries. Garne et al. (2013) shows that among Nordic insurers different rainfall thresholds are being used to define "extreme rainfall", which may lead to different censoring of damage data. Moreover, river flooding is not commonly covered by home insurance policies in the Netherlands, whereas in some other European countries private insurers provide insurance against river flooding (Botzen and Van den Bergh, 2008; Seifert et al., 2013). This is one of the reasons that damage related to this type of flooding was not part of the analyses in this thesis.

3. More research can be carried out to investigate ways to automate classification of insurance claim data. The manual classification of claims calls for a considerable amount of work. Further research needs to explore the applicability of text recognition algorithms to insurance data to automate the classification process and reduce classification efforts considerably. An automated system enables analyses to be done on larger data sets that cover different cities and regions and more extreme rainfall events than those covered in current work. The challenge will be to train algorithms to retrieve information from telegram style texts containing much professional jargon and abbreviations.

4. Further research needs to explore other multiparameter statistical approaches to model insurance damage data, as well as ways to reduce computational efforts. It is expected that non-linearity and intercorrelation effects occur frequently in insurance damage data. In this thesis, decision-tree techniques were tested to account for these behaviours and the large number of variables involved with rainstorm damage. This approach, however, is computationally demanding and therefore currently impractical when dealing with large damage databases. The need to consider multiparameter statistical models to analyse damage data was also emphasized by Thieken et al. (2005); Merz et al. (2013). Because of the rarity of damage events it is also recommended to further explore methods that can deal with rare events data, such as rare events logistic regression that was used in Chapter 5. In this thesis, possible biases related to spatial clustering in data were not considered. More sophisticated statistical models that account for spatial correlations can improve damage functions. Furthermore, a number of studies found decorrelation lengths for short-duration rainfall that are shorter than the lengths assumed in this thesis (Moreau et al., 2009; Janssen et al., 2013; Gregersen et al., 2013). Future research with high-resolution spatial data is required to investigate if this can possibly explain the low explanatory power of the damage models that were found in this thesis.

5. It is recommended to study the effects of local urban drainage characteristics, such as sewer capacity, sewer type, soil type and percentage of impervious area, and topography on the claim probability. This type of information is often not available at a nationwide scale, it is, therefore, recommended to use a case study design for a city or region where high-quality geographic data are available and of sufficient size to collect a meaningful data set.

6. A future study investigating the factors influencing claim size is recommended. Current work failed to identify explanatory factors for claim size. Possibly, a research on this subject requires an interview survey among affected households to collect data on building-specific and household-specific variables, such as door threshold level, floor type and age, presence of a basement, level of precaution, presence of building occupant at the time of the damaging event and level of self-reliance of the homeowner.

References

André, C., Monfort, D., Bouzit, M., and Vinchon, C. (2013). Contribution of insurance data to cost assessment of coastal flood damage to residential buildings: insights gained from Johanna (2008) and Xynthia (2010) storm events. *Natural Hazards and Earth System Science*, 13(8):2003–2012, doi:10.5194/nhess-13-2003-2013. (cited on pp. 2, 3, 5, 36, 64, 85).

Arnbjerg-Nielsen, K., Willems, P., Olsson, J., Beecham, S., Pathirana, A., Bülow Gregersen, I., Madsen, H., and Nguyen, V.-T.-V. (2013). Impacts of climate change on rainfall extremes and urban drainage systems: a review. *Water Science and Technology*, 68(1):16–28, doi:10.2166/wst.2013.251. (cited on p. 2).

Ashley, R. M., Balmforth, D. J., Saul, a. J., and Blanskby, J. D. (2005). Flooding in the future–predicting climate change, risks and responses in urban areas. *Water Science and Technology*, 52(5):265–73. (cited on p. 2).

Barnett, B. J. and Mahul, O. (2007). Weather Index Insurance for Agriculture and Rural Areas in Lower-Income Countries. *American Journal of Agricultural Economics*, 89(5):1241–1247, doi:10.1111/j.1467-8276.2007.01091.x. (cited on p. 4).

Berne, A., Delrieu, G., Creutin, J.-D., and Obled, C. (2004). Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology*, 299(3-4):166–179, doi:10.1016/j.jhydrol.2004.08.002. (cited on p. 14).

Blanc, J., Hall, J., Roche, N., Dawson, R., Cesses, Y., Burton, A., and Kilsby, C. (2012). Enhanced efficiency of pluvial flood risk estimation in urban areas using spatial-temporal rainfall simulations. *Journal of Flood Risk Management*, 5(2):143–152, doi:10.1111/j.1753-318X.2012.01135.x. (cited on p. 36).

Borisov, A. (2009). Zero-Inflated Boosted Ensembles for Rare Event Counts. In *Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis*, pages 227–228, Lyon, France. Springer. (cited on p. 60).

Bortoluzzo, A. B., Claro, D. P., Ceatano, M. A. L., and Artes, R. (2011). Estimating total claim size in the auto insurance industry: A comparison between tweedie and zero-adjusted inverse Gaussian distribution. *Brazilian Administration Review*, 8(1):37–47. (cited on p. 3).

Botzen, W. J. W., Bergh, J. C. J. M., and Bouwer, L. M. (2009). Climate change and increased risk for the insurance sector: a global perspective and an assessment for the Netherlands. *Natural Hazards*, 52(3):577–598, doi:10.1007/s11069-009-9404-1. (cited on p. 6).

Botzen, W. J. W., Bouwer, L. M., and Van den Bergh, J. C. J. M. (2010). Climate change and hailstorm damage: Empirical evidence and implications for agriculture and insurance. *Resource and Energy Economics*, 32(3):341–362, doi:10.1016/j.reseneeco.2009.10.004. (cited on p. 4).

Botzen, W. J. W. and Van den Bergh, J. C. J. M. (2008). Insurance against climate change and flooding in the Netherlands: present, future, and comparison with other countries. *Risk Analysis*, 28(2):413–26, doi:10.1111/j.1539-6924.2008.01035.x. (cited on p. 87).

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Wadsworth, Belmont, California. (cited on pp. 47, 48, 49).

Brunetti, M. T., Peruccacci, S., Rossi, M., Luciani, S., Valigi, D., and Guzzetti, F. (2010). Rainfall thresholds for the possible occurrence of landslides in Italy. *Natural Hazards and Earth System Science*, 10(3):447–458, doi:10.5194/nhess-10-447-2010. (cited on p. 1).

Busch, S. (2008). Quantifying the risk of heavy rain: its contribution to damage in urban areas. In *Proceedings of the 11th International Conference on Urban Drainage*, Edinburgh, Scotland, UK. (cited on pp. 2, 5, 64).

Caradot, N., Granger, D., Chapgier, J., Cherqui, F., and Chocat, B. (2011). Urban flood risk assessment using sewer flooding databases. *Water Science and Technology*, 64(4):832, doi:10.2166/wst.2011.611. (cited on p. 5).

Castañeda Vera, A., Barrios, L., Garrido, A., and Mínguez, I. (2014). Assessment of insurance coverage and claims in rainfall related risks in processing tomato in Western Spain. *European Journal of Agronomy*, 59:39–48, doi:10.1016/j.eja.2014.05.005. (cited on pp. 4, 6).

Changnon, S. A., Changnon, D., Fosse, E. R., Hoganson, D. C., Sr, R. J. R., and Totsch, J. M. (1996). Effects of Recent Weather Extremes on the Insurance Industry: Major Implications for the Atmospheric Sciences. *Bulletin of the American Meteorological Society*, pages 425–435. (cited on p. 6).

Changnon, S. A., Pielke Jr, R. A., Changnon, D., Sylves, R. T., and Pulwarty, R. (2000). Human Factors Explain the Increased Losses from Weather and Climate Extremes. *Bulletin of the American Meteorological Society*, 81(3):437–442. (cited on p. 3).

Cheng, C. S. (2012). Climate Change and Heavy Rainfall-Related Water Damage Insurance Claims and Losses in Ontario, Canada. *Journal of Water Resource and Protection*, 04(02):49–62, doi:10.4236/jwarp.2012.42007. (cited on pp. 2, 5, 6, 24, 37, 64).

Chiocchio, C., Iovine, G., and Parise, M. (1997). A proposal for surveying and classifying landslide damage to buildings in urban areas. In *Proc. Int. Symp. Engineering Geology and the Environment*, pages 553–558, Athens. (cited on p. 36).

City of Rotterdam (2011). Sewer Plan Rotterdam 2011-2015 (in Dutch), URL: http://www.rotterdam.nl/GW/Document/Waterloket/GRP%20rapport%202011-2015%20juni2011.pdf. Technical report. (cited on p. 65).

City of Rotterdam (2014). Open Data Centre Rotterdam: Shape file of Rotterdam's sewer sytem. (cited on p. 65).

Climate Service Center (2013). Machbarkeitsstudie "Starkregenrisiko 2050", URL: http://www.climate-service-center.de/imperia/md/content/csc/workshopdokumente/extremwetterereignisse/csc_machbarkeitsstudie_abschlussbericht.pdf. Technical report. (cited on pp. 2, 6, 37, 64).

Coulthard, T. and Frostick, L. (2010). The Hull floods of 2007: implications for the governance and management of urban drainage systems. *Journal of Flood Risk Management*, 3(3):223–231, doi:10.1111/j.1753-318X.2010.01072.x. (cited on pp. 1, 10, 36).

De Jong, P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press, New York. (cited on p. 28).

De Man, H. (2014). *Best urban water management practices to prevent waterborne infectious diseases under current and future scenarios*. PhD thesis, University of Utrecht. (cited on p. 6).

De Moel, H. and Aerts, J. C. J. H. (2010). Effect of uncertainty in land use, damage models and inundation depth on flood damage estimates. *Natural Hazards*, 58(1):407–425, doi:10.1007/s11069-010-9675-6. (cited on p. 10).

De'ath, G. and Fabricius, K. E. (2000). Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology*, 81(11):3178–3192, doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2. (cited on pp. 37, 47, 48, 49).

Deletic, A., Dotto, C. B. S., McCarthy, D. T., Kleidorfer, M., Freni, G., Mannina, G., Uhl, M., Henrichs, M., Fletcher, T. D., Rauch, W., Bertrand-Krajewski, J. L., and Tait, S. (2012). Assessing uncertainties in urban drainage models. *Physics and Chemistry of the Earth*, 42-44:3–10. (cited on pp. 4, 36).

Dick, W., Stoppa, A., Anderson, J., Coleman, E., and Rispoli, F. (2011). Weather Index-based Insurance in Agricultural Development. Technical report, International Fund for Agricultural Development. (cited on p. 4).

Dorland, C., Tol, R. S. J., and Palutikof, J. P. (1999). Vulnerability of the Netherlands and Northwest Europe to storm damage under climate change: A model approach based on storm damage in the Netherlands. *Climate Change*, 43(3):513–535, doi:10.1023/A:1005492126814. (cited on p. 4).

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, doi:10.1111/j.1600-0587.2012.07348.x. (cited on p. 71).

Douglas, I., Garvin, S., Lawson, N., Richards, J., Tippett, J., and White, I. (2010). Urban pluvial flooding: a qualitative case study of cause, effect and nonstructural mitigation. *Journal of Flood Risk Management*, 3(2):112–125, doi:10.1111/j.1753-318X.2010.01061.x. (cited on pp. 2, 10, 36, 64).

Einfalt, T., Hatzfeld, F., Wagner, A., Seltmann, J., Castro, D., and Frerichs, S. (2009). URBAS: forecasting and management of flash floods in urban areas. *Urban Water Journal*, 6(5):369–374, doi:10.1080/15730620902934819. (cited on p. 2).

Einfalt, T., Pfeifer, S., and Burghoff, O. (2012). Feasibility of deriving damage functions from radar measurements. In *9th International Workshop on Precipitation in Urban Areas*, pages 245–249, St. Moritz (Switzerland). (cited on pp. 2, 24, 37, 64).

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *The Journal of Animal Ecology*, 77(4):802–13, doi:10.1111/j.1365-2656.2008.01390.x. (cited on p. 60).

Elmer, F., Seifert, I., Kreibich, H., and Thieken, A. H. (2010a). A delphi method expert survey to derive standards for flood damage data collection. *Risk Analysis*, 30(1):107–24, doi:10.1111/j.1539-6924.2009.01325.x. (cited on pp. 2, 10, 64).

Elmer, F., Thieken, A. H., Pech, I., and Kreibich, H. (2010b). Influence of flood frequency on residential building losses. *Natural Hazards and Earth System Science*, 10(10):2145–2159, doi:10.5194/nhess-10-2145-2010. (cited on pp. 4, 5).

Ernst, J., Dewals, B., Archambeau, P., Detrembleur, S., Erpicum, S., and Pirotton, M. (2008). Integration of accurate 2D inundation modelling, vector land use database and economic damage evaluation. In *Flood Risk Management: Research and Practice*, pages 286–. (cited on p. 10).

Falconer, R., Cobby, D., Smyth, P., Astle, G., Dent, J., and Golding, B. (2009). Pluvial flooding: new approaches in flood warning, mapping and risk management. *Journal of Flood Risk Management*, 2(3):198–208, doi:10.1111/j.1753-318X.2009.01034.x. (cited on p. 3).

Freni, G., La Loggia, G., and Notaro, V. (2010). Uncertainty in urban flood damage assessment due to urban drainage modelling and depth-damage curve estimation. *Water Science and Technology*, 61(12):2979–93, doi:10.2166/wst.2010.177. (cited on pp. 3, 5, 6, 10, 36).

Gall, M., Borden, K. a., and Cutter, S. L. (2009). When Do Losses Count? *Bulletin of the American Meteorological Society*, 90(6):799–809, doi:10.1175/2008BAMS2721.1. (cited on pp. 6, 64).

Garne, T. W., Ebeltoft, M., Kivisaari, E., and Moberg, S. (2013). Weather related damage in the Nordic countries – from an insurance perspective, URL: http://www.fkl.fi/materiaalipankki/tutkimukset/Dokumentit/Weather_related_damage_in_the_Nordic_countries.pdf. Technical report. (cited on pp. 1, 3, 36, 64, 85, 87).

Gersonius, B., Zevenbergen, C., Puyan, N., and Billah, M. M. M. (2008). Efficiency of private flood proofing of new buildings - adapted redevelopment of a floodplain in the Netherlands. *WIT Transactions of Ecology and the Environment*, 118:247–259. (cited on p. 3).

Gregersen, I. B., Sørup, H. J. D., Madsen, H., Rosbjerg, D., Mikkelsen, P. S., and Arnbjerg-Nielsen, K. (2013). Assessing future climatic changes of rainfall extremes at small spatio-temporal scales. *Climatic Change*, 118(3-4):783–797, doi:10.1007/s10584-012-0669-0. (cited on p. 87).

Grigg, N. S. and Helweg, O. J. (1975). State-of-the-Art of Estimating Flood Damage in Urban Areas. *Journal of the American Water Resources Association*, 11(2):379–390, doi:10.1111/j.1752-1688.1975.tb00689.x. (cited on p. 3).

Hartmann, D., Tank, A. K., Rusticucci, M., Alexander, L., Brönnimann, S., Charabi, Y., Dentener, F., Dlugokencky, E., Easterling, D., Kaplan, A., Soden, B., Thorne, P., Wild, M., and Zhai, P. (2013). Observations: Atmosphere and Surface. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter 2. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. (cited on p. 2).

Hauger, M. B., Mouchel, J.-M., and Mikkelsen, P. S. (2006). Indicators of hazard, vulnerability and risk in urban drainage. *Water Science and Technology*, 54(6-7):441–450. (cited on p. 3).

Hess, K. R., Abbruzzese, M. C., Lenzi, R., Raber, M. N., and Abbruzzese, J. L. (1999). Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. *Clinical Cancer Research*, 5(11):3403–10. (cited on p. 37).

Hoen, A. and Van Leeuwen, M. (2012). Naar een nieuwe indeling statistiek (brand)schadeoorzaken. *De Beursbengel*, 75(817):10–12. (cited on p. 84).

Hoen, A. and Van Leeuwen, M. (2013). Nieuwe indeling schadeoorzaken zakelijke brandverzekeringen. *De Beursbengel*, 76(826):4–6. (cited on p. 84).

Hohl, R., Schiesser, H.-H., and Aller, D. (2002). Hailfall: the relationship between radar-derived hail kinetic energy and hail damage to buildings. *Atmospheric Research*, 63(3-4):177–207, doi:10.1016/S0169-8095(02)00059-5. (cited on pp. 2, 4, 36, 64).

Horn, B. (1981). Hill shading and the reflectance map. *Proceedings of the IEEE*, 69(1):14–47, doi:10.1109/PROC.1981.11918. (cited on pp. 41, 46).

Hurford, A., Parker, D., Priest, S., and Lumbroso, D. (2012). Validating the return period of rainfall thresholds used for Extreme Rainfall Alerts by linking rainfall intensities with observed surface water flood events. *Journal of Flood Risk Management*, 5(2):134–142, doi:10.1111/j.1753-318X.2012.01133.x. (cited on pp. 3, 11, 24, 36).

Hurford, A. P., Priest, S. J., Parker, D. J., and Lumbroso, D. M. (2011). The effectiveness of extreme rainfall alerts in predicting surface water flooding in England and Wales. *International Journal of Climatology*, doi:10.1002/joc.2391. (cited on pp. 3, 10).

Imai, K., King, G., and Lau, O. (2007). relogit: Rare Events Logistic Regression for Dichotomous Dependent Variables. In Imai, K., King, G., and Lau, O., editors, *Zelig: Everyones Statistical Software*. (cited on p. 71).

"Insurance Against Rain" (1911). Insurance Against Rain. *Nature*, 86(2161):144–144, doi:10.1038/086144d0. (cited on p. 84).

Jaffrain, J. and Berne, A. (2012). Influence of the Subgrid Variability of the Raindrop Size Distribution on Radar Rainfall Estimators. *Journal of Applied Meteorology and Climatology*, 51(4):780–785, doi:10.1175/JAMC-D-11-0185.1. (cited on p. 59).

Jak, M. and Kok, M. (2000). A database of historical flood events in the Netherlands. In *Flood Issues in Contemporary Water Management. NATO Science Series 2, Environmental Security*, pages 139–146, Delft. Kluwer Academic Publisher, The Netherlands. (cited on pp. 1, 5, 9, 36, 64).

Janssen, N., Overeem, A., and Uijlenhoet, R. (2013). Assessing the quality of disdrometer data for measuring rainfall in the Rotterdam region and modelling the spatial correlation for short-term rainfall accumulation intervals, BSc thesis, Wageningen University. (cited on p. 87).

Jongman, B., Kreibich, H., Apel, H., Barredo, J. I., Bates, P. D., Feyen, L., Gericke, a., Neal, J., Aerts, J. C. J. H., and Ward, P. J. (2012). Comparative flood damage model assessment: towards a European approach. *Natural Hazards and Earth System Science*, 12(12):3733–3752, doi:10.5194/nhess-12-3733-2012. (cited on pp. 3, 36, 64).

Jonkman, S. N., Bockarjova, M., Kok, M., and Bernardini, P. (2008). Integrated hydrodynamic and economic modelling of flood damage in the Netherlands. *Ecological Economics*, 66(1):77–90, doi:10.1016/j.ecolecon.2007.12.022. (cited on pp. 1, 10).

Kadaster (2013). Online viewer of the National Building Register held by Kadaster, URL: http://bagviewer.pdok.nl/. (cited on p. 65).

Kennedy, P. (2003). *A Guide to Econometrics*. MPG Books, Bodmin, Cornwall, 5th edition. (cited on p. 20).

King, G. and Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis*, 9:137–163. (cited on pp. 70, 71).

Kirtman, B., Power, S., Adedoyin, J., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schär, C., Sutton, R., van Oldenborgh, G., Vecchi, G., and Wang, H. (2013). Near-term Climate Change: Projections and Predictability. In Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter 11. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. (cited on p. 2).

KNMI (2000). Handbook for the meteorological observation. Chapter 6: Precipitation. Technical report, KNMI, De Bilt. (cited on p. 11).

Koot, A. C. J. (1977). *Inzameling en transport van rioolwater*. Waltman, ISBN: 9021230658. (cited on p. 22).

Kreibich, H., Thieken, A. H., Petrow, T., Müller, M., and Merz, B. (2005). Flood loss reduction of private households due to building precautionary measures - lessons learned from the Elbe flood in August 2002. *Natural Hazards and Earth System Sciences*, 5(1):117–126. (cited on pp. 3, 36).

Lawson, N. and Carter, J. (2009). Greater Manchester Local Climate Impacts Profile (GMLCIP) and assessing Manchester City Councils vulnerability to current and future weather and climate. Technical Report May, University of Manchester. (cited on pp. 3, 5).

Lee, S.-K. and Jin, S. (2006). Decision tree approaches for zero-inflated count data. *Journal of Applied Statistics*, 33(8):853–865, doi:10.1080/02664760600743613. (cited on p. 48).

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks: Sage Publications. (cited on pp. 16, 71).

Lozano, F. J., Suárez-Seoane, S., Kelly, M., and Luis, E. (2008). A multi-scale approach for modeling fire occurrence probability using satellite data and classification trees: A case study in a mountainous Mediterranean region. *Remote Sensing of Environment*, 112(3):708–719, doi:10.1016/j.rse.2007.06.006. (cited on p. 4).

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models (Second Edition)*. Chapman and Hall, Chicago and Londen. (cited on pp. 16, 70).

Merz, B., Kreibich, H., and Lall, U. (2013). Multi-variate flood damage assessment: a tree-based data-mining approach. *Natural Hazards and Earth System Science*, 13(1):53–64, doi:10.5194/nhess-13-53-2013. (cited on pp. 2, 3, 4, 36, 37, 60, 87).

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A. (2010). Review article "Assessment of economic flood damage". *Natural Hazards and Earth System Science*, 10(8):1697–1724, doi:10.5194/nhess-10-1697-2010. (cited on pp. 1, 3, 36).

Merz, B., Kreibich, H., Thieken, A., and Schmidtke, R. (2004). Estimation uncertainty of direct monetary flood damage to buildings. *Natural Hazards and Earth System Science*, 4(1):153–163, doi:10.5194/nhess-4-153-2004. (cited on pp. 3, 10, 36).

Middelmann-Fernandes, M. (2010). Flood damage estimation beyond stage-damage functions: an Australian example. *Journal of Flood Risk Management*, 3(1):88–96, doi:10.1111/j.1753-318X.2009.01058.x. (cited on p. 10).

Ministry of Transport Public Works and Water Management (2003). Insurability of damages related to extreme rainfall and pluvial flooding (in Dutch). Technical report, RIZA, Lelystad, The Netherlands. (cited on pp. 13, 66).

Moisen, G. G. (2008). Classification and Regression Trees. In Jørgensen, S. E. and Fath, B. D., editors, *Encyclopedia of Ecology*, pages 582–588. Elsevier, Oxford, UK. (cited on pp. 48, 49).

Moreau, E., Testud, J., and Le Bouar, E. (2009). Rainfall spatial variability observed by X-band weather radar and its implication for the accuracy of rainfall estimates. *Advances in Water Resources*, 32(7):1011–1019, doi:10.1016/j.advwatres.2008.11.007. (cited on p. 87).

Norbiato, D., Borga, M., and Dinale, R. (2009). Flash flood warning in ungauged basins by use of the flash flood guidance and model-based runoff thresholds. *Meteorological Applications*, 16(1):65–75, doi:10.1002/met.126. (cited on p. 3).

Overeem, A., Holleman, I., and Buishand, A. (2009). Derivation of a 10-Year Radar-Based Climatology of Rainfall. *Journal of Applied Meteorology and Climatology*, 48(7):1448–1463, doi:10.1175/2009JAMC1954.1. (cited on pp. 24, 25, 38, 43).

Overeem, A., Leijnse, H., and Uijlenhoet, R. (2011). Measuring urban rainfall using microwave links from commercial cellular communication networks. *Water Resources Research*, 47(12):1–16, doi:10.1029/2010WR010350. (cited on pp. 14, 25, 38).

Parker, D. J., Priest, S. J., and McCarthy, S. (2011). Surface water flood warnings requirements and potential in England and Wales. *Applied Geography*, 31(3):891–900, doi:10.1016/j.apgeog.2011.01.002. (cited on pp. 11, 58).

Peleg, N., Ben-Asher, M., and Morin, E. (2013). Radar subpixel-scale rainfall variability and uncertainty: lessons learned from observations of a dense rain-gauge network. *Hydrology and Earth System Sciences*, 17(6):2195–2208, doi:10.5194/hess-17-2195-2013. (cited on p. 59).

Pielke, R. A. and Downton, M. W. (2000). Precipitation and Damaging Floods: Trends in the United States, 193297. *Journal of Climate*, 13(20):3625–3637, doi:10.1175/1520-0442(2000)013<3625:PADFTI>2.0.CO;2. (cited on pp. 2, 64).

Pistrika, A. K. and Jonkman, S. N. (2009). Damage to residential buildings due to flooding of New Orleans after hurricane Katrina. *Natural Hazards*, 54(2):413–434, doi:10.1007/s11069-009-9476-y. (cited on pp. 3, 10, 36).

Pitt, M. (2008). Learning lessons from the 2007 floods, and independent review by Sir Michael Pitt. Technical report, The Pitt Review Cabinet Office, London, UK. (cited on pp. 1, 10, 36).

Poussin, J. K., Botzen, W. W., and Aerts, J. C. (2014). Factors of influence on flood damage mitigation behaviour by households. *Environmental Science and Policy*, pages 1–9, doi:10.1016/j.envsci.2014.01.013. (cited on p. 3).

Priest, S. J., Parker, D. J., Hurford, a. P., Walker, J., and Evans, K. (2011). Assessing options for the development of surface water flood warning in England and Wales. *Journal of environmental management*, 92(12):3038–48, doi:10.1016/j.jenvman.2011.06.041. (cited on pp. 11, 58).

Rejwan, C., Collins, N. C., Brunner, L. J., Shuter, B. J., and Ridgway, M. S. (1999). Tree Regression Analysis on the Nesting Habitat of Smallmouth Bass. *Ecology*, 80(1):341, doi:10.2307/177003. (cited on p. 37).

Ririassa, H. and Hoen, A. (2010). Rainfall and damage: a study on relationships between rainfall and claims in relation to climate change (in Dutch). Technical report, Dutch Association of Insurers. (cited on pp. 10, 37, 38).

Risk Management Solutions (2013). RMS White Paper: The 2012 U.K. Floods, URL: http://www.rms.com/resources/publications/natural-catastrophes. Technical report. (cited on p. 2).

Rodríguez, J. P., McIntyre, N., Díaz-Granados, M., and Maksimović, C. (2012). A database and model to support proactive management of sediment-related sewer blockages. *Water Research*, 46(15):4571–86, doi:10.1016/j.watres.2012.06.037. (cited on p. 5).

Segoni, S., Rosi, A., Rossi, G., Catani, F., and Casagli, N. (2014). Analysing the relationship between rainfalls and landslides to define a mosaic of triggering thresholds for regional scale warning systems. *Natural Hazards and Earth System Sciences Discussions*, 2(3):2185–2213, doi:10.5194/nhessd-2-2185-2014. (cited on p. 1).

Seifert, I., Botzen, W. J. W., Kreibich, H., and Aerts, J. C. J. H. (2013). Influence of flood risk characteristics on flood insurance demand: a comparison between Germany and the Netherlands. *Natural Hazards and Earth System Science*, 13(7):1691–1705, doi:10.5194/nhess-13-1691-2013. (cited on pp. 38, 87).

Septer, D. and Schwab, J. (1995). *Rainstorm and flood damage: Northwest British Columbia 1891–1991*. B.C. Ministry of Forest, Victoria, B.C. (cited on p. 5).

Smith, C. and Lawson, N. (2012). Identifying extreme event climate thresholds for greater Manchester, UK: examining the past to prepare for the future. *Meteorological Applications*, 19(1):26–35, doi:10.1002/met.252. (cited on pp. 2, 5, 64).

Smith, D. I. (1994). Flood damage estimation - A review of urban stage-damage curves and loss functions. *Water SA*, 20(3):231–238. (cited on p. 3).

Spekkers, M. H., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2014a). On the occurrence of rainstorm damage based on home insurance and weather data. *Natural Hazards and Earth System Sciences Discussions*, 2(8):5287–5313, doi:10.5194/nhessd-2-5287-2014. No citations.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2013a). A spatial analysis of rainfall damage data using C-band weather radar images. In Butler, D., Chen, A. S., Djordjevic, S., and Hammond, M. J., editors, *Proceedings of the International Conference on Flood Resilience: Experiences in Asia and Europe*, Exeter, UK. Centre for Water Systems, University of Exeter. No citations.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2013b). A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences*, 17(3):913–922, doi:10.5194/hess-17-913-2013. No citations.

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E. (2014b). Decision-tree analysis of factors influencing rainfall-related building structure and content damage. *Natural Hazards and Earth System Science*, 14(9):2531–2547, doi:10.5194/nhess-14-2531-2014. No citations.

Statistics Netherlands (2012). StatLine online database, URL: http://statline.cbs.nl (viewed on August 2012). (cited on p. 27).

Statistics Netherlands (2013). Demographic statistics by municipality 2013, URL: http://www.cbs.nl. (cited on p. 65).

Statistics Netherlands (2014). StatLine online database, URL: http://statline.cbs.nl (viewed on July 2014). (cited on pp. 65, 66).

Stichting RIONED (2004). Sewer guideline (module C2100) "Hydrodynamic calculations and hydraulic design" (in Dutch). Technical report. (cited on p. 22).

Stichting RIONED (2008). Sewer guideline (module B2200) "Functional design: collection and transport of stormwater" (in Dutch). Technical report. (cited on p. 43).

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4):323–48, doi:10.1037/a0016973. (cited on p. 60).

Sušnik, J., Strehl, C., Postmes, L., Vamvakeridou-Lyroudia, L., Savić, D., Kapelan, Z., and Mälzer, H.-J. (2014). Assessment of the Effectiveness of a Risk-reduction Measure on Pluvial Flooding and Economic Loss in Eindhoven, the Netherlands. *Procedia Engineering*, 70:1619–1628, doi:10.1016/j.proeng.2014.02.179. (cited on p. 4).

Ten Veldhuis, J. A. E. (2010). *Quantitative risk analysis of urban flooding in lowland areas*. PhD thesis, Delft University of Technology. (cited on pp. 1, 86).

Ten Veldhuis, J. A. E. (2011). How the choice of flood damage metrics influences urban flood risk assessment. *Journal of Flood Risk Management*, 4(4):281–287, doi:10.1111/j.1753-318X.2011.01112.x. (cited on pp. 2, 10, 36, 64).

Ten Veldhuis, J. A. E., Clemens, F. H. L. R., Sterk, G., and Berends, B. R. (2010). Microbial risks associated with exposure to pathogens in contaminated urban flood water. *Water Research*, 44(9):2910–8, doi:10.1016/j.watres.2010.02.009. (cited on p. 6).

Ten Veldhuis, J. A. E., Clemens, F. H. L. R., and Van Gelder, P. H. A. J. M. (2011). Quantitative fault tree analysis for urban water infrastructure flooding. *Structure and Infrastructure Engineering*, 7(11):809–821, doi:10.1080/15732470902985876. (cited on pp. 4, 5, 36, 43).

The Center for Neighborhood Technology (2014). The Prevalence and Cost of Urban Flooding, URL: http://www.cnt.org/media/CNT_PrevalenceAndCostOfUrbanFlooding.pdf. Technical Report May. (cited on p. 6).

Therneau, T. M. and Atkinson, E. J. (2014). An Introduction to Recursive Partitioning Using the RPART Routines. Technical report. (cited on pp. 47, 48, 49).

Thieken, A. H. (2011). Methods for the Documentation and Estimation of Direct Flood Losses. In Zenz, G. and Hornich, R., editors, *Urban Flood Risk Management Conference*, Graz (Austria). (cited on pp. 4, 10).

Thieken, A. H., Müller, M., Kreibich, H., and Merz, B. (2005). Flood damage and influencing factors: New insights from the August 2002 flood in Germany. *Water Resources Research*, 41(12):1–16, doi:10.1029/2005WR004177. (cited on pp. 3, 5, 36, 46, 87).

Van der Sande, C., Soudarissanane, S., and Khoshelham, K. (2010). Assessment of relative accuracy of AHN-2 laser scanning data using planar features. *Sensors*, 10(9):8198–214, doi:10.3390/s100908198. (cited on pp. 38, 46).

Van der Zon, N. (2013). Background information about AHN2 (in Dutch). Technical report, Actueel Hoogtebestand Nederland, Amersfoort. (cited on pp. 38, 43, 46).

Van Luijtelaar, H. and Rebergen, E. (1997). Guidelines for hydrodynamic calculations on urban drainage in the Netherlands: Backgrounds and examples. *Water Science and Technology*, 36(8-9):253–258, doi:10.1016/S0273-1223(97)00590-8. (cited on p. 22).

Van Mameren, H. and Clemens, F. H. L. R. (1997). Guidelines for hydrodynamic calculations on urban drainage in the Netherlands: Overview and principles. *Water Science and Technology*, 36(8-9):247–252, doi:10.1016/S0273-1223(97)00591-X. (cited on p. 22).

Visser, F. (2014). Rapid mapping of urban development from historic Ordnance Survey maps: An application for pluvial flood risk in Worcester. *Journal of Maps*, 10(2):276–288, doi:10.1080/17445647.2014.893847. (cited on p. 5).

Weiss, A. D. (2001). Topographic Position and Landforms Analysis, Poster Presentation, ESRI User Conference, San Diego, CA. (cited on pp. 41, 46).

Willems, P., Arnbjerg-Nielsen, K., Olsson, J., and Nguyen, V. (2012). Climate change impact assessment on urban rainfall extremes and urban drainage: Methods and shortcomings. *Atmospheric Research*, 103:106–118, doi:10.1016/j.atmosres.2011.04.003. (cited on p. 2).

Wilson, M. F. J., O'Connell, B., Brown, C., Guinan, J. C., and Grehan, A. J. (2007). Multiscale Terrain Analysis of Multibeam Bathymetry Data for Habitat Mapping on the Continental Slope. *Marine Geodesy*, 30(1-2):3–35, doi:10.1080/01490410701295962. (cited on p. 46).

Wind, H. G., Nierop, T. M., De Blois, C. J., and De Kok, J. L. (1999). Analysis of flood damages from the 1993 and 1995 Meuse Floods. *Water Resources Research*, 35(11):3459–3465, doi:10.1029/1999WR900192. (cited on p. 5).

Zhou, Q., Mikkelsen, P. S., Halsnæ s, K., and Arnbjerg-Nielsen, K. (2012). Framework for economic pluvial flood risk assessment considering climate change effects and adaptation benefits. *Journal of Hydrology*, 414-415:539–549, doi:10.1016/j.jhydrol.2011.11.031. (cited on pp. 4, 10, 36).

Zhou, Q., Panduro, T. E., Thorsen, B. J., and Arnbjerg-Nielsen, K. (2013). Verification of flood damage modelling using insurance data. *Water Science and Technology*, 68(2):425–32, doi:10.2166/wst.2013.268. (cited on pp. 2, 4, 5, 6, 10, 24, 37, 64).

Glossary

building content

Portable good inside a building or semi-permanently attached to a building.

building structure

Permanent and unmovable components of a building and its foundation.

claim frequency

Rate at which claims occur in a given sample.

claim probability

Probability that a claim occurs at a risk address on a day.

claim size

Monetary costs associated with an individual claim.

content

See building content.

damage assessment

Procedure of estimating damage, either by expert judgements or by using damage models.

damage data

Data reporting characteristics about the adverse consequences of a damage event, collected during or in the aftermath of an event.

damage mechanism

Process by which damage is generated.

damage model

Mathematical model to estimate damage to an object or a spatially aggregated unit based on a set of explanatory variables.

depression filling

Process of stormwater running down a slope and filling up depressions at the bottom if no drainage facilities are available.

deviance

Goodness-of-fit measure of a model being evaluated compared to the null model.

intangible damage

Damage that cannot be expressed in monetary values.

pluvial flooding

Flooding caused by stormwater being unable to enter urban drainage systems or flowing out of urban drainage systems when capacity is exceeded.

probability of claim occurrence

See claim probability.

property

See building structure.

rainstorm

Weather condition with heavy rainfall.

rainstorm damage

Damage that results from pluvial flooding or rainwater intrusion through defects in the building envelope.

rainwater

Water that has fallen as rain.

risk address

Location of insured property.

river flooding

Flooding as a result of high river discharges causing the water to overflow riverbanks.

stage-damage function

Relationship between flood damage and flood depth (i.e. stage), typically developed for a specific building class or land use.

stormwater

Rainwater that has fallen on a built-up area.

tangible damage

Damage that can be expressed in monetary values.

urban drainage system

Sequence of facilities designed to drain wastewater and stormwater with the aim to minimize problems caused to human life and the environment.

water authority

Organisation responsible for the water management, including the prevention of flooding.

water-related damage

Damage caused by physical contact with water, independent of the source of water.

Summary

In this thesis, insurance data related to the impact of local rainstorms to building structure and content are analysed to gain knowledge on causes of variability in damage data. This is of importance, as rainstorms cause considerable damage to urban societies all over the world. Moreover, there is strong evidence that rainstorm damage will likely increase in the future as a consequence of climate change and urbanisation. So far, little research on this topic focused on the collection and analysis of damage data, which hampers the development of prediction models for rainstorm damage. Yet, damage data and models have a high potential of providing valuable information to homeowners, water authorities and insurers to support damage prevention and reduction.

The general objective of this thesis is to explain the variability of rainstorm damage based on statistical analyses of home insurance data and a wide range of explanatory data, including weather, building-related, topographic and socioeconomic data. The current work particularly focuses on damage that results from small-scale pluvial flooding and rainwater intrusion due to defects in the building envelope. The research data are drawn from two home insurance databases from Dutch insurance industry: a large, nationwide insurance database and a detailed, property level insurance database for a case study in Rotterdam.

Statistical analysis of the data set for the Rotterdam case study shows that the majority of rainfall-related claims are caused by water intrusion due to defects in the building envelope, mainly cases of leaks in the building's roof. A relatively small number of claims relate to public system failures, such as sewer flooding and depression filling. Although rare, sewer-related claims are associated with significantly larger claims sizes than the average rainstorm damage claim, because of the large costs involved in cleaning sewer spills. Crossing insurance data with weather radar data reveals that public system failures contribute to the occurrence of claims only when a minimum rainfall intensity threshold is exceeded.

Of all weather variables studied, the most important predictor for claim probability is the storm event maximum hourly rainfall intensity. Furthermore, decision-tree analysis of the nationwide insurance database shows that other statistically significant contributors to claim probability are real estate value, ground floor area, building age, building type, household income and ownership structure. Together, these variables

only explain part of the damage variability. No meaningful relationships between weather and other contextual variables and claim size were found. Future studies based on larger data sets and other explanatory variables are recommended to provide more insights into causes of claim size variability.

Decision-tree models outperform global regression model in terms of variance explained, because many variables are involved in the prediction of rainstorm damage, most of them to some extent intercorrelated. The decision tree approach asks for fewer assumptions about the data; a disadvantage of this method is, however, that it requires at least thousands of records to sufficiently populate tree subgroups. This may be inconvenient because rainstorm damage events are rare and only small subsets of claim data may be available for research purposes. Alternatively, a rare events logistic regression model can be applied to insurance damage data when the number of claims is limited and claim probability is low.

It is recommended, from a scientific perspective, to automate, standardize and extend the recordings of damage data and the claim classification process, with the aim of improving future analysis of damage data. The processing of insurance databases calls for a considerable amount of work, because insurance databases are prone to human errors and usually do not contain information of weather-related, building-related and socioeconomic variables. Complementary data need to be collected and appropriately converted to formats that allow analysis of variables influencing damage. Moreover, information on actual causes of damage cannot be easily retrieved from insurers' data archives without much data classification efforts.

The results of current work on rainstorm damage data demonstrate the potential for better understanding of damage mechanisms using insurance claims. Damage data from more insurance companies are required, as well as damage data related to regions with different characteristics, to compare and validate results of this study. An automated claim classification system enables analyses to be done on larger data sets that cover different cities and more extreme rainfall events than those covered in current work. Future studies can also investigate the co-occurrence of different kinds of weather-related hazards and their relationships with weather variables. Damage data from other insurance industry domains, such as crop, car and business insurance, can be analysed to have a more complete view on the impact of rainstorms. Together, this will contribute to better prediction models for rainstorm damage, which in turn can help to enhance the resilience of urban societies to heavy rainfall.

Samenvatting

In dit proefschrift staan de analyses van verzekeringsgegevens over neerslagschade aan gebouwen en inboedels centraal. Het doel van deze analyses is om een beter begrip te krijgen van de oorzaken die variabiliteit in schade verklaren. Dit is van belang omdat schade door regenval wereldwijd aanzienlijk is en naar verwachting zal gaan toenemen als gevolg van klimaatsverandering en verstedelijking. Tot nu toe is er weinig wetenschappelijk onderzoek verricht, gericht op het verzamelen en analyseren van neerslagschadegegevens. Dit terwijl schadegegevens en -modellen belangrijke informatie kunnen leveren aan huiseigenaren, waterbeheerders en verzekeraars om uiteindelijk schade te verminderen of te voorkomen.

De hoofddoelstelling van dit proefschrift is om variabiliteit in neerslagschade te verklaren door statistische relaties te onderzoeken tussen schadegegevens van opstal- en inboedelverzekeringen en verklarende factoren zoals weergerelateerde, gebouwspecifieke, topografische en sociaaleconomische variabelen. Dit onderzoek richt zich in het bijzonder op schade als gevolg van regenwateroverlast en binnendringend regenwater door defecten aan het gebouw. De onderzoeksgegevens zijn gebaseerd op twee verzekeringsdatabases: een grote landsdekkende verzekeringsdatabase en een gedetailleerde verzekeringsdatabase op het niveau van individuele adressen voor de gemeente Rotterdam.

Uit statistische analyses blijkt dat de meerderheid van de neerslaggerelateerde claims in Rotterdam veroorzaakt wordt door binnendringend regenwater door defecten aan het gebouw, zoals daklekkages. Een relatief klein aantal claims is in verband te brengen met het falen van publieke systemen, zoals de overbelasting van riolen en de accumulatie van regenwater in lager gelegen gebieden. Hoewel rioolgerelateerde claims zeldzaam zijn, ligt de gemiddelde hoogte van deze claims significant hoger dan de gemiddelde hoogte van een neerslaggerelateerde claim. Dit is met name omdat het verontreinigd water uit een riool tot hoge schoonmaakkosten leidt. In een vergelijking tussen verzekeringsgegevens en regenradargegevens is te zien dat alleen als de neerslagintensiteit een minimum drempelwaarde overschrijdt, het falen van publieke systemen gaat bijdragen aan het vóórkomen van claims.

Van alle onderzochte neerslagkarakteristieken is de maximum uurneerslag van de neerslaggebeurtenis de belangrijkste voorspeller van de kans op een schadeclaim. Op basis van analyses met beslisbomen op de landsdekkende verzekeringsdatabase kan

verder worden geconcludeerd dat ook de WOZ-waarde, het vloeroppervlak en de leeftijd van het gebouw, het gebouwtype, het huishoudinkomen en het eigenaarsbelang statistich significante factoren zijn. Samen met de maximale uurneerslag verklaren deze variabelen slechts een deel van de variabiliteit in schade. Er zijn geen betekenisvolle relaties gevonden tussen weersgerelateerde en andere verklarende variabelen en de gemiddelde hoogte van een claim. Het strekt tot de aanbeveling dat toekomstig onderzoek zich richt op grotere databases en een groter aantal verklarende variabelen die mogelijk de hoogte van claims kunnen verklaren.

Beslisbomen hebben een beter voorspellend vermogen dan globale regressiemodellen, onder meer omdat een groot aantal variabelen betrokken is bij het voorspellen van neerslagschade en deze variabelen tot een zekere hoogte onderling gecorreleerd zijn. De benaderingswijze van beslisbomen stelt weinig eisen aan de gegevens; het nadeel van deze methode is echter dat minstens duizenden waarnemingen vereist zijn om in voldoende mate subgroepen in de gegevens te kunnen onderscheiden. Dit kan problematisch zijn omdat neerslagschade een zeldzame gebeurtenis is en vaak slechts kleine deelverzamelingen van schadegegevens beschikbaar zijn voor onderzoek. Als alternatief kan logistische regressie voor zeldzame gebeurtenissen toegepast worden bij een beperkt aantal schadeclaims of bij een lage kans op een claim.

Vanuit een wetenschappelijk oogpunt wordt het aangeraden om de opslag en classificatie van schadegegevens te automatiseren, te standaardiseren en uit te breiden, om op die manier toekomstige analyses van schadegegevens te verbeteren. Het verwerken van verzekeringsdatabases is namelijk tijdrovend omdat ze kwetsbaar zijn voor menselijke fouten en vaak geen informatie bevatten over weersgerelateerde, gebouwspecifieke en sociaaleconomische variabelen. Het is daarom nodig om aanvullende bestanden te verzamelen en te converteren naar geschikte formatten zodat onderzoek naar de effecten van variabelen mogelijk is. Het is bovendien vaak lastig om informatie over de directe oorzaken van schade uit verzekeringsdatabases te halen zonder de claims eerst te moeten classificeren.

Het resultaat van dit proefschrift over neerslagschade laat zien wat de potentie is van het gebruik van verzekeringsgegevens om meer inzicht te krijgen in schademechanismen. Schadegegevens van meer verzekeraars zijn nodig, alsmede schadegegevens gebasseerd op gebieden met andere karakteristieken, om de resultaten van dit onderzoek te vergelijken en te valideren. Een geautomatiseerde classificatie van claims kan in de toekomst bijdragen aan analyses op grotere bestanden die betrekking hebben op verschillende steden en die extremere neerslaggebeurtenissen omvatten dan de neerslaggebeurtenissen die in dit proefschrift zijn onderzocht. Toekomstig onderzoek kan zich ook richten op het samenvallen van verschillende weersinvloeden en hun relaties met weergerelateerde variabelen. Schadegegevens van andere verzekeringssectoren kunnen ook beschouwd worden om een completer beeld te krijgen van neerslagschade, zoals oogst- en autoschade en zakelijke schade. Samen zal dit bijdragen aan betere voorspellende modellen voor neerslagschade, die op hun beurt nuttig kunnen zijn voor het versterken van de veerkracht van een samenleving als het gaat om de gevolgen van extreme regenval.

Acknowledgements

As a Dutchman, I have been cycling to Delft for the past few years, rain or no rain, to work on my thesis and here it is. I would like to thank my colleagues of the Urban Drainage group for the many discussions and pleasant coffee breaks. I especially would like to thank the following persons. Arco Krijgsman and Marcel Fekkes, my good old athletics friends, for assisting me as paranymphs during the defence. Rob Bakker and Wil Wijshoff, from the Achmea insurance group, for their enthusiasm and great help to make available the research data of Chapter 5. Alex Hoen for giving me the opportunity to work with the data from the Dutch Association of Insurers and the helpful comments on many draft versions of publications. Emiel Verstegen and Martijn Koole for the hundreds of hours they spent on classifying the claim data of Chapter 5 – impressive! Elsbeth Ciesluk for the awesome cover design. Matthijs Kok for helping me to get into contact with insurance industry and the fruitful cooperation. François Clemens for all the good conversations and support, which helped me to finish the book, and his enthusiasm for the topic. Marie-claire ten Veldhuis for the many lunch walks and talks, the fruitful cooperation and the many pleasant business trips we had together that shaped our research topics. I would like to acknowledge the Royal Netherlands Meteorological Institute (KNMI) for their support and making available the rainfall data.

About the author

Matthieu Spekkers was born in Purmerend, the Netherlands, on 15 February 1982. He received pre-university education at the "Da Vinci College" from 1994 to 2000. In 2000, he started his study Civil Engineering and Management at the University of Twente, Enschede, the Netherlands. His final master's project was about roughness of riverbeds, for which he conducted experiments in a laboratory flume at the Leichtweiß-Institut für Wasserbau, University of Braunschweig, Germany. After finishing his studies in 2007, Matthieu continued working at the University of Twente for another year as a junior researcher, where he investigated the development of pavement layers beneath river dunes. In 2009, he worked as a water management consultant at Nelen & Schuurmans, Utrecht, the Netherlands, on flood modelling and safety. In March 2010, he started as a Ph.D. candidate at the Delft University of Technology in the Urban Drainage group, within the EU-funded project SMARTeST. His research resulted in this thesis. Presently, he is working in the same group as postdoc.


Matthieu Spekkers