Nat. Hazards Earth Syst. Sci. Discuss., doi:10.5194/nhess-2017-7, 2017
Manuscript under review for journal Nat. Hazards Earth Syst. Sci.
Published: 12 January 2017

Natural Hazards and Earth System Sciences
Discussions
Open Access

Data-mining for multi-variable flood damage modelling with limited data

Dennis Wagenaar[1], Jurjen de Jong[1], Laurens M. Bouwer[1]

[1] Deltares, Delft, The Netherlands
Correspondence to: Dennis Wagenaar (dennis.wagenaar@deltares.nl)

Abstract. Flood damage assessment is usually done with damage curves only dependent on the water depth. Recent studies have shown that data-mining techniques applied to a multi-dimensional dataset can produce significantly better flood damage estimates. However, creating and applying a multi-variable flood damage model requires an extensive dataset, which is rarely available and this can limit the application of these new techniques. In this paper we enrich a dataset of residential building and content damages from the Meuse flood of 1993 in the Netherlands, to make it suitable for multi-variable flood damage assessment. Results from 2D flood simulations are used to add information on flow velocity, flood duration and the return period to the dataset, and cadastre data is used to add information on building characteristics. Next, several statistical approaches are used to create multi-variable flood damage models, including regression trees, bagging regression trees, random forest, and a Bayesian network. Validation on data points from a test set shows that the enriched dataset in combination with the data-mining techniques delivers a significant improvement over a simple model only based on the water depth. We find that with our dataset, the trees based methods perform better than the Bayesian Network.

1 Introduction

Because flood risk management becomes increasingly risk-based, flood damage estimation is increasingly important in flood risk assessment. Flood risk assessment supports policy makers to decide which flood risk management measures are most efficient in reducing flood risks and how much investment is cost-efficient. With the European Union Floods Directive (EC, 2007) now fully in place, national flood risk assessment are being developed with the final aim to support flood risk management plans. In the Netherlands, such flood damage assessment has been used to derive the optimal protection standard for flood protection (Kind, 2013; van der Most, 2014), using the current Dutch standard method for damage modelling (Kok et al., 2005). Also for insurance applications, more precise estimates of flood damages are required.

Flood risk assessments require flood damage models. These models typically predict the fraction of damage based on the water depth, and average building repair and replacement costs for different types of buildings (Messner et al., 2007; Jonkman et al., 2008). When validated, such simple flood damage models often don’t perform well (e.g. Jongman et al., 2012). This is because water depth alone cannot explain the full complexity of the flood damaging processes and several studies have only found low correlation coefficients (typically below 0.5) between the water depth and the flood damage (e.g. Merz et al., 2013, Pistrika&Jonkman, 2009). Furthermore, often no local data is available on flood damage and therefore a relationship between the water depth and damage either needs to be estimated or transferred from other areas (Wagenaar et al., 2016). This can cause errors as simple models hold many implicit assumptions that may not be valid for the situation the model is transferred to, for instance when the transferred function is only valid for specific flood durations. Transferability however could be improved, when a model describes more variations of the damaging process, and when more variables are included in the damage models.

Current approaches suffer from two main limitations: first, they rely on limited information and usually only take into account water depth as a predictor, and use a deterministic relation between water depth and some fraction of average maximum damages; secondly, they are deterministic in nature, while it has been shown that uncertainties in this approach are large, but generally not quantified e.g. in the Dutch standard method (Egorova et al., 2008). Some of the multi-variable methods are able to provide probability distributions, rather than deterministic estimates of damages.

Recently, multi-variable flood damage models have been created based on a German dataset based on telephone interviews. Using information from this database, Merz et al. (2013) used regression and bagging trees and Vogel et al. (2014) used Bayesian Networks to predict the flood damage. Spekkers et al. (2014) did something similar for pluvial floods. These multi-variable flood damage models have been shown to perform better than simple flood damage models in terms of explained variability, both tested on their own dataset and on datasets from other floods (Schröter et al., 2014). Also, some multi-variable approaches (Bayesian Networks, Bagging trees and Random Forests) generate probability distributions of estimated damages, and thus provide information on uncertainties of the estimates. Therefore, multi-variable flood damage models look like a promising approach to improve flood damage modelling.

The application of multi-variable flood damage models for flood risk management studies is still difficult because of the large data requirements. Running a multi-variable flood damage model for a new area requires for every object several variables on the flood hazard and building characteristics that are not yet typically collected. Also creating new multi-variable flood damage models is currently rarely done because they also require records of flood damages at building level.

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). Such datasets may have been created for compensation purposes or to build simple flood damage models but may miss most of the desired variables. An example of such a dataset is the flood damage dataset collected after the Meuse flood of 1993 in the Netherlands (WL Delft, 1994) that is used here, and previously described in Wind et al. (1999).

In this paper we will explore the use of data-mining techniques to build flood damage models based on a dataset that is very different from the datasets used so far (fewer variables, different sources of variables and different country). Methods will be applied to enrich the Meuse 1993 flood damage dataset with extra flood hazard and building characteristic variables. We will answer the question of whether this enriched dataset from a different source then previous studies is suitable to build a multi-variable flood damage model. The expectation is that the multi-variable models perform better than a model based on a single variable (water depth) and that even data with limited quality will improve the results.

2D flood simulations of the 1993 situation on the Meuse are used to enrich the dataset with additional flood characteristics. Cadastre data is used to enrich the Meuse dataset with extra building characteristics. Four different data-mining techniques are then applied to this enriched dataset: a regression tree, bagging regression trees, random forest and a Bayesian network. A part of the dataset will be held back and will only be used for validation. This validation is then used to determine whether the enriched dataset combined with data-mining techniques performs better than a traditional damage function based on the original dataset of water depths.

2 Method

2.1 Datasets

2.1.1 Meuse 1993 damage dataset

The dataset available for this research is based on the Meuse flood of 22 December 1993 in the Province of Limburg in the Netherlands and is described in WL Delft (1994). The flood caused a total of 254 million guilder (price level 1993) in direct damages, which is approximately 180 million euros today (price level 2016). The flood inundated 180 km2, which is about 8% of the Province of Limburg. 32% of the damage pertains to residential buildings and content, for this study only the damage to this category is used. Other major damage categories were business (29%), government (24%) and agriculture (8%) (WL Delft, 1994).

Damage information was collected in the context of compensation by the national government. The damage to citizen households was collected by an organisation called “Stichting Watersnood 1993” , the damage to companies was collected by another organisation called “Stichting Watersnood Bedrijven 1993”. These organisations collected the data by sending damage experts from insurance companies to the affected buildings, several weeks after the flood event had occurred. The building structure of a rental residential building belonged to a company. So, for rental residential buildings the damage was collected by a different organisation than the damage to the content for the same structure. For privately owned residential buildings all data was collected by the same organisation.

Directly after the damage data was collected in 1994, the data was shared with WL Delft (now Deltares) to create a flood damage model. WL Delft received 5780 records for damage to residential buildings. This dataset however did not include the building structure damage to all rental houses. Estimates of building structure damages were available for some rental houses and the aggregate building structure damage to rental houses was available. Several manual actions were undertaken in 1994 to repair this dataset and produce a complete dataset of building structure and building content damage.

The original data was not available for this study and only the final product of the WL Delft study was available. It is also not completely clear what manual actions happened to which records by WL Delft in 1994. The building structure damage records may therefore be of inconsistent quality. Another issue with the dataset is that for privacy reasons the exact locations of the buildings were not shared with WL Delft. Only the 6 digit postal code was available for this study, which makes it difficult to enrich the dataset, as between 1 and 20 buildings share the same 6 digit postal codes in the dataset.

In the original dataset the water depth (relative to the ground floor level) was estimated by the experts that surveyed the damage. The quality of water depth estimate is questioned by WL Delft (1994; report 9, appendix A) because it was not the main aim of the survey and the experts visited several weeks after the water had receded. A plot of the water depth (see Fig. 1) and the damage doesn't show an obvious relation. The correlation between the water depth and the damage is weak (Pearson correlation coefficient = 0.18).

The final dataset also contains information on the number of inhabitants per building, whether the house has a basement and whether the house was attached to other houses. This data is however not described in any of the available reports so the collection methods are not known, but the recorded values are clear enough to incorporate in this study. Two more variables are also included in the WL Delft dataset and also not described in any available report. These are emergency actions and ownership of the house. The meaning of the values found in the dataset for these variables is however not sufficiently clear, and could unfortunately not be taken into account in this study.

2.1.2 Upgraded Meuse 1993 dataset

To improve the dataset, information is required on both the flood hazard and additional exposure variables. The results of a 2D flood simulation and cadastre data were used to upgrade the dataset, in terms of hazard and exposure information, respectively. Because no observational data is available on flood characteristics other than the water depth, a simulation of the flood event was done. In the 2D flood simulation tool WAQUA (Rijkswaterstaat, 2013), a verified model of the state of the Meuse during the 1993 flood was available (Becker, 2012) and applied in this study to get extra variables. Using this model, a new simulation was run using a discharge boundary condition at Eijsden and the a water level boundary condition at Keizersveer for the period 1 November 1993 to 31 Januari 1994. This simulation was used to create a maximum water depth map, a flood duration map, a flood return period, and a flow velocity map at a spatial resolution varying between 10 and 40 meter.

The maximum water depth and flow velocity are standard outputs of WAQUA. Flood duration is however not a standard output and is more difficult to get from a 2D flood simulation because the drainage also needs to be included in the schematisation (Wagenaar, 2012). During the 1993 Meuse flood, most drainage occurred because of the natural slope in terrain and therefore the 2D flood simulation implicitly includes most of the drainage because the discretised bed level is included. The flood duration can then be calculated by analysing the time-varying maps of the water depth and calculating for every cell the time between the moment a cell is inundated and the moment the cell is dry again. However, some cells in the digital elevation map in WAQUA are surrounded by cells that have a higher elevation. These cells do not drain in the 2D flood simulation and are still inundated at the end of the simulation. For these cells the flood duration has been calculated based on the change in water depth. If the water depth in a cell stays the same in the simulation for 24 subsequent hours the cell is considered dry at the moment this stable water depth is first reached.

Simulations were also ran with the same Meuse 1993 schematisation for design discharges with 1, 10, 50, 100, 250 and 1250 return periods. These discharges are based on HR2006 (Diermanse, 2004) and have discharges of respectively 1300, 2260, 2869, 3109, 3431 and 4000 m^3/s. The results of these simulations were combined to create a flood return period map for the Meuse 1993 situation. This map shows for each cell at what return period it first floods. Figure 2 shows that large water depths occurred and that most of the area floods frequently. The majority of the houses is however located in the safest areas with the lowest water depths and highest return periods.

These maps (water depth, flow velocity, flood duration and return periods) were linked to the original damage records using cadastre data. The data of the cadastre has exact building locations, postal codes, living area within the residential buildings, the building footprint area and the construction year. The building year was used to filter the data to find the building stock of 1993. Then, based on the building locations the 2D flood simulation results were linked to the cadastre data.

This combination of cadastre data and 2D flood simulation data is then used to make the link with the original flood damage records. First per postal code a list is made of the damage records in the postal code area and ranked based on the water depth in the original damage records. Then another list is made of the objects per postal code according to the cadastre and also ranked based on the simulated water depth. The cadastre objects combined with the 2D flood simulation data is then linked per postal code based on the water depth rank. This results in a join between the original damage records, cadastre data and 2D flood simulation results. Table 1 gives an overview of the available records in this combined dataset.

The method of joining cadastre objects with damage records within a postal code area based on water depth rank is error prone. The modelled water depth is on average 30 cm larger than then the recorded water depth. This is possibly because the difference in reference level of both data sources as the recorded water depth is relative to the floor level and the modelled water depth is relative to the digital elevation map. Not all houses have the same floor elevation and both the recorded and the modelled water depth are uncertain, because of recording and model imprecisions. It is therefore likely that some damage records have been linked to the wrong object. However, errors will likely be limited, because the join on postal codes is accurate. Object and flood variables are generally similar for buildings within the same postal code area (e.g. houses within a street are typically similar to each other) so these errors are expected not to significantly disturb the general trends in the data. The errors are therefore considered acceptable given that the purpose of the dataset is only to build a flood damage model. If significant errors are present this would result in a reduced performance of the data-mining algorithms on the test set.

2.2 Data-mining algorithms

Several data-mining (sometimes called machine learning) techniques have been applied to the enriched dataset to build multi-variable flood damage models. The different data-mining techniques all have different ways to generalize the training data in such a way that it can give useful predictions of the total damage, based on all independent variables (thus excluding total, content and structure damage).

These multi-variable flood damage models are be compared to a reference model to assess the value of the enriched dataset and to assess the value of multi-variable flood damage models in general. Table 2 provides an overview of the different data-mining algorithms applied to the different datasets.

2.2.1 Regression: Root function

In order to assess the multi-variable flood damage models, a reference flood damage function was constructed. For this a root function was chosen because many damage functions in the literature have this shape. Merz et al. (2012) applied the same method to get a reference damage function. The root function (1) is fitted to the dataset in such a way that the coefficients c_1 and c_2 are optimised to get the smallest possible error based on the total damage (td) and water depth (df) data. The optimisation of the coefficients is done with the Python package SciPy.

$$td = c_1 + c_2\sqrt{df} \tag{1}$$

2.2.2 Regression tree learning

Decision trees are a way to represent complex relationships between data and classes in a tree structure. A decision tree can be seen as a series of binary questions (nodes) leading to an answer in the form of a class (leaf). A question can be related to any variable at any value (e.g. is the water depth smaller than 0.5m).

A regression tree is similar to decision trees but instead of classes it results in real numbers. In theory, regression trees can be very large and have a separate leaf for each unique value in the dataset. However, more common is to combine several similar unique values inside the same leaf and represent it with a summary statistic number (mean). In such a case the regression tree is an approximation of the relationship.

Regression tree learning algorithms can create optimal regression trees based on a dataset. In this paper the dataset consists of 4398 flood damage records (incomplete records are discarded) with for each damage record 11 variables (see table 1). The regression tree algorithm aims to split the dataset into subsets in such a way that the mean squared error (MSE) of the predicted total damage for all observations reduces maximally compared to the observed data. It does this by calculating the MSE error reduction for all candidate splitting variables according to their value and then picking the combination that maximises the mean square error (MSE) reduction (ΔI) as shown in (2). The regression tree is grown by repeating this process at each node of the tree. This has been done with the Scikit learn library in Python (Pedregosa et al. 2011).

$$\Delta I = \frac{1}{n}\left(\sum(y_n - \bar{y})^2 - \sum(y_{nL} - \bar{y}_L)^2 - \sum(y_{nR} - \bar{y}_R)^2\right) \tag{2}$$

Where ΔI is the Reduction in MSE error of total damage for a particular split variable and value, n is the total number of observations in the node, y_n is the vector of observed target values in the node and $\bar{y}$ is the mean of the target values in the node. y_{nL} and y_{nR} are vectors with the observed target values of the left and right group after the split and $\bar{y}_L$and $\bar{y}_R$ are the mean observed target value for the left and the right group.

A regression tree algorithm keeps splitting the dataset into new branches until no more reductions in the MSE can be made. This can result in overfitting, which results in very large trees with only one data point per leaf. These very large trees are not a realistic representation of reality, and they typically perform badly when they have to predict the damage for a new data point that wasn't used for building the tree. There are several methods to prevent overfitting. The simplest methods require a minimum number of data points in a leaf or set a maximum number of nodes that the tree is allowed to contain. The disadvantage of these methods is that they sometimes don't build out a branch within the tree which at first doesn't look promising but which can make valuable homogeneity improvements deeper in the tree. A method called pruning is a more sophisticated method, in which the entire tree is first build with a subset of the data points, and then cut back based on its performance on data points that were not used for building the tree. This method was investigated in this research. This was done using the Matlab Statistical Toolbox (Matlab website), based on the work by Breiman et al. (1984), because the Python libraries do not support pruning. The performance of the pruning algorithm on this dataset was similar to a regression tree built with a combination of a minimum data point requirement per leaf and a maximum number of leaves. Therefore, the rest of the study was performed without pruning in the Scikit learn library in Python (Pedregosa *et al.2011*). Accordingly, the results shown do not include pruning.

2.2.3 Bagging regression trees

Another method to avoid overfitting and generally improve the accuracy of decision/regression trees is bootstrap aggregating, also called bagging. The idea behind the method is to resample the dataset multiple times and to build a new regression tree for each resampled dataset. This results in an ensemble of regression trees. The resulting flood damage is then the average of the ensemble of regression trees. Resampling is done by building several datasets by randomly picking records from the original dataset (each record is allowed to be used multiple times in the same dataset). Every resampled dataset therefore randomly leaves out a fraction of the observations and puts more weight on other observations because they are picked multiple times. Bagging regression trees also lead to probabilistic outcomes because the ensemble of trees can be seen as a probability distribution of the outcome.

2.2.4 Random Forest

A random forest is a more advanced variation of bagging regression trees. Apart from building multiple trees with resampled datasets it also randomly excludes a subset of variables at each decision split. This will result in an ensemble of regression trees each based on a different set of damage records and each leaving out a different number of variables at each decision split. For this paper the default settings of Scikit learn are applied, in our case this means 8 variables are left out at each decision split.

2.2.4 Bayesian Network

A Bayesian Network is a type of Probabilistic Graphical Model that represents a set of random variables and their conditional dependencies in a directed acyclic graph (DAG) structure. Each variable in the network may be observed or represented as a prior probability distribution and dependencies between variables are represented with edges representing joint probability distributions. The edges in a Bayesian Network are directed which means there is a direction in which the influence of one variable flows to the other. From this network, inference can be done in order to use knowledge of one variable to make predictions about other variables.

Bayesian Networks and Probabilistic Graphical Models in general are used in many different fields, such as bioinformatics, image processing, speech recognition and decision support systems. Recently, they have also been applied to flood damage modelling (Vogel et al., 2014; Schröter et al. 2014; Van Verseveld, 2014). Schröter et al. (2014) found that their performance is often better than that of the different types of tree methods. Furthermore, a Bayesian Network can give its result as a probability distribution and does not require information about each variable in order make predictions. If fewer variables are available, the Bayesian Network handles this by adjusting the probability distribution of the outcome. This makes it ideal for transfer of models to other locations where less data is available than for the location where the model was originally based on. Furthermore, it returns (for each object) probability distributions rather than deterministic values, which is valuable for assessing uncertainties within the damage model estimates.

A Bayesian Network can be discrete, continuous or a combination. In this paper fully discrete Bayesian Networks are used, in which all variables are discretized into bins. Given a network the probability of particular set of discrete variable values can be calculated with the following formula:

$$P(X_i, \dots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i)) \tag{3}$$

Where X_i are the variables and $parents(X_i)$ is the set of variables directed to X_i. The probability of a single variable value can be obtained by taking the sum of all the probabilities that contain the variable value of interest. The conditional probabilities are stored in conditional probability tables (CPTs). These tables show, for each combination of parent variable values, the probability of each possible output value.

A data-driven Bayesian Network can derive all its CPTs from the data and even derive its graph structure from the data. For this paper, two Bayesian Networks were made: A data-driven Bayesian Network with both the graph structure and the CPTs derived from the dataset and an expert network where the graph structure was estimated in an expert session but the CPTs were derived from the dataset. All calculations were done with a Python library called libpgm developed at CyberPoint Labs in 2012 by Charles Cabot (http://pythonhosted.org/libpgm). This library follows the methodology described in Koller and Friedman (2009).

The CPTs are learned with maximum likelihood estimation. This method estimates the (joint) probability distributions based on the number of observations. The discretisation assumptions have an impact on the maximum likelihood estimation. If the variables are discretised into a large number of bins more possible combinations of states are possible. These combinations of states grow exponentially with the number of bins of the parent variables. A too fine discretisation therefore quickly leads to more possible states than available data points. This results in a poor performance of the maximum likelihood estimation. Koller and Friedman (2009) call this one of the key limiting factors in learning Bayesian Networks from data. A too coarse discretisation on the other hand is also not desirable because it limits the precision of the Bayesian Network. For this study a balance was found by trying several discretisation resolutions in order to gain the best results.

Discretisation was done by splitting the data into bins with an equal number of data points in each bin. This works better than making equal sized bins because of the large extremes in especially the damage data. Equal sized bins would either increase the number of bins, which is detrimental to the maximum likelihood estimation (having bins that contain no observations), or the bins would be so large that a majority of the data points would end up in the same bin, which would limit the Bayesian Network performance. The result of using an equal number of data points is that most bins are fairly equal in size with only the highest bin being much larger. The number of bins per variable was chosen based on the performance of a test set. This was done manually by varying the discretisation of the most important variables until the smallest error was found. For the Bayesian Network with the data-driven structure the number of bins chosen was slightly larger, because the network is less complex than the expert network.

The performance of the Bayesian Network on the testing data can be sensitive for discretisation. There are two possible alternatives for the discretisation method applied in this paper: An optimisation algorithm could be applied to determine the optimal discretisation, or a continuous Bayesian Network could be used (Friedman and Goldszmidt, 1996). Apart from solving the discretization problem the advantage of a continuous Bayesian Network is that it would probably perform better in predicting extreme values but a disadvantage is that the Bayesian Network is restricted to specific families of parametric probability distributions (Friedman and Goldszmidt, 1996). An optimization algorithm for the discretization can minimize the error produced by the discretizing but does not solve the fundamental problem of having too few data points.

The data-driven structure is also learned with the libpgm Python library. This library is using a constrained-based approach for structure learning, as is described in Koller and Friedman (2009). In a constrained based approach the structure is learned by calculating dependencies and conditional dependencies among the variables. When two variables are dependent regardless of what they are conditioned by, an edge (connection) is formed. The algorithm follows this procedure to create the entire network. The result is shown in figure 4 (left).

As an alternative to the data-driven structure a structure was also made in an expert meeting involving the following Deltares flood damage/Bayesian Network experts: Karin de Bruijn, Marcel van der Doef, Kathryn Roscoe, Laurens Bouwer and Dennis Wagenaar. In the expert meeting the network was constructed based on a combination of expert judgement/logic and with the knowledge of figure 3 in this paper. The experts focused mainly on edges that they thought are relevant for estimating the flood damage. The result is shown in Figure 4(right).

The total damage is the sum of the structure damage and the content damage. Therefore in the expert network it has been decided not to use the total damage variable. Instead the total damage is calculated as the sum of the expected value of the structure and the content damage. In the data-driven network the structure damage was not included by the algorithm. Therefore, the total damage variable itself is used for the data-driven network.

3 Results

3.1 Model comparison

The different models are tested on a test set that was not used for training the models. Two indicators are used to rate the performance of the models: Root Mean Square Error (RMSE) and the Pearson correlation coefficient. The RMSE is the average absolute error divided by the average damage, so a smaller RMSE is a better model. The Pearson correlation coefficient is a measure of the linear dependence between two variables. This measure is used to compare the predicted damages with the actual damages in the test set. A Pearson correlation of one means a perfect correlation, zero means no correlation and minus one a perfect inverse correlation. Table 3 shows the results for the different models.

Table 3 shows that given that the models can use all data, random forest and bagging regression trees perform equally well. Bagging regression trees and Random Forest do perform significantly better than normal regression trees, as was also noted by Merz et al. (2013) for flood damages in Germany. Random Forest and Bagging regression trees also outperform the Bayesian Networks. The normal regression trees also works better than the Bayesian Networks. This contradicts earlier findings by Schröter et al. (2014), who found that in most cases (with different better training data) that Bayesian Networks outperformed the regression trees.

Many explanations are possible for the relatively poor performance of the Bayesian Networks. The discretization of the data is a possible problem. Some trends could be too subtle to be captured by the rough discretization, but not enough data points are available for a more precise discretization. Perhaps there still is some space for improving the discretization, for example by applying an optimization algorithm to pick bin definitions in such a way that the available information is applied optimally (Vogel et al. 2012 applied such an algorithm). Another possible reason is that Bayesian Networks might be more sensitive to low quality data in combination with a small dataset. Some of the CPTs applied in the Bayesian Networks here are large and conditional probabilities are based on a relatively small number of observations. Some wrong observations may then have a relatively large impact on the damage prediction.

In the data-driven network the variable of interest (total damage) in our test is only influenced by the water depth. This is because the water depth relative to the ground floor is known while the content damage is not known, so the known water depth blocks all the influence of other variables and the unknown content damage has no influence because it is unknown (it is a target variable). The data-driven Bayesian Network is therefore in our test in practice only dependent on the water depth. So the structure learning decides to ignore the other variables when the water depth relative to the ground floor is available. This is probably because the data-driven structure algorithms finds all variables equally important and therefore draws only the most important edges (connections) regarding the total damage. Other methods (e.g. as described by Riggelsen, 2008) for structure learning might be able to give better results.

3.2 Benefits of more data

The models were trained with different numbers of variables to see whether the additional data is valuable. As expected, the best performing model with a high number of variables always performs significantly better than the best performing model with fewer variables. More data therefore seems to add potential value to the damage prediction despite the possible quality issues in the additional data.

The relatively good performance of regression trees compared to the fitted root function based on only the water depth is striking. A likely explanation for that is shown in the relation of water depth to the average damage in Figure 5. It seems that a damage function based on only the water depth should be downward sloping after 90cm, a root function therefore has the wrong shape. The shape that is suggested by Figure 5 would make physically no sense, if there are no other variables related to both the water depth and the damage that explain this downward sloping. However, in this case the return period might play such a role as it is negatively correlated with the water depth. Areas with large water depth are more frequently experiencing floods. Areas that experience floods more often are possibly better prepared and experience therefore possibly less damage.

Any data-driven flood damage model only based on water depth that learns that at greater water depths the damage is smaller than at lower water depths is not having a sufficiently good generalization about the relationship between water depth and flood damage. In a way the regression tree in this case is therefore overfitting on this specific flood event. This does not show up in the test set because the test set was based on the same event. Therefore, when the aim is to make a generally applicable flood damage model, a data-driven flood damage model should be tested on flood events that are not included in the training data (such as was done in Schröter et al., 2014). Overfitting on a flood event also shows the importance of using multiple variables in a model, because it shows that floods are different from each other in more ways than just the water depth. In this case using more variables could help produce a model that can explain the reduction in damage for higher water depths. However, for such a model to be generally applicable the training data for the model would need sufficient variety to be able to model the entire spectrum of possible events.

4 Conclusion and discussion

Additional data improves flood damage modelling relative to a test set, even if this data comes from a collection of different sources and is of limited quality (error prone). The data-mining algorithm is also important. Given the same data there are large differences between the algorithms. Random Forests and bagging regression trees perform significantly better than normal regression trees and the Bayesian Networks perform poorly compared to any of the tree based methods.

Our current approach doesn't show that the additional variables are beneficial for the Bayesian Networks. However, because the tree methods can benefit from the additional data it is likely that in some cases Bayesian Networks could also. The poor performance of the Bayesian Networks contradicts earlier studies (Schröter et al., 2014) and could be due to the discretization method or problems with data quantity or quality.

This paper trained flood damage models on just a single flood event. Ideally training data should consist of multiple events so that the spectrum of possible damages which the model is trained upon is larger. Especially for the transfer to other areas this would be important.

The test set that was applied in this paper for the validation of the model, was randomly selected from the data and consistently applied among all models. The indicators for model performance would have been more accurate if cross-validation was used instead of a single test set. Expectations are that this would cause minor shifts in results but that it would not influence the conclusions of this paper.

This paper did not address another benefit of Bayesian Networks, Random Forest and Bagging trees, which is the incorporation of uncertainty. Bayesian Networks do this explicitly in the method and for Bagging Trees or Random Forest each tree can be seen as a possible damage estimate and together the trees represent a probability distribution.

The methods applied in this manuscript provide an uncertainty estimate for a single object. For policy decision making it is often useful to aggregate these uncertainty estimates to a total uncertainty for the entire flood event. This can be done with the assumption that all objects are perfectly correlated to each other (one tree will apply to the entire event but what tree is uncertain), or with the assumption that all objects are independent of each other (each object will have a different tree but what tree is uncertain). Both assumptions are however not completely correct (Wagenaar et al., 2016). The Bayesian Network framework might offer a middle way to model this correctly. If each object has a copy of the original Bayesian Network, and these Bayesian Networks are linked together based on the location of the objects, it can be explicitly taken into account that nearby objects are more likely to have similar damages. This could be an argument to prefer Bayesian Networks over tree based methods.

Bayesian Networks do have more potential advantages that were not yet used in this study. They are flexible and expert knowledge can be added to the conditional probability tables or the network. Furthermore, Bayesian Networks are designed to have more than one unknown variable and deal with them correctly by increasing the uncertainty. This is an advantage when transferring models to other areas where less data is available. Tree methods also provide options to deal with missing values (Kreibich et al., 2016).

Data-mining and multi-variable statistical techniques can help to create and improve flood damage models and have many theoretical advantages over deterministic damage functions based only on the water depth. The application of these techniques however remains difficult in practice, because of the limited number of data points, as well as acquisition of that data. 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. One possible way forward is to merge available datasets from different events or countries, and use expert knowledge in order to make a more generally applicable model which also works in circumstances outside areas for which flood damage data is available.

Acknowledgments

We thank our colleagues Kathryn Roscoe for advice on the Bayesian Networks. This research has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement number 641811 (Improving predictions and management of hydrological extremes – IMPREX), see also http://www.imprex.eu.

References

Becker, A., 2012. Maas-modellen 5de generatie: Modelopzet, kalibratie en verificatie. Deltares rapport 1204280-000-ZWS-0011 (in Dutch).

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

Diermanse, F.L.M., 2004. HR2006 – herberekening werklijn Maas. Delft Hydraulics Q3623.00 (in Dutch).

EC, 2007. DIRECTIVE 2007/60/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 23 October 2007 on the assessment and management of flood risks. Official Journal of the European Union. L 288/27

Egorova, R., J. van Noortwijk, en S. Holterman. „Uncertainty in flood damage estimation.” International Journal River Basin Management 6, nr. 2 (2008): 139-148.

Friedman, N. and Goldszmidt, M., 1996. Discretizing continuous attributes while learning bayesian networks. In Proc. ICML, page 157–165

HOWAS 21, website accessed 03-08-2016: http://dx.doi.org/10.1594/GFZ.SDDB.HOWAS21

Jongman, B., Kreibich, H., Apel, H. B., Bates, P., Feyen, L., Gericke, A., Neal, J., Aerts, J.C.J.H., Ward, P. Comparative flood damage model assessment: towards a European approach. Natural Hazards and Earth Sciences, 12, 3733-3752. 2012

Jonkman, S.N., M. Bockarjova, M. Kok, P. Bernardini (2008). Integrated hydrodynamic and economic modelling of flood damage in the Netherlands. Ecological Economics, 66, 77-90.

Kadaster website. Accessed 25-10-2016: https://www.kadaster.nl/bag

Kind, J.M. 2013. Economically efficient flood protection standards for the Netherlands. Journal of Flood Risk Management, 7(2), 103-117.

Kok, M., H.J. Huizinga, A.C.W.M. Vrouwenvelder, en W.E.W van den Braak. Standaardmethode 2005, Schade en Slachtoffers als gevolg van overstroming. HKV, TNObouw, Rijkswaterstaat DWW, 2005.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. The MIT Press. ISBN: 978-0-262-01319-2.

Kreibich, H., Botto, A., Merz, B., Schröter, K., 2016. Probabilistic, Multivariable Flood Loss Modeling on the Mesoscale with BT-FLEMO. Risk Analysis. DOI: 10.1111/risa.12650

Messner, F., E. Penning-Rowsell, C. Green, V. Meyer, S. Tunstall, A. van der Veen (2007). Evaluating flood damages: guidance and recommendations on principles and methods. Floodsite report T09-06-01.

Pedregosa F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit learn: Machine Learning in Python. Journal of Machine Learning Research 12 (2011) 2825-2830

pythonhosted.org/libpgm, website accessed 10-08-2016: http://pythonhosted.org/libpgm/

Pistrika A.K., Jonkman S.N., 2009. Damage to residential buildings due to flooding of New Orleans after hurricane Katrina. Natural Hazards. Vol. 54 Issue 2, pp. 413-434

Matlab website: accessed on 25-10-2016. https://nl.mathworks.com/products/statistics/

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.

Schröter, K., Kreibich, H., Vogel, K., Riggelsen, C., Scherbaum, F., Merz, B., 2014. How useful are complex flood damage models? Water Resour. Res. 50, 3378–3395. doi:10.1002/2013WR014396

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., and ten Veldhuis, J. A. E.: Decision-tree analysis of factors influencing rainfall-related building structure and content damage, Nat. Hazards Earth Syst. Sci., 14, 2531-2547, doi:10.5194/nhess-14-2531-2014, 2014.

Riggelsen, C. 2008. Learning Bayesian Net-works: A MAP Criterion for Joint Selection ofModel Structure and Parameter. In ICDM, 2008 Eighth IEEE International Conference on Data Mining, pages 522-529.

Rijkswaterstaat, 2013. User’s Guide WAQUA: General Information. Version 10.59, October 2013 (in Dutch).

van der Most, H., Tanczos, I., De Bruijn, K.M., Wagenaar, D.J., 2014. New, Risk-Based standards for flood protection in the Netherlands. 6th International Conference on Flood Management (ICFM6), September 2014, Sao Paulo, Brazil.

Van Verseveld, H., 2014. Impact Modelling of Hurricane Sandy on the Rockaways. Relating high-resolution storm characteristics to observed impact with use of Bayesian Belief Networks. MSc thesis Delft University of Technology - Deltares.

Vogel, K., Riggelsen, C., Kreibich, H., Merz, B., Scherbaum, F., 2012. Flood Damage and Influencing Factors: A Bayesian Network Perspective, in: Proceedings of the 6th European Workshop on Probabilistic Graphical Models (PGM 2012). Presented at the 6th European Workshop on Probabilisit Graphical Models, Granada, Spain, pp. 347–354.

Wagenaar, D.J., 2013. The significance of flood duration for flood damage assessment. Master Thesis, Delft University of Technology.

Wagenaar, D.J., De Bruijn, K.M., Bouwer, L.M. & De Moel, H., 2016. Uncertainty in flood damage estimates and its potential effect on investment decisions. Natural Hazards and Earth System Sciences, 16(1), 1-14.

Wind, H.G., T.M. Nierop, C.J. de Blois, J.L. de Kok, 1999. Analysis of flood damages from the 1993 and 1995 Meuse floods. Water Resources Research, 35, 3459-3466.

WL Delft, 1994. Onderzoek watersnood Maas, Deelrapport 1: Wateroverlast December 1993. WL Delft, Delft (in Dutch).

WL Delft, 1994. Onderzoek watersnood Maas, Deelrapport 9: Schade. WL Delft, Delft (in Dutch).

Table 1: Description of the variables in the flood damage dataset for the Meuse flood of 1993.

	Variable	**Unit**	**Source**	**Pearson correlation on damage**
td	Total damage	Guilder (1993 value)	Original dataset[a]	1
sd	Structure damage	Guilder (1993 value)	Original dataset[a]	0.85
cd	Content damage	Guilder (1993 value)	Original dataset[a]	0.83
df	Water depth relative to floor	m	Original dataset[a]	0.18
dg	Water depth relative to DEM	m	Flood simulation[b]	0.18
bs	Basement	1=Yes, 2=No	Original dataset[a]	-0.04
dh	Detached house	1=Yes, 2=No	Original dataset[a]	0.08
hs	Household size	Number	Original dataset[a]	0.17
fv	Flow velocity	$m\ s^{-1}$	Flood simulation[b]	0.04
fd	Flood duration	h	Flood simulation[b]	0.05
rp	Return period	year	Flood simulation[b]	-0.09
ba	Building age	year	Cadastre[c]	0.01
la	Floor area for living	m^2	Cadastre[c]	0.04
fa	Footprint area building	m^2	Cadastre[c]	-0.02

[a] WL Delft, 1994

[b] 2D flood simulation data using WAQUA

[c] Basisregistraties Adressen en Gebouwen (BAG), version 2011 (Kadaster website).

Table 2: Overview of the applied data mining algorithm.

Data-mining algorithm	**Applied on:**			
	Water depth	**Original variables** [a]	**All variables**	**Purpose**
Regression: Root function	x			Reference, representing traditional method.
Regression tree	x	x	x	Comparison of the methods and to see whether adding extra variables results in better predictions.
Bagging regression trees	x	x	x	
Random Forest	x	x	x	
Data-driven Bayesian Network	x	x	x	
Expert Bayesian Network	x	x	x	

[a] Only data recorded directly after the flood (the variables of WL Delft, 1994)

Table 3: Results of different models for two indicators: RMSE and correlation coefficient

Calculation	RMSE	Correlation coefficient
Root function	0.612	0.152
Regression tree	0.561	0.313
Bagging regression tree	0.504	0.388
Random forest	0.508	0.394
Data-driven Bayesian Network	0.629	0.208
Expert Bayesian Network	0.607	0.206

Table 4: The best performing model with different number of variables for two indicators RMSE and the correlation coefficient.

Variables	**Method**	**RMSE**	**Correlation coefficient**
Only water depth	Regression tree	0.564	0.306
Only original variables (waterdepth, household size, detached house, basement)	Bagging trees	0.551	0.345
All variables	Random Forest	0.508	0.394

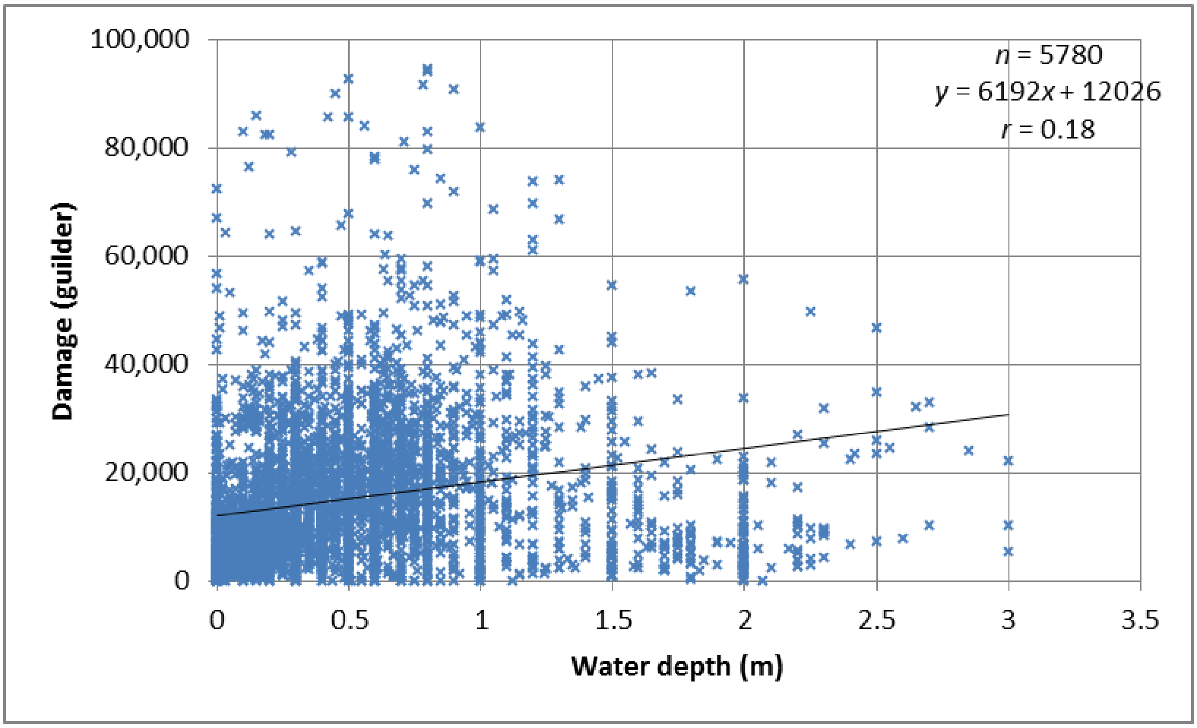

Figure 1: The water depth and the damage in the original dataset.

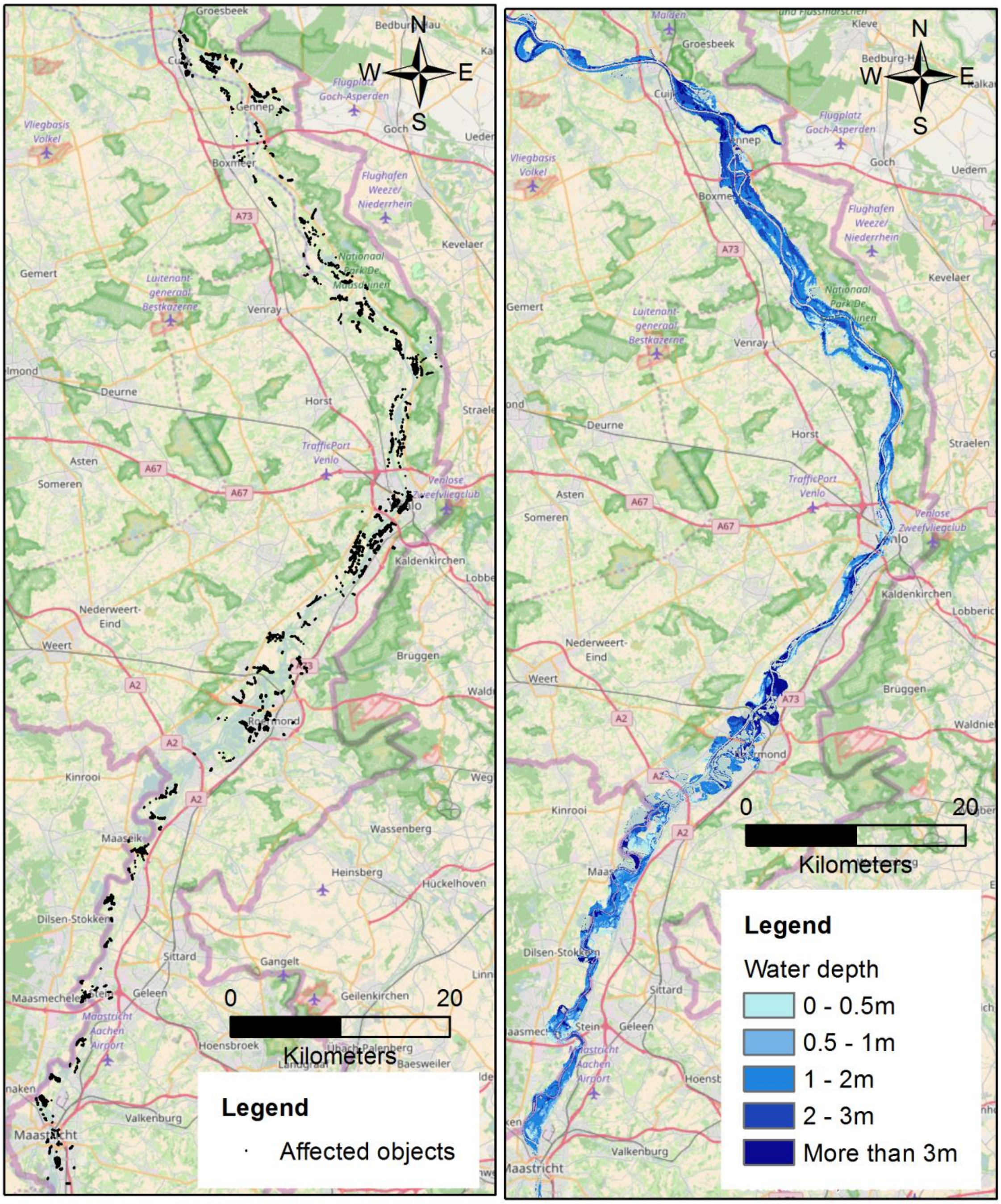

Figure 2: Maps of the affected objects and the simulated water depth (similar maps were made for return period, flow velocity and flood duration).

5

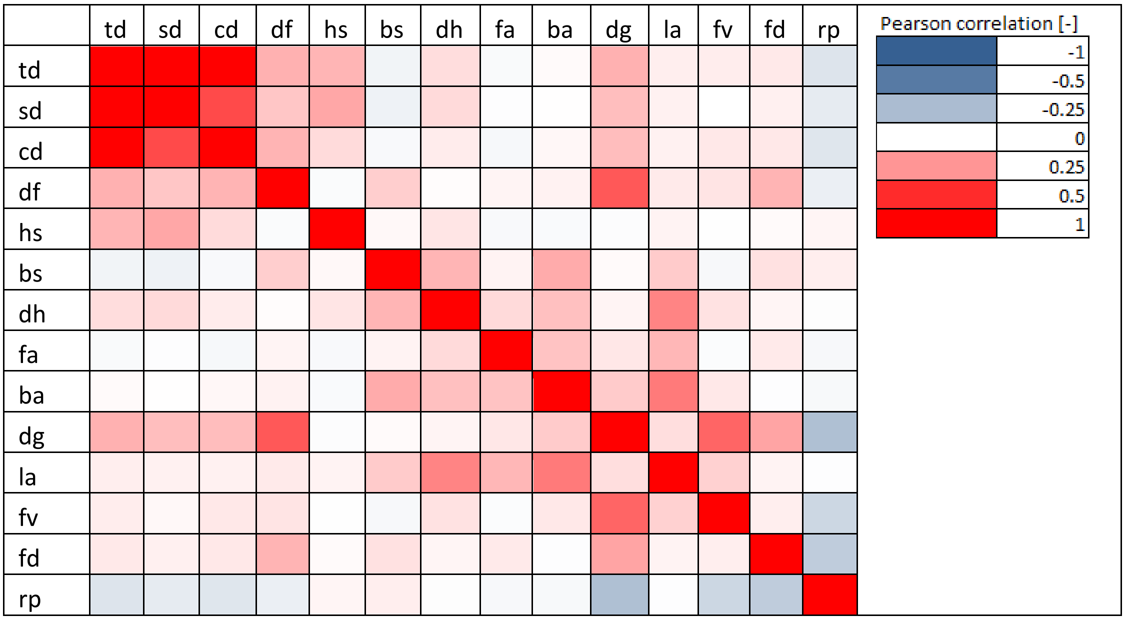

Figure 3: Correlation coefficients between the different predictors. See Table 1 for a description of the abbreviations).

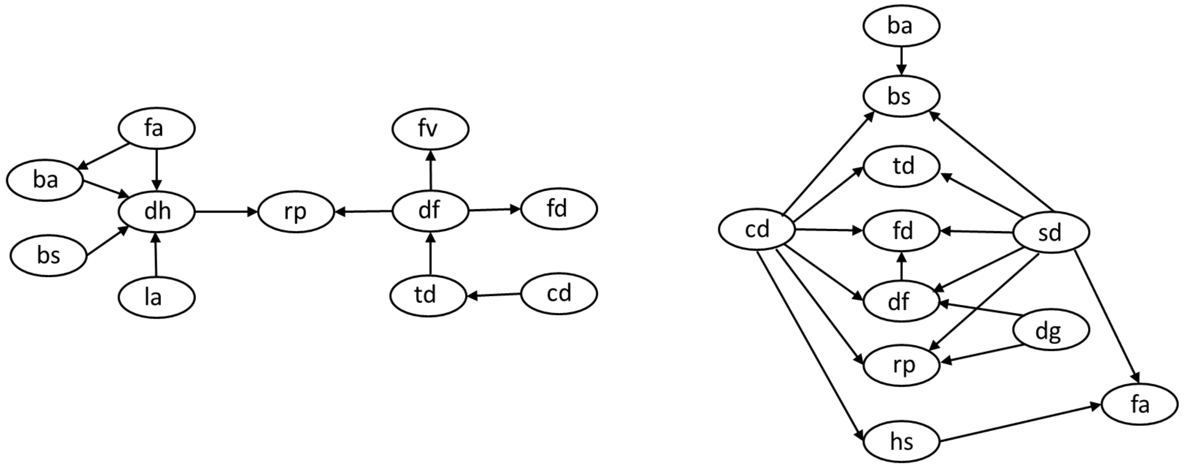

Figure 4: Right the Bayesian Network constructed by experts, left the Bayesian Network learned from the data (see Table 1 for the definitions of the abbreviations). Note that not all variables have to be used for the networks.

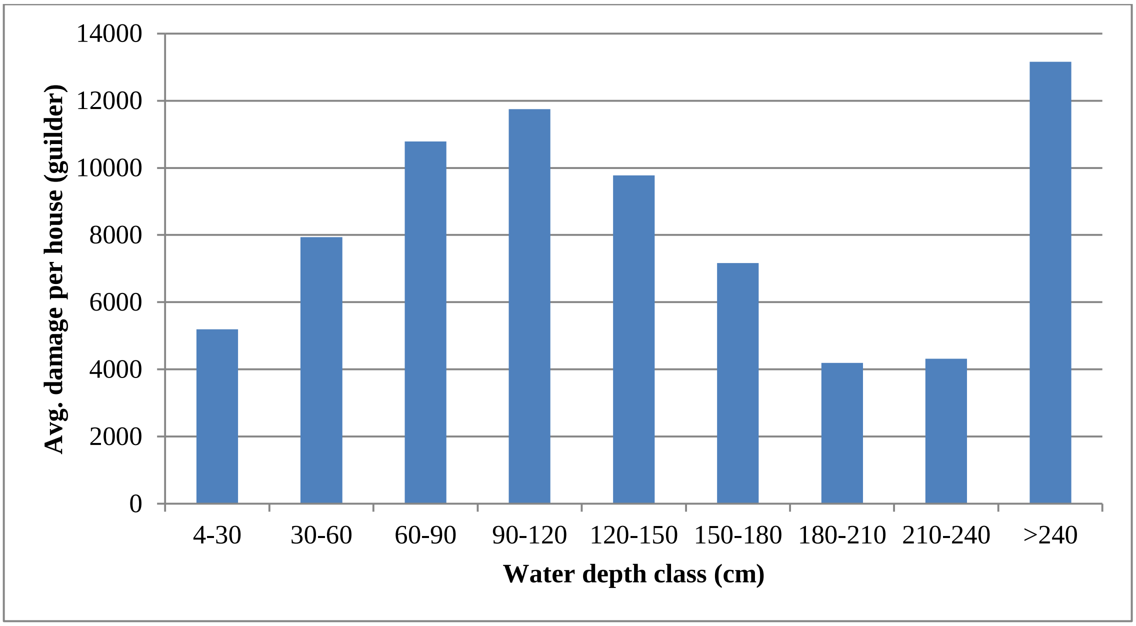

Figure 5: Average damage during the Meuse flood of 1993 per water depth class.

5