Nat. Hazards Earth Syst. Sci. Discuss., 2, C522–C524, 2014
www.nat-hazards-earth-syst-sci-discuss.net/2/C522/2014/
© Author(s) 2014. This work is distributed under
the Creative Commons Attribute 3.0 License.

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.

Anonymous Referee #1

Received and published: 28 April 2014

The paper analyses factors influencing building damage caused by rainfall using decision trees. The objective is to identify relationships between rainfall and building related data, socio economic factors, topographic indicators and to investigate their usefulness to explain rainfall related damage in urban areas. For this purpose a nationwide data set for The Netherlands is set up on the district level. The data analyses is conducted by application of decision trees as an established data mining approach. The comupations are carried out using the rpart library of the R software. The decision tree models are cross-validated and compared to a global Poission regression model. Model evaluation is carried out using R^2 as a measure for the variance explained by the models. Main conclusions are that decision trees perform better than global regression models suposedly because decision trees are able to capture non-linear and local relationships

in the data. As a large fraction of variance remains unexplained the authors recommend to improve the damage data bases and to collect data on explanatory variables on the level of individual objects at risk, i.e. individual buildings.

Reading the paper is interestin and provides new insight to a topic which has not received much attention in reseach yet, but is, without doubt, of relevance both for the insurance industry and in a broader context for the risk analysis and management of pluvial hazards. The main contributions of the paper are i) the structured compilation of an area-wide (The Netherlands) consistent data base of potential explanatory variables covering various domains (rainfall, building, socio-economic and topographic) and ii) the acquired knowledge concerning the importance of the individual variables from the different domains and. Accordingly, in principle I recommend the paper for publication. However, I see several aspects which need to be complemented or which require a more detailed explanation to make the paper stronger of which the major ones are the following. Please find further minor comments on the attached marked up manuscript.

1. 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?

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

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

4. 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).

5. 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.

Please also note the supplement to this comment:
http://www.nat-hazards-earth-syst-sci-discuss.net/2/C522/2014/nhessd-2-C522-2014-supplement.pdf

---

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