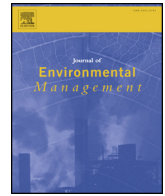




ELSEVIER

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Application of statistical techniques to proportional loss data: Evaluating the predictive accuracy of physical vulnerability to hazardous hydro-meteorological events

Candace Chow^{a,*}, Richard Andrášik^b, Benjamin Fischer^c, Margreth Keiler^a^a University of Bern, Geography Institute, Hallerstrasse 12, 3012, Bern, Switzerland^b CDV Transport Research Centre, Líšeňská 33a, 63600, Brno, Czech Republic^c Geoformer Igp AG, Sebastiansplatz 1, 3900, Brig-Glis, Switzerland

ARTICLE INFO

Keywords:

Multivariate analysis
 Predictive accuracy
 Dimension reduction
 Proportional loss
 Empirical physical vulnerability functions
 Hydro-meteorological hazards

ABSTRACT

Knowledge about the cause of differential structural damages following the occurrence of hazardous hydro-meteorological events can inform more effective risk management and spatial planning solutions. While studies have been previously conducted to describe relationships between physical vulnerability and features about building properties, the immediate environment and event intensity proxies, several key challenges remain. In particular, observations, especially those associated with high magnitude events, and studies designed to evaluate a comprehensive range of predictive features are both limited. To build upon previous developments, we described a workflow to support the continued development and assessment of empirical, multivariate physical vulnerability functions based on predictive accuracy. Within this workflow, we evaluated several statistical approaches, namely generalized linear models and their more complex alternatives. A series of models were built 1) to explicitly consider the effects of dimension reduction, 2) to evaluate the inclusion of interaction effects between and among predictors, 3) to evaluate an ensemble prediction method for applications where data observations are sparse, 4) to describe how model results can inform about the relative importance of predictors to explain variance in expected damages and 5) to assess the predictive accuracy of the models based on prescribed metrics. The utility of the workflow was demonstrated on data with characteristics of what is commonly acquired in ex-post field assessments. The workflow and recommendations from this study aim to provide guidance to researchers and practitioners in the natural hazards community.

1. Introduction

Hydro-meteorological hazards that occur in mountainous areas can have devastating consequences on local communities. In Switzerland, natural hazards that occurred between 1972 and 2016 amounted to average damages of approximately CHF 305 million per year. A major proportion of these damages were caused by a limited number of high magnitude events; for instance, the 2005 floods alone resulted in CHF 3 billion in damages (Bundesamt für Umwelt, 2018). Furthermore, spatial patterns of risk to natural hazards in mountain regions are expected to change in the future due to climatic and environmental factors (Mazzorana et al., 2012; Papathoma-Köhle, 2016). Weather extremes in Europe are expected to result in increasingly more frequent and higher magnitude precipitation events, which have been associated with the onset of hazardous occurrences such as floods and debris flows (Toreti

et al., 2013; Volosciuk et al., 2016). Additionally, developments resulting in changes to land use patterns are expected to alter the vulnerability of elements at risk (Thieken et al., 2016). However, there is still an incomplete understanding of the independent and joint effects of hazard driving forces and events often engender highly variable consequences to affected elements (Vogel et al., 2014). Given the destructive potential of these events, there is justifiable interest in gaining a better understanding of the drivers and the prediction of expected damages. Post-event vulnerability and consequence analyses about the causes and impacts of hazards can support future decisions on integrated risk management, ranging from the spatial planning of communities at risk, optimized coordination of emergency efforts and resources, to the assessment of how effective protection measures are.

Of the types of potential consequences of hydro-meteorological hazards on elements at risk, physical vulnerability is defined as the

* Corresponding author.

E-mail addresses: candace.chow@giub.unibe.ch (C. Chow), andrasik.richard@gmail.com (R. Andrášik), beni.fischer@bluewin.ch (B. Fischer), margreth.keiler@giub.unibe.ch (M. Keiler).

<https://doi.org/10.1016/j.jenvman.2019.05.084>

Received 28 November 2018; Received in revised form 9 May 2019; Accepted 21 May 2019

0301-4797/© 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

predisposition of a building to be susceptible to any degree of damage as a result of a specific hazard impact (Uzielli et al., 2008) and as a function of its profile and environment (Ettinger et al., 2016). Hazard impact is determined by the interaction of two factors - the intensity of a given hazard and the susceptibility of the elements at risk (Uzielli et al., 2008). The consequence of hazard impact on buildings may be expressed as proportional loss, $y = d/v$, where d is the amount of damage sustained in monetary units and v is the insurance value of a specific structure (Rheinberger et al., 2013). In other cases, damage grades have been used to indicate ranked degrees of structural and non-structural damages with respect to a given hazard intensity (Charvet et al., 2015; Ettinger et al., 2016; Laudan et al., 2017).

Charvet et al. (2017) identified types of physical vulnerability functions based on the data collection source. Empirical functions use data from post-hazard assessments; judgement-based functions derive damage estimates from expert elicitation; analytical functions are based on the results of numerical simulations of structural damage and hybrid functions employ a combination of the aforementioned approaches (Charvet et al., 2017). In most literature found in the natural hazards domain, physical vulnerability functions (or curves) are defined as quantitative, investigative approaches used to evaluate the physical vulnerability of buildings to natural hazard events (Papathoma-Köhle et al., 2017). More specifically, empirical physical vulnerability functions mathematically relate hazard intensity to the damage response of a building (Tarbotton et al., 2015) and consist of two main types. Damage functions typically represent the response in terms of absolute damage (i.e. the cost to completely restore an affected building) or relative loss (i.e. a percentage that represents the damaged proportion of a building). Fragility curves describe the conditional probability that a damage state will be reached or exceeded with respect to a given hazard intensity level (Choi et al., 2004). Empirical vulnerability functions have been developed to assess damage responses of buildings to different types of hazards, including but not limited to tsunamis (e.g. Charvet et al., 2017), floods (e.g. Büchele et al., 2006), fluvial sediment transport (e.g. Totschnig et al., 2011a, b) and debris flows (e.g. Papathoma-Köhle et al., 2012). Empirical-based analyses that relate damages from hazard processes to intensity and susceptibility features are considered to be more limited to the use of other investigative methods (Rheinberger et al., 2013). Consequently, this study focuses on the further development and continued assessment of empirical vulnerability functions based on building damages sustained from the occurrence of hydro-meteorological hazards.

Inferences and a better understanding of damages require the collection of relevant data following the occurrence of hazard events and the continued development and application of empirical vulnerability functions; this is not without its challenges. Firstly, empirically-based analyses require large quantities of accurate and complete ex-post data records at building level to be considered reliable (Ciurean et al., 2017; Papathoma-Köhle et al., 2017). Ex-post data is defined in this context as data collected about a given event following its occurrence. From an analytical standpoint, this type of assessment is difficult to conduct for certain types of hazards such as debris flows, where the number of directly affected buildings is notably less than those typically affected by the wider spatial extent of hazards such as floods and earthquakes. Furthermore, object-specific observations are often sparse (Vogel et al., 2014; Laudan et al., 2017). This is partially attributed to the rapid post-event intervention of authorities to restore the functionality of affected communities, which effectively reduces the amount of time field-based damage assessments can be conducted (Ettinger et al., 2016). An alternative to using observed damage data involves numerically modelled hazard intensities (e.g. Quan Luna et al., 2011). However, the outputs are associated with model and parameter uncertainties that warrant further investigation (e.g. Chow et al., 2018).

Secondly, data collected on object-specific damages have high dimensionality due to a large number of contributing factors to the pre-conditions and consequences of a given hazard event. In this context,

pre-conditions refer to the combined status and characteristics of the object and its surrounding environment prior to the occurrence of an event. Examples can include, but are not limited to, the number of surrounding buildings (i.e. sheltering effect), the proximity of a building to the main channel or preferential pathways, the implementation of local protection measures and building construction type. As a general rule of thumb, at least ten records should be available per feature variable (factor), also referred to as events per predictor or EPV (Concato et al., 1995; Peduzzi et al., 1995, 1996). The studies upon which the basis of this rule was founded were designed to evaluate the effects of varying the numbers of events with respect to a constant number of predictor variables. Results from these studies highlighted a range of concerns as EPV was reduced below 10 events. In particular, any conclusions drawn from results with < 10 EPV could be challenged on the basis of increased bias and variability, unreliable confidence interval coverage and problems with model convergence. Vittinghoff and McCulloch, 2007 also demonstrated that between 5 and 10 observation records per feature variable is enough, especially if results are statistically significant. In cases where there are less than five records per factor, dimension reduction prior to conducting multivariate analysis is requisite. The resultant dataset should contain sufficient instances of all unique combinations of feature variable values so that any findings that result would be subject to less contention associated with the use of low EPV.

Thirdly, vulnerability functions developed in the past did not consider the full range of hazard process characteristics (i.e. focusing primarily on flow or sediment deposition depths as intensity proxies) and did not consider the influence of building properties (e.g. construction type). In Papathoma-Köhle et al. (2011), these are referred to as functional relationships, which are limited to relating hazard intensity to the proportional loss of elements at risk. In recent studies (e.g. Table 1), additional building and surrounding area characteristics, in addition to multiple hazard intensity proxies and their interactions, have been considered. Certain features have already been identified as important or advantageous to consider. Specific examples are cited under three categories as:

- building resistance features (e.g. exposition in the flow direction, Laudan et al., 2017; susceptibility of building elements to intrusion, Laudan et al., 2017; building characteristics or structural adaptation to the local environment as a means to minimize hazard impacts, Charvet et al., 2015; Margreth and Romang, 2010; Ettinger et al., 2016);
- surrounding area features (e.g. shadowing effects of neighbouring buildings that retain process materials from the building in question, Laudan et al., 2017; distance to channels or bridges, Ettinger et al., 2016); and
- damage pattern features (e.g. process intensity proxies, Charvet et al., 2015; Rheinberger et al., 2013; pairwise interactions between intensity proxies, Rheinberger et al., 2013).

In this study, we refer to the products of these developments as multivariate vulnerability functions. Table 1 summarizes four recently conducted, empirically-based studies with the objective of predicting building damages from the occurrence of natural hazards with multivariate data. These studies include but are not limited to floods, tsunamis and debris flow events. In general, these four studies considered a lower number of feature variables (p) compared to a relatively higher number of observations (n). Of the feature variables included in the statistical models, there is a differentiation between building resistance and surrounding area profile attributes (pre) and hazard intensity proxies or damage patterns ($post$). In three of the studies, the expected value of y -response was ordinal damage grades, whereas, in one of the studies, the y -response was expressed as bounded proportional loss values.

Despite conducting evaluations with available databases and more

Table 1
Overview of past studies that focus on developing and evaluating empirically-based vulnerability functions.

	number of events or sites (hazard)	location	number of observations (n)	number of feature variables (p)		y-response variable	y-response values
				pre			
				pre	post		
Laudan et al. (2017)	1 (flash flood)	Germany	94	13	4	damage grade (D; ordinal)	1 to 5
Ettinger et al. (2016)	1 (flash flood)	Peru	898	8	1	damage levels (DO; ordinal)	1 to 5
Charvet et al. (2015)	1 (tsunami)	Japan	19,815	2	3	damage states (DS; ordinal)	0 to 5
Rheinberger et al. (2013)	5 (debris flows)	Switzerland	132	6	3	proportional loss ratio (numeric)	[0,1], logit-transformed to $[-\infty, \infty]$

advanced statistical estimation methods in previous studies, existing models have not been able to explain all systematic variability in the data, especially at higher levels of damage (Charvet et al., 2014). The unexplained residual variability in the observed damages may be resolved by considering additional or different explanatory variables describing hazard intensities, building resistance, environmental features and/or their interactions. The aggregated effects of the aforementioned factors and challenges render hydro-meteorological hazard assessments inherently complex. While it is possible to perform analyses and develop vulnerability functions with currently available data and models, the amount of confidence assigned to the results and their transferability to other locations and future scenarios must be critically reviewed. Vulnerability functions are developed with damage data caused by a hazard event with certain intensities, spatial and temporal distributions (Totschnig et al., 2011a, b) and affected buildings with specific characteristics. The specificity, with which these functions are developed, has implications for transferability. Consequently, Papathoma-Köhle (2016) recommended that vulnerability assessments be revised and constantly adjusted and Ettinger et al. (2016) cautioned that vulnerability indicators are too site-specific to be applied operationally to other locations. The authors of the latter study highlighted differential, site-specific building structures and channel settings as reasons for exercising discretion. Additionally, site-specific triggering processes and upstream-downstream evolution of hazard processes over time and space should be taken into consideration (Di Baldassarre and Montanari, 2009). Given this context, what can we learn from past hazard events and how can this insight be used to inform decisions in the future?

In light of the aforementioned challenges and research gaps, an updatable workflow is described to support further development and evaluation of empirically-based, multivariate vulnerability functions. Moreover, the multi-step procedure considers the option of updating methods at each of the steps and with respect to the nature of available data. The study is conducted on an empirical dataset that consists of hazard, building and surrounding area characteristics of three debris flow and sediment-laden flood events that resulted in heavy building damages. These events occurred in 2005 in the Swiss Alps. Furthermore, the study describes a procedure to pre-process ex-post damage data, which is often subject to the challenges of data sparsity and high dimensionality. High dimensionality occurs when there are a greater number of feature variables to observation records, where each feature represents a dimension. Data sparsity refers to the treatment of missing data entries. While the explicit exploration of missing data in natural hazards studies is limited, very high proportions of missing data have been observed in other fields and treatment methods have been evaluated (e.g. Albrecht et al., 2010 and Nguyen et al., 2017). Different methods, also referred to as matrix completion approaches can be considered, however, there is currently no consensus on the best approach to handle missing data. Often, domain and/or data specific best practices are prescribed after experiment-based trials are validated to determine whether resultant solutions are realistic. In the field of natural hazards, only one known study, conducted by Macabuag et al. (2016), demonstrated the use of multiple imputation (MI) techniques on a dataset with missing entries.

Empirical vulnerability functions are derived by applying statistical model fitting techniques on building damage data. The type of model that is applied is dependent on the expected outcome (i.e. nature of the y-response variable). When choosing models to investigate a particular problem, several aspects should be considered, including the objective, underlying assumptions, model structure and how parameters are estimated. Linear regression models and generalized linear models (GLMs) have been applied in previous studies. Table A (Supplementary material) compares the key differences between the two types of models under the aforementioned considerations. Kawano et al. (2016) recognized the importance of being able to detect and represent non-linear and non-monotonic dependencies in data describing complex

phenomena, especially with regards to damage modelling and associated uncertainties. Compared with linear models, GLMs relax assumptions of normality for both the y-response variable and errors (McCullagh and Nelder, 1989) and assumed linearity between response and feature variables (Charvet et al., 2017). Additionally, Rheinberger et al. (2013) highlighted an additional advantage to applying double generalized linear models (DGLM; Smyth, 1989), a type of GLM that adjusts for overdispersion, which is commonly associated with proportional loss data.

While some models predict single outcomes, others have been further developed to predict an ensemble or range of outcomes. The application of an ensemble predictor, random generalized linear model (RGLM; Song et al., 2013), was first evaluated by (Laudan et al., 2017) for damage modelling. The RGLM is comprised of a set of models (i.e. bags), each containing a sufficiently different subset of the original feature variables, so that variability is maximized. RGLM incorporates elements of randomness and instability, feature independency and forward variable selection based on the evaluation of a model fit metric (i.e. AIC; Akaike, 1974). Consequently, the RGLM performs both dimension reduction and model fitting simultaneously.

The study brings together these multiple lines of investigation on statistical models to building damage data in an end-to-end workflow. Furthermore, the effect of input data on the predictive accuracy of vulnerability functions that are developed are evaluated. In particular, the study considers original input data, datasets with reduced dimensions and the inclusion of interaction terms, with the aforementioned model structures.

2. Data and methodology

2.1. Data

Data acquisition of both structural and non-structural features about buildings affected by hazards supports a better understanding of the contributing factors to specific damage processes. However, there are recognized challenges associated with data availability, quality and existing collection methods. Firstly, direct, real-time observations at appropriate temporal and spatial resolutions that are required for vulnerability assessments are difficult, if not impossible to obtain (Gaume et al., 2009). Furthermore, data to support the accurate characterization of structural failures is often unavailable or incomplete (Papathoma-Köhle et al., 2011). In light of these challenges, supplementary data about object-specific damages, building resilience and properties of the immediate environment were collected for this study, based on findings and recommendations of comparable studies (e.g. Ettinger et al., 2016; Laudan et al., 2017; Rheinberger et al., 2013).

2.1.1. Data acquisition in swiss communities

In the summer of 2005, torrential rains across extensive areas of the Swiss Alps caused large-scale floods and numerous landslides. The highest losses were recorded in the Canton of Bern, at 805 million CHF (DETEC, 2008), mainly caused by three local events: a debris flow occurred in Brienz and sediment-laden floods affected both Diemtigen and Reichenbach (Bezzolo and Hegg, 2008; Scheidl et al., 2008). Feature variables describing the pre-conditions and consequences of these events were organized under three categories of interest: building profile and resistance, surrounding area profile and damage patterns. In some studies, vulnerability is defined as the set of features describing pre-existing conditions (i.e. building design and site-specific environmental characteristics) that increase their susceptibility to the impacts of hazard processes (Ettinger et al., 2016; Papathoma-Köhle et al., 2011). This corresponds to the building resistance and surrounding area profile categories. Collectively, the pre-condition features are represented in a “pre” dataset, while a “post” dataset includes the pre-condition and damage pattern features. Data was compiled from three sources: the cantonal insurance provider (Gebäudeversicherung Bern;

GVB), responses to the field-based survey conducted with local residents and data derived from remotely sensed products (i.e. ortho-photographs, photographs, digital elevation model; Swisstopo, 2005).

Ground-survey based data acquisition was conducted at the three affected sites in 2018 to collect supplementary data about pre-condition attributes, in addition to responses about damages, beyond the conventional consideration of flow or sediment deposition depths. Engaging residents involves care in survey design and careful choice of questions that can be answered with confidence by non-experts while providing representative data for further analyses. Generally, data on a relatively limited and partially representative sample of affected buildings can be obtained through surveys. The selection of buildings for sampling was largely a function of the residents’ willingness to grant access. Considering the time since the event occurred and when the survey was conducted, questions about damages were mainly limited to binary (yes or no) or categorical responses to minimize uncertainty attributed to memory and acknowledging that respondents may not necessarily have the expertise to provide more detailed responses. Data collection was conducted through three means to maximize response rate: in-person interviews, if residents were immediately available for consultation, delivery of a hard copy of the survey with return postage paid and an online version of the survey coded on the third party Ona platform (Ona, 2018). Details about the types of features that were collected are presented in 2.1.3.

2.1.2. Proportional loss

The GVB shared data on proportional loss (relative damages), for buildings located in the three affected communities. Proportional loss values are continuous, bounded fractional response values [0,1]. The values describe the consequence of a specific hazard process on site-specific buildings, where 0 represents no loss or damage sustained and 1 represents a total loss or complete structural failure.

Exploratory assessments of the y-response variable are important to determine which model structure is suitable for subsequent analyses of the expected value. In the case of proportional loss, since it is defined on a bounded interval that is not concentrated within the interval, the values follow a non-normal distribution. A logit-transformation is applied to proportional loss values with the logit function from the R package car (Fox et al., 2011). Logits are real numbers, which range from minus infinity to infinity. The transformation effectively increases the resolution of values distributed towards both bounded ends. A correction was applied to remap the boundary values to 0.025 and 0.975, respectively, which resulted in logit-transformed proportional loss values between -3.66 and 3.66 (Fig. 1). The nature of the proportional loss data informs the types of models that are built later on to support further analyses; details are provided in 2.2. Both proportional loss and logit-transformed proportional loss are distributed in a bounded domain and their empirical probability density functions are bimodal (Fig. 1). Therefore, normality cannot be assumed.

2.1.3. Feature variables

Immediately following the occurrence of the events, photographs depicting object-specific damages from hazard processes were compiled from local residents (Bezzolo and Hegg, 2008). Flow conditions at each affected building are required to derive physical vulnerability functions. In this study, the average sediment deposition height represents the main hazard intensity proxy of both types of hazard events. The values were estimated from the object-specific photographs by examining where visible debris or water marks were left on the facades of affected buildings. Additional data that indicate certain degrees of impact of the specific hazard processes on buildings was collected as binary responses through the field surveys.

Exploratory assessments of the feature variables to the y-response were conducted to gain insight into the strength of pairwise relationships. Three types of statistical tests were conducted to detect the strength of association between the continuous y-response and the type

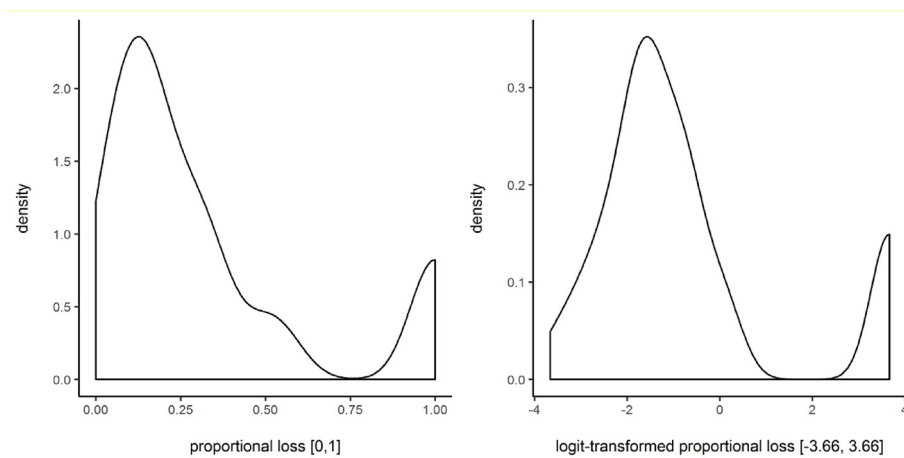


Fig. 1. Visualization of proportional loss value distributions with density plots of the original proportional loss values [0,1] (left) and logit-transformed proportional loss values [-3.66, 3.66] (right).

of feature variable. Spearman's correlation coefficients were computed for continuous variables, the Wilcoxon test was conducted on binary variables (i.e. with 2 levels) and Kruskal-Wallis test was conducted on categorical variables (i.e. with > 2 levels). A correlation coefficient $\rho \geq 0.50$ and p-values < 0.05 from the tests indicate that the feature variable has a significant correlation with proportional loss. A summary of the feature variables considered in this study are presented in Table B (Supplementary material); the bivariate significance detected with the original data set is indicated in light blue.

2.2. Data pre-processing

A dataset, with n -observations and p -feature variables, resulted from the compilation and coding of collected survey data; this dataset requires further pre-processing before modelling. This pre-processing step follows two main objectives: to address data sparsity (2.2.1) and high dimensionality (2.2.3).

Data were pre-processed and analyzed in RStudio (R Core Team, 2018). The pre-processing workflow is separated into two sections (Fig. 2). Section A prepared data for models that take original variables as inputs, while section B prepared data for models built with alternative predictors in the form of principal components. Sections 2.2.1–2.2.4 provide further details about each of the pre-processing steps.

2.2.1. Data sparsity

Firstly, missing data must be addressed before further analysis is possible. Figure A (Supplementary material) illustrates the extent and distribution of the missing data (i.e. indicated in red). The figure provides an impression about the prevalence of missing values, highlighting the challenge of collecting data at building level and the importance of this pre-processing step.

The effectiveness of any missing data treatment method is strongly affected by the ratio between missing data and available observations. Consequently, the efficacy of treatment methods is expected to decrease when applied to higher numbers of missing cases (Munguía and Armando, 2014). As the missing to available data ratio is minimized, the results are expected to improve, since the availability of actual observations provides a more precise estimate of the real distribution. Otherwise, without access to more observations, the choice of imputation method should be specific to a given data set and should not be generalized to other data sets without thorough data exploration. Ideally, the design and application of a missing data treatment plan should be customized to each feature variable within a dataset to properly address 1) the nature of the missing data pattern, 2) the percentage of

missing cases and 3) whether the actual range of observations is known and represented by the available observations (Munguía and Armando, 2014). Table 2 summarizes the three types of missing data and the prescribed treatment.

While there is some general understanding about the missing data mechanisms for each of the feature variables, a more precise idea about whether the ranges of observed values represent reality may be largely unknown. Based on the aforementioned classifications, all three types of missing data were observed in the compiled dataset used in this study. For instance, average sediment deposition height is a feature variable found in this dataset; its values generally increase with proportional loss, the y-response variable. However, the variance in proportional loss is incompletely explained by this feature. Furthermore, there is the possibility of interaction effects among variables. Consequently, deposition heights that correspond with highly or completely damaged buildings cannot be assumed to be comparable or higher than the heights observed with low to medium proportional losses. In other words, if higher damages can be attributed to a combination of factors (e.g. boulder impact and/or large volumes of sediment intrusion in a given building), it is impossible to determine which magnitude of sediment deposition height contributes to the overall damage. Therefore, the use of MNAR-specific imputation methods would require a better understanding about the missing data values themselves. However, without additional opportunities to revisit the data collection process, only techniques that assume MCAR and MAR missing data patterns can be applied; this is recommended when there is incomplete knowledge about the nature of missing values (Lazar et al., 2016).

Imputation methods aim to optimize data retention by assigning a plausible value to each missing observation. This is accomplished by preserving the characteristics of each feature variable while simultaneously considering the impact of relationships between feature variables in a given dataset. In this study, we evaluated three missing data treatment techniques under assumptions of MCAR/MAR, namely, mean-based imputations (Meyer, 2018), k-nearest neighbour (kNN; Templ and Alfons, 2009) and multiple imputation based on principle component analysis (MIPCA; Josse and Husson, 2012). Figure B (Supplementary material) visually compares the data distributions of original and imputed average sediment deposition height values, with respect to the three treatment techniques. From this example, it is evident that the baseline approach (mean-based imputations) generates results with zero variability. Although the other two methods that were assessed (kNN and MIPCA) did not fully capture the same distribution as the observed values, the variability and maximum/minimum ranges of imputed values were in higher agreement than values imputed with

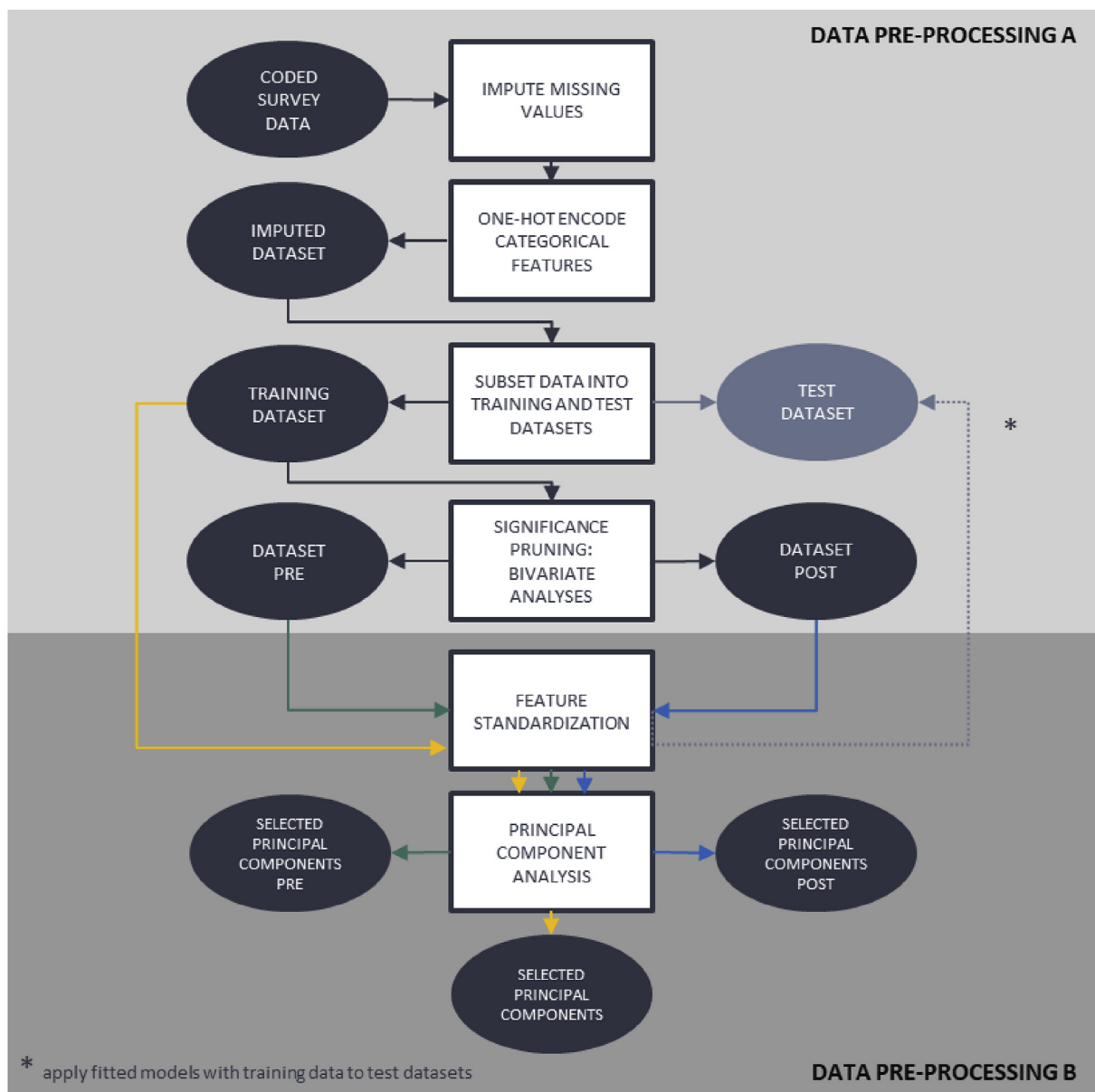


Fig. 2. Data pre-processing workflow to address the challenges of sparsity and high dimensionality in a given input dataset prior to model building.

Table 2
Overview of different classifications of missing data and prescribed treatments (after Macabuag et al. (2016)).

classification	method of identification	recommended action
missing completely at random (MCAR)	Determine if missing data distribution is consistent for the complete dataset (Kolmogorov-Smirnoff test for disaggregated data or χ^2 -test for aggregated data)	Conduct complete-case analysis (i.e. exclude observations with any instances of missing data and perform regression analysis on the remaining dataset) or estimate missing data with MI techniques
missing not at random (MNAR)	Determine if missing data from another feature is related to the reason that data from the target feature is missing	Vulnerability analysis cannot be conducted without introducing bias; revisit data collection process
missing at random (MAR)	Neither MCAR or MNAR	Estimate missing data with MI techniques

averages alone. Furthermore, using only highly correlated features to impute missing values in the target variable improves the agreement between the distribution of imputed values and that of the observed values. However, due to aggregated sources of uncertainty, it is not definitely clear in this example which method (i.e. kNN or MIPCA) performs better. Nevertheless, it is of interest to have an idea about how different imputation results compare, since more representative imputation results support the ability to draw conclusions from subsequent analyses with greater confidence.

For this particular dataset, a combination of the aforementioned imputation techniques was applied (Fig. 3). In particular, kNN was used to impute both numeric and categorical missing entries and MIPCA was used to impute the average sediment deposition heights for the subset of debris flow data. With reference to the findings presented in Figure B, while the results associated with mean-based imputation were found to be less inadequate than those engendered with the other two methods, closer inspection of the values imputed with kNN revealed some unsatisfactory joint distributions. This is indicated by the vertical

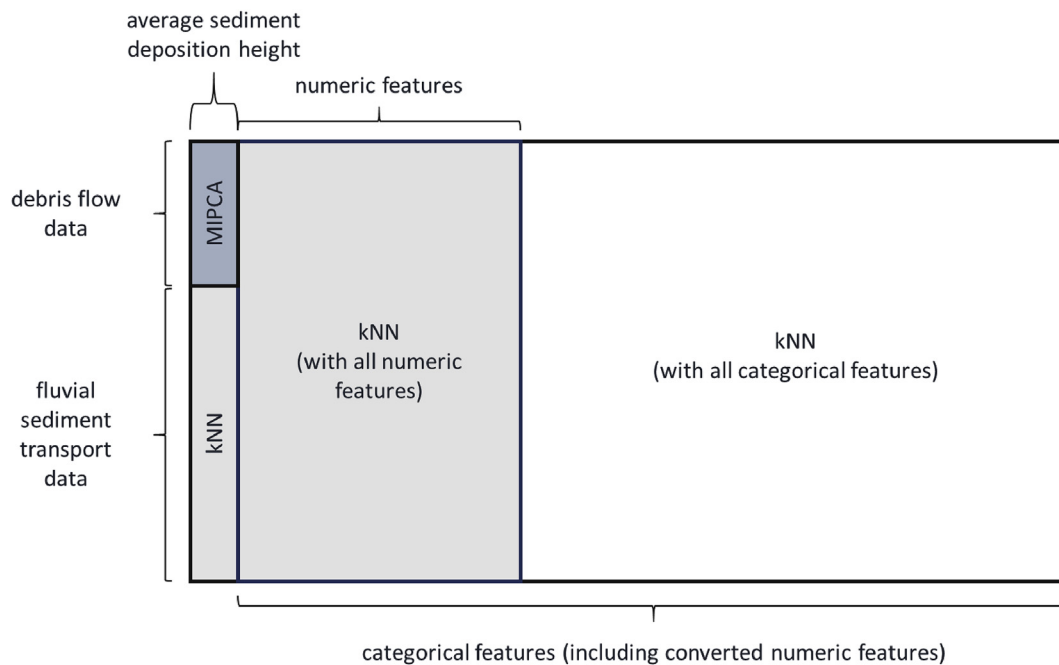


Fig. 3. A hybrid dataset with complete observation records after the application of multiple types of imputation techniques.

alignment of data clusters in the kNN margin plot. Additionally, Fig. 3 indicates that imputations were conducted for subsets of debris flow and fluvial sediment transport data separately. Imputation with the MIPCA approach resulted in a stronger agreement when applied to the DF subset, using only selected features that were highly correlated with the target variable. Consequently, MIPCA was used to impute missing cases of average sediment deposition height for debris flow observations and all other numeric feature variables were imputed with the kNN method. In general, these data distribution comparisons provided a basis to select for particular sets of estimates to build a complete set of observation records for further analyses, in lieu of additional observations in field.

2.2.2. One-hot encoding

With the hybrid dataset of complete observations, all categorical feature variables (nominal and ordinal) that were label encoded with multiple factor levels were one-hot encoded. Label encoding often carries a misleading assumption that the higher the categorical value, the more significant the level or class. For example, a feature variable

representing different types of building materials may contain classes label encoded as 1 = masonry, 2 = concrete, 3 = wood (Fig. 4). From this example, the numbers associated with the codes do not correspond to additional information about the building's structural vulnerability and interpolations between classes (e.g. building material with a value of 1.5) are meaningless. Furthermore, label encoding of ordinal data carries an additional, often invalid assumption that levels are equidistant from each other. To exclude the introduction of these sources of errors in model predictions, all categorical features were one-hot encoded as binary values before the features were used to train models.

2.2.3. Dimension reduction

This pre-processing step addresses two main concerns – the inclusion of feature variables that are not highly correlated to the y-response and high correlations between variables, which are undesirable in subsequent regression modelling. By comparing the performance of models built with these subsets of pre-selected features, we investigated whether the variance in proportional loss can be better explained with fewer features. Two dimension reduction approaches were evaluated.

observations (n)	feature variables (p)				
	building materials	X ₂	X ₃	...	X _p
building 1	2 (concrete)
building 2	3 (wood)
...
building n	1 (masonry)

	concrete	wood	masonry	X ₂	X ₃	...	X _p
building 1	1	0	0
building 2	0	1	0
...
building n	0	0	1

Fig. 4. Example of one-hot encoding of a label coded feature variable (e.g. building materials).

Table 3
Overview of datasets after significance pruning.

dataset	number of observations (n)	number of feature variables (p)	inclusion of damage pattern features	all available feature variables, including one-hot encoded categorical data selected features based on the strength of pairwise correlations with proportional loss selected features based on the strength of pairwise correlations with proportional loss
all	81	84	Y	
pre	81	21	N	
post	81	48	Y	

The first form of evaluation, bivariate analyses, was conducted to detect relationships between feature variables (x) and proportional loss (y), and instances of high correlations among features. The results of these analyses were summarized in Table B (Supplementary materials) and used to prepare input datasets based on significance pruning. This excluded redundant or relatively low importance features from further modelling. Two subsets (Table 3) were prepared as a result; the choice of features in each of the subsets also corresponds to the pre-conditions (pre) and post-event conditions (post) contributing to vulnerability assessment. The null hypothesis is that neither subset of features has an effect on the explanation of proportional loss when compared with a baseline model that includes all available features (all).

The second way to reduce dimensionality is by conducting principal component analysis (PCA; Hotelling, 1933; Pearson, 1901). This approach is attractive as the analysis returns uncorrelated variables in the form of principal components and mitigates against overfitting. While techniques for the exploration and visualization of 2- and 3-dimensional problems are commonly applied, a different approach is required to explore a high-dimensional dataset. The overall aim is to retain a minimal number of principal components (PCs) to reduce input feature variable dimensionality. PCA can be considered as a type of multi-dimensional scaling that returns a linear transformation of a higher number of feature variables (i.e. > 3) into a lower dimensional space in the form of PCs, while retaining as much information about the original feature variables (X_1, X_2, \dots, X_p) as possible. Consequently, these components are the result of the optimization problem and are linear combinations of the original feature variables, generated based on maximum variances (Aguilera et al., 2006). A standardized input dataset of interest (n x p) is decomposed into two orthogonal output matrices containing PC loadings and scores, respectively (Figure C; Supplementary material). The loadings output matrix stores eigenvector coefficients, which can be interpreted as weights associated with each of the feature variables to a specific PC. The weights are calculated according to the degree of variance in each variable and indicate the relative contribution or importance of a variable to a particular PC. The scores output matrix can be interpreted as a new measurement value that is the sum of the product of normalized values and the relative contribution of the particular value (i.e. eigenvector coefficient). Two main PCA approaches exist: eigenvalue decomposition and singular value decomposition. The latter is preferred over the former method for numerical stability. In this study, the prcomp routine from the R package stats (R Core Team, 2018) was used to perform PCA.

Truncation or exclusion of PCs beyond the top-ranked number of components informs about the complexity of the input dataset; the dataset is less complex if a greater percentage of variance can be explained by a lower number of components. Furthermore, the noise in the data is reduced in the process. Three metrics (stopping rules) are commonly applied to guide the optimal number of PCs retained, namely the number of components 1) with eigenvalues greater than 1, under unit variance, 2) prior to the inflection point observed in scree plots and 3) that explain at least a user-defined threshold of cumulative variance (e.g. 75%) (James et al., 2013). An eigenvalue > 1 or s/p indicates that the associated PCs account for more variance than by one of the original feature variables, on the condition that the data is standardized. If y-aware standardization has been applied to the features (further explanation in 2.2.4), variables may no longer have unit variance; eigenvalue > s/p accounts for this change, where s represents the sum of variances across all feature variables included in the PCA and p is the number of features. Scree plots illustrate the number of PCs that correspond with a proportion of variance explained (%) in descending order of magnitude. It is a visual heuristic that is commonly used to support the selection of a certain number of PCs based on relative importance. The inflection point at the base of the steeply descending slope is assumed to be an indicative point where subsequent PCs have a limited to negligible contribution to explaining residual variance in the y-response variable.

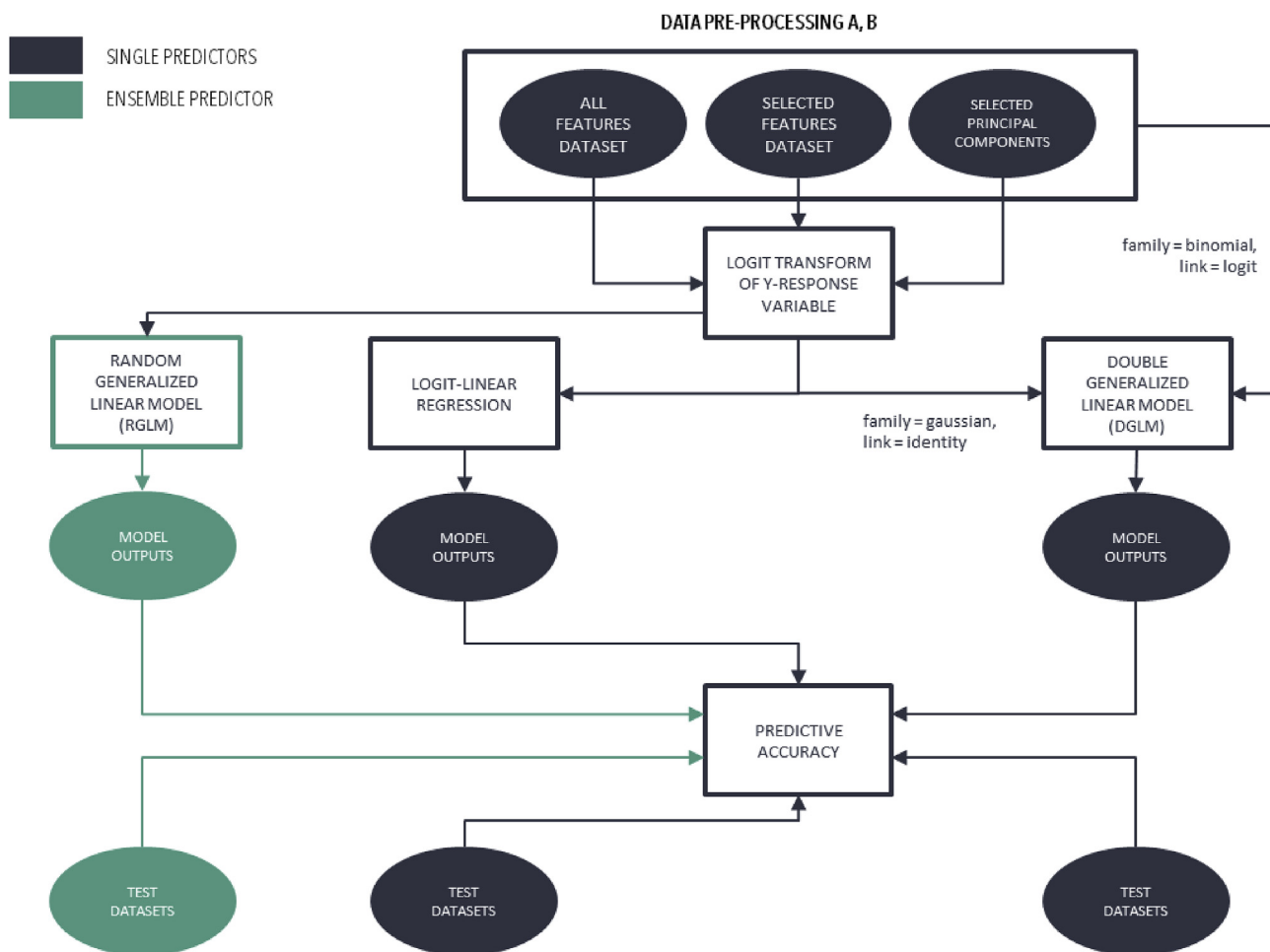


Fig. 5. Model building and assessment workflow.

In this study, all three of the aforementioned metrics were applied to retain an optimal number of components. Variable importance could then be identified and more easily interpreted by rotating the PC loadings matrix. The portions of the scores matrix corresponding to PCs that were retained were used as alternative explanatory variables to build models (Figure D - left; Supplementary material).

2.2.4. Feature standardization

Standard PCA is an unsupervised approach, since the y-response variable is not directly considered; PCs that capture the maximum variance in the feature data set are assumed to explain most of the variance in the y-response (James et al., 2013). Results are only as reliable as the quality of data introduced to a model. PCA results are sensitive to input feature pre-selection based on significance and if features are standardized. Standardizing feature variables prior to further analyses with PCA address concerns about the interpretability and credibility of results. Model inputs are often collected at different scales or measured in different units. The effective variance of each feature (e.g. the variance of a count of occurrences per 100,000 buildings will likely be greater than a measurement in centimeters expressed as meters) cannot be accurately compared. Consequently, PCA results would erroneously assign very high loadings to features with the highest unscaled variances.

Typically, X-aware standardization, which involves centering each feature variable on its mean and dividing by the standard deviation, is applied to feature variables prior to conducting PCA. Without standardization, variables with high variance would be associated with larger resultant loadings. This would erroneously lead to the

dependence of a PC on variables with high, unscaled variances. Consequently, while the actual values of the predictors are modified in the standardization process, the loss in interpretability of the features is counterbalanced by the increased interpretability of resultant model coefficients as changes from low to high values (Gelman, 2008).

However, since the y-response variable is not directly related to the variance of the feature variables, it is possible that PCs with low eigenvalues, constructed on features with low variance but high explanatory power with respect to y, are excluded from further analyses. In these cases, the underlying assumption that high variance in feature variables explains the most variance in the y-response is invalid. Exclusion of PCs with low eigenvalues from further analyses can be problematic if the y-response variable has a close relationship with these components; this adversely impacts the predictive accuracy of the associated model. To address this concern, y-aware standardization has been recommended by Zume & Mount (2016) and was applied to the three test sets of data prepared in the previous step. The standardization of feature variables to the y-response requires a model that reflects the nature of the data. The choice of standardization model will impact subsequent PCA results. However, the magnitude of this impact can only be assessed when reviewing the results at the end of the modeling stage. A logit model, rather than the standard linear model, is more appropriate to transform each column of feature data values with respect to the bounded y-response variable. Assessments with y-logit and standard linear transformations were conducted to support the evaluation of this pre-processing step.

Performing y-aware standardization at this stage of the workflow with all available observations (e.g. 81) means that the entire set of

response variables will be used. Consequently, prior to conducting this step, 21 observations are retained as a test set. Subsetting the available data at this point ensures that information from the test set is not used to train the models. Furthermore, y-logit standardization was only applied to datasets used to fit two of the single predictor models. These include the logit-linear regression and double generalized linear models, which are introduced in the next section (2.3).

At the end of this step, x- and y-aware standardized training datasets were prepared for PCA (Table C – first three columns on the left; Supplementary materials). The x-aware datasets established the baseline or null hypothesis that y-aware standardization has no impact on results. The y-aware datasets supported the investigation of the alternative hypothesis. When y-aware standardization is applied, binary features take on numeric values in the process.

Prior to performing PCA, training datasets may contain feature variables with zero variance (i.e. constant values) and should be excluded from further analyses. This is pertinent to datasets with low observation records or rarely occurring levels. As a result, the full range of levels per feature may be under-represented, even with stratified random sampling to generate training and test datasets.

2.3. Statistical models

Considering past developments (e.g. studies presented in Table 1), logit-linear models (i.e. linear regression applied to datasets with logit-transformed dependent variable) and GLMs (e.g. double and random generalized linear models) were built and their performance in terms of predictive accuracy compared in this study. A workflow describing the evaluation of selected models of interest is illustrated in Fig. 5, which continues with the outputs prepared by the end of the pre-processing steps (Fig. 2). Table 4 summarizes the characteristics of the models that were applied to specific datasets and the corresponding R packages.

2.4. Model building

Four types of models were built in this study: with original feature variables, with pre-selected features based on significance pruning, with principle components (PCs) and the aforementioned models with additional consideration of interaction terms. Ettinger et al. (2016) reported that while preliminary analysis of second-order (pairwise) interactions between features did not identify significant results, interdependencies should be considered in future investigations. Moreover, the effects of interactions between principal components on predictive accuracy have yet to be examined in past developments of vulnerability functions. The inclusion of interaction terms from PCs can be interpreted as a way to evaluate a more comprehensive set of interactions among weighted linear combinations of feature variables, rather than pairwise interactions of original variables. In this study, series of logit-linear regressions were built with original variables (LLR1 and LLR5), PCs (LLR3, LLR4, LLR7 and LLR8), original variables with pairwise interaction terms between original variables (LLR2 and LLR6) and PCs with interactions between PCs (LLR4C and LLR8C). Two sets of DGLMs were built to model expected damages as Gaussian (DGLM_G0 to G5) and binomial distributions (DGLM_B0 and B1). Additionally, two RGLMs were built (RGLM0 and RGLM1). Altogether, 23

models were built and evaluated (Table 5).

2.5. Model diagnostics

Once the different models were fitted on training data, the model objects were passed on to the test data. Diagnostics applied to modelled results provide a basis for comparison based on predictive accuracy; the model with the highest predictive accuracy was then chosen. In this study, three metrics were calculated to support this assessment. The first metric, AICc, assesses model fit against added complexity in terms of the number of features included in a model. It is an adaptation that is recognized to return more accurate results than AIC (Akaike, 1974) when modelling with small sample sizes. Both AIC and AICc depends on the goodness-of-fit (likelihood function) and considers an extra penalty term that prevents selecting overfitted models with too many parameters. In this way, the resultant AICc or AIC represents a compromise between model fit and complexity.

Both root mean squared error (RMSE) and mean absolute error (MAE) are metrics that evaluate predictive accuracy (James et al., 2013). Since errors are squared prior to averaging in the RMSE calculation, it can be used to detect the presence of large errors, whereas errors in the MAE calculation are averaged. If both RMSE and MAE scores calculated for models of interest are relatively lower than the baseline model, there is an improvement in the predictive power of the alternative models. The model with the highest predictive accuracy is associated with the lowest RMSE and MAE scores.

Model selection in this study was based on the highest relative predictive accuracy. Significant or important features were then identified with the selected model. In the case where PCs were used as alternative explanatory variables in the models, important features were identified by ranking the absolute value of variable loadings associated with the first PC (Figure D - right; Supplementary material).

3. Results

3.1. Model selection

Table 6 summarizes the ranked scores of the three model diagnostic metrics calculated for each of the 23 models that were built. Based on these results, we observed that models built with all available feature variables all failed to converge (i.e. indicated by the -Inf or NA values under AICc) due to the high number of explanatory variables with respect to the number of observations.

The AICc metric favoured a combination of linear regression and GLMs based on model fit against added complexity, whereas RMSE and MAE metrics favoured GLMs only based on predictive accuracy. Since the objective of the study is to optimize the latter, the models with the highest predictive accuracy were identified to be DGLM_G4 or DGLM_G5, and RGLM0, as single and ensemble predictors, respectively. Due to the added complexity of the RGLM approach, the DGLM_G4 or G5 models were considered to be the best performing models. Both of the DGLM_G4 and G5 models were built with reduced dimensionality, more specifically, with a subset of the first four PCs as alternative explanatory variables (Table 5). Additionally, the G5 model accounts for overdispersion by considering the two hazard types.

Table 4

Overview of selected models of interest – logit-linear regression (R Core Team, 2018), DGLM (Corty, 2018; Smyth, 1989) and RGLM (Langfelder, 2018; Song et al., 2013).

model	type of predictor	R package::function	input data		
			pre	post	all
logit-linear regression	single	stats::lm()	✓	✓	✓
double generalized linear model (DGLM)	single	dglm::dglm()	✓	✓	✓
random generalized linear model (RGLM)	ensemble	randomGLM::randomGLM()			✓

Table 5
Combinations of model inputs and models that were built and evaluated in this study; for models built with PCs, the number of PCs retained as alternative explanatory variables correspond to the results presented in Table C

model	n	p	pre-selected features	original feature variables	principal components	interaction terms	dispersion
LLRBL	60	0	N	N	N	N	N
LLR0	60	87	N	Y	N	N	N
LLR1	60	20	Y	Y	N	N	N
LLR2	60	20	Y	Y	N	Y	N
LLR3	60	20	Y	N	Y (PC1-5)	N	N
LLR4	60	20	Y	N	Y (PC1,3,4)	N	N
LLR4B	60	20	Y	N	Y (PC1,3)	Y	N
LLR4C	60	20	Y	N	Y (PC1,3)	Y (PCs)	N
LLR5	60	55	Y	Y	N	N	N
LLR6	60	55	Y	Y	N	Y	N
LLR7	60	55	Y	N	Y (PC1-4)	N	N
LLR8	60	55	Y	N	Y (PC1-3)	N	N
LLR8B	60	55	Y	N	Y(PC1-3)	Y	N
LLR8C	60	55	Y	N	Y(PC1-3)	Y (PCs)	N
DGLM_G0	60	22	Y	Y	N	N	N
DGLM_G1	60	22	Y	N	Y(PC1-5)	N	N
DGLM_G2	59	55	Y	Y	N	N	N
DGLM_G4	60	55	Y	N	Y(PC1-4)	N	N
DGLM_G5	60	55	Y	N	Y(PC1-4)	N	Y
DGLM_B0	60	24	Y	Y	N	N	N
DGLM_B1	60	55	Y	Y	N	N	Y
RGLM0	60	86	N	Y	N	N	N
RGLM1	60	86	N	Y	N	Y	N

Table 6
Ranked relative performance of models, from highest to lowest, based on a conventional model selection metric (AICc) and predictive accuracy metrics (RMSE, MAE). The single predictor (DGLM_G4) and ensemble predictor (RGLM0) models that performed best based on the defined objective function of predictive accuracy in this study are indicated in bold.

model fit and complexity		predictive accuracy			
AICc		RMSE		MAE	
LLR4B	-8242.39	RGLM0	0.11	DGLM_G4	0.08
LLR2	-655.69	DGLM_G4	0.11	RGLM0	0.08
DGLM_G5	-109.22	RGLM1	0.16	RGLM1	0.12
DGLM_G4	-81.75	DGLM_G0	0.24	DGLM_G5	0.15
LLR8C	-38.81	DGLM_G5	0.27	DGLM_G0	0.19
LLR7	-32.43	DGLM_G1	0.40	DGLM_G1	0.31
LLR8	-27.24	LLR5	1.12	LLR5	0.72
DGLM_G1	-18.24	LLR7	1.19	LLR4	0.95
LLR4C	35.37	LLR8	1.22	LLR7	0.96
DGLM_G0	36.57	LLR4C	1.36	LLR8	1.00
LLR3	37.68	LLR4	1.37	LLR4C	1.05
LLR4	37.88	LLR3	1.50	LLR3	1.07
LLR1	89.22	LLR1	1.54	LLR8C	1.09
DGLM_B1	887.59	LLRBL	1.88	LLR1	1.16
DGLM_B0	889.39	LLR8C	1.93	LLRBL	1.40
LLR5	1492.15	LLR4B	2.05	LLR4B	1.63
LLR0	-Inf	LLR8B	2.05	LLR8B	1.63
LLR6	-Inf	LLR0	2.60	LLR0	2.00
LLR8B	-Inf	DGLM_B1	5.49	DGLM_B1	5.48
LLRBL	NA	DGLM_B0	5.68	DGLM_B0	5.67
DGLM_G2	NA	LLR6	53.41	LLR6	32.09
		LLR2	78.40	LLR2	33.03
		DGLM_G2	NA	DGLM_G2	NA

While the inclusion of interaction terms appears to improve the model fit, improvements to predictive accuracy could not be conclusively demonstrated in this study with the working datasets. Finally, modelling expected damages as a binomial distribution resulted in lower predictive accuracy than modelling proportional loss as a Gaussian distribution.

3.2. Feature importance

Features that significantly contributed to models with the highest

relative predictive accuracy (either DGLM_G4 or G5) were identified from the ranked absolute value of variable loadings associated with the first PC and reported in Table 7. Based on the evaluation of the models with the available dataset, the two features were identified to contribute the most to explaining proportional loss (i.e. with higher variable loadings). These include the highest level of a target building (e.g. upper floors to ground floor or basement) that was affected by water and/or sediment intrusion and the points of weakness(es) that materials entered into the structure (e.g. windows, doors, walls). To a lesser extent, building resistance and surrounding area profile features helped to explain residual variability in proportional loss when included in the model. The features identified as important contributors included the insurance value, the type of wall material it was constructed with and the process pathway. The latter included data on whether the area around the building is open or if the building is located beside preferential conduits and the number of neighbouring buildings sheltering hazard process materials from the building in question.

4. Discussion

The study indicates that several challenges exist for researchers and practitioners in risk management communities, especially with regards to identifying best practices to analyse data and to deduce vulnerability functions. In particular, 1) critical decisions often need to be made in this domain with the most prevalent information available and 2) new data is continually being collected. The utility of the workflow (Figs. 2 and 5) was demonstrated on a dataset that was compiled from post-ex field assessments. Doing so highlighted the challenges of working with real data, while showing what kinds of insights can be derived from findings and how they can be interpreted. The reliability, and therefore transferability, of the chosen vulnerability function is dependent on multiple factors, including the quality and quantity of empirical data used to derive it, the statistical approach applied to the data and the manner in which damages were appraised (Papathoma-Köhle et al., 2017). The following sections discuss insights from the modelling results using the prescribed workflow.

Table 7

Ranked summary of features that contributed most to explaining the variance in proportional loss associated with the model with highest relative predictive accuracy, DGLM_G4.

feature variables	categories	variable loadings
level of building affected by intrusion: 2. OG	damage patterns	0.835
pathway(s) of sediment intrusion: throughout building	damage patterns	0.480
total destruction (Y/N)	damage patterns	0.177
insurance value	building resistance	0.080
damage due to boulder (> 1 m) intrusion or large woody debris (Y/N)	damage patterns	0.073
building frame shifted (Y/N)	damage patterns	0.070
wall materials: masonry	building resistance	0.070
process pathway: street/preferential pathway	surrounding area profile	0.060
process pathway: open	surrounding area profile	0.054
number of neighbouring buildings	surrounding area profile	0.040
hazard type: sediment-laden flood	damage patterns	0.036
hazard type: debris flow	damage patterns	0.036
average sediment deposition height	damage patterns	0.034
distance to channel	surrounding area profile	0.032
estimated volumes of sediment inside of building: none	damage patterns	0.032
level of building affected by intrusion: EG	damage patterns	0.031
pathway(s) of sediment intrusion: through windows or doors	damage patterns	0.023
sediment in building interior (Y/N)	damage patterns	0.022
local protection measures: vegetation	surrounding area profile	0.021
damage claim	damage patterns	0.020
pressure damage to openings from impact of process materials (Y/N)	damage patterns	0.018

4.1. Critical discussion of data pre-processing

Results from the exploratory analyses highlighted the high dimensionality (i.e. a number of feature variables is greater than the number of observations) and sparsity (i.e. data contains a significant number of missing entries) of the input dataset. The problem of dimension reduction with PCA on this type of data has been found to share comparable characteristics typically associated with non-linear models, particularly the challenges of overfitting, inadequate locally optimal solutions and inefficient execution of traditional PCA algorithms (Ilin and Raiko, 2010). In particular, an analytical solution cannot be reached when a data covariance matrix is non-trivial to estimate. Furthermore, the objective function contains multiple local minima. This is in contrast to classic PCA, where solutions return a single global minimum and it is difficult to verify if the output of an optimization problem with missing data is the true solution. Therefore, matrix regularization steps are imperative to prevent overfitting a model if the data is intended for further analyses.

It is highly recommended to collect a larger number of actual observations when possible. However, this may not be viable, especially in natural hazard studies, and listwise deletion of partially incomplete observation records would result in too much data loss and emphasize biases in the remaining data. Matrix completion methods are a type of applied regularization to address the problem of missing data. In this study, selected data imputation methods were assessed to demonstrate how cases of missing data can be treated before further analyses are conducted. These methods included mean-based imputation, k-nearest neighbour and MIPCA. The result of this pre-processing step is a hybrid dataset with complete observation records, which is based on the combination of survey or observed data and statistically imputed data. As such, it reflects the realities in the three study sites to a limited extent and any conclusions using this data should be drawn with caution since many of the records no longer link back to actual buildings. Consequently, there are implications on the subsequent analysis of results based on the use of hybrid data. It is important to open this discussion to provide proper guidance to researchers and practitioners when specifying an imputation model, predicated on the risk of introducing estimates that do not accurately reflect the nature of the missing data.

While the data sparsity problem can be addressed with imputation, the challenge of high dimensionality persists. For datasets with fewer observation records than measured feature variables (i.e. $p > n$), PCA

overfits to noise and is an inconsistent estimator of the subspace of maximal variance. This means that the estimator fails to converge in probability to the true value of the parameter of interest. This type of problem also requires regularization, which involves including additional information to reach a viable solution. It is imperative to resolve this problem before results can be used to inform decisions with acceptable levels of confidence. Two solutions may be considered. The first involves the collection of additional observations so that $n > p$. Consequently, an adequate number of observations can support the differentiation of signal from noise. The second solution is predicated on an underlying assumption that the available data is well represented in a sparse basis (Johnstone and Lu, 2009). This approach involves reducing the dimensionality of the dataset prior to applying PCA-based methods. In particular, a simple asymptotic model was proposed in the study to verify the consistency of the main PC identified with standard PCA, if and only if $p(n)/n \rightarrow 0$ (Johnstone and Lu, 2009). Furthermore, it has been demonstrated that if PCA is conducted on a selected subset of coordinates that represent the largest sample variances, then consistency can be recovered, even if $p(n) \gg n$.

Although the model proposed by Johnstone and Lu (2009) was not applied in this particular study, the idea of dimension reduction was achieved by significance pruning via bivariate analyses, which resulted in two subsets of feature variables (i.e. pre and post) that are highly correlated to the y-response. However, it is important to note that significance detection is dependent on sample size. Therefore, analyses conducted with the current dataset can only provide general guidance about features of interest. The evaluation should be rerun with the acquisition of complete observations without imputation and with a sufficient number of observations per feature level (Ettinger et al., 2016). From the results, it was evident that the number of overall dimensions to explain at least 75% of the cumulative variance in the y-response was significantly reduced from using original feature variables with 1) pre-selected features and 2) using PCs identified from PCA results, where the input is features that have been standardized to the y-response with an adapted logit transformation. Consequently, dimension reduction and y-aware feature standardization prior to conducting PCA are recommended as data pre-processing steps.

It was observed that while dimensions were consistently reduced from the number of feature variables prior to pre-processing, the recommended number of PCs to retain varied (Table C; Supplementary materials). An additional metric that may be considered involves the exclusion of all components below a threshold that is defined on the

level of noise in a given dataset (Gavish and Donoho, 2014). A combination of visual heuristics and information that can be readily obtained from the PCA model object were used in this study.

The use of the outputs from PCA is twofold: firstly, they can provide insight about relative feature importance to explain variance in the y-response variable. Secondly, the retained PCs can be used as alternative explanatory variables to build models for further analyses. In summary, PCA is a useful method that can handle instances of multicollinearity and facilitate data dimension reduction. However, the results are only considered to be consistent if the underlying problem involving sparse and wide input datasets is regularized.

4.2. Predictive accuracy of proportional loss

Overall, of all the models that were built and evaluated, the GLMs were associated with higher predictive accuracy scores, compared with the linear regression models. Based on the results, prediction of the expected value with a DGLM with a Gaussian family distribution performed better than formulating the problem with a binomial distribution. Definition of hazard types (i.e. debris flows and sediment-laden floods) in the optional overdispersion sub-model resulted in comparative predictive accuracy to the model without. However, some improvement in the model fit against added complexity could be observed (i.e. reduction in AICc score associated with DGLM_G5 compared with G4). It may be of interest to consider other sources in the future; for instance, Rheinberger et al. (2013) indicated that overdispersion detected across building levels is possible, especially for residences. DGLMs built with reduced dimensions (i.e. preselected variables based on significance pruning and subsequently retained PCs) returned the highest predictive accuracy and model fit among all of the models that were evaluated. RGLMs were also associated with high predictive accuracy and warrant further investigation.

In general, models built with PCs and interaction between PCs resulted in higher predictive accuracy than models built with the same subsets of feature variables without interactions. Second-order interaction effects between original variables may have performed relatively worse given a low number of observation records. The results may be associated with overfitting when the additional pairwise interaction terms are added to the model.

In general, higher predictive accuracy was returned with models built from predictors identified via information gained from dimension reduction. This finding suggests that information from certain pre-processing steps is helpful and joint application of variable selection with pre-processing and conventional model selection approaches can further improve the predictive accuracy of GLMs. These observations should be investigated further with a larger set of observation records to determine if results are consistent across other comparable datasets.

In terms of model selection based on model fit or predictive accuracy, conventional model selection approaches based on the identification of the most parsimonious model (i.e. AICc scores that identify solutions predicated on model fit while minimizing complexity). This may not necessarily identify models with the highest predictive accuracy. The application of AICc and other similar metrics may be limited to inferential or exploratory analyses rather than predictive (Leek and Peng, 2015), such that selecting for the most parsimonious model can be inconsistent with the objective of maximizing predictive accuracy. This can account for the discrepancy between the combination of linear regression and GLMs selected for based on model fit alone and the selection of only GLMs based on predictive accuracy; the observation is consistent with observations reported in Li et al. (2017). This may also have implications for RGLM results, since some form of AIC is used in the process to evaluate model fit; further investigation is recommended.

Given the critical review of existing challenges associated with the input data used to derive the multivariate vulnerability functions, the models should be accepted and used to support decision making with caution. It is highly recommended that more data will be collected in

the future to support the continued derivation and evaluation of these functions with greater reliability before transferability to other considered scenarios. Additional points of discussion are elaborated on in section 4.4.

4.3. Feature importance

Predictive models using proxy predictors may support the identification of causal variables, which can provide guidance in future data collection and investigative efforts (Li et al., 2017). Models built on feature variables that are proxies are often referred to as black boxes. This is because there is a recognized limitation of proxies to directly inform how the dependent variable is related to causal variables or drivers. Consequently, application of information gained from fitted models, such as the results from this study, may be limited to providing an indication of importance when explaining expected damages and defining the scope of future investigations. These results can be used to recommend certain aspects to focus on, especially when resources are limited.

While it has been recognized in past studies (see Introduction) that hazard intensity proxies are strongly correlated to expected damages, the inclusion of pre-condition features describing both the building resistance and surrounding area may help to explain residual variability in the data. In general, the features identified as important in this study are consistent with the aforementioned features identified in past studies.

4.4. Applications and challenges

Insights from vulnerability functions fitted with available data communicate theoretical possibilities predicated on what was observed and collected in the past. There is also the possibility that observed values are not fully representative of reality. Furthermore, while predictive accuracy can inform about the past, there are limitations when applying the predictions to future scenarios or conditions at another location in the present with the same degree of accuracy. Transferability is extended if the model can be continuously updated to correct for past errors and if new observations, including site- and hazard-specific data, can be used to retrain the model. This recommendation takes into consideration findings reported by Charvet et al. (2017), which showed that developed fragility functions cannot typically be generalized or applied to comparable buildings in different geographic locations. However, it should be noted that the occurrence of unpredictable extreme events is always possible. Models are unable to anticipate them and the magnitude of associated errors is unknown. For this reason, direct application of the fitted model with highest predictive accuracy from this study should not be directly applied to future scenarios without further investigation.

Furthermore, structural variability among buildings means that hazard processes will have differential consequences, which also has implications for model transferability to different locations where building construction varies. In several studies with sufficiently large numbers of observations describing a range of building structure differences, separate vulnerability functions were generated for each (e.g. Charvet et al., 2015). From an engineering perspective, the primary question is whether the structural integrity of a building has been compromised. Consequently, there should be an emphasis on resistance-based investigations conducted by experts with access to building plans to determine the amount of pressure the structure can sustain for a given natural hazard. This perspective may be more actively integrated in future field assessments and survey design. In particular, questions need to be pertinent to the deformation or movement of building structures (e.g. shifting or destruction of building frames), whereas questions about water or flood proofing or about the impact of pressure on building elements can only reasonably provide partial impressions. Comprehensive assessment about building strength against

impact pressures and high velocities should be conducted by engineers; it is generally not reasonable to expect that affected residents be able to provide this type of information.

Nevertheless, the prescribed workflow and evaluation metrics highlighted in this study can be a useful starting point. Emphasis is placed on the ability to update inputs and components of the workflow and to continually remodel and to reassess vulnerability functions with respect to the nature of the input data. The treatment of wide and sparse datasets that are common in this field was demonstrated with the pre-processing part of the workflow. A range of multivariate models were built in a stepwise manner, so that the relative importance of feature contributions and interaction effects to explaining the variance observed in the *y*-response could be demonstrated. The workflow also includes selected metrics to quantitatively assess model fit against complexity and predictive accuracy. Furthermore, the study provides guidance on how to obtain information on relative feature importance from model objects. Pertinent issues associated with each of the factors contributing to the reliability of derived vulnerability functions are unlikely to be resolved at once but rather iteratively with time and upon data availability. Consequently, the results from this study are descriptive rather than prescriptive.

5. Conclusions and future work

In summary, the workflow can be used to assess the potential of statistical models to predict proportional loss associated with buildings affected by natural hazards. The findings can be used as guidance to collect additional data in the future to maximize information gain about pre-hazard event conditions that contribute to losses. In such a way, even if the exact nature of hazards to come cannot be known with full certainty, it may still be possible to minimize vulnerability to these hazards by reducing building susceptibility and exposure. A number of recommendations stem from the different components and lines of investigations that an end-to-end workflow involves. Questions arise from outstanding challenges and curiosity along the way. This section provides a summary of ways the continued development and assessment of vulnerability functions may be improved.

5.1. Data quality

The quality of data that is defined as inputs affects the quality of the resultant vulnerability function. Firstly, the nature of missing data must be accurately identified and treated prior to further analyses. In this study, three data imputation techniques were applied as a first assessment; the results demonstrated that there can be stronger agreement between original and imputed data (Figure B; Supplementary material). Additionally, performing imputations under MCAR/MAR assumptions raises valid questions about the impact of effectively assuming random missingness for a data set with mixed missing data patterns. Due to the nature of MNAR data, the missing data has a different distribution than the observed data. Since missing MNAR data values can only be estimated from available information, it is important to emphasize that bias is introduced to the predicted values (Munguía and Armando, 2014). This should be weighed against the degree of bias and limited prediction power that would otherwise be associated with the use of a lower number of complete observations in subsequent modeling stages. Future work should involve more extensive investigations into appropriate methods to handle different types of non-observed responses, including numeric (continuous), categorical (ordinal, nominal) and hybrid data (Munguía and Armando, 2014). Furthermore, Lazer et al. (2016) described the development of more advanced diagnostic tools that would be capable of categorizing instances of missing data based on both the mechanisms of generation and at varying resolutions or on multiple levels. Future work in matrix regularization may apply an alternative approach based on a probabilistic formulation of PCA. Ilin and Raiko (2010) proposed a computationally efficient algorithm that is an

extension of variational Bayesian learning (VB). In particular, the study demonstrated the effects of regularization and the modelling of posterior variance. The availability of such tools and approaches would be instrumental in exploring hybrid solutions to address the different underlying natures of missing cases and more accurately capture the properties of real distributions of interest with data imputation. As a result, hybrid datasets could represent reality more closely, especially given the challenges of acquiring complete observation records in the field of natural hazards.

Furthermore, data aggregation should generally be avoided (Charvet et al., 2017). Since the number of observation records is often limited, it may be tempting to collate and analyze data from multiple sources as a single dataset (e.g. from different events or hazard types). If aggregated data is used, care needs to be taken to select for an appropriate model structure. In this study, the aggregation of 81 observations from three hazard events were fitted with logit-linear and GLMs was found to be a good starting point. Further investigation with generalized linear mixed models have been recommended, where a random intercept is introduced for each subgroup within the dataset to explicitly account for the subgroup as an explanatory variable (Rossetto et al., 2014). It may also be possible to account for the effects of combining subgroups of data by modelling for overdispersion in the DGLM.

A dataset with a limited number of observation records may result in overfitted models. In general, a minimum number of observations must be used to generate reliable results and the number required to develop a vulnerability function varies with the degree of uncertainty that users of the function are willing to accept. For example, Laudan et al. (2017) expressed that the 94 observations that their study was based on was considered to be small and results have low transferability and should not be generalized. Furthermore, the issue of unbalanced datasets is also prevalent, especially given low numbers of observations. This is when there is an over or underrepresentation of certain types of building feature combinations and/or buildings that sustained a certain amount of damage. Consequently, a representative dataset should have a minimum number of observations that represent a relatively equal range all possible unique feature combinations.

The call for more comprehensive and systematic data collection is imperative to support the ongoing verification of modelled results and to be able to apply associated findings to inform risk management strategies with greater confidence. In particular, objective criteria to accurately document building resistance should be defined in collaboration with experts in the building engineering domain and consistently applied to acquire data in the future. This would effectively minimize bias that may otherwise be introduced due to variations in individual understanding and interpretations and contribute to aggregated uncertainties and errors (Laudan et al., 2017). In terms of accounting for potentially important damage driving features, variables identified to be strongly correlated to proportional loss or important with respect to relative variable loadings should be considered in subsequent data acquisition campaigns (Laudan et al., 2017).

5.2. Models and model selection

Model structures of interest should be chosen with respect to the nature of the expected value and whether underlying assumptions about their distribution are satisfied. The comparison of predictive accuracy between linear regressions and GLMs in this study with non-normally distributed proportional loss demonstrated the importance of choosing an appropriate model structure. In this example, the lower predictive power is the result of linear regressions fitted with data that violated the assumptions of normality.

Joint applications (i.e. model building based on information learned from pre-processing) was found to be advantageous, particularly steps prescribed to reduce high dimensional datasets prior to further analyses. Furthermore the evaluation of models based on predictive

accuracy metrics, rather than conventional diagnostics based on the assessment of model fit is recommended for selecting predictive models (Li et al., 2017).

The workflow described in this study is based on a multi-step procedure that accounts for dimension reduction prior to model fitting and these models were found to be associated with greater predictive accuracy. More advanced modelling techniques, such as the sparse principal component regression for generalized linear models (SPCR-glm), may be of interest. It is comprised of a basic loss function that is based on a combination of the regression squared loss and PCA loss (Kawano et al., 2016). By considering both loss functions simultaneously in a single-stage procedure, sparse PC loadings that are directly related to a response variable are identified. This effectively streamlines the main challenges addressed with the workflow and may be of interest to investigate further. A sensitivity analysis may be conducted to determine the most optimal turning parameter values with predictive accuracy as the target objective.

5.3. Sources of uncertainty associated with results

The study did not explicitly examine sources of uncertainty but acknowledges that such an assessment should be conducted and reported with results. Uncertainty contains information beyond that which is contained in a single prediction and failing to communicate this can result in adverse consequences. Charvet et al. (2017) recommended that uncertainty in both the explanatory and response variables should be quantified. Merz et al. (2013) emphasized that contributions may be attributed to data sparsity and generally limited understanding of damaging processes, among many other sources of uncertainty at the interface of natural and built environments. The introduction of aggregated uncertainty that is invariably introduced in quantitative evaluations can potentially be significant (Vogel et al., 2014) and, therefore, should be clearly communicated with results. From a practical perspective, the inclusion of uncertainty information with damage analyses can serve as an instrumental way to discuss the cost-benefits of investing in particular risk management strategies and the consequences of insufficient preparedness.

This study described an end-to-end workflow that can provide guidance on the development, evaluation and interpretation of empirically-based, multivariate physical vulnerability functions for buildings affected by hazardous processes. The workflow was complimented with a review of outstanding challenges and potential solutions. The final section highlighted recommendations and new lines of investigation that may be of interest to researchers and practitioners in risk management. The recommendations stemming from this work can serve as a basis upon which critical and continuous review of vulnerability functions is possible. Furthermore, as new data becomes available, first insights gained about drivers of damage and modelling techniques can be applied to build models that may better capture their relationships to loss.

Declaring funding sources

This study was funded by the Swiss National Science Foundation [SNSF number 159,899] and supported by the Czech Ministry of Education, Youth and Sports within the National Sustainability Programme I, project of Transport R&D Centre [LO1610], on the research infrastructure acquired from the Operation Programme Research and Development for Innovations [CZ.1.05/2.1.00/03.0064].

Acknowledgements

The authors kindly thank three anonymous reviewers for their comments and suggestions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jenvman.2019.05.084>.

References

- Aguilera, A., Escabias, M., Valderrama, M., 2006. Using principal components for estimating logistic regression with high-dimensional multicollinear data. *Comput. Stat. Data Anal.* 50 (8), 1905–1924. <http://doi.org/10.1016/j.csda.2005.03.011>.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19 (6), 716–723. <http://doi.org/10.1109/TAC.1974.1100705>.
- Bezzolo, G., Hegg, C., 2008. Ereignisanalyse Hochwasser 2005, Teil 2 - Analyse von Prozessen, Massnahmen und Gefahrengrundlagen.
- Büchle, B., Kreibich, H., Kron, A., Thielen, A., Ihringer, J., Oberle, P., Merz, B., Nestmann, F., 2006. Flood-risk mapping: contributions towards an enhanced assessment of extreme events and associated risks. *Nat. Hazards Earth Syst. Sci.* 6 (4), 485–503. <http://doi.org/10.5194/nhess-6-485-2006>.
- Bundesamt für Umwelt, 2018. Schäden und Lehren aus Naturereignissen.
- Charvet, I., Suppasri, A., Imamura, F., 2014. Empirical fragility analysis of building damage caused by the 2011 Great East Japan tsunami in Ishinomaki city using ordinal regression, and influence of key geographical features. *Stoch. Environ. Res. Risk Assess.* 28 (7), 1853–1867. <http://doi.org/10.1007/s00477-014-0850-2>.
- Charvet, I., Suppasri, A., Kimura, H., Sugawara, D., Imamura, F., 2015. A multivariate generalized linear tsunami fragility model for Kesenuma City based on maximum flow depths, velocities and debris impact, with evaluation of predictive accuracy. *Nat. Hazards* 79 (3), 2073–2099. <http://doi.org/10.1007/s11069-015-1947-8>.
- Charvet, I., Macabuag, J., Rossetto, T., 2017. Estimating Tsunami-Induced Building Damage through Fragility Functions: Critical Review and Research Needs. *Frontiers in Built Environment*. <http://doi.org/10.3389/fbuil.2017.00036>.
- Choi, E., DesRoches, R., Nielson, B., 2004. Seismic fragility of typical bridges in moderate seismic zones. *Eng. Struct.* 26 (2), 187–199. <http://doi.org/10.1016/j.engstruct.2003.09.006>.
- Chow, C., Ramirez, J., Keiler, M., 2018. Application of sensitivity analysis for process model calibration of natural hazards. *Geosciences* 8 (6), 218. <http://doi.org/10.3390/geosciences8060218>.
- Ciurean, R.L., Hussin, H., van Westen, C., Jaboyedoff, M., Nicolet, P., Chen, L., Frigerio, S., Glade, T., 2017. Multi-scale debris flow vulnerability assessment and direct loss estimation of buildings in the Eastern Italian Alps. *Nat. Hazards* 85 (2), 929–957. <http://doi.org/10.1007/s11069-016-2612-6>.
- Concato, J., Peduzzi, P., Holford, T., Feinstein, A., 1995. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. *J. Clin. Epidemiol.* 48, 1495–1501.
- Corty, R., 2018. Dglm. Retrieved from: <https://www.rdocumentation.org/packages/dglm/versions/1.8.3/topics/dglm>.
- Di Baldassarre, G., Montanari, A., 2009. Uncertainty in river discharge observations: a quantitative analysis. *Hydrol. Earth Syst. Sci.* 13 (6), 913–921. <http://doi.org/10.5194/hess-13-913-2009>.
- Ettlinger, S., Mounaud, L., Magill, C., Yao-Lafourcade, A., Thouret, J., Manville, V., Negulescu, C., Zuccaro, G., De Gregorio, D., Nardone, S., Uchuchoque, J., Arguedas, A., Macedo, L., Llerena, N., 2016. Building vulnerability to hydro-geomorphic hazards: estimating damage probability from qualitative vulnerability assessment using logistic regression. *J. Hydrol.* 541, 563–581. <http://doi.org/10.1016/j.jhydrol.2015.04.017>.
- Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbut, M., Bateman, A., Blaškovičová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., Sempere-Torres, D., Stancalie, G., Szolgay, J., Tسانis, I., Velasco, D., Viglione, A., 2009. A compilation of data on European flash floods. *J. Hydrol.* 367 (1–2), 70–78. <http://doi.org/10.1016/j.jhydrol.2008.12.028>.
- Gavish, M., Donoho, D., 2014. The Optimal Hard Threshold for Singular Values Is $4\sqrt{3}$. <http://doi.org/10.1103/PhysRevB.73.195113>.
- Gelman, A., 2008. Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* 27 (15), 2865–2873. <http://doi.org/10.1002/sim.3107>.
- Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *American Psychological Association* 24 (6), 417–441.
- Ilin, A., Raiko, T., 2010. Practical approaches to principal component analysis in the presence of missing values. *Jmlr* 11, 1957–2000. <http://doi.org/10.1109/TPAMI.2010.46>.
- James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Introduction to Statistical Learning. Springer Retrieved from: <https://www.springer.com/de/book/9781461471370>.
- Johnstone, I., Lu, A., 2009. On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* 104 (486), 682–693. <http://doi.org/10.1198/jasa.2009.0121>.
- Josse, J., Hussion, F., 2012. Handling missing values in exploratory multivariate data analysis methods. *J. Soc. Fr. Stat.* 153 (2), 79–99. Retrieved from: <http://journal-sfids.fr/index.php/J-Sfids/article/view/122>.
- Kawano, S., Fujisawa, H., Takada, T., Shiroishi, T., 2016. Sparse principal component regression for generalized linear models. *Comput. Stat. Data Anal.* 124, 180–196. <http://doi.org/10.1016/j.csda.2018.03.008>.
- Langfelder, P., 2018. randomGLM. Retrieved from: <https://www.rdocumentation.org/packages/randomGLM/versions/1.02-1/topics/randomGLM>.
- Laudan, J., Rözer, V., Sieg, T., Vogel, K., Thielen, A., 2017. Damage assessment in Braunsbach 2016: data collection and analysis for an improved understanding of

- damaging processes during flash floods. *Nat. Hazards Earth Syst. Sci.* 17 (12), 2163–2179. <http://doi.org/10.5194/nhess-17-2163-2017>.
- Lazar, C., Gatto, L., Ferro, M., Bruley, C., Burger, T., 2016. Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *J. Proteome Res.* 15 (4), 1116–1125. <http://doi.org/10.1021/acs.jproteome.5b00981>.
- Leek, J., Peng, R., 2015. What is the question? *Science* 347 (6228), 1314–1315. Retrieved from. <http://umli.dml.oclc.org/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=lxh&AN=24665041&site=ehost-live>.
- Li, J., Alvarez, B., Siwabessy, J., Tran, M., Huang, Z., Przeslawski, R., Radke, L., Howard, F., Nichol, S., 2017. Application of random forest and generalised linear model and their hybrid methods with geostatistical techniques to count data: predicting sponge species richness. *Environ. Model. Softw* 97, 112–129. <http://doi.org/10.1016/j.envsoft.2017.07.016>.
- Macabuag, J., Rossetto, T., Ioannou, I., Supparsi, A., Sugawara, D., Adriano, B., Imamura, F., Eames, I., Koshimura, S., 2016. A proposed methodology for deriving tsunami fragility functions for buildings using optimum intensity measures. *Nat. Hazards* 84 (2), 1257–1285. <http://doi.org/10.1007/s11069-016-2485-8>.
- Margreth, S., Romang, H., 2010. Effectiveness of mitigation measures against natural hazards. *Cold Reg. Sci. Technol.* 64 (2), 199–207. <http://doi.org/10.1016/j.coldregions.2010.04.013>.
- Mazzorana, B., Comiti, F., Scherer, C., Fuchs, S., 2012. Developing consistent scenarios to assess flood hazards in mountain streams. *J. Environ. Manag.* 94 (1), 112–124. <http://doi.org/10.1016/j.jenvman.2011.06.030>.
- Merz, B., Kreibich, H., Lall, U., 2013. Multi-variate flood damage assessment: a tree-based data-mining approach. *Nat. Hazards Earth Syst. Sci.* 13, 53–64. <http://doi.org/10.5194/nhess-13-53-2013>.
- Meyer, D., 2018. Impute. Retrieved from. <https://www.rdocumentation.org/packages/e1071/versions/1.7-0/topics/impute>.
- Munguía, T., Armando, J., 2014. Comparison of imputation methods for handling missing categorical data with univariate pattern. *Revista de Metodos Cuantitativos Para La Economía y La Empresa* 17 (1), 101–120.
- Papathoma-Köhle, M., 2016. Vulnerability curves vs. Vulnerability indicators: application of an indicator-based methodology for debris-flow hazards. *Nat. Hazards Earth Syst. Sci.* 16 (8), 1771–1790. <http://doi.org/10.5194/nhess-16-1771-2016>.
- Papathoma-Köhle, M., Kappes, M., Keiler, M., Glade, T., 2011. Physical vulnerability assessment for alpine hazards: state of the art and future needs. *Nat. Hazards* 58. <http://doi.org/10.1007/s11069-010-9632-4>.
- Papathoma-Köhle, M., Keiler, M., Totschnig, R., Glade, T., 2012. Improvement of vulnerability curves using data from extreme events: debris flow event in South Tyrol. *Nat. Hazards* 64 (3), 2083–2105. <http://doi.org/10.1007/s11069-012-0105-9>.
- Papathoma-Köhle, M., Gems, B., Sturm, M., Fuchs, S., 2017. Matrices, curves and indicators: a review of approaches to assess physical vulnerability to debris flows. *Earth Sci. Rev.* 171, 272–288. November 2016. <http://doi.org/10.1016/j.earscirev.2017.06.007>.
- Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11), 559–572.
- Peduzzi, P., Concato, J., Feinstein, A., Holford, T., 1995. Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J. Clin. Epidemiol.* 48, 1503–1510.
- Peduzzi, P., Concato, J., Kemper, E., Holford, T., Feinstein, A., 1996. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* 49, 1373–1379.
- Quan Luna, B., Blahut, J., Van Westen, C., Sterlacchini, S., Van Asch, T., Akbas, S., 2011. The application of numerical debris flow modelling for the generation of physical vulnerability curves. *Nat. Hazards Earth Syst. Sci.* 11 (7), 2047–2060. <http://doi.org/10.5194/nhess-11-2047-2011>.
- R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria URL. <https://www.R-project.org/>.
- Rheinberger, C., Romang, H., Bründl, M., 2013. Proportional loss functions for debris flow events. *Nat. Hazards Earth Syst. Sci.* 13 (8), 2147–2156. <http://doi.org/10.5194/nhess-13-2147-2013>.
- Rossetto, T., Ioannou, I., Grant, D., Maqsood, T., 2014. Guidelines for Empirical Vulnerability Assessment, vol. 108 GEM Foundation. <http://doi.org/10.13117/GEM.VULN-MOD.TR2014.11>.
- Scheidl, C., Rickenmann, D., Chiari, M., 2008. The use of airborne LiDAR data for the analysis of debris flow events in Switzerland. *Nat. Hazards Earth Syst. Sci.* 8 (5), 1113–1127. <http://doi.org/10.5194/nhess-8-1113-2008>.
- Smyth, G., 1989. Generalized linear models with varying dispersion. *Journal of the Royal Statistical Society Series B Methodological* 51 (1), 47–60.
- Song, L., Langfelder, P., Horvath, S., 2013. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinf.* 14. <http://doi.org/10.1186/1471-2105-14-5>.
- Swisstopo, 2005. swissALTI3D. Retrieved from. https://shop.swisstopo.admin.ch/en/products/height_models/alti3d.
- Tarbotton, C., Dall'Osso, F., Dominey-Howes, D., Goff, J., 2015. The use of empirical vulnerability functions to assess the response of buildings to tsunami impact: comparative review and summary of best practice. *Earth Sci. Rev.* 142, 120–134. <http://doi.org/10.1016/j.earscirev.2015.01.002>.
- Templ, M., Alfons, A., 2009. An application of VIM, the R package for visualization of missing values, to EU-SILC data. *Statistics* 1–10.
- Thieken, A., Cammerer, H., Dobler, C., Lammel, J., Schöberl, F., 2016. Estimating changes in flood risks and benefits of non-structural adaptation strategies - a case study from Tyrol, Austria. *Mitig. Adapt. Strategies Glob. Change* 21 (3), 343–376. <http://doi.org/10.1007/s11027-014-9602-3>.
- Toreti, A., Schneuwly-Bollscheiler, M., Stoffel, M., Luterbacher, J., 2013. Atmospheric forcing of debris flows in the southern Swiss Alps. *Journal of Applied Meteorology and Climatology* 52 (7), 1554–1560. <http://doi.org/10.1175/JAMC-D-13-077.1>.
- Totschnig, R., Sedlacek, W., Fuchs, S., 2011a. A quantitative vulnerability function for fluvial sediment transport. *Nat. Hazards* 58 (2), 681–703. <http://doi.org/10.1007/s11069-010-9623-5>.
- Totschnig, R., Sedlacek, W., Fuchs, S., 2011b. A quantitative vulnerability function for fluvial sediment transport. *Nat. Hazards* 58 (2), 681–703. <http://doi.org/10.1007/s11069-010-9623-5>.
- Uzielli, M., Nadim, F., Lacasse, S., Kaynia, A., 2008. A conceptual framework for quantitative estimation of physical vulnerability to landslides. *Eng. Geol.* 102 (3–4), 251–256. <http://doi.org/10.1016/j.enggeo.2008.03.011>.
- Vittinghoff, E., McCulloch, C., 2007. Relaxing the rule of ten events per variable in logistic and cox regression. *Am. J. Epidemiol.* 165 (6), 710–718. <http://doi.org/10.1093/aje/kwk052>.
- Vogel, K., Riggelsen, C., Korup, O., Scherbaum, F., 2014. Bayesian network learning for natural hazard analyses. *Nat. Hazards Earth Syst. Sci.* 14 (9), 2605–2626. <http://doi.org/10.5194/nhess-14-2605-2014>.
- Volosciuk, C., Maraun, D., Semenov, V., Tilinina, N., Gulev, S., Latif, M., 2016. Rising mediterranean sea surface temperatures amplify extreme summer precipitation in central Europe. *Sci. Rep.* 6 (August), 1–7. <http://doi.org/10.1038/srep32450>.
- Zumel, N., Mount, J., 2016. Vtreat: a data.Frame Processor for Predictive Modeling. Retrieved from. <http://arxiv.org/abs/1611.09477>.