3. CHAPTER 3: EVALUATING THE EFFECTIVENESS OF HOUSEHOLD LEVEL FLOOD RISK REDUCTION MEASURES BY THE APPLICATION OF PROPENSITY SCORE MATCHING[1]

Abstract

The employment of damage mitigation measures by individuals is an important component of integrated flood risk management. In order to promote efficient damage mitigation measures, accurate estimates of their damage mitigation potential are required. That is, for correctly assessing the damage mitigation measures' effectiveness from survey data, one needs to control for sources of bias. A biased estimate can occur if risk characteristics differ between individuals who have, or have not, implemented mitigation measures. This study removed this bias by applying an econometric evaluation technique called Propensity Score Matching to a survey of German households along three major rivers that were flooded in 2002, 2005 and 2006. The application of this method detected substantial overestimates of mitigation measures' effectiveness if bias is not controlled for, ranging from nearly €1,700 to €15,000 per measure. Bias-corrected effectiveness estimates of several mitigation measures show that these measures are still very effective since they prevent between €6,700-14,000 of flood damage per flood event. This study concludes with four main recommendations regarding how to better apply Propensity Score Matching in future studies, and makes several policy recommendations.

[1] This chapter is based on: Hudson, P., Botzen, W.J.W., Kreibich, H., Bubeck, P., Aerts, J.C.J.H., 2014. Evaluating the effectiveness of flood damage mitigation measures by the application of propensity score matching. *Natural Hazards and Earth System Sciences*, 14, 1731-1747.

3.1 Introduction

Chapter 1 discussed several studies that project increasing flood risk. Moreover, globally natural disaster losses are increasing. Therefore, this trend of increasing flood risk means that the potential benefits from investing in a public or private damage risk reduction (DRR) are also increasing. Think of, for example, privately implemented DRR measures, such as sealing cellars to flood waters, or elevating buildings above expected inundation depths. The movement towards integrated flood risk management (Kron, 2005; Kreibich et al. 2007) places greater weight on the responsibilities of private agents to limit flood risk, for instance by mitigating possible levee effects (IPCC, 2012). A levee effect can occur when individuals feel safer after flood protection infrastructure has been installed. A reduction in flood risk lowers the expected costs of living or doing business in the area owing to a lower flood frequency, which promotes greater exposure, increasing potential flood damage. DRR measures implemented by households could help to mitigate these effects. For integrated flood risk management to be successful, an important research question that needs to be answered is 'which private DRR measures are most effective at reducing flood damage?' This chapter focuses on private DRR measures because integrated flood risk management requires all stakeholders in a flood risk area to play a role in managing risk. The potential of government investment in this area is relatively more known than that of private households. This chapter seeks to add to the nascent literature on this topic.

There are several studies that investigate potential flood damage reduction that can be achieved by various DRR measures. For example, Holub and Fuchs (2008) investigate the cost- effectiveness of risk reductions using a cost-benefit analysis approach, where, if benefits are larger than costs, the measure is regarded as an efficient DRR. Holub and Fuchs (2008) estimate the natural hazard risk posed in their sample area. Once the level of risk is known, the sample area is divided into different risk zones, and the level of exposure within a risk band is used to estimate damage. Holub and Fuchs (2008) then proceed to calculate the benefits of the measures by assuming that a DRR prevents all damage up to a certain severity of hazard. Poussin et al. (2012) employ a similar method by modelling the risk reducing effect of DRR measures, such as wet-proofing a house, by assuming that the effectiveness of a measure is a percentage reduction in flood damage simulated by a flood risk model. Other studies, such as De Moel et al. (2013), Dutte et al. (2003), and DEFRA (2008), also apply similar methodologies. While these methods are useful, they are not able to

empirically evaluate DRR measures because they assume, on the basis of expert judgement, that the DRR measures are effective to a predetermined degree.

Damage models, however, do not provide empirical proof that the DRR measures are able to prevent damages up to the assumed degree. Therefore, studies are undertaken that use household survey data, empirically ground the evaluation in specific cases, such as Kreibich et al. (2005; 2009; 2011) and Bubeck et al. (2012). Bubeck et al. (2012) use a repeated-measure design to compare the amount of flood damage suffered by the same households during two consecutive flood events along the German part of the Rhine in 1993 and 1995. To avoid possible bias due to differences in flood hazard characteristics, the most important damage-influencing facto: namely, inundation depth (Thieken et al., 2005), was controlled for. Only those households were included in the comparison that reported identical water levels in the cellar and ground floor during both flood events. This comparison reveals a central tendency towards lower flood damage in 1995. Moreover, less extreme damage values were recorded for the later event. This trend towards lower flood damage in 1995 is attributed to a considerable increase in DRR measures implemented by households between the 1993 and the 1995 flood event. Those households that increased the level of DRR measures, showed the largest reduction in flood damage suffered. However, this method still may not produce an accurate estimate of the effectiveness of a DRR for several reasons. The first is that an explicit value for the effectiveness per DRR has not been provided. The second is that other possible differences in hazard characteristics were not controlled for, such as flow velocity or contamination of floodwater. Also possible changes in household characteristics, such as an increase in the value of household contents for example, between the floods were not taken into account.

A different survey data methodology is that of Kreibich et al. (2005; 2011). In these studies a more direct estimate of effectiveness was provided. In Kreibich et al. (2005), for the various DRR measures, households were divided into those who have employed a particular DRR and those who did not. Once the sample has been divided into two groups based on the use of a DRR, the average damage suffered in each group is calculated, and the difference between these averages forms the estimated effectiveness. These results are important initial steps regarding the evaluation of DRR measures. However, a drawback of this approach is that the difference in average damage suffered between those who installed a DRR and those

who did not install a DRR may still not provide an accurate estimate of the damage savings obtained by the DRR. This is because other factors could have influenced the difference in damage, such as inundation depth, flow velocity, or differences in household characteristics. This is because in effect the respondents are able to 'treat' themselves by employing a DRR measure.

Kreibich and Thieken (2009) employ a similar method to examine the success of DRR measures in Dresden. In particular, they estimate the mean difference in damage between individuals who suffered roughly similar natural hazard risks, and refine the DRR effectiveness estimate by removing a source of bias, but still leaving several factors uncontrolled for, and creating problems due to very small sample sizes (groups of 3-5 households). Finally, the later study of Kreibich et al. (2011) had an additional benefit to its micro-scale cost-benefit analysis due to using a sample consisting of structurally identical households. The identical household construction removes some sources of bias. However, the approaches employed have meant that potential sources of bias have been independently controlled for. These issues result in a direct effectiveness estimate, but one that is potentially inaccurate due to the presence of selection bias.

Angrist and Pischke (2009) states that the difference in observed means contains two effects: the treatment effect (employing a DRR, the effect of interest[2]) and a selection bias. This is due to the traits that drive both outcomes (flood damage) and employing the DRR measure. A method for controlling for many sources of bias simultaneously, Propensity Score Matching (PSM), will be applied to the data used by Kreibich et al. (2005; 2011). The application of PSM can create a more refined and reliable estimate of the protective qualities of a DRR by removing the selection bias that may be present in previous studies that used a mean comparison evaluation methodology. Selection bias arises because survey data is observational, and both the outcome of interest (damage reduction) and the employment of risk reduction measures can be driven by individual traits and their own decision process. This means that the two groups are systematically different, and cannot form the counter-factual observations needed for an unbiased effectiveness estimate. For example, suppose that

[2] While it can be argued that as there is no external treatment (due to the self-treatment) that using the term treatment is not appropriate. However, this Chapter uses the term 'treatment' of the econometric literature to indicate the effect of interest: the possible damage prevented; and 'treatment group' to indicate the group that employs a given DRR measure.

the control group faces a higher flood hazard than the treatment group, and then the treatment effect may be overestimated by a mean comparison methodology.

The general idea of PSM is that in the absence of an experimental design (i.e. a randomised controlled trial), assignment to treatment is frequently nonrandom (i.e. that the respondents are 'self-treated') which can generate sources of statistical biases due to systematically different characteristic (Heinrich et al., 2010). PSM aims to find 'treatment' and 'control' group members that are sufficiently similar to another, so that an estimate of the mean impact can be provided. PSM uses the probability that a respondent employs a given risk reduction measure, given a set of observed variables (Heinrich et al., 2010). A suitable PS will produce valid matches for estimating the impact of a respondent's decision to employ a DRR measure. Therefore, rather than attempting to match on all values of the variables, cases can be compared on the basis of propensity scores alone (Heinrich et al., 2010), which reduces possible selection bias.

D'Agostino (1998) notes that PSM has been applied to a wide range of topics, for instance in medicine it is commonly used to study the effectiveness of drugs or surgical methods. For example Vincent et al. (2002) investigate the effectiveness of blood transfusions when the patient is critically ill and suffering from anemia. In economics, PSM is applied to a wide variety of economic issues. For example, Dehejia and Wahba (2002) provide an evaluation of the effects of taking part in a government-training programme on incomes. In the above cases, PSM is used because the most reliable method of estimating the treatment effect, a controlled randomised trial, is unfeasible due to practical and or ethical concerns, and therefore a different technique is needed.

The objectives of the current chapter are two-fold. The first is to remove selection bias that may be present in previous DRR effectiveness estimates, in order to produce a more accurate estimate of DRR effectiveness. The second is to judge the applicability of PSM to wider natural hazards research. To the best of this chapter's knowledge, this chapter is the first study to use PSM to evaluate the installation of flood DRR measures. Furthermore, only one other study has applied PSM to natural hazards; being Butry (2009) who investigates the success of wildfire mitigation programmes. The current chapter seeks to apply PSM to provide a bias-free estimate of the flood damages prevented due to DRR measures, which will

be useful in guiding integrated flood risk management strategies and the role individuals can play in mitigating flood risk.

3.2 The propensity score matching method

To evaluate the potential reduction in damage due to the use of a private DRR (referred to as 'treatment'), it is required to make an estimate of the difference between what occurred and what would have occurred if the agent had not employed the DRR measure. This is the average treatment effect on the treated (ATT), which is defined in eq. (3.1). Below, E(.) is the expectations operator; T is a binary variable for employing the DRR measure or not; y_1 is the outcome under treatment; while y_0 is the outcome under non-treatment;

$$ATT = E(y_1 - y_0|T = 1) = E(y_1|T = 1) - E(y_0|T = 1) \quad (3.1)$$

A positive ATT indicates that participation in the treatment is expected to increase the outcome variable (damage suffered) while a negative value indicates a reduction (in damage suffered). For a DRR measure, a highly negative ATT would indicate that it was effective at mitigating flood damage. However, either the outcome under treatment ($E(y_1|T = 1)$) or under non-treatment ($E(y_0|T = 0)$) is observed. Therefore, for individual i the ATT cannot be constructed, as only the first half of eq. (3.1) is known. The intuitive method of recreating the counterfactual observation is to use the respondents who did not take part in the treatment. Angrist and Pischke (2009) provide a general expression for the difference between sample sub-group averages, showing the potential combination of the ATT and selection bias (SB) in eq. (3.2):

$$E(y_1|T = 1) - E(y_0|T = 0) = ATT + SB \quad (3.2)$$

Selection bias is present ($SB \neq 0$) if there are traits that explain both participation in the treatment and outcomes, and where these traits differ across these two groups. These traits are confounders, and their influence on outcomes and participation masks the true value of the ATT, which PSM attempts to minimise. If there were random entry into the control and treatment group, treatment participation would no longer be tied to individual traits. This means that the difference in mean damage between the two groups would provide an unbiased estimate of the ATT, which is the rationale behind a controlled randomised trial. However, while randomised trials will provide an unbiased estimate of the effect, Grossman and Mackenzie (2005) argue that a controlled randomised trial can only provide a reliable estimate if behaviour can be monitored and outcomes

observed during the trial period. A trial for a DRR measure is in practice unfeasible due to the organisational requirements, costs and unpredictable nature of flood events. There are also ethical concerns about forcing the control group to remain unprepared for potential disasters. Therefore, survey data and observational outcomes must be used, which, in turn, means that entry into the treatment group is non-random and driven by traits such as total exposure or perceived risk, making selection bias a potential problem. In such a case PSM can be used to estimate the ATT.

PSM, developed by Rosenbaum and Rubin (1983), is based on the intuition that, by conditioning on the confounders, it is possible to find agents who are similar enough to form each other's counterfactual observation, and can, therefore be matched together. If the individuals in a match are similar, enough selection bias can be removed and average difference in outcomes between the matches is a reliable estimate of the ATT. Originally, matching was based on covariates, where a researcher would attempt to find individuals who have the same values of the confounding covariates, and match these individuals. However, this can be problematic or even impossible with large numbers of confounders. Identical individuals are easy to find if there are only two binary confounders, and thus only four possible combinations to group respondents. But this becomes more difficult, the more relevant confounders there are that must be included in the matching process, and even more complicated if these variables are continuous rather than binary variables. For example, if matching takes place based on whether the building is located in an urban area, then urban treatment group members can be matched with urban control group members. If household content values are additionally matched upon, then control and treatment group members who are both urban and have an equal contents value must be found. Then if house size is also matched, matches must be identical in all three respects. This dimensionality issue can greatly reduce the possible sample size. Matching on the PSV removes this dimensionality issue as the estimated propensity score compresses the relevant information into a single value. PSM allows a match to be made by finding two, or more, agents with a sufficiently close PSV.

For PSM to be valid the following three conditions are required to hold, where $\perp$ represents independence; X is the set of observable traits; p(X) is the propensity score as a function of the observable traits; T is a binary variable for participation in the treatment group or not; and y_1 is the outcome under treatment, while y_0 is the outcome under non-treatment:

- Condition 1: Unconfoundedness –$(y_1, y_0) \perp T|p(X)$;
- Condition 2: Balancing – $T \perp X|p(X)$;
- Condition 3: Overlap – The probability distributions for the control and treatment group share a common support, as in Figure 3.1.

Condition 1 means that treatment participation and potential outcomes are independent of one another, conditional on the PSV, in effect achieving y_1 or y_0 is as good as random. The role of condition 1 is that, by conditioning on the set of confounders, the selection bias in the treatment is removed. Unconfoundedness holds when all the confounders have been included in generating the PSV.

Condition 2 is that, when conditioned on p(X), treatment participation and individual traits are independent of one another. When condition 2 holds, the PSV is a balancing score, and then matching on the value of the PSV achieves the same as conditioning on each individual confounder value.

Condition 3 implies that the observations have a similar enough PSV to create a good match of individuals. Heckman et al. (1996) provide a formulation of the bias introduced due to matching, showing that the smaller the common support, the greater the possible bias in the final estimate (an example of a common support is displayed in Figure 3.1). The reason is that outside this range the matched participants are potentially too different from one another. Heckman et al. (1996) then proceed to state that by only matching over PSV located in the common support this matching quality bias is removed. Matching quality bias is introduced when matched individuals are too different from each another.

Taken together conditions 1 and 3 remove bias from the estimate, while condition 2 allows for matching based on a single value constructed from all the confounders.

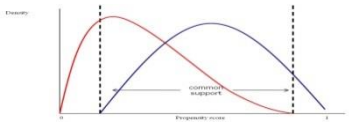

Figure 3.1 An example of a common support

In most cases, a probit or logit model will estimate the PSV. It has been found that using an estimate of the PSV rather than the true PSV (the actual probability for an individual to employ a DRR) can increase efficiency (Rosenbaum, 1987; Robins et al., 1995; Rubin and Thomas, 1996; Heckmen et al., 1998; Hirano et al., 2003). The variables to be included in the PSV

model need to meet the aims of Conditions 1 and 2. Brookhart et al. (2006) find that including variables that are only connected to outcomes tends to reduce the variance of the final estimate, while variables that only affect participation tend to increase the variance. Taken together, this implies that variables connected to outcomes should be included; their inclusion reduces bias or at least reduces the variance of the model. However, there is a trade-off because the more variables in the PSV function, the smaller the potential overlap between the probability distributions.

The evaluation of the PSV is not focused on the quality of the PSV estimates, in the sense that the estimated PSV is close to the true PSV, or that the regression used to estimate the PSV is consistent (unbiased). The role of the PSV is solely to collapse the relevant information into a single value, which is achieved upon balancing (Rosenbaum, 2002). Furthermore, the actual estimated coefficients of the probit or logit model are also unimportant; evaluation is based solely on achieving the conditions of balancing, Unconfoundedness and sufficient overlap.

Once Conditions 1-3 are deemed to hold, a matching algorithm must be selected. The algorithm will find for each agent in the treatment group (a) member(s) of the control group who has (have) a similar enough PSV, and these two are matched; the average difference between the outcomes (flood damage) of the matches is an estimate of the ATT. There are several methods for the matching process:

1. Nearest-neighbour matching: A match is the person with the closest PS to the observation of interest, but located in the control group. However, it may be that the nearest neighbour is very far away, in terms of the PSV, increasing the potential bias of the estimate, due to poor quality matches. With this method, matching with, or without, replacement can have a large effect. This is because by matching without replacement, an individual is out of the sample once it has been matched. If this individual would have been a good match for another agent, then a worse match for that agent must be made. Matching with replacement solves this issue, and the ordering of data will no longer be important, but the use of less unique information can increase the variance. The trade-off to be made is between bias-reduction (matching with replacement) and precision (matching without replacement).
2. Caliper/Radius matching: Caliper matching creates a match by accepting any PSV as viable if it lies within a bandwidth around the

PSV in which we are interested in for example +/-2%. The benefit of this method is that the number of bad matches will be reduced due to the bandwidth. However it is possible that fewer matches may be made compared with nearest-neighbour matching, as, if no agent is located inside the caliper then there is no match. Radius matching is an extension of this approach because it matches all the observations found inside the bandwidth. There is no strong reason to select one bandwidth over another, a priori, as there is a trade-off between the total matches that can be made against the bias of the matches.

3. Stratification Matching: The area of PSV overlap is partitioned into intervals or strata. Each stratum is defined over a specific range of the PSV, e.g., [0.1, 0.3], and within each strata there are no statistically significant differences between the traits of the treatment and those of the control groups. The overall ATT is estimated by first solving for the ATT within each stratum, and then using a weighted average of the strata ATT. These strata are commonly the same as those used to test the balancing assumption.
4. Kernel Matching: Kernel models use a weighted average of all of the observations in the control group to create matches for the members of the treatment group, where the greater the distance between the PSVs, the lower the weight. As such models use all the members of the control group to create a counterfactual observation for a treatment group member, bad matches will be included in the process. However, the weighting process reduces the influence of bad matches, mitigating their influence. The bandwidth of the kernel is very important, as it determines the degree of smoothing, and large bandwidths may introduce bias into the estimated ATT. While the bandwidth is important, it is unclear what the correct bandwidth is before the investigation begins. Selection of the bandwidth should be treated as a trade-off between bias and variance.

Caliendo and Kopeinig (2005) state that there is no single preferred matching method, as the suitability of each matching method is dependent on the features of the data concerned, but, as the number of possible matches increases, the estimates of each matching method will tend towards the same value. Nevertheless, in small samples, matching with replacement is clearly preferred in order to maximise the number of possible matches. Additionally, if there are a large number of unmatched

control group members, then a kernel matching method may be useful (Caliendo and Kopeinig, 2005) to capture this otherwise lost information. The various measures do allow for a robustness check of the estimated PSV (and, as such, the estimated ATT), as if Conditions 1 and 2 hold they should provide an equally consistent estimate of the ATT. If there is a large difference between estimates, then a detailed investigation to find the missing confounder will be required, so that the estimated PSV can be made more reliable. As such, in small samples, a set of consistent results may indicate that a suitable set of confounders has been found. All of the above matching methods are used here to act as a robustness test and to provide an average estimate of the bias due to selection bias.

The following PSV function will be estimated, where ϕ(.) is the standard normal distribution CDF, θ is a vector of coefficients, x_{it} and is a vector of explanatory variables, and ε_{it} is the error term:

$$T_{it} = \varphi(\theta_0 + \boldsymbol{\theta}' \boldsymbol{x_{it}} + \varepsilon_{it}) \quad (3.3)$$

In eq. (3.3) $\mathbf{x_{it}}$ consists of the confounding variables that explain both participation and outcomes, or at the least outcomes. Variables that only explain participation are to be avoided. Once this model has been estimated for a given vector of $\mathbf{x_{it}}$, the balancing assumption will be tested, as a series of t- or F-tests within each supposed PSV strata. In effect, the sample is stratified by PSV and tested for a lack of systematic differences between the control and treatment group members of that stratum. When balance is achieved, the matching process will be carried out. If balancing is not achieved additional variables will be added to the $\mathbf{x_{it}}$ vector until balance is achieved. The fitted value of eq. (3.3) is the PSV that is used to create matches.

The vector of confounding variables will be guided by economic intuition. The economic incentive to undertake a DRR is the savings due to the installation of the measure over the measure's lifetime. The damage generated by a flood can be viewed as coming from the following process:

$$Damage_{it} = F(Hazard_{it}, Exposure_{it}, Vulnerability_{it}) \quad (3.4)$$

Chapter 1 noted that each element of eq. (3.4) is positively related to the damage outcome. The incentive to employ a DRR is based more on expected damage; the individual's perception of the risk faced, but expected and actual damage may be similar in construction. This economic framework means that there will be a large overlap between the incentive

to employ a DRR and the final outcome, and, as such, the major confounders can be found by focusing on the elements of eq. (3.4). Therefore, the aim of eq. (3.4) is to capture the aspects of the decision making process for employing DRR measures that potentially overlap with damage outcomes (in order to minimise selection bias).

3.3 Data

3.3.1 Survey description

The data was collected via two surveys, one after the flood in 2002, and another one after the floods in 2005 and 2006 in both the Elbe and the Danube river catchments in Germany (Kreibich et al., 2005; 2011; Thieken et al., 2005; Kreibich and Thieken, 2009). On the basis of building specific random samples of private households in flood-affected areas, computer-aided telephone interviews were undertaken in April and May 2003 and in November and December 2006. These surveys resulted in 1697 and 461 completed interviews with private households, respectively. These were large magnitude flood events, as the 2002 flood and caused an estimated total direct damage of €11.6 billion in Germany (Kron, 2004). The flood history of the two catchment areas is quite different. Before 2002, the last major flood that had occurred along the Elbe was in the 1950's while along the Danube a major flood had occurred in 1999 (Thieken et al., 2005). Figure 3.2 presents a map of the catchment areas, as well as an indication of the areas surveyed to provide information on household flood preparedness and consequences of the floods.

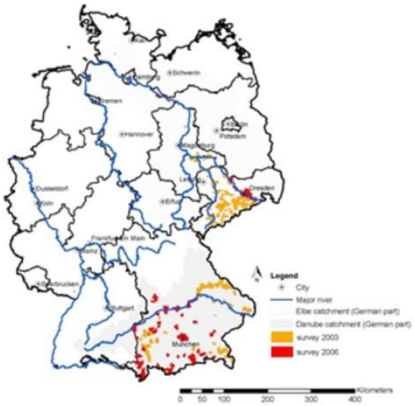

Figure 3.2 A map of the survey locations and river catchment areas

The questionnaires addressed the following topics: emergency and precautionary measures; flood experience; flood parameters (e.g., contamination, water level); socio-economic parameters; and flood

damage. The sample provided by the surveys is trimmed in two respects. The first is that any observations with damage over €100,000 are removed when investigating contents damage and over €300,000 when investigating building damage, as these respondents are strong outliers, and there are few of these observations regarding the sample that can be matched. Furthermore, if these individuals are included in the sample the balancing assumption could not be achieved, and the methodology could not be applied, as described in Section 3.2.

3.3.2 Variables

A brief description of the DRR measures investigated in this study is provided in Table 3.1; a more detailed description can be found in Kreibich et al. (2005; 2011).

The confounding variables are described in Chapter 9.2, but the intuition behind their inclusion is explained here. The variables have been divided into categories based on the elements of the risk triangle and eq. (3.4). The category assigned to each variable is not important for the PSM model. Rather, the categories are used to determine the variables derived from the survey that can influence flood damages. To control for exposure, the value of household contents (for contents damage) or the house price (for building damage) has been included. House prices and contents values fully capture exposure, as they represent the value at risk, where greater values indicate greater potential losses from a flood.

Vulnerability is a more complicated concept. In this study, the focus is on physical vulnerability. The following confounding variables have been used: if the household has a cellar as these houses generally experience higher flood damage (Kreibich et al., 2011), the age of the building, the quality of the building materials and whether the building is located in an urban environment. Floor space is used to proxy the size of the building, as larger buildings may be more likely to come into contact with floodwaters. Where required, to either reduce an ATT's variance or to achieve the balancing assumption, the quality and duration of a flood warning was also included. A warning provides time to make sure that static DRR measures are used correctly or allows mobile measures, like 'Dry flood-proofing' for example, to be employed.

Table 3.1 Flood risk reduction measures

DRR	Description
Flood adapted use	Use in a low value way the flood endangered floors, to keep possible flood damage low, e.g.,

	storing only low value items in flood prone areas
Wet flood-proofing	Avoid valuable, fixed units as interior fitting in the flood endangered floors, but use water resistant or easy replaceable materials for interior fitting,
Adapted building structure	Adapting the building structure, e.g., had an especially stable building foundation, or waterproof sealed cellar walls
Dry flood-proofing	Mobile barriers to prevent water entering the building, e.g., sandbags or local small flood protection walls.

The following variables are used to control for the hazard that the respondents faced: flood water height inside the building; flood duration; contamination of flood water; the return period of the flood; velocity; flood experience; if the building could not be used while flooded; and whether the building is located along the Elbe river (as compared to the Danube river).

It should be noted that above variables might not be useable for all potential PSV functions that analyse a DRR. This is because a variable should be reasonably unaffected by the use of the particular DRR, and certain measures are aimed at directly altering these variables. This problem occurs with 'Dry flood-proofing', which potentially affect water height, flow velocity, and the duration that a building was flooded. In principal the same set of confounders is used to construct the PSV for each DRR. However, when certain variables were included it proved impossible to achieve the balancing assumption. Therefore, not every variable could be included in each PSV function. The list of variables included in each PSV function is displayed in Table 9.3 in Appendix B (Chapter 9.2).

In order to retain as much information as possible and to achieve the balancing assumption, the survey variables were coded in the following manner. Where a variable was categorical, the categories were treated as separate binary variables, and a binary dummy variable was created for each category. Variables such as water height and duration were left as continuous variables.

However, occasionally one of the categorical confounding variables was dropped from the PSV function. Removing a categorical confounding variable will not completely remove all of the information contained by this variable, possibly only altering the variance of the model. However, there is a core set of variables included in each PSV function based on: housing type and quality; whether the building has a cellar; total floor space;

building/contents value; building age; experiences relating to the 2002 (or later) flood(s); warning duration; and how often the individual has been affected by flooding in the past. These variables as a whole capture the elements of eq. (3.4) quite well. For example, contents value would capture the level of contents exposure completely. The methodological approach followed was that the core variables are included in every PSV model and additional variables are added as required to achieve balance or to improve the variance of the estimates.

Once the PS has been estimated, only observations with PSVs within the common support are retained. The common support is determined by removing any observation that has a PSV that lies outside the over lapping areas of PSV.

3.4 Estimation results

3.4.1 Treatment effect estimates

The ATT estimates are presented in Table 3.2 for the five matching methods used. Several methods were used to test the consistency of the ATT estimates, and infer the validity of the confounding variable vector. In particular, the ratio of the standard deviation to the mean of a set of ATT estimates was calculated as a consistency indicator (Table 3.2). This indicator ranges in value from 0.04-0.54, where the smaller the value, the smaller the spread of ATT estimates. Some of the DRR measures have ATT estimates that are very strongly concentrated around a central value. As an illustration, for the significantly effective DRR measures (the effective measures, though 'Dry flood-proofing' are only partial successful), the above consistency indicator ranges from 0.04-0.08. However, for the ineffective measures the indicator ranges from 0.12-0.54. This indicator is especially large for 'Adapted building structure', namely 0.33-0.54, implying that a confounding variable may be missing from the PSV function due to the greater spread of estimated ATT values.

In order to have an overview of the potential bias in a DRR's estimated effectiveness a mean comparison is also carried out similar to that of Kreibich et al. (2005; 2009; 2011). However, the results are not directly comparable with Kreibich et al. (2005; 2011), as these previous studies used (slightly) different data, and the dependent variable here is the absolute value of damage suffered rather than the flood damage proportional to exposure. The former is reported here since it improves the interpretability of the results by generating an explicit value for damage prevented.

The estimated ATTs show that once PSM has removed the sources of bias originating from exposure, vulnerability and hazard, several DRR measures are still effective at reducing flood damage (Table 3.2). Four sets of ATT estimates are highly significantly different from 0 (past the 1% level): these are the DRR measures 'Flood adapted use' with respect to contents and building damage, and 'Wet flood-proofing' with respect to contents and building damage. A fifth ATT set is marginally significant at the 10% level: this is 'Dry flood-proofing' with respect to building damage. This indicates that these DRR measures are the most effective ones out of those investigated. Furthermore, it appears that a large bias is introduced by the elements of eq. (3.4) into mean comparison estimates (Table 3.2). The (average) bias is always negative, and ranges from €387-€14,515, across all measures and types of damage investigated. This implies that a simple comparison of means may result in a substantial overestimate of the damage reduction potential of DRR measures.

After comparing the distributions of the confounders and other descriptive statistics, the main reason for the bias appears to be due to the control group having had, on average, a greater proportion of households suffering from contaminated flood waters, higher water levels, and floods with higher return periods over 1 in 200 years. This may seem to be counter intuitive in that households who did not employ DRR measures face a greater hazard it may be that these householders are less risk averse, underestimate risks, are more myopic or suffer from charity hazard due to the possibility of being compensated by the German government. The possibility of government compensation is mentioned in Seifert et al. (2013). It may also be simply an idiosyncratic feature of these flood events and for a different series of flood events the potential bias may be reversed. Exposure and vulnerability indicators seem rather similar across the two groups.

Table 3.2 Estimates of the effectiveness of private disaster risk reduction [in euros]

	Flood adapted use (contents damage)	Flood adapted use (building damage)	Wet flood-proofing (contents damage)	Wet flood-proofing (building damages)	Water barrier (contents damage)	Water barrier (building damages)	Adapted building structure (contents damage)	Adapted building structure (building damages)
Nearest Neighbour Matching	-6386*** (2364)	-13943** (6694)	-5255* (3099)	-10276* (6030)	4099 (4162)	-8543 (6675)	-2608 (3470)	-1032 (9036)
Radius Matching	-6923*** (2059)	-13574*** (4853)	-4536*** (1919)	-10660*** (4237)	4837 (2964)	-8404* (4465)	-1521 (2629)	-2281 (6236)
Stratification Matching	-6649*** (1660)	-16042*** (5519)	-5217*** (1889)	-11478*** (3354)	4034 (2760)	-8263* (4987)	-1211 (3078)	-3856 (5828)
Kernel Matching (Gaussian)	-7092*** (1599)	-16035*** (4469)	-5830*** (1814)	-12630*** (3347)	3408 (2749)	-9438** (4402)	-1885 (2653)	-5235 (5577)
Kernel Matching (Epanechnikov)	-6608*** (1581)	-14793*** (4644)	-5170*** (1598)	-11466*** (3659)	4110 (2800)	-8108* (4373)	-1339 (2670)	-2478 (5726)
Mean Comparison	-8415*** (1361)	-21968*** (374)	-9063*** (459)	-25817*** (3915)	-713 (1594)	-15,486*** (4315)	-1326 (1760)	-13888*** (4564)
Matches	85	93	80	88	68	80	55	60
Bias	-1683	-7583	-3861	-14515	-4811	-6935	-387	-10912
Average ATT estimate	-6732	-14385	-5202	-11302	4098	-8551	-1713	-2976
Spread of ATT estimates	0.04	0.07	0.09	0.08	0.12	0.06	0.33	0.54
Effective DRR	Yes	Yes	Yes	Yes	No	Yes	No	No

Notes:*, **, *** stand for statistical significance at the 10%, 5% and 1% levels, respectively. The numbers in parenthesis are standard errors. Where analytical standard errors are not available, they have been calculated via bootstrapping with 2000 repetitions. The ATT estimates above have been rounded to the nearest whole euro. The ATT is change in expected flood damages due to a DRR is given, i.e. the more negative the ATT, the more effective is the DRR in mitigating flood damage. The spread of ATTs is measured by the ratio of the standard deviation to the mean of ATT estimates; bias has been estimated as the difference between the 'mean comparison' estimate and the average of the PSM estimates.

While exposure and vulnerability indicators are required to remove bias in the estimated ATT, because they are important confounders at an individual level, the larger degree of difference in hazard seems to be the major source of bias in this application. The reason is that these distributions are most divergent across the groups. Therefore, a simple mean difference in damage fails to account for the differing severity of the floods affecting the control and treatment group.

The DRR measures 'Flood adapted use' and 'Wet flood-proofing' are still very effective when bias has been removed, as these measures have prevented, respectively, about €6,700 and €5,200 of contents damage. The selection bias present in mean comparisons is rather substantial, as for 'Flood adapted use' the bias is 25% of the size of the estimated ATT, while, for 'Wet flood-proofing', the selection bias is 74% of the size of the ATT. Selection bias appears to be a very powerful masking force in a mean comparison.

It appears that 'Flood adapted use' (e.g., storing only low value items in flood prone storeys) is more effective than 'Wet flood-proofing' (e.g., using flood resistant materials to construct interior fittings) for reducing contents damage, which is most likely because the former is a direct measure for limiting the impacts of floods on contents, while 'Wet flood-proofing' would be an indirect way of reducing contents damage due to storage units being more flood safe. The two measures work by altering different aspects of eq. (3.4); 'Flood adapted use' alters the effective level of exposure, while 'Wet flood-proofing' would reduce the vulnerability of household storage units.

The measures that are effective at reducing building damage, i.e. 'Flood adapted use', 'Wet flood-proofing' and 'Dry flood-proofing' again suffer from a substantial bias of €7,583, €14,515, and €6,935, respectively. As a percentage of the ATT, this bias is 55%, 128%, and 81%. The bias regarding building damage as a proportion of the ATT is, on the whole, larger than that present in the estimated ATTs relating to content damage. 'Flood adapted use', 'Wet flood-proofing' and 'water barrier' are still potentially very effective DRR measures preventing €14,385, €11,302 or €8,551 of building damages respectively. 'Wet flood-proofing' is more effective than 'Dry flood-proofing' at reducing building damage because it has reduced the vulnerability level of the building. 'Dry flood-proofing' would reduce the amount of water entering the house, but, dependent on the magnitude of the flood may be overtopped and then would not work at all. Considering the magnitude of the floods suffered, which was up to a 1 in 500 year return period in some cases (Risk Management Solutions, 2003), it may be that 'Dry flood-proofing' may be more effective at reducing building damages incurred from smaller magnitude flood events. The series of strategies represented by 'Flood adapted use' would have caused its reduction in damages due to lower levels of exposure in floodable areas.

'Adapted building structure' was, via a mean comparison, detected to have no significant effect of reducing contents damage, and, even controlling for bias via PSM it is still ineffective. A further finding is that, if nearest-neighbour matching is ignored, then the average bias is only about €150. Such a remarkably close estimate in a small sample could mean that, for this measure, its implementation could be almost as good as random. If all estimated ATTs are included, then the bias increases to 16% of the average ATT for this DRR. This observation reinforces the misleading nature of mean comparisons because sometimes a mean comparison is an accurate estimation technique, while in other cases it is not. The results for 'Adapted building structure' regarding building damage are the most inconsistent set of ATT estimates. This may indicate that there is a missing confounder in the relationship between 'Adapted building structure' and building damage, as, if the whole set of confounders was found then the estimated ATTs should be closer together in value. This inconsistency means that any inference about 'Adapted building structure' and building damage (and, a smaller, degree contents damage) should be treated with caution.

The measure that seems most effective is 'Flood adapted use' as it has a substantial impact on both building and contents damages, while being closely followed by 'Wet flood-proofing'. 'Wet flood-proofing' may be less effective because damage in this chapter is measured via replacement values;

'Flood adapted use' aims at reducing this value while 'Wet flood-proofing' does not. An interesting observation from the ATT estimates is that 'Dry flood-proofing' had a very different effect on contents and building damage. The same measure had a positive (insignificant) ATT regarding contents damage and a negative ATT for building damage, from which it can be inferred that 'Dry flood-proofing' is effective at protecting the building but not its contents. This could be an artefact of an incomplete set of confounders, but this argument fails to explain why 'Dry flood-proofing' protects the building.

3.4.2 Sensitivity analysis

Table 3.3 presents the results of a sensitivity analysis using the methodology suggested in Rosenbaum (2002), who attempts to provide an indication of the possible strength that an excluded confounder would require to alter the results qualitatively. It must be kept in mind that the results of this investigation cannot be viewed as a test of the unconfoundedness assumption. The two areas of sensitivity presented are the bounds on possible statistical significance and the potential 95% confidence interval around the ATT estimate. The way to understand the sensitivity results is as follows: for example, suppose $= 3$, then an excluded confounder would have to change the participation odds by threefold for the observed result to become statistically insignificant at the selected level. This would indicate an ATT estimate that is very insensitive to possibly excluded confounders, and that inference based on the estimated ATT is more reliable than for lower values of Γ. Sensitivity to potential excluded confounders, in this study, will be judged upon what strength of confounder would be required to remove statistical significance at the 10% level. In addition, it is examined what would be required for the possible 95% confidence interval of estimates to include 0. The 95% confidence interval always contains 0 for the results found to be statistically insignificant, i.e. 'Dry flood-proofing' with respect to contents damage and 'Adapted building structure' with respect to contents and building damage. Thus, Table 3.3 only presents the results of this sensitivity analysis for the DRR measures that were found to have a statistically significant effect (up to and including the 10% level).

On the whole 'Flood adapted use (contents damage)' and 'Wet flood-proofing (contents and building damage)' ATT estimates are fairly robust to the possible presence of a missing confounder since, except for nearest-neighbour matching, to remove the statistical significance of the results would require a possible confounder to alter the odds ratio by over 200%. As all relevant and applicable variables from the original survey were included, it is not likely to be the case that such a powerful confounder would have been excluded. 'Dry flood-proofing' on the other hand is less robust as a possibly excluded confounder would have to alter the odds ratio by only 20% to significantly change results. It must be kept in mind that when the ATT for 'Dry flood-proofing' was estimated, because of survey design, it was not able to have a large range of confounders for the hazard component of risk. A large number of hazard variables would be directly affected by the measure and the use of these particular variables would confuse the causal direction of the estimates. It is likely that the negative effect on building damage is still an overestimate, judging from the previously found importance of hazard characteristics.

Table 3.3 Sensitivity analysis

Measure	Matching Method	Statistical significance	95% confidence interval of the ATT includes 0
Flood adapted use (contents damage)	Nearest neighbour	1.4	1.2
	Kernel Matching (Gaussian)	3.4	2.8
	Kernel Matching (Epanechnikov)	3.2	2.6
	Radius Matching	3.3	2.6
Flood adapted use (building damage)	Nearest neighbour	1.2	1.1
	Kernel Matching (Gaussian)	2.5	2.1

	Kernel Matching (Epanechnikov)	2.4	2.0
	Radius Matching	1.6	1.3
Wet flood-proofing (contents damage)	Nearest neighbour	1.7	1.4
	Kernel Matching (Gaussian)	2.5	2.1
	Kernel Matching (Epanechnikov)	2.3	1.9
	Radius Matching	2.2	1.8
Wet flood-proofing (building damage)	Nearest neighbour	1.4	1.2
	Kernel Matching (Gaussian)	3.4	2.7
	Kernel Matching (Epanechnikov)	3.2	2.6
	Radius Matching	3	2.5
Dry flood-proofing (building damage)	Kernel Matching (Gaussian)	1.5	1.2
	Kernel Matching (Epanechnikov)	1.4	1.2
	Radius Matching	1.4	1.2

Notes: The sensitivity to excluded confounders can be estimated for each matching method used, but to save on space only three matching methods have been selected. For statistical significance the number presented refers to the gamma required to reduce significance to past the 10% level.

As the complete range of hazard variables seems to be a major source of bias, it is likely that if the complete range of hazard variables could be included in the confounding vector for 'Dry flood-proofing' it would alter participation odds by more than 20%. Therefore, although 'Dry flood-proofing' seems to reduce building damage, this result should be treated with caution. 'Flood adapted use' appears to be more sensitive to missing confounders regarding building damage compared with contents damage. It is difficult to judge how robust this measure is compared with 'Dry flood-proofing'. For kernel matching 'Flood adapted use (building damages)' seems to be fairly sensitive and more so than 'Dry flood-proofing'. While for nearest-neighbour and radius matching the results are less sensitive than those for 'Dry flood-proofing'. However, compared to 'Dry flood-proofing' 'Flood adapted use (building damages)' contains a more complete range of variables (mainly regarding the hazard) making it less likely that a confounder has been excluded from the model. The results of Table 3.3 may indicate that certain DRR measures are quite sensitive to missing confounders. However, this finding must be balanced against the smaller likelihood that relevant confounders are actually missing from the model.

3.5 Discussion

3.5.1 Discussion of the effectiveness of the studied natural disaster risk reduction measures

The application of PSM to flood damage survey data is able to remove the substantial bias present in estimates of damage reduction via DRR measures based on simple mean comparisons. The bias removed is large, as for the statistically significant content-related measures the bias is around €1,700 to €3,900, while for building-damage related measures the bias is around €6,900 to €14,500. In all cases, the biases are a substantial proportion of the ATT. PSM allows us to provide a more accurate estimate of a DRR's effectiveness, while maintaining as wide a sample as possible. The estimated ATT estimates displayed in the previous section are a refinement of previous estimates of DRR measures in Germany (Kreibich et al., 2005; 2011).

Once bias-corrected estimates have been produced the effectiveness of private DRR measures was found to be less than previously estimated by a comparison of mean damage. Nevertheless, the overall picture of effective DRR measures has not altered substantially as only one previously detected effective measure: namely 'Adapted building structure' in respect to building damage (Kreibich et al. 2005) has been reduced to marginal effectiveness. The most effective DRR is 'Flood adapted use', followed by 'Wet flood-proofing'. This is due to their ability to significantly reduce both contents and building damages. 'Flood adapted use' may also be more favourable, because as a series of coping strategies it may involve smaller installation costs than other measures. The reasons

for the effectiveness of the various measures are described in detail in Kreibich et al. (2005; 2011). Kreibich et al. (2011) also provides indications of the costs of installing various DRR measures, estimated for a model building, i.e. for a detached, solid single-family house with a property area of 750m^2, from which the cost-benefit ratios of some of the currently investigated DRR measures can be calculated. The successful measure common to this study and Kreibich et al. (2011) is 'Dry flood-proofing'. Kreibich et al. (2011) provide a cost estimate of €6,100 for installing 10m of 'Dry flood-proofing'. Assuming that a flood affects a building every year, the expected lifetime discounted (discounted at 3%) cost-benefit ratio is 22.3. The less often a flood is expected to occur, the smaller the cost-benefit ratio, until the breakeven point is reached with an expected flood frequency of around once every 22 years.

The first implication for future flood risk management is that 'Flood adapted use' and 'Wet flood-proofing' should be expanded due to their double dividend return for only one set of installation costs. The next implication is that, while individual level DRR measures do still seem to be powerful tools for limiting flood risk, the role of DRR, as part of current risk management strategies should be altered to take into account the finding that they are less effective than previously believed. This reduction in effectiveness confirms the importance of multiple stakeholders undertaking action as a part of a risk management strategy. A related implication is that, as selection bias was a prominent feature of this study, the possible presence of selection bias in evaluations of non-randomly-employed flood risk management strategies (e.g., the success of a flood warning system) is a concern. Therefore, evaluation techniques that control for many sources of bias simultaneously are required to produce accurate evaluations to guide more productive risk management policies.

It should be noted that the above policy implications are based on the experience of three floods with high overall return periods and water depths. For instance, the average water depth for the treatment group (averaged over all DRR measures) it is approximately 30cm, while for the control group (averaged over all DRR measures) is approximately 80cm. The largest gap is for DRR at nearly 70cm. The investigated DRR measures might respond differently if average floodwater heights were systematically lower across the sample population. While PSM controlled for many sources of bias, it would be useful to analyse in more detail how well the investigated measures perform under a wider range of flood events and in different regions. For instance 'Dry flood-proofing' may be more effective in limiting the damage of more frequent flood events with shallow water depths. Conducting an investigation of the effectiveness of DRR measures that covers a wider range of flood events and geographical areas, while using PSM, could create more readily generalizable results and policy implications.

3.5.2 Discussion of the application of propensity score matching

The value added of PSM in the current application is dependent on the inferred size of selection bias. The estimates of selection bias contained in mean comparison estimates range from 16%-128% of the size of the ATT. Therefore, selection bias can create quite misleading inferences about the ATT as in one case ('Dry flood-proofing' for building damages) the bias is larger than the ATT estimate. The wide range of selection bias indicates a strong possibility for misleading inferences to be made from simple evaluation techniques. Therefore, evaluation techniques that provide a way of controlling for the possibility of large selection bias effects are required. PSM is a technique that is able to achieve the possible removal of selection bias.

The applicability of PSM is strengthened by the ability to employ many different ways of creating a match. This is because the more consistent the results of several matching methods are, the more likely it is that unconfoundedness holds. This becomes apparent from the results of different matching methods for 'Flood adapted use' (contents damage) and 'Adapted building structure (building damage)'. The estimates for 'Flood adapted use' are very closely scattered together. However, the ATT estimates for 'Adapted building structure (building damage)' is about 13 times as wide as that of 'Flood adapted use'. Additionally, by using several matching methods, patterns in the ATT estimates can be revealed. These patterns can allow inference about the true value of the ATT in

a way that a single estimate may not. For example, 'Adapted building structure (contents damage)' provides four estimates that seem to be centred around a value of -1,500, while the fifth is -2,600. This could indicate that the true value is more closely centred on -1,500.

It appears that direct measures of exposure performed better than indirect measures; e.g., contents value is preferred to income. Furthermore, it appears that differences in hazard were a major source of bias. Therefore, a wide range of questions relating to hazard characteristics should be asked. This study successfully applied the following core variables to each PSV function: contents or building value; flood experience; flood water depth and duration; water contamination; flow velocity; building age; and housing material quality. A related recommendation is that the survey must contain not only all of the relevant confounders, but additionally variables that explain outcomes. Relevant confounders can be difficult to identify, as they require a synthesis of the literature that investigates flood damage outcomes and the use of DRR measures. The survey questions should also be presented in a way that allows for the easy construction of dummy variables based on variables that only explain damage outcomes. These variables would provide ample scope for meeting the balancing assumption and reducing the models' variance.

The application of PSM also indicated that large samples are very useful. Large samples are useful, as it is possible that in a flood-affected area the treatment group could be relatively small, simply because few people in the area have chosen to employ a particular DRR. Sampling highly flood-prone areas may also solve this issue, as there is a stronger incentive in these areas to employ a DRR. However, this potentially makes the sample less representative of the larger population at risk. While it is difficult to judge the smallest number of matches that produces a reliable estimate of the ATT, Prirracchio et al. (2012) note that using nearest-neighbour matching (without replacement) and a sample size (total participants) of 40 resulted in a maximum relative bias of 10%. From Prirracchio et al. (2012), it can be inferred that a sample of 100 has a relative bias of 3%, while with a sample of 600 (the total sample in this chapter's application was approximately 640) it is approximately 1.5%. It is difficult to generalise this, but, when combined with the arguments of Holmes and Olsen (2010) and Caliendo and Kopeinig (2005), if several matching methods produce similar results, even in small samples, these results appear to be robust. However, while this study has focused on the applicability of PSM there are additional matching techniques that can be applied, such as Coarsened Exact Matching (see Blackwell et al., 2010 for a sample algorithm). The idea of this matching technique is to temporarily coarsen the data into meaningful groups, then to conduct an exact match on these coarsened data and then only retain the original (uncoarsened) values of the matched data (Blackwell et al., 2010). Such matching techniques can be used in future studies to further support PSM methods (Blackwell et al., 2010) in order to further identify the suitability of the matches made .

The application of PSM seemed to indicate that the relationship between different DRR measures and the confounders may be different between measures. For instance, receiving a flood warning can be a confounder for the use of mobile flood barriers, but not for static DRR measures. Moreover, a variable may allow for balancing in one equation, while in another its presence may invalidate this assumption. Both of these problems mean that an inflexible approach to selecting PSV variables is to be avoided in order to increase the number of situations where PSM can be applied. The principle concern, however, should always be the strength of the unconfoundedness assumption.

3.6 Conclusion

The literature that evaluates DRR measures using survey data is limited. Simple evaluation methodologies and small sample numbers of observational data have the potential to create misleading inferences regarding the success of various DRR measures. This is due to confounding variables, which are variables that explain both the outcomes and the use of a DRR, thereby introducing bias into the estimated effectiveness. The current study sought to remove confounding bias by applying Propensity Score Matching (PSM) to a sample of German households living along the Elbe and Danube rivers who were surveyed in response to floods occurring in 2002, 2005 and 2006.

PSM was applied in order to meet the first objective of this study of more precisely evaluating the effectiveness of various DRR measures. PSM removes confounding bias by matching every individual who uses a DRR with a sufficiently similar individual who did not employ the DRR in order to form the required counter-factual observation. Once PSM has been applied it was found that previous research using mean comparisons of flood damage could result in very inaccurate estimates of the effectiveness of a DRR, due to the presence of confounding variables. However, once PSM has refined previous evaluation estimates by removing the large selection bias, it is found that several DRR measures are still very effective measures for reducing flood risk at an individual level. Moreover, the overall image of successful DRR measures is broadly the same as revealed under previously used methods, only their damage reducing effect is less than may have been previously inferred.

The refined estimates of the damage prevention potential of various DRR measures resulted in several policy recommendations for integrative flood risk management. This study indicates that the most effective measure to extend would be 'Wet flood-proofing' due to the double dividend that this DRR offers and its robustness to excluded confounders. 'Flood adapted use' may be an even more effective DRR to expand, but it is more sensitive to excluded confounders. However, while employing 'Dry flood-proofing' seems to be effective, this result is highly sensitive and should be treated with care. The next implication is that, because selection bias was detected to be strongly present, future evaluation of the success of flood risk