Nat. Hazards Earth Syst. Sci. Discuss.,
doi:10.5194/nhess-2017-7-RC1, 2017

Natural Hazards
and Earth System
Sciences
Open Access
Discussions

Interactive comment on "Data-mining for multi-variable flood damage modelling with limited data" *by* Dennis Wagenaar et al.

Anonymous Referee #1

Received and published: 10 February 2017

The present manuscript about flood damage modeling with limited data pursues an interesting approach. The topic fits well to the scope of NHESS. However, there are several issues that need to be clarified or reconsidered:

1 General Comments

1. The authors may want to reconsider the title of the manuscript. "Data mining" is very prominent in the title, but I think this does not reflect the main focus of this paper. As a matter of fact, the aim of this manuscript is not primarily to do classical data mining on a huge data set (i.e. clustering, anomaly detection, classification), but rather to employ various unsupervised learning algorithms with the

aim of finding the best model to explain flood damage with a couple of independent variables (which of course is a part of data mining). Thus, the aim is to compare methodologies for a specific application example, rather than discovering patterns in a huge data set. To emphasize the focus of this work (multivariate flood damage modeling, limited data), I would suggest to rephrase the title to "Multi-variable flood damage modeling with limited data using supervised learning approaches."

2. Results and conclusions should be pointed out more clearly in the abstract. Please provide concise information regarding the improvements instead of pointing out a "significant improvement" and mentioning that some models "perform better".

3. With respect to the presentation quality, I advise to work on the language, on the structure of the manuscript and on the presentation of results. The "common thread" and the main take-home messages are not fully clear and concise throughout the manuscript. The discussion is relatively short, even though there are plenty of interesting aspects in this research that would be worth discussing, and that need to be discussed against the background of uncertain input data. Some formulations are too difficult to understand from a linguistic point of view. For instance p2, l23: "More commonly available (although still rare) are simple datasets that hold records with the flood damage that occurred for each building with sometimes a few other variables (such as location or water depth)."

4. The authors might want to think about the reference function. The root function is a simple, univariate function, which serves as a reference for sophisticated multivariate methods. Total damage and water depth are correlated with r = 0.18; I would assume that this value doesn't change significantly when calculating the correlation between total damage and the square root of the water depth. So the reference model is actually a rather bad model, possible improvements regarding

the GOF when using more advanced methods seem natural. It might be interesting to include a more sophisticated regression model as a reference (e.g. using LASSO, as this includes both variable selection and regularization).

5. Results with respect to the most important variables should be reported in greater detail. It is not clear to me which of the variables are actually beneficial for modeling total damage. The correlation coefficients in Table 1 do not provide any information on that, neither do the other tables or the results section. Variable selection is not discussed at all in this manuscript. For instance, total damage in the Bayesian networks is apparently (c.f. p25, l25-32; Figure 3) influenced by water depth (data-driven network) or water depth and structure damage (expert network), implying that there is no added value of using additional data. Table 4 somehow indicates that the increase in GOF is primarily dependent on the algorithms applied, rather than on additional data.

6. Given that the benefits of additional data are emphasized in this manuscript (p11, line 5; p11, line 26-27), it is worth discussing that care has to be taken when introducing additional data to a model. Even though pruning and bagging is mentioned, this topic is not really emphasized in this manuscript. Showing awareness of regularization and penalization of additional variables is of prime importance.

7. Uncertainty estimation is mentioned as one of the main merits of the methods used in this manuscript (p2, l16). However, uncertainties are neglected in the discussion section of this paper. Even though a number of sources for uncertainty are pointed out (e.g. data collected by different organizations; exact locations of buildings are not known; water depth is only based on estimates and has been questioned by experts; collection methods for variables "inhabitants", "basement" and "attached buildings" are unknown; uncertain join of data based on water depth rank), implications are not discussed.

8. Tables:

(a) Table 1: last column should read “Pearson correlation on total damage” (there are 3 different damage variables in the data set).

(b) Table 2: caption: “...algorithms”. Column names for col 2 and 3 need to be more specific, “water depth” is also part of the “original variables”. However, the authors may wish to reconsider if this table is really needed, all information presented is the text.

(c) Table 3: It is not clear to which dataset (water depth only / original data set / all variables) these values refer to. I guess it is the data set containing all variables, except for the root function? In addition, the authors may wish to consider adding a GOF-measure for the explained variation (i.e. R^2) to the table.

(d) Table 4: It is not really clear how the “best performing” models have been selected – seemingly on the correlation coefficient? Table 3 indicates that RMSE and correlation coefficient of bagging regression tree and random forest show almost identical GOF.

(e) May I propose to combine Tables 3 and 4 by reporting all values (i.e. RMSE, r, and maybe R^2 for all 6(5) methods, structured by input dataset). This would also incorporate the information from Table 2.

9. Figures:

(a) Figure 1: please rephrase caption, e.g. “Scatter plot showing the relation between water depth and damage in the original data set”.

(b) Figure 2: It might be interesting to check if plotting the affected houses atop the water depth is easier to understand. It seems that some houses on the left map are not located in the inundated area at all, albeit they are labelled as “affected objects”. A comparison is quite difficult, because the map sections are not identical (right map is shifted slightly towards north-west).

(c) Figure 3: The caption should be rephrased – td, sd and cd are no “predictors” (as mentioned at p5, l32).

(d) Figure 4: please rephrase the capture, e.g.: “Bayesian Network learned from data (left) and Bayesian Network constructed by experts (right). Note that not all variables are used in the network.”

(e) Figure 5: The authors might reconsider plotting only the mean value for each class – boxplots for each bin would be more informative. Please include the number of observations to for each category.

10. Please adhere to the journal standards concerning references (see guidelines for authors). References should be formatted accordingly and consistently, and references should be sorted alphabetically.
11. References to relevant literature are sparse, while the number of references to gray literature is relatively high. Especially sections 1 and 2 would benefit from some additional references.

12. Please adhere to the journal standards concerning references (see guidelines for authors). References should be formatted accordingly and consistently, and references should be sorted alphabetically.
13. References to relevant literature are sparse, while the number of references to gray literature is relatively high. Especially sections 1 and 2 would benefit from some additional references.

2 Specific Comments

1. p1, line 8: “Flood damage assessment is usually done with damage curves only dependent on the water depth.” – I would agree that most assessments include water depth as the main determinant of direct damage, but against the background of recent research, I would disagree that it is still state of the art to build flood damage assessments solely on water depth (c.f. Dutta et al., 2003; Kreibich et al., 2005; Thieken et al., 2005; Apel et al., 2009; Elmer et al. 2010; Merz et al. 2013; van Ootegem et al. 2015; Gerl et al. 2016). I would advise to slightly rephrase this sentence, indicating that more sophisticated, multivariate approaches (including hydrological modeling) are on the rise.

2. p1, line 20–21: “Because flood risk management becomes increasingly risk-based, flood damage estimation is increasingly important in flood risk assessment.” Please rephrase, this is unclear.
3. p1, line 23: “...flood risk assessments are...”
4. p1, line 27: “These models typically predict the fraction of damage...” – the authors may wish to clarify what the denominator of the fraction is by adding e.g. “...as percentage of total possible damage”.
5. p2, line 11: “... based on a German dataset based on ...” – please rephrase.
6. p2, line 11ff: The authors might want to add some additional references to their literature overview about multi-variate flood damage models. In addition, it might be of interest for the reader to know about the types of covariates used in these studies.
7. p2, line 13: “Spekkers et al. (2014) did something similar ...” – please specify.
8. p2, line 14: “These multi-variable flood damage models have been shown to perform better...” – the authors may want to provide some quantitative indication regarding how much the performance of these multi-variate models exceeded the performance of simple flood damage models.

9. p2, line 27: “...that is used here, and previously described...” – please rephrase

10. p2, line 29: “...very different from the datasets used so far (fewer variables, different sources of variables and different country).” – please explain in more detail. What is meant by “different sources” and why is data from the Netherlands expected to be “very different” from data from Germany? Also, this seems to refer only to the data set used by Merz et al. (2013) and Vogel et al. (2014).

11. p3, line 11: “The dataset used in this study is based on...”

12. p3, line 12: “...in the Netherlands (WL Delft, 1994).”

13. p3, line 13: 180 km^2

14. p3, line 14: “32 % of the damage pertains to residential buildings and content, for this study only the damage to this category is used”. – please rephrase.

15. p3, line 14: Please explain briefly why you decided not to consider damage to business and government buildings.

16. p3, line 17: I think the term “citizen household” is not very common. Maybe replace with “private households”?

17. p3, line 17ff: Please use a consistent, clear terminology. Distinguishing between “citizen households” (p3, l17), “companies” (p3, l18), “rental residential buildings” (p3, l21) “residential buildings” (p3, l25) and “rental houses” (p3, l26) and “privately owned residential buildings” (p3, l22) is confusing.

18. p3, line 20–23: “The building structure ... content for the same structure.” – please rephrase these two sentences to make this more clear.

19. p3, line 23: What is meant by “building content”? Furnishings?

20. p3, line 25: “The dataset did not include the building structure damage to all rental houses” – It is not clear to me until now, if the data have simply been collected by two different companies (as p3, l17ff imply) or if these two companies have also collected different types of data? Based on the text I assumed that structural damage to rental houses has been collected by “Stitching Watersnood Bedrijven 1993”?

21. p3, line 27: “Several manual actions were undertaken...” – please explain/provide some insight into what type of actions this could have been.

22. p3, line 30–31: So, apparently the “manual actions” are not known at all. Please refer to possible impacts of these manual actions on the results in the discussion.

23. p4, line 6: as a matter of fact, this correlation between water depth and damage is almost negligible. Other studies have found more obvious relationships between water depth and damage (e.g. Merz et al., 2003; Pistrika et al., 2009; Prettenthaler et al., 2010). The assumption that water depth as the main determinant of direct damage does not seem to hold in this case. Please discuss possible reasons for this weak correlation in the discussion (is this only due to the questionable quality of the water depth data mentioned at p4, l4?).

24. p4, line 9: “However, this data is not described...”

25. p4, line 23: “... and 40 meters.”

26. p5, line 1: The authors may wish to explain shortly how return levels are computed.

27. p5 line 5: I do not understand why Figure 2 would show that most of the area floods frequently. Isn’t this just a map about water depth?

28. p5, line 17: “The method of joining cadastre objects with damage records within a postal code area based on water depth rank is error prone.” – This is a quite straightforward approach, which is understandable given the lack of further information. However, this join is probably linked with relatively high uncertainty, depending on the spatial resolution of the DEM used and the (uncertain) expert estimation of water depth in the first place. It was mentioned that between 1 and 20 buildings share the same 6 digit postal code (p4, l2), so mismatches are likely to occur in postal codes with a larger number of buildings. The authors are probably right that houses within a postal code area are similar to some extent, but I am not sure if this is true for variables like “household size” or “floor area for living”. Are water depths within a postal code area similar, too, or are the ranks clearly distinguishable? The problem in case of a large number of mismatches is, that this just seemingly increases precision of the analysis. It might be worth testing if results change when simply using a mean/median value for all buildings within one postal code.

29. p5, line 29: “Several data mining (sometimes called machine learning) ...” – please rephrase. Even though these are closely linked and often used as synonyms, data mining and machine learning are not exactly identical. Rather, machine learning is a sub-field of data mining, i.e. data mining is not only restricted to machine learning methods.

30. p5, line 31: “...based on all independent variables (thus excluding total, content and structure damage).” – please, rephrase. This might be confusing to some readers, as the BN (Figure 4; p9, l34) includes content damage and structure damage.

31. p6, line 6: “...because many damage functions in the literature have this shape”. – please provide references, additional to Merz et al. (2012).

32. p6, line 8: may I suggest to use different variable names (variable names with

subscripts, e.g. d_t for total damage) in the formula? df is a common abbreviation for degrees of freedom.

33. p6, line 8: “...to get the smallest possible error based on the total damage and water depth data. The optimization of the coefficients is done with the Python package SciPy” – please rephrase and clarify (e.g. “...are optimized using ordinary least squares estimation from the Python package SciPy”).

34. p6, line 15: “However, it is more common to ...”

35. p6, line 19: “...with 11 variables for each damage record.”

36. p6, line 21 “...reduces maximally...” replace with “...is minimized...”

37. p6, line 22 and p6, line 25: “...by calculating the MSE reduction for all...” and “...is the reduction in MSE of total damage ...” (“MSE error” is redundant).

38. p6, line 23: abbreviation MSE is already explained in p6, l20.

39. p6, l. 24ff: please try to integrate the formula and the explanation of variables more naturally into the flow text. The sentence “The regression tree... (Pedregosa et al. 2011).” might be added at the end of the page.

40. p7, line 10: “the Matlab Statistical Toolbox (Matlab website)” – replace with “Matlab’s ‘Statistics and Machine Learning Toolbox’”

41. p7, line 11: ”Python libraries do not support pruning” – I think there are custom implementations of pruning in Python. The authors might want to look at sgenoud’s fork of the scikit-learn package at github.

42. p7, line 11: “performance of pruning was similar” – can you provide some information about the method and results of the comparison?

43. p8, line 9: the authors might want to add references to these fields of application.

44. p8, line 27: please cite the URL as a normal reference, i.e. "All calculations were done using the Python library libpgm (Cabot 2012)."

45. p9, line 6: "...balance was found by trying several discretization resolutions in order to gain the best results." – please rephrase and add more concise information ("...trying several discretization results until the best solution was found based on xxxxx criterion")

46. p9, line 13: "This was done manually by varying the discretization of the important variables until the smallest error was found" – this is rather vague. What is "manually"? What do you mean by "important variables" and what is the "smallest error"?

47. p9, line 7–15: please make this paragraph more concise

48. p9, line 28–32: please rephrase, focus on methodology and advantages/disadvantages of a manually established network. Contributing experts other than the authors should be added in an "Acknowledgements" section rather than in the text.

49. p9, line 33 – p10, line 1: "The total damage is ... and the content damage." Please explain in more detail, this is not fully conclusive to me.

50. p10, line 5: please provide information about important independent variables within the results section.

51. p10, line 15 : "(with different better training data)" – please rephrase

52. p11, line 1: the authors may wish to put section 3.2 into the discussion section.

53. p11, line 7–8: "The relatively good performance ... is striking." – the authors may wish to replace "striking" with "worth noting". Actually, given the rather bad fit of the root function (as explained by the authors in the following paragraph) and the concern about overfitting with regression trees, I assume that both the authors and the reader would have expected this behavior, at least to some extent.

54. p11, line 10: I think boxplots would be a more informative representation for Figure 5. Also, the conclusion of a "downward slope" based only on the means for each class should be interpreted with care. It has to be noted that variance/number of outliers gets smaller for data points with water depth > 1.3 m.p11, line 12: This is an interesting peculiarity of this data set. While it seems to be plausible that preparedness effects might mitigate total damage (note the very weak correlation of -0.09 in this case), it is counter-intuitive that return period is negatively correlated with water depth. Basically, events associated with high return periods are rare events with high water depth, i.e. the higher the water depth, the greater the return period. Under the assumption that values for return periods are relatively homogeneous for the Meuse flood (which was one actual event with a certain return period), this would mean that areas with a high water depth get flooded more frequently *at relatively higher water depths*. Yet, I would assume that they get flooded more frequently, but at lower level. So, in the case of the Meuse flood, areas with high water depth showed lower return periods. Does this indicate possible inaccuracies of the flood return period maps?

55. p11, line 21: "...are different from each other in more ways than just the water depth" – please rephrase.

56. p11, line 22: While overfitting based on a single variable is a valid concern, concluding to use multiple variables to avoid overfitting might be erroneous if the use of extra variables is not penalized.

57. p11, line 25: rephrase as "Discussion and conclusion"

58. p11–p12: please work on the discussion section, a large portion of page 12 (l10-l26) is is mainly about potential advantages of BN that are not visible in the results of this study.

59. p12, line 15–17: "but what tree is uncertain" – please rephrase (two times).

60. p12, last paragraph: please rephrase your final conclusions, this is somewhat clumsy from a linguistic point of view; e.g. split the first sentence into two sentences at the third "and": "In this paper we utilized different data sources compared to previous studies to acquire this data and showed that also on this dataset the methods are beneficial, especially the tree-based methods" – simplify, rephrase; "One possible way forward is to..." replace with "Future work may include ..."; etc.

Interactive comment on Nat. Hazards Earth Syst. Sci. Discuss., doi:10.5194/nhess-2017-7, 2017.