Nat. Hazards Earth Syst. Sci. Discuss., 2, 2263–2305, 2014
www.nat-hazards-earth-syst-sci-discuss.net/2/2263/2014/
doi:10.5194/nhessd-2-2263-2014

Natural Hazards and Earth System Sciences
Open Access
Discussions

Decision tree analysis of factors influencing rainfall-related building damage

M. H. Spekkers[1], M. Kok[2], F. H. L. R. Clemens[1], and J. A. E. ten Veldhuis[1]

[1]Delft University of Technology, Department of Water Management, Delft, the Netherlands
[2]Delft University of Technology, Department of Hydraulic Engineering, Delft, the Netherlands

Received: 24 February 2014 – Accepted: 17 March 2014 – Published: 1 April 2014

Correspondence to: M. H. Spekkers (m.h.spekkers@tudelft.nl)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Flood damage prediction models are essential building blocks in flood risk assessments. Little research has been dedicated so far to damage of small-scale urban floods caused by heavy rainfall, while there is a need for reliable damage models for this flood type among insurers and water authorities.

The aim of this paper is to investigate a wide range of damage-influencing factors and their relationships with rainfall-related damage, using decision tree analysis. For this, district-aggregated claim data from private property insurance companies in the Netherlands were analysed, for the period of 1998–2011. The databases include claims of water-related damage, for example, damages related to rainwater intrusion through roofs and pluvial flood water entering buildings at ground floor. 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 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), ownership structure (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, which suggest that variability in average claim size is related to explanatory variables that cannot be defined at the district scale. 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). 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.

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

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 (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). Meanwhile, reliable damage models for this type of floods 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 from Nordic insurance industry show that damage to residential buildings due to heavy rainfall is of the order of tens of Euros per year per capita (Garne et al., 2013).

The objective of a damage model is to predict damage related to single objects (e.g., buildings) or spatially aggregated units (e.g., postal districts, neighbourhoods), based on a set of explanatory variables. Traditional models, developed for river floods, 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 (Pistrika and Jonkman, 2009; Merz et al., 2004, 2010; Thieken et al., 2005; 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 well understood (Deletic et al., 2012) to make reliable flood depth calculations. A source of uncertainty relates to incomplete knowledge of failure mechanisms that lead to flood-

ing. 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 the present paper 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., De'ath and Fabricius, 2000; Rejwan et al., 1999) 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 paper, 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 whole the Netherlands. In exploratory studies based on the same database, relationships between various characteristics of rainfall events and various damage variables were investigated (Spekkers et al., 2013a, b; Ririassa and Hoen, 2010). These studies found that rainfall characteristics explain only part of the variance in water-related damage data. Similar conclusions were drawn by Zhou et al. (2013); Cheng (2012); Einfalt et al. (2012) 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. First, global regression models were used in the aforementioned studies, but they may not be the most appropriate model choice given the complexity of the problem. Second, 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 paper aims to investigate a wide range of damage-influencing factors, defined at 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 and content damage data. The paper is structured as follows. First, an overview of the data sources and a description of how response and explanatory variables were derived from the data is given (Sect. 2). In Sect. 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, followed by a discussion in Sect. 5. Finally, Sect. 6 summarizes conclusions and recommendations.

2 Data

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 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. 1). In the Netherlands, almost all buildings are insured for water-related property damage, be-

cause it is commonly obliged to take out property insurance in the case of a mortgage. The data are aggregated at the level of 4-position 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 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 adjusted for inflation according to the correction indices in Table 3. Extensive checks on missing and incorrect values (e.g., blanks, zeros and incorrect dates) and inconsistencies in the data are discussed in Spekkers et al. (2013a). Figure 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 ≤ 10 %). This is mainly the case for property insurance, 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 2 for definitions. The next section discusses the explanatory variables.

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 (Spekkers et al., 2013a). These non-rainfall-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 Spekkers et al. (2013a) 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):

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

where ζ is the probability of a non-rainfall-related claim on a day for an individual, insured household. Figure 4 shows the estimated ζ per year for content and property claims, based on claim and policyholder counts of 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. 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 policyholders 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 (~ 15 500 claims) and around 6300 cases (~ 19 000 claims) respectively. Figure 3 shows the distributions of the response variables for the subsets; the distribution are skewed to the right.

2.3 Damage-influencing variables

2.3.1 Rainfall-related variables

For each case in the subset, rainfall-related variables were extracted from weather radar data (Table 2). The following approach is also described in Spekkers et al. (2013b). A database of C-band weather radar images was used, provided by the Royal Netherlands Meteorological Institute (Table 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.

First, 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 after a rainfall event. 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.

Second, 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 Sect. 2.3.4). 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. 5).

2.3.2 Topographic variables

A digital terrain model (DTM) of the Netherlands was used to characterise districts in terms of their steepness (Table 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 × 5 m tiles (Van der Zon, 2013). Data gaps in the DTM were filled using linear interpolation. More background 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 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. The derived values were then spatially aggregated to obtain median variable values per district. Median values were used rather than mean values, to reduce the effect of outliers. The housing stock in the year 2011 (reference date: 31 December) was used for the calculations. 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.

2.3.3 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 damage may be less strong 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 1 and 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. Due to privacy restrictions, aggregations are based on at least 10 records (i.e., households or persons), otherwise the variable value was marked not available. For household income, the minimum number of observations is set stricter, namely 100. Because only house owners can take property insurance, the variable "ownership structure" is only relevant for content-related response variables.

2.3.4 Building-related variables

Building-related variables were based on the National Building Register (NBR), a geo-database of all buildings and addresses in the Netherlands (Table 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". For each case, three variables were derived: fraction of low-rise buildings, build-

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

ing age and ground floor area (Table 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.

2.3.5 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.

3 Methodology

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 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 best splits the data into two groups, or nodes, in terms of the response variable. A data set is split into two groups by a chosen reference value of an explanatory variable: a group for which values are less than the chosen reference value and a group for which values are greater than or equal to the chosen reference value. From all possible splits of all explanatory variables, the one that maximizes homogeneity of the response variable in the resulting groups, is selected. This process is recursively repeated on each subgroup until a large tree is learned. 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 cross-validation, which is discussed in Sect. 3.2.

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):

$$\Phi = D_P - D_L - D_R \tag{2}$$

A split is sought that maximizes Eq. (2). The expression of the within-node deviance is specified depending on the type of response data. For continuous data, which 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). 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.

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 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) (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), 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 the discussion section, Sect. 5.

When the splitting variable is missing for an individual case, a surrogate splitting variable is used instead. A surrogate splitting variable is a variable that splits the data most similar compared to the primary splitting variable (Breiman et al., 1984). There are a number of sources for missing data, most notable are: missing rainfall data because weather radars were not operational and values of socioeconomic variables that were censored because of privacy regulations if values were based on small samples (i.e., small number of persons or objects).

Content and property claim data were analysed separately, as it is not possible to link both databases. Consequently, 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 2 were used as model input, except ownership structure in the case of property claim data.

3.2 Determining size of tree and variable importance

One way to determine the right tree size is to use 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 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 determining the exact tree size at which the plateau is reached is imprecise, 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.

Decision trees can also be used to identify important variables. Variable importance is defined as the sum of the goodness-of-split measure (Eq. 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 *rpart* library for R 2.15.3 was used for this study, developed by Therneau and Atkinson (2014).

3.3 Comparison with global multiple regression model

Results of decision tree analysis were compared to results of global multiple regression analysis. Only regression models for claim frequency were considered. A Poisson regression model was used to explain claim frequency as a function of various combinations of explanatory variables. Poisson trees and Poisson regression models were compared in terms of variance explained by the models.

4 Results

4.1 Explorative analysis

To explore data, pairwise correlations between explanatory and response variables were analysed (Table 5). Spearman's correlation coefficients were calculated to account for the non-normal distributions of response data (Fig. 3). Note that the categorical variable “season” is not listed in Table 5. 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). Moreover, there are a large number of significant links between explanatory variables and claim frequency than between explanatory variables and average claims size. Other significant factors associated with claim frequency (with $|\rho| > 0.20$) include household income, real estate value, ownership structure (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 house owners only, whereas content-related data include tenants and house owners. 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 2). 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.

4.2 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. 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 mm h^{-1}, 15–37 mm h^{-1} and > 37 mm h^{-1}, with most claims (67 %) falling into the lowest rainfall in-

tensity group. Figure 7 illustrates the splitting method for the top split: the claim frequency is plotted against maximum rainfall intensity (see top, Fig. 7) and a split value for maximum rainfall intensity is sought that maximizes the goodness-of-split measure (see bottom, Fig. 7). For cases associated with rainfall intensities larger than 37 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 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 26 % of the variance in validation data (Fig. 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, ownership structure, ground floor area and fraction of low-rise buildings (Fig. 9). Similar to the previous tree, maximum rainfall intensity is the top splitting variable and also the value of the split (16 mm h^{-1} vs. 15 mm h^{-1}) is consistent between trees. Maximum rainfall intensity appears another two times lower in the tree (node 4 and 6), which emphasizes the importance of this variable in explaining claim frequency. For low-intensity rainfall events (rmax < 16 mm h^{-1}), ownership structure is a significant variable: districts with relatively many owner-occupied buildings (own > 0.52) receive more claims than districts with relatively many rented buildings (own < 0.52). Highest claim frequencies are observed for cases with high rainfall intensities (rmax ≥ 16 mm h^{-1}), relatively large and mostly low-rise buildings (floor ≥ 86 m^2, low ≥ 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.

4.3 Variable importance

The importance of variables in predicting claim frequency are listed in Table 6. 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 ownerships structure 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. 9.

4.4 Comparison with Poisson regression models

Table 7 summarizes the regression results after fitting various Poisson regression models to the same data that was used to learn the Poisson trees. Various combinations of explanatory variables were attempted to explain claim frequency. Note that the categorical variable "season" was not included in 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.

Best fits were found for the models that were based on the combination of variables that were actually used in the decision trees (variant 3 in Table 7): $r^2_{\text{cv}} = 0.18$ and $r^2_{\text{cv}} = 0.11$ for property-related and content-related data respectively. Adding more variables (variant 4 in Table 7) 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 Poisson trees (22–26 %).

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 6) 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.

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 more efficiently. Moreover, knowledge on customer groups associated with high claim frequencies may give hints on where damage prevention programs are most likely to have impact. Insights in 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 at 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. Particularly for average claim size, explanatory variables may need to be gathered at individual building level, such as floor type, value of content, door threshold level and level of precaution. Future data collection efforts should focus on collecting data at the subdistrict scale, preferably at the level of individual households. Another possible explanation is that important explanatory variables are missing. As mentioned in the introduction, variables related to urban drainage systems (e.g., storage capacity and sewer type) 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 present study 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.

Although not researched in detail in this paper, 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 “expects” zero counts in the data; however, only non-zero counts were considered in this study. A split criterion based on a deviance function of a distribution that does not allow zero counts, 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). Parameters of this deviance function cannot be estimated explicitly and requires an iterative process. As a consequence, computational times to learn trees increased tremendously (∼ 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 paper 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 which 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. 2.2). Moreover, the choice of splitting criterion effects the way data is partitioned. There may be more appropriate splitting criteria for rate data than the ones tested in present paper. 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). In such approaches, results are aggregated over an ensemble of trees, where each tree is based on random but realistic changes in the training data (Borisov, 2009; Elith et al., 2008; Strobl et al., 2009).

Similar to the conclusions by Merz et al. (2013), this study 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.

This study investigated tree models for claim frequency and average claim size given a likelihood of 99 % of rainfall-related damage. It did not consider tree models for the probability of occurrence of rainfall-related damage, while it is worthwhile to study this too as part of a wider, risk-based approach.

6 Conclusions and recommendations

In this paper, a wide range of factors associated with rainfall-related damage were investigated, using decision tree analysis. For this, district-aggregated claim data from private property insurance companies in the Netherlands were analysed, considering claim frequency and average claim size per day. Analyses 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), ownership structure (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, which suggests that variability in average claim size is related to explanatory variables that cannot be defined at the district scale. 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.

Acknowledgements. This work has been funded by the EU 7th Framework Programme project Smart Resilience Technology, Systems and Tools (SMARTeST 2010–2013). The authors would like to thank the Dutch Association of Insurers, Royal Netherlands Meteorological Institute, Statistics Netherlands and TU Delft's Map Room for their support and making available the data.

References

Blanc, J., Hall, J., Roche, N., Dawson, R., Cesses, Y., Burton, A., and Kilsby, C.: Enhanced efficiency of pluvial flood risk estimation in urban areas using spatial-temporal rainfall simulations, J. Flood Risk Manage., 5, 143–152, doi:10.1111/j.1753-318X.2012.01135.x, 2012. 2265

Borisov, A.: Zero-inflated boosted ensembles for rare event counts, in: Advances in Intelligent Data Analysis VIII: 8th International Symposium on Intelligent Data Analysis, 227–228, Springer, Lyon, France, 2009. 2283

Breiman, L., Friedman, J., Olshen, R., and Stone, C.: Classification and Regression Trees, Wadsworth, Belmont, California, 1984. 2274, 2275, 2276

Cheng, C. S.: Climate change and heavy rainfall-related water damage insurance claims and losses in Ontario, Canada, J. Water Resour. Protect., 04, 49–62, doi:10.4236/jwarp.2012.42007, 2012. 2266

Climate Service Center: Machbarkeitsstudie "Starkregenrisiko 2050", Tech. rep., available at: http://www.climate-service-center.de/imperia/md/content/csc/workshopdokumente/extremwetterereignisse/csc_machbarkeitsstudie_abschlussbericht.pdf, 2013. 2266

Coulthard, T. and Frostick, L.: The Hull floods of 2007: implications for the governance and management of urban drainage systems, J. Flood Risk Manage., 3, 223–231, doi:10.1111/j.1753-318X.2010.01072.x, 2010. 2265

De'ath, G. and Fabricius, K. E.: Classification and regression trees: a powerful yet simple technique for ecological data analysis, Ecology, 81, 3178–3192, doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2, 2000. 2266, 2273, 2274, 2275

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.: Assessing uncertainties in urban drainage models, Phys. Chem. Earth, 42–44, 3–10, doi:10.1016/j.pce.2011.04.007, 2012. 2265

Douglas, I., Garvin, S., Lawson, N., Richards, J., Tippett, J., and White, I.: Urban pluvial flooding: a qualitative case study of cause, effect and nonstructural mitigation, J. Flood Risk Manage., 3, 112–125, doi:10.1111/j.1753-318X.2010.01061.x, 2010. 2265

Einfalt, T., Pfeifer, S., and Burghoff, O.: Feasibility of deriving damage functions from radar measurements, in: 9th International Workshop on Precipitation in Urban Areas, 245–249, St. Moritz (Switzerland), 2012. 2266

Elith, J., Leathwick, J. R., and Hastie, T.: A working guide to boosted regression trees, J. Anim. Ecol., 77, 802–13, doi:10.1111/j.1365-2656.2008.01390.x, 2008. 2283

Freni, G., La Loggia, G., and Notaro, V.: Uncertainty in urban flood damage assessment due to urban drainage modelling and depth-damage curve estimation, Water Sci. Technol., 61, 2979–93, doi:10.2166/wst.2010.177, 2010. 2265

Garne, T. W., Ebeltoft, M., Kivisaari, E., and Moberg, S.: Weather related damage in the Nordic countries, Tech. rep., available at: http://www.fkl.fi/materiaalipankki/tutkimukset/Dokumentit/Weather_related_damage_in_the_Nordic_countries.pdf, 2013. 2265

Hess, K. R., Abbruzzese, M. C., Lenzi, R., Raber, M. N., and Abbruzzese, J. L.: Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma, Clinical Cancer Research: an official journal of the American Association for Cancer Research, 5, 3403–10, available at: http://www.ncbi.nlm.nih.gov/pubmed/10589751, 1999. 2266

Horn, B.: Hill shading and the reflectance map, Proc. IEEE, 69, 14–47, doi:10.1109/PROC.1981.11918, 1981. 2271, 2291

Hurford, A., Parker, D., Priest, S., and Lumbroso, D.: Validating the return period of rainfall thresholds used for Extreme Rainfall Alerts by linking rainfall intensities with observed surface water flood events, J. Flood Risk Manage., 5, 134–142, doi:10.1111/j.1753-318X.2012.01133.x, 2012. 2265

Jaffrain, J. and Berne, A.: Influence of the subgrid variability of the raindrop size distribution on radar rainfall estimators, J. Appl. Meteorol. Clim., 51, 780–785, doi:10.1175/JAMC-D-11-0185.1, 2012. 2281

Jak, M. and Kok, M.: A database of historical flood events in the Netherlands, in: Flood Issues in Contemporary Water Management, NATO Science Series 2, Environmental Security, 139–146, 2000. 2265

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.: Comparative flood damage model assessment: towards a European approach, Nat. Hazards Earth Syst. Sci., 12, 3733–3752, doi:10.5194/nhess-12-3733-2012, 2012. 2265

Kreibich, H., Thieken, A. H., Petrow, Th., Müller, M., and Merz, B.: Flood loss reduction of private households due to building precautionary measures – lessons learned from the Elbe flood in August 2002, Nat. Hazards Earth Syst. Sci., 5, 117–126, doi:10.5194/nhess-5-117-2005, 2005. 2265

Lee, S.-K. and Jin, S.: Decision tree approaches for zero-inflated count data, J. Appl. Stat., 33, 853–865, doi:10.1080/02664760600743613, 2006. 2275

Merz, B., Kreibich, H., Thieken, A., and Schmidtke, R.: Estimation uncertainty of direct monetary flood damage to buildings, Nat. Hazards Earth Syst. Sci., 4, 153–163, doi:10.5194/nhess-4-153-2004, 2004. 2265

Merz, B., Kreibich, H., Schwarze, R., and Thieken, A.: Review article "Assessment of economic flood damage", Nat. Hazards Earth Syst. Sci., 10, 1697–1724, doi:10.5194/nhess-10-1697-2010, 2010. 2265

Merz, B., Kreibich, H., and Lall, U.: Multi-variate flood damage assessment: a tree-based data-mining approach, Nat. Hazards Earth Syst. Sci., 13, 53–64, doi:10.5194/nhess-13-53-2013, 2013. 2265, 2266, 2267, 2283

Moisen, G. G.: Classification and Regression Trees, in: Encyclopedia of Ecology, edited by: Jørgensen, S. E. and Fath, B. D., Elsevier, Oxford, UK, 582–588, 2008. 2274, 2275

Overeem, A., Holleman, I., and Buishand, A.: Derivation of a 10-Year Radar-Based Climatology of Rainfall, J. Appl. Meteorol. Clim., 48, 1448–1463, doi:10.1175/2009JAMC1954.1, 2009. 2270, 2290

Overeem, A., Leijnse, H., and Uijlenhoet, R.: Measuring urban rainfall using microwave links from commercial cellular communication networks, Water Resour. Res., 47, 1–16, doi:10.1029/2010WR010350, 2011. 2290

Parker, D. J., Priest, S. J., and McCarthy, S.: Surface water flood warnings requirements and potential in England and Wales, Appl. Geogr., 31, 891–900, doi:10.1016/j.apgeog.2011.01.002, 2011. 2281

Peleg, N., Ben-Asher, M., and Morin, E.: Radar subpixel-scale rainfall variability and uncertainty: lessons learned from observations of a dense rain-gauge network, Hydrol. Earth Syst. Sci., 17, 2195–2208, doi:10.5194/hess-17-2195-2013, 2013. 2281

Pistrika, A. K. and Jonkman, S. N.: Damage to residential buildings due to flooding of New Orleans after hurricane Katrina, Nat. Hazards, 54, 413–434, doi:10.1007/s11069-009-9476-y, 2009. 2265

Pitt, M.: Learning lessons from the 2007 floods, and independent review by Sir Michael Pitt, Tech. rep., The Pitt Review Cabinet Office, London, UK, 2008. 2265

Priest, S. J., Parker, D. J., Hurford, a. P., Walker, J., and Evans, K.: Assessing options for the development of surface water flood warning in England and Wales, J. Environ. Manage., 92, 3038–48, doi:10.1016/j.jenvman.2011.06.041, 2011. 2281

Rejwan, C., Collins, N. C., Brunner, L. J., Shuter, B. J., and Ridgway, M. S.: Tree Regression Analysis on the Nesting Habitat of Smallmouth Bass, Ecology, 80, 341, doi:10.2307/177003, 1999. 2266

Ririassa, H. and Hoen, A.: Rainfall and damage: a study on relationships between rainfall and claims in relation to climate change (in Dutch), Tech. rep., Dutch Association of Insurers, 2010. 2266, 2290

Seifert, I., Botzen, W. J. W., Kreibich, H., and Aerts, J. C. J. H.: Influence of flood risk characteristics on flood insurance demand: a comparison between Germany and the Netherlands, Nat. Hazards Earth Syst. Sci., 13, 1691–1705, doi:10.5194/nhess-13-1691-2013, 2013. 2268

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and ten Veldhuis, J. A. E.: A statistical analysis of insurance damage claims related to rainfall extremes, Hydrol. Earth Syst. Sci., 17, 913–922, doi:10.5194/hess-17-913-2013, 2013a. 2266, 2268, 2269

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and Ten Veldhuis, J. A. E.: A spatial analysis of rainfall damage data using C-band weather radar images, in: International Conference of Flood Resilience, Exeter, UK, available at: http://repository.tudelft.nl/view/ir/uuid:f0c17744-0609-4e93-b2fe-c5b8a24b7e4a/, 2013b. 2266, 2270

Strobl, C., Malley, J., and Tutz, G.: An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests., Psychol. Methods, 14, 323–48, doi:10.1037/a0016973, 2009. 2283

Ten Veldhuis, J. A., Clemens, F. H., and van Gelder, P. H.: Quantitative fault tree analysis for urban water infrastructure flooding, Struct. Infrastruct. E., 7, 809–821, doi:10.1080/15732470902985876, 2011. 2266, 2271

Ten Veldhuis, J. A. E.: How the choice of flood damage metrics influences urban flood risk assessment, J. Flood Risk Manage., 4, 281–287, doi:10.1111/j.1753-318X.2011.01112.x, 2011. 2265

Therneau, T. M. and Atkinson, E. J.: An Introduction to Recursive Partitioning Using the RPART Routines, Tech. rep., available at: http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf, 2014. 2274, 2276

Thieken, A. H., Müller, M., Kreibich, H., and Merz, B.: Flood damage and influencing factors: New insights from the August 2002 flood in Germany, Water Resour. Res., 41, 1–16, doi:10.1029/2005WR004177, 2005. 2265, 2272

Van der Sande, C., Soudarissanane, S., and Khoshelham, K.: Assessment of relative accuracy of AHN-2 laser scanning data using planar features., Sensors, 10, 8198–214, doi:10.3390/s100908198, 2010. 2271, 2290

Van der Zon, N.: Background information about AHN2 (in Dutch), Tech. rep., Actueel Hoogtebestand Nederland, Amersfoort, available at: http://www.ahn.nl/\protect\T1\textdollar1b6l/page/, 2013. 2271, 2290

Weiss, A. D.: Topographic Position and Landforms Analysis, Poster presentation, ESRI Users Conference, San Diego, CA, 2001. 2271, 2291

Wilson, M. F. J., O'Connell, B., Brown, C., Guinan, J. C., and Grehan, A. J.: Multiscale terrain analysis of multibeam bathymetry data for habitat mapping on the continental slope, Marine Geodesy., 30, 3–35, doi:10.1080/01490410701295962, 2007. 2271

Zhou, Q., Mikkelsen, P. S., Halsnæs, K., and Arnbjerg-Nielsen, K.: Framework for economic pluvial flood risk assessment considering climate change effects and adaptation benefits, J. Hydrol., 414–415, 539–549, doi:10.1016/j.jhydrol.2011.11.031. 2012. 2265

Zhou, Q., Panduro, T. E., Thorsen, B. J., and Arnbjerg-Nielsen, K.: Verification of flood damage modelling using insurance data., Water Sci. Technol., 68, 425–32, doi:10.2166/wst.2013.268, 2013. 2266

Table 1. Overview of data sources used in this study.

#	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 of 2007–2012.	Online viewer: http://ahn.geodan.nl/ahn/. More background: Van der Sande et al. (2010); Van der Zon (2013).

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

Table 2. Model variables and variable definitions. Value ranges (column 3) are related to subsets of property and content claim data respectively.

Variable name	Definition	Min – Max (Median)	Source
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 (Euro)	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 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[illegible]–5.2 (3.7)	
Age of breadwinner (age1)	Median age of main breadwinner per district (yr)	24–68 (51); 27–72 (50)	3
Ownership structure (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 (Euro)	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 (yr)	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

Table 3. Inflation adjustment according to the online database of Statistics Netherlands (http://statline.cbs.nl). The average inflation per year for the Netherlands is used (second column). 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

Table 4. 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(Average claim size) $= y_i$	Normal ($\mu;\sigma$)	$D = \sum \left[w_i (y_i - \hat{\mu})^2 \right]$	$\hat{\mu} = \frac{\sum w_i y_i}{n}$
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)	$D = 2 \sum \left[k_i \log\left(h^{-1}(k_i) \right) - h^{-1}(k_i) - \log\left(1 - \exp(-h^{-1}(k_i)) \right) - k_i \log\left(\hat{\lambda} K_i \right) + \hat{\lambda} K_i + \log\left(1 - \exp(-\hat{\lambda} K_i) \right) \right]$ where $h(x) = \frac{x}{1 - \exp(-x)}$	Iterative procedure

Table 5. 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	–

Table 6. 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 7. Poisson regression: claim frequency modelled as a function of (1) the most important variable according to Table 6, (2) all rainfall-related variables, (3) the variables actually used in the decision tree and (4) the variables with importance score > 0.02. The cross-validated coefficient of determination, r^2_{cv}, is calculated using a similar approach as discussed in Sect. 3.2.

Response variable ∼ Explanatory variables	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
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
4: rmax + rmean + rvol + own + floor + low + inc + rev + edu	0.26	0.12

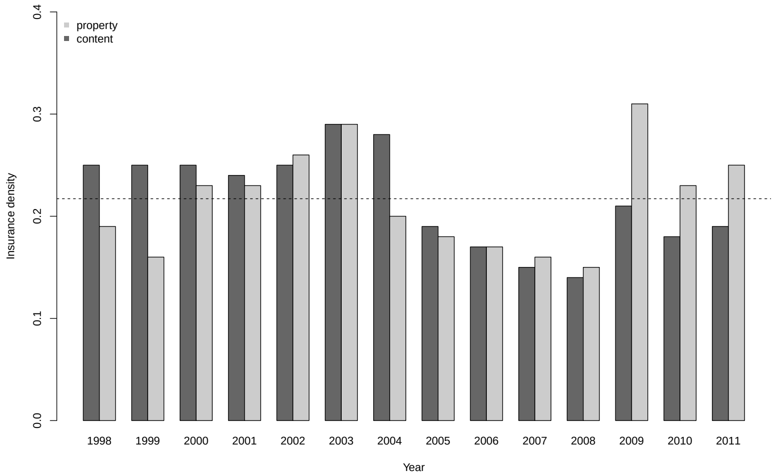

Fig. 1. Insurance density per year: the number of insured households in the database from the Dutch Association of Insurers per year divided by the total number of households in the Netherlands per year. Light bars represent property insurance, dark bars content insurance. The dashed horizontal line (= 22 %) represents the average insurance density for the period 1998–2011 (the same percentage for content and property insurance).

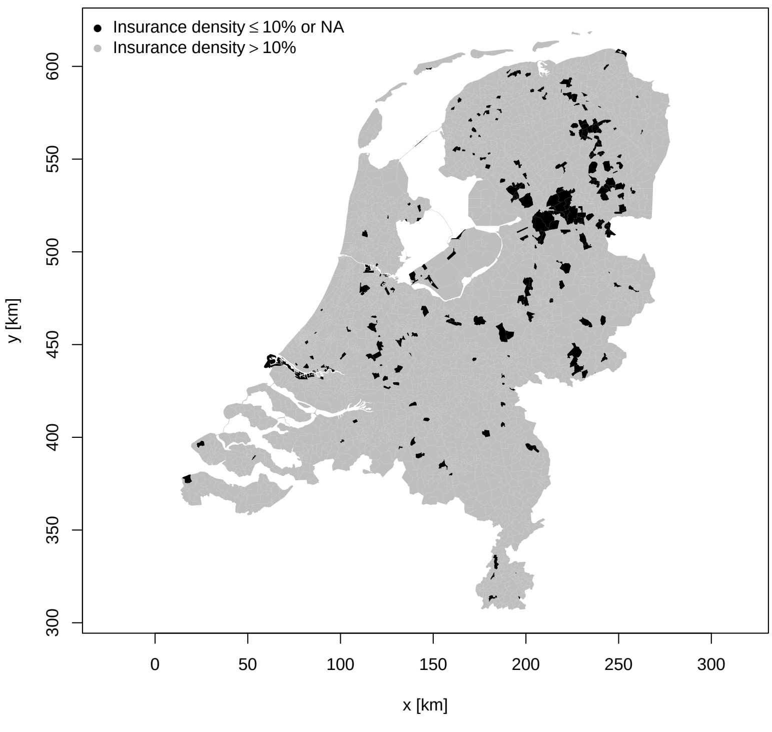

Fig. 2. Property insurance density: the fraction of house owners included in the database from Dutch Association of Insurers, averaged over the years 1998–2011. Dark areas denote districts that have an insurance density of less than 10 % or where values are not available. Note that this figure is slightly different for individual years.

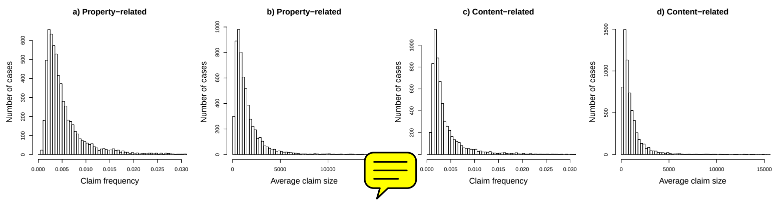

Fig. 3. Histograms of response variables in subsetted data: **(a)** claim frequency of property-related cases, **(b)** average claim size of property-related cases, **(c)** claim frequency of content-related cases and **(d)** average claim size of content-related cases. Histograms of claim frequency and average claim size have a bin size of 0.0005 and 250 Euros respectively.

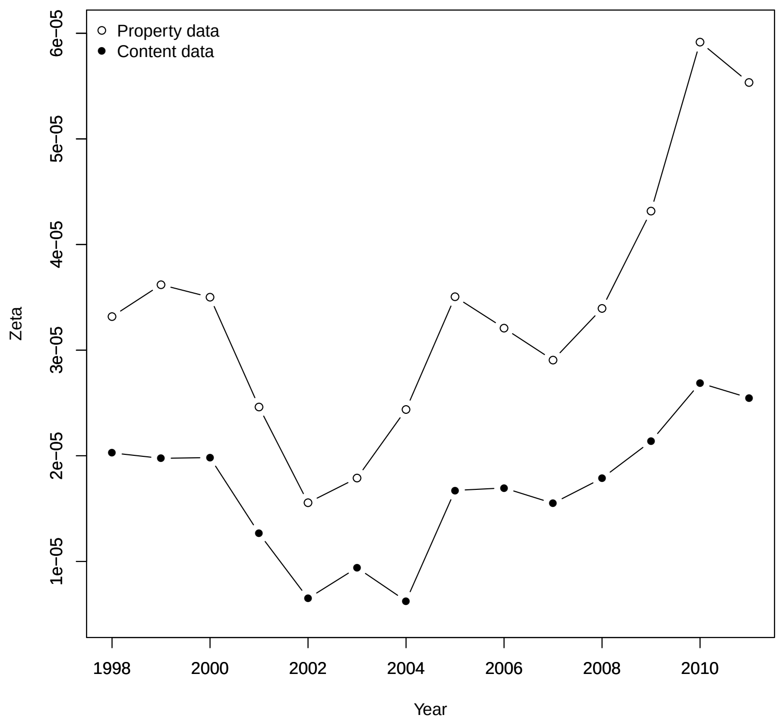

Fig. 4. Average probability of a non-rainfall-related claim per day per policyholder for the years 1998–2011. The white dots are related to property claim data, the black dots to content claim data.

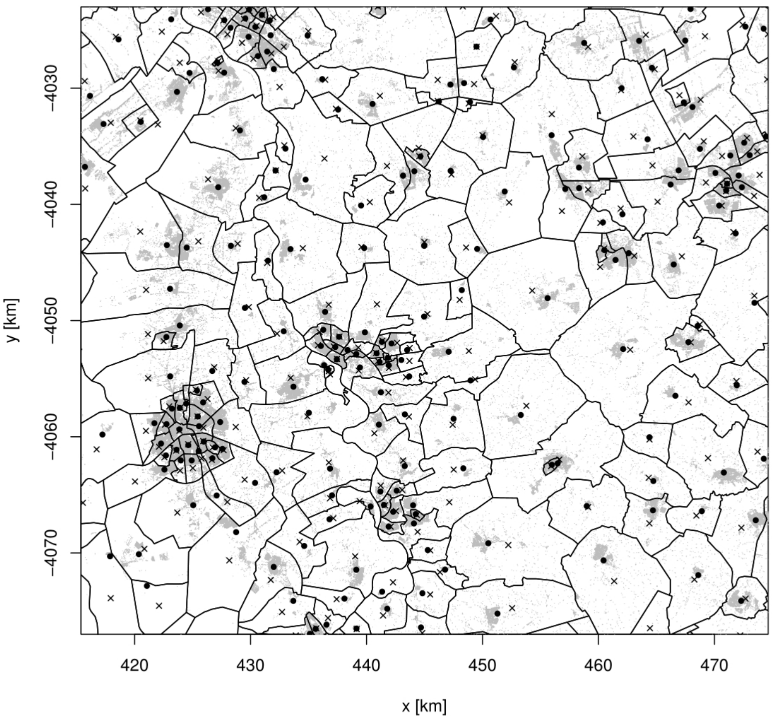

Fig. 5. An example map showing postal districts (polygons), their geometric centroid (crosses) and their building-weighted centroids (dots). The grey dots are residential areas used in the weighting.

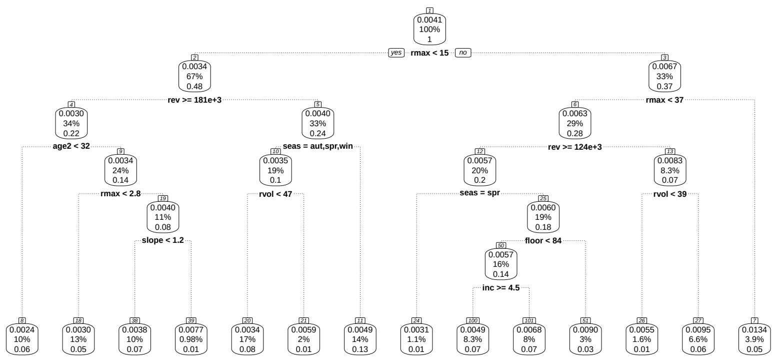

Fig. 6. Poisson tree explaining the property claim frequency as a function of rainfall-related, building-related, socioeconomic and topographic variables (tree size = 14). The values at nodes are, from top to bottom: (1) node index, (2) claim frequency (i.e., Poisson group mean), (3) percentage of claims falling into the group and 4) remaining deviance relative to the deviance of the undivided data.

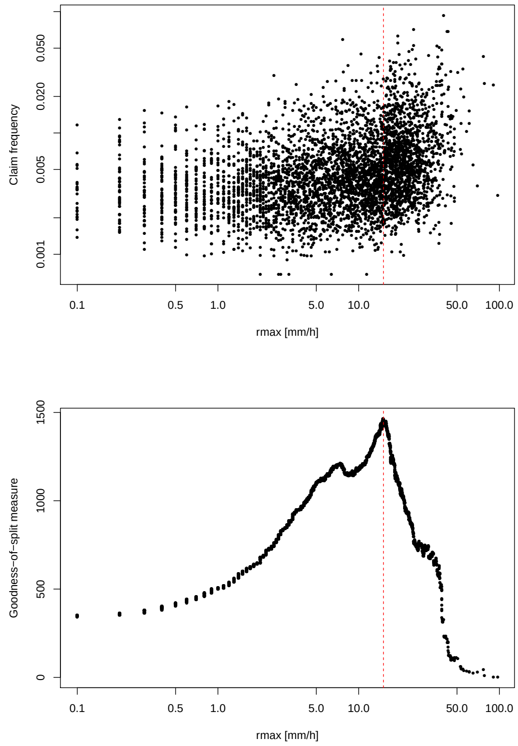

Fig. 7. Scatter plot of claim frequency against maximum rainfall intensity, for the undivided data (top figure). The dashed vertical line represent splitting value that maximizes the goodness-of-split measure (bottom figure).

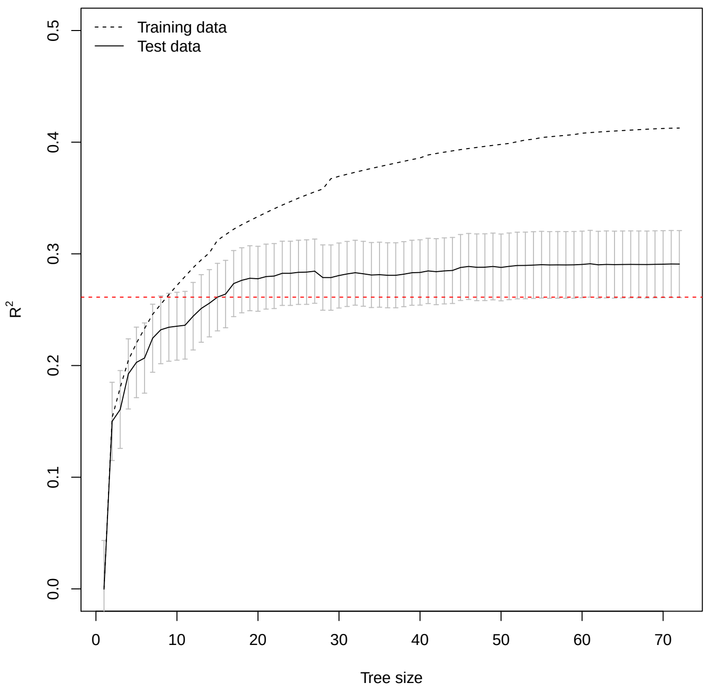

Fig. 8. Performance of the Poisson tree for property claim frequency: the fraction of variance explained by the tree as a function of tree size, based on training data (dashed line) and validation data (solid line). The error bars represent one standard deviation of uncertainty. The horizontal dashed line is based on the 1-SE rule (see Sect. 3.2). The intersection of this line with the solid one determines the size of the tree.

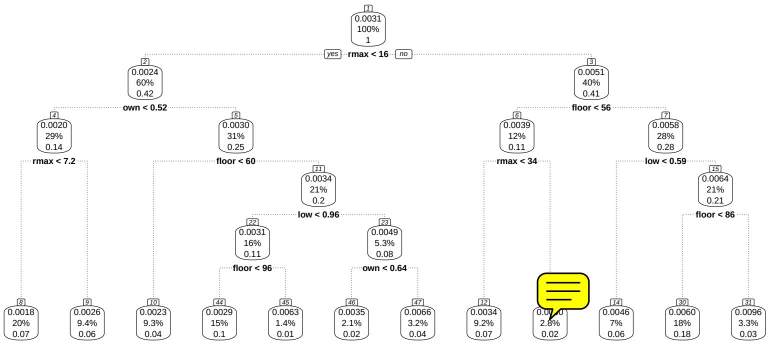

Fig. 9. Poisson tree explaining the content claim frequency (tree size = 12). The values at nodes are, from top to bottom: (1) node index, (2) claim frequency (i.e., Poisson group mean), (3) percentage of claims falling into the group and (4) remaining deviance relative to the deviance of the undivided data.