Nat. Hazards Earth Syst. Sci. Discuss., 2, C1368–C1385, 2014
www.nat-hazards-earth-syst-sci-discuss.net/2/C1368/2014/

Natural Hazards and Earth System Sciences
Discussions

Interactive comment on "Decision tree analysis of factors influencing rainfall-related building damage" *by* M. H. Spekkers et al.

M. H. Spekkers et al.

m.h.spekkers@tudelft.nl

Received and published: 9 July 2014

We thank the reviewer for his/her time and effort in commenting our manuscript. Our response:

Major comments

RC1: *For the main part, Section 3 on the Methodology needs revision. This concerns the concept of surrogate variables which is mentioned as an important feature of Decision Trees to cope with the problem of missing values and therefore should be explained. How many cases in the data base are affected?*

AC1: Lines 15–16 on page 2275 needs revisions. Not only because of the valid point the reviewer is making, but also because the paragraph incorrectly describes how we (finally) dealt with missing data. The use of surrogate splits to impute missing data is a common approach in decision tree learning (see e.g. Breiman et al. (1984)). Initially, we applied this approach to our study. In the course of the research, however, we realised that it was not the most suitable approach given our data. In our case, the main source of missing data is rainfall data. Because (obviously) rainfall-related variables do not correlate well with any of the other explanatory variables, none of explanatory variables can act as a proper surrogate when rainfall data is missing. (Note that if rainfall data is missing, none of rainfall-related variables are available, thus they cannot substitute each other.) Alternatively, we discarded the cases without rainfall data. Originally, missing socioeconomic data was an issue, but after we discarded cases where the number of policyholders was less than 100 (line 19–20, page 2269), none of the records contained missing values for socioeconomic variables.

We would, therefore, like to rewrite this paragraph as follows: "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 present study can act as a suitable surrogate. Alternatively, we discarded the cases without rainfall data (8–11% of the cases). Still, surrogate variables were recorded at each node for the purpose of calculating variable importance (see Sect. 3.2)."

RC2: *Further, please explain what is meant with 'training data' in comparison to cross validation data. Is training date the complete data set?*

AC2: Training data indeed refers to the complete data set. To clarify this, we will add the following after line 24 at page 2273 "(...) until a large tree is learned. Trees are trained based on the complete data set." The remainder of the paragraph will be moved to Sect. 3.2. Section 3.2 will then start as follows: "The large tree is trimmed back to a simpler tree that still contains most of the predictive power of the large tree. The right size of tree is determined using 10-fold cross-validation. The following (...)". Moreover, line 13–15 on page 2278 will be changed into: "The tree explains 32 % of the variance in training data (i.e., $R^2 = 1 - \frac{sum of deviance at terminal nodes}{deviance of undivided data}$) and, on average, 26 % of the variance in the cross-validation data sets (Fig. 8)."

RC3: *Why is global regression only conducted for claim frequency and not for claim size? Please add a description of Poisson regression models.*

AC3: Our primary focus was on a global regression model for claim frequency, because only trees could be derived for claim frequency and not for claim size. We agree with the reviewer (see also RC5) that it is good scientific practice to also report claim size results. Section 3.3 will be rewritten as follows: "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:

$$\log(k_i) = \log(K_i) + \beta_0 + \beta_1 x_{1i} + \ldots + \beta_n x_{ni}, \quad (1)$$

where k_i is the number of claims observed for case i, K_i is the number of insured households for case i and $\beta_0, , \beta_n$ 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:

$$\log(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_n x_{ni} + \varepsilon_i, \quad (2)$$

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." Table 7 is updated with "claim size results". (Updated tables can be found at the end of the document; updates figures can be found in the supplement to this document.)

RC4: *For the evaluation of model performance it could be interesting to include additional performance criteria which represent the precision of model predictions (e.g. mean bias, root mean square error) or which also reflect the complexity of the model (e.g. BIC, AIC).*

AC4: The performance indices the reviewer mentions can be useful, particularly when the aim is to develop reliable models and various model structures have to be compared (in terms of precision/complexity). The focus of this study is, however, on the identification of the important factors that influence claim frequency and size and less on model development. Therefore, we have limited ourselves to evaluation of R^2 and the variable importance index.

RC5: *Further, the discussion of results should be more detailed concerning the failure to derive models for claim size. It is likewise important to understand why the model approaches did not work for the data at hand and to identify possible approaches to overcome these problems.*

AC5: Good point. We will elaborate on this a bit more in discussion. Line 23–25 on page 2281 will be moved to line 11, page 2282 (starting a new paragraph) and

rewritten as follows: “There can be a number of explanations for the failure to derive tree models for average claim size. First, 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, independently of the flood depth. Thus, rainfall-related variables may be less informative in this context. Second, damage assessments are inherently uncertain, because of interpretation errors of insured and damage experts, which are difficult to capture in a model. Third, claim size directly relates to the value of the damaged materials and goods. This type of information will be lost when aggregating data to a district level.”

Minor comments

RC6: *p. 2265, line 14: a total number would make the comparison to the above numer easier.*

AC6: Good suggestion. The line will be changed to: “Recent figures from Nordic insurance industry show that damage to residential buildings due to heavy rainfall was around 300 million euros per year in Denmark, for the years 2009–2011 (Garne et al., 2013).”

RC7: *p. 2267, line 24: given all buildings being insured, is there room for expanding the data base by including mor than included witin the 22% coverage of households?*

AC7: Not without much efforts. The database used has not been built for the purpose of this study. The Dutch Association of Insurers uses the database for general insurance statistics (e.g. yearly reports for Dutch insurance industry). For that purpose a coverage of 22% is sufficient. To extent the database, many more, often small-sized

companies have to be reached. Not only is the data that can potentially be gained from these companies small, it requires much manual work to standardize various data formats insurers use. Ideally, data steams from insurers should be automated, standardized and centrally stored, but that, of course, requires huge investments.

RC8: *p. 2268, line 5: the term ’damages’ refers to compensation. Please use ’damage’ if you refer to the adverse consequences (also in the remainder of the paper).*

AC8: Good point by the reviewer. In this example, however, we do refer to adverse consequences of water intrusion/flooding. To clarify this, we will rephrase the sentence as follows: “Damage to building structure and content can have a wide range of causes, (...)”. We will change “damage” to “compensation” where appropriate.”

RC9: *p. 2269, line 19: reference not clear*

AC9: The “$>$ 0.1” means “a claim frequency of more than 1 per 10 policyholders”. This value remains an arbitrary choice, but most cases associated with claim frequency above this threshold could not be explained by rainfall and are likely related to data errors that could not be solved during the preprocessing of the data.

RC10: *p. 2270, line 3: to improve readability it would be good to state all the variables also in the text, not only in table 2.*

AC10: OK. We will change the first paragraph of Sect. 2.3.1 to: “For each case in the subset, rainfall volume, rainfall duration, maximum rainfall intensity and mean rainfall intensity were extracted from weather radar data. Definitions of these variables can be

found in Table 2. (...) The rainfall-related variables were obtained using the following steps, as is also described in Spekkers et al., (2013b)."

RC11: *p. 2270, line 13–15: What is the reasoning behind this assumption? What about event and drainage system specific differences? Please cite an appropriate reference to substantiate this assumption.*

AC11: Due to the flatness of the country, storage in urban drainage systems (below the level of the overflow weir) is important. The time it takes for an urban drainage system to restore to equilibrium (i.e., a state with only dry weather flow) depends on the static storage capacity of the system and the pump overcapacity (i.e., capacity installed at the treatment plant minus average dry weather flow). For Dutch sewers systems, design criteria are available that state that urban drainage systems should restore to an equilibrium state in around 10 to 24 hours (Stichting RIONED., 2008). We agree with the reviewer that drainage system properties may vary from location to location. For practical reasons (i.e., data on system properties not being available at a nationwide scale), we have selected a fixed time of 12 hours. We will add the aforementioned reference to the text after line 15 on page 2270.

RC12: *p. 2271, line 2: what is the basic resolution of the DEM?*

AC12: It is not entirely clear what the reviewer means with "basic resolution". The data we have been using was available at a 5 m x 5 m spatial resolution, see line 7 on page 2271 and Table 1.

RC13: *p. 2272, line 12–15: how many cases are lost due to the privacy restrictions?*

AC13: See also AC1. Although privacy regulations mentioned in line 12–15 did affect the collection of "raw" data, none of the cases in the subset that was used for tree building contained missing socioeconomic data. Therefore, line 12–15 might as well be removed.

RC14: *p. 2273, line 18: what defines a best split? insert the explanations given below here.*

AC14: We have rewritten the paragraph (starting at line 17 on page 2273) to integrate RC14 and RC15. We have avoided the term "homogeneity" and used the more common statistical term "variance" instead: "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 minimizes the variance of the response variable in the resulting groups, is selected."

RC15: *p. 2273, line 22: homogeneity in terms of?*

AC15: See AC14.

RC16: *p. 2273, line 24 to p. 2274, line 2: move sentences to sec. 3.2*

AC16: See AC2.

RC17: *p. 2274, line 25: what is meant by rate data?*

AC17: We mean event rate data, data that report how often events occur within a certain unit of time or space. We will add "event" before "rate" to be more precise.

RC18: *p. 2276, line 17: why?*

AC18: See AC3.

RC19: *p. 2276–2277, section 4.1: claim size related results are not discussed! please comment on data given in Table 5*

AC19: The following will be changed in Sect 4.1: line 4–6 on page 2277 ("Moreover, there are a large number of significant links between explanatory variables and claim frequency than between explanatory variables and average claims size.") will be moved to the end of line 15. After that, we will add the following lines: "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 ownership structure were significantly correlated with average claim size, for property-related and content-related claims."

RC20: *p. 2278, line 13–15: these values should be marked in Figure 8. Explain better the reasoning behind the intersection of horizontal line and tree results. Why not also include results for content related claim frequency? What are the corresponding*

optimum tree structures in Figures 6 and 9?

AC20: Good suggestion to improve visualisation of Fig. 8. A new version of Fig. 8 can be found in the supplement to this document. The caption of the figure now contains a better explanation of how tree size was selected.

The figure is just to illustrate the 10-fold cross-validation process. Adding the figure for content claim frequency does not add much to this explanation. They more or less show the same thing.

Figure 6 and 9 already show the optimum, or pruned, tree structure. To be more clear on this, we will add the term "pruned" to the captions: "Pruned poisson tree explaining (...)". Since we have not defined the term "pruned" yet, we will add the following to line 9 on page 2276: "This tree is referred to as the pruned tree."

RC21: *p. 2279, line 6: What are the reasons for this? It is likewise important to understand why the model approaches did not work for the data at hand and to identify possible approaches to overcome these problems.*

AC21: See AC5.

RC22: *p. 2279, line 20–22: it is not shown which variables are used for surrogates where in the decision trees.*

AC22: See first AC1. Surrogate variables were not used to split data. All splits that are shown in the decision trees are primary splits. Surrogates were used, however, for the calculation of variable importance (Table 6). That is why mean rainfall intensity is still reported in Table 6, although it does not appear in the decision trees.

RC23: *p. 2280, line 2: the methodology of Possion regression modesl is not described.*

AC23: Good point. In AC3 we propose a better description of the global models. The lines 4–8 on page 2280 ("Note that the categorical variable (...) the explanatory variables.") will be moved to section 3.3.

RC24: *p. 2280, line 12–14: For clarity it would be helpful to compile all results in one table and not to provide them distributed over text, figures and tables.*

AC24: Helpful suggestion. We generated a new table that includes all results, see Table 7 at the end of the document, replacing the original Table 7. The new table also takes into account RC3. Note that caption has been changed too.

RC25: *p. 2280, line 22: Actually in this paper two different damage pathways are jointly considered: direct impact of rainfall and pluvial damage. Could a separation of the data base according to damage pathway provide an improvement? I would expect that the different factors will be of varying importance in each case.*

AC25: The reviewer makes an interesting point here. We also expect improved relationships when the two damage mechanisms are analysed separately. Unfortunately, the database lacks information on these mechanisms. So, we cannot make separate analyses. In fact, we are working in another study on a much more detailed insurance database. This database contains communication transcripts of calls and reports between insurer, client and damage experts, which allows us to study individual damage mechanisms.

RC26: *p. 2282, line 16: what are zero counts? please be more detailed*

AC26: It is an uncommon term and should therefore be corrected. The term can be avoided by slightly rephrasing line 14–16 on page 2282: "The Poisson deviance function that was used allows responses to be zero (i.e., no claim); however, only cases with claims were considered in this study. 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."

RC27: *p. 2283, line 11–12: please name some examples*

AC27: To the best of our knowledge, literature on splitting criteria for event rate data is very limited, and no examples were found of an alternative splitting criterion. The one we used is based on the Poisson distribution, which is a commonly used distribution to model count data. Other distributions for count data that are commonly used are the binomial and negative binomial distributions. Similarly to the splitting criteria that is based on the Poisson distribution, splitting criteria may be developed based on other distributions for count data. We would like to extend line 11–12 on page 2283: "There may be more appropriate splitting criteria for rate data than the ones tested in present paper, for example, splitting criteria based on other distributions for count data, such as the binomial or the negative binomial distribution."

RC28: *p. 2284, line 10–12: or rather need further investigation? What about the size of the underlying data set? Expanding the data base could help to carve out non-linear and/or local relationships.*

AC28: This comment also related to RC15 by Bouwer. We agree with both reviewers that there may be more reasons that explain the poor relationships with claim size as found in this study. We would, therefore, like to change line 10–12 on page 2284 and line 21–23 on page 2264 (abstract) to: "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. This includes the inclusion of other explanatory factors that were not used in present study, 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." We think that the size of the data set will not matter that much as it is already considerably large. Larger sample size (getting closer to actual population size) will probably result in more significant relationships, however, their effect size may not have any importance in a practical context.

RC29: *p. 2291, table 2: separate columns for property and content claims for improved readability*

AC29: We have updated Table 2, see end of the document.

RC30: *p. 2291, table 2: has education level a continuous scale?*

AC30: Education level is not continuous with respect to individuals. There are seven main levels, defined on an ordinal scale. Here, we have, for practical reasons, aggregated levels to districts by averaging the levels, assuming equal intervals between the scales.

RC31: *p. 2291, table 2: why mean and not median?*

AC31: No significant differences were found between the two. We, therefore, only reported one of them.

RC32: *p. 2293, table 4: with which objective?*

AC32: Good point. The bottom right cell of the table will be replaced by: "$\hat{\lambda}$ using maximum likelihood estimation" We will also add the following footnote to the table: "$h^{-1}(x)$ needs to be calculated numerically, which is inconvenient for decision tree learning where deviance needs to be evaluated for every split."

RC33: *p. 2295, table 6: some of them are not significant as given in Table 5*

AC33: Table 5 and 6 should be treated separately. Table 5 presents global regression results. Non-significant variables in this table can still be significant in the tree approach (e.g., because a variable may be significant in a certain node).

RC34: *p. 2299, figure 3: hard to read!*

AC34: See updated Fig. 3 in the supplement to this document.

RC35: *p. 2304, figure 8: I think a non-parametric approach to quantify variation in results is more reasonable, e.g. IQR. What about uncertainty in the DT based on training data?*

AC35: The reviewer proposes an interesting alternative to calculate standard errors. The approach we have been using, where standard errors are calculated using a set of cross-validation results, is, however, a well-established approach in the field of decision tree learning, which we therefore prefer.

The main use of the standard errors here is to select an appropriate tree size. Tree size selection is only based on the set of cross-validation trees and therefore standard errors were not derived for training data.

RC36: *p. 2305, figure 9: combine Figures 6 and 9 in one multi panel Figure. Indicate the nodes and splits included in optimum pruned trees.*

AC36: We will leave this suggestion to the editor. It is a good idea to combine figures, but is should fit the layout of the journal too. Would it be possible to make the figures spanning two columns? The trees in Fig 6 and 9 are already the optimal/pruned versions, see also AC20.

References

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

Stichting RIONED: Module B2200 Functional desgin: collection and transport of stormwater (in Dutch), Technical Report, http://www.riool.net/nl_NL/leidraad-riolering, 2008.

Please also note the supplement to this comment:
http://www.nat-hazards-earth-syst-sci-discuss.net/2/C1368/2014/nhessd-2-C1368-2014-supplement.pdf

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., 2, 2263, 2014.

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) Property data	Min – Max (Median) Content data	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)	2.6–5.2 (3.7)	
Age of breadwinner (age1)	Median age of main breadwinner per district (yr)	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 (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 7. 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 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. 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 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 claim size ∼				
1: rmax	0.02	0.02	-	-
2: rmax + rmean + rvol + rdur	0.02	0.02	-	-
4: all variables	0.05	0.05	-	-