GIMA
Geographical Information Management and Applications

ESTIMATING POST-EARTHQUAKE AID PRIORITY AREAS

EMPIRICAL DECISION SUPPORT FOR POST-EARTHQUAKE HUMANITAIRIAN AID DISTRIBUTION

MSc Geographical Information Management and Applications | UU WUR UT TUD

Student	Evelien Bulte
Email	e.c.c.bulte@students.uu.nl
Supervisor	Dr. Derek Karssenberg, UU
Professor	Prof. Dr. Steven de Jong, UU
Date	7 June 2017

Universiteit Utrecht
TU Delft
University of Twente.
Wageningen UR

510
PERSONALIZED HUMANITARIAN AID THROUGH BIG DATA

The Netherlands Red Cross

Preface

This thesis is the results of seven months of research in the context of the Geographical Information Management and Applications Master's program. During this period I have gained a lot of new skills and knowledge and got a look into the humanitarian sector.

The Netherlands Red Cross data initiative 510 hosted my research and helped me to define the research problem. I would like to thank Maarten van der Veen for giving me the opportunity to get to know the humanitarian aid sector through the team he initiated. His dedication to this cause motivated and at the same time challenged me to produce a useful output. I admire all 510 team members who voluntarily help to improve humanitarian operations by putting in their effort and knowledge. I want to thank them and other team members for discussing my research progress with me and for providing me with useful advices. Brenda Bastiaensen taught me a lot about data analysis and machine learning in the short period of time that we had. Without this knowledge I would not have been able to accomplish my objectives.

Finally, I would like to thank my supervisor Derek Karssenberg for his useful feedback on my thesis. Also the responsible professor and second reviewer helped to complete this research process.

Abstract

In the first days following a disaster, humanitarian decision makers often deal with a scarcity of information on the spatial aspects of the event's impact, and thus the need for humanitarian aid of the affected population. By learning from data of past events Priority Index Models (PIM's) can rapidly produce an estimate of a disaster's impact, which can help decision makers to identify aid priority areas. This enables empirically-based decision support, in contrast to the more subjective models that are currently used. The main objective of this study is to explore the usability of pre- and post-event open data to train a model to rapidly estimate post-earthquake aid neediness for any earthquake prone area on earth. As far as known, machine learning algorithms have not been applied before to predict aid priority areas after seismic hazards specifically. To achieve the research objective the Gorkha earthquake of 2015 in Nepal was used as a test case. Country- and hazard-specific open data related to this earthquake were used to predict aid-neediness. Damage to residential buildings was select as the most suitable aid-neediness indicating variable. Three different statistical models were fitted to the data: a multivariate linear regression model and two random forest regression models (one predicting completely damaged houses and the other predicting a combination of completely and partially damaged houses). 24 variables in four different categories (hazard, exposure, physical vulnerability and socio-economic vulnerability) were identified as predictors of post-earthquake structural damage. All three models could successfully produce an output on administrative level 4 (VDC) for the 16 most affected districts. Statistically, the random forest model predicting bot partially and completely damaged houses performed best with an R-squared of 0.63 on an independent test dataset. However, the random forest model predicting only completely damaged is favourable because the output is more intuitive and extendable. Also, the R-squared is not much lower with 0.60 and two-third of the highest priority areas were identified correctly. The linear model prediction resulted in an R-squared of 0.53. Additionally, this model's output gave reason to suspect that the identified relationship between 'school attendance', 'toilet presence' and 'foundation type' and damage might not be applicable to other events or countries. The mean Macroseismic intensity and total population were most important in all models and are considered to be indispensable model components. For a future event within Nepal a model output of similar accuracy can be expected, but the presence of case- and country-specific relationships in the current model makes a useful estimation for a future event in another country very unlikely. However, after training the model on events in different countries the model is expected to be able to produce an output that is useful for aid prioritisation decision making. The extent to which the model can be successfully applied to different countries and cases can be improved by excluding secondary hazard susceptibility variables, finding an alternative uniform socio-economic vulnerability variable and using composite building quality variables. Model simplicity and data preparedness are key aspects in the successful further development of these models.

Contents

1 Introduction

Data of the EM-DAT database, containing emergency events since 1988, show that while only 3% of all people affected by a natural disaster are affected by an earthquake, earthquakes are responsible for 55% of all direct deaths caused by natural disasters. Earthquakes have claimed more lives than all other types of geophysical disasters together, killing nearly 750,000 and affecting 121 million people globally (CRED, 2015). Besides an increase of the world population, another cause for these high numbers is that urbanization within earthquake-prone areas has increased significantly in the last years. This has increased the likelihood that a seismic hazard will turn into a major catastrophe (CRED, 2015; Smolka et al., 2004). Especially in development countries natural disasters have both more macroeconomic and social impact. This is a result of insufficient mitigation and prevention measures such as seismic proof building and efficient warning systems (Ortuño et al., 2013).

In order to minimize the impact of such a deadly event in these regions national, international and transnational organizations take all sorts of pre- and post-disaster measures. One of the actors to assist in post-disaster response and recovery are humanitarian aid organizations. They deliver material and logistic assistance in order to save lives and reduce human suffering. They usually operate under exceptional and turbulent circumstances. Often they have to plan complex disaster response based on forecasts or without reliable field assessments (Pedraza-Martinez, 2013). Making decisions is a challenge because of the constantly changing situation and scarcity of information (ACAPS, 2016). While time is limited aid workers must be able to zoom in on local situations as well as zoom out to see the bigger picture.

Because time and quantity of the organization's relief resources are limiting factors, emergency managers have to find an optimal schedule for assigning resources to the affected areas (Fiedrich et al., 2000). To make decisions about where, what and how much of their goods and services should be distributed, humanitarian organizations rely upon assessments carried out by NGO's and local institutions present in the affected area. The Inter-Agency Standing Committee (IASC) coordinates multiple actors to perform needs assessments in fixed formats after a humanitarian emergency. As a joint effort of key stakeholders a Multi-Cluster Initial Rapid Assessment (MIRA) is performed in the immediate aftermath of a sudden-onset disaster. Though they do not always succeed, they aim to produce a Preliminary Scenario Definition which provides a situation overview and indicates the estimated number of affected people in the impacted area within the first 72 hours. A more detailed MIRA output is presented in the MIRA Report within two weeks (IASC, 2012a, p. 5). In the initial phase however, when these assessments are not available yet, decision makers rely upon other secondary qualitative information sources. One aid worker mentioned that his idea of the spatial distribution of a disaster's impact in the first days is formed mostly by driving around a lot in the area (Becks, 2016). Depending on how much exposure an area receives in both the humanitarian community and mass media this can lead to over and under serving of places (Johnson, 2015).

1.1 Problem Statement

Humanitarian aid organizations are thus often faced with problems of resource allocation decision making. The main cause for this is the absence of needs assessment results and scarcity of credible information in the immediate aftermath of a disaster. As a consequence of this, it is possible that relief resources are distributed unequally. As indicated in the Sphere Standards, internationally recognized sets of common principles and universal minimum standards in life-saving areas of humanitarian response, the first step in humanitarian response is to assess the needs of the affected population, and design a prioritized plan of action based on those needs (The Sphere Project, 2011). One key objective of a needs assessments is thus to identify immediate humanitarian priorities (IASC, 2012b). In this process "setting priorities is part of strategic response planning" (Benini, 2015). A supportive tool which rapidly produces an accurate overview of the overall aid neediness in an affected area could help to identify priority areas. An increasing amount of research concerns mathematical models and systems which help in the decision aid processes developed when trying to respond to the consequences of a disaster (Ortuño et al., 2013). Numerous tools have been developed to support prioritisation, but a universally suitable algorithm to establish priority indices has not been established (Benini 2015).

1.2 Towards a Solution

Such a rapid needs assessment that helps aid distributers to prioritize can be performed by means of a Priority Index Model (PIM). As Figure 1.1 explains, this model produces an estimation of the spatial distribution of post-disaster human priorities based on event-specific data and a pre-composed risk database.

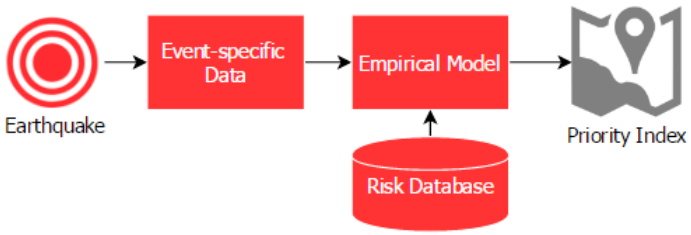


Figure 1.1: Basic principle of a Priority Index Model.

The empirical model quantifies the relationship between on the one hand event-specific and country-specific data (in the risk database), and on the other hand aid neediness, the output. The relationship between those factors is derived from the analyses of pre- and post-event data of past events, hence called empirical. By requiring only little event-specific data the model can produce rapid estimations for future events immediately after they take place.

The concept of PIM's is relatively new and the possibilities that models such as the one described above offer are only recently being discovered by humanitarian decision makers. Nevertheless, the subject has been explored before which provides some preliminary knowledge. To build on to the existing knowledge and experiences, this study proposes a method for developing an Earthquake PIM that is empirically underpinned and adapted to the current data environment, characterized by a large quantity of openly available datasets. To assess the possibilities of this methodology a PIM will be developed for Nepal. Information of the 7.8 M_w Gorkha earthquake of 25 April 2015 will be used to train the model. This earthquake killed nearly 9,000 people and destroyed more than 500,000 homes (OCHA, 2015b). The main reason for the selection of this event is the relatively extensive amount of assessment data that is openly available.

The fact that PIM's are a relatively new study subject also means that there is little known about what statistical model best to use to define the empirical relationships. A quantitative comparison of two different statistical models can provide new theoretical insights and practical recommendations. On the one hand a more traditional multivariate linear regression model (LM) is applied, which is a logical choice given the fact that there are multiple factors prediction one outcome. On the other hand a random forest regression (RF) algorithm is applied. This algorithm could be described as the machine learning version of the LM. Machine learning models originated in the artificial intelligence domain and are computational models based on more complex algorithms that can learn from the training data and improve themselves, therefore they generally produce more accurate model predictions than classical LM's, but they do come with some of their own limitations regarding insight in relationships between variables. Also the way in which both models are built requires different data preparations. Therefore, the models will be compared not only in terms of predictive accuracy, but also in terms of general usability and suitability for implementation in humanitarian decision making processes. Both models and their applications are explained in more detail in Chapter 4 Methodology.

Current rapid impact and severity prediction models mostly generate output by creating composite measures of hazard-, exposure- and vulnerability related variables. Often these are then multiplied using equal or subjective weighing. Both pre-composing measures and equal weighing are questionable because of a lack of empirical evidence to do so. By making use of automated predictor variable selection methods the model will largely refrain from making subjective assumptions, pre-selection of variables and weights assignment.

As mentioned, the model to be developed aims to be adapted to the current data environment. One reason for this aim is reproducibility of the model for other earthquake prone areas on earth. To develop one model that is applicable to many areas over time has many advantages over the development of separate models for individual countries. Some of these advantages are standardizations of data collection processes, multiplication of training cases and a larger target area. This implies that the model will run on open data as much as possible. The fact that national authorities in development regions often keep less detailed and frequent country datasets can form a challenge. Data availability can thus steer rather important modelling decisions. In this perspective the method can be viewed as a data-driven approach.

It is important to mention that PIMs are not intended to be a replacement for other early stage damage and needs assessment tools, but rather to support general aid distribution prioritisation and comparative analysis based on common indicators in the first (and optionally the second) post-disaster phase as defined in the IACS's Operational Guide for Coordinated Assessments in Humanitarian Crises (IASC, 2012b). Therefore the model output will present multi-cluster aid needs and will not specify about the specific type(s) of aid needed. Several challenges are faced in this study:

Challenge #1: Quantifying aid neediness

Ideally, to quantify the relationships between aid inducing variables and aid neediness of a past event a quantitative measure of the level of multi-cluster aid neediness after the event would be the response variable. However, since aid neediness is such a broad notion it is not something that is often measured quantitatively in surveys. Because of the absence of an objective measure a proxy indicator can be used. No studies focussed on the measure of aid neediness specifically for the application of a decision support tool have been performed yet. However, structural damage, human losses or economic losses caused by an earthquake have been frequently modelled. Logically, these variables show similarities with the ones that explain aid neediness in a certain area. But both economic losses and casualties will

not be very useful. Damages in monetary value will not represent communities' needs correctly, since for example heavy damages to valuable governmental structures do not imply any urgent need for basic provisions. Neither are human losses in the direct interest of humanitarian aid workers since they do not participate in urban search and rescue activities. The selection of a final aid neediness proxy indicator is one of the research objectives.

Challenge #2: Defining predictor variables

A lot has been written about what variables induce structural, human or economic losses after an earthquake. Explanatory variables are often physical vulnerability features and exposure. Severity indices or impact models often also include social vulnerability features, such as poverty, age and gender. As was explained by Johnson (2015), especially poverty is expected to be an important indicator for aid neediness. Although the final selection of indicators will result from the model training, it will be necessary to make a data pool of preselected datasets for which an algorithm can be developed. The content of the preselected data pool will to a large extent be determined by the availability and accuracy of the data. The fact that authorities in development regions keep less detailed and frequent country datasets adds another challenge.

Challenge #3: Present output on a low granularity

The level of geographical aggregation on which the output is presented is very important, since this will enable decision makers to better target relief resources to the right locations. If the output is to be presented on a low administrative level, so must the data in the risk database be. Especially the aggregation of the response variable measure is determining the output level. For the case of Nepal this brings an extra challenge, since there is a relatively big difference between administrative levels 3 (district) and 4 (VDC: Village Development Committee). After a distinction between 75 districts, there is a jump to 3,157 VDCs (see Figure 2.1). Many open datasets are available on district level but much less on VDC level. Additional problems are faced because of the decreasing number of VDCs since Nepalese authorities are continuously merging VDCs in an attempt to increase urbanized settlements (Techsansar, 2016).

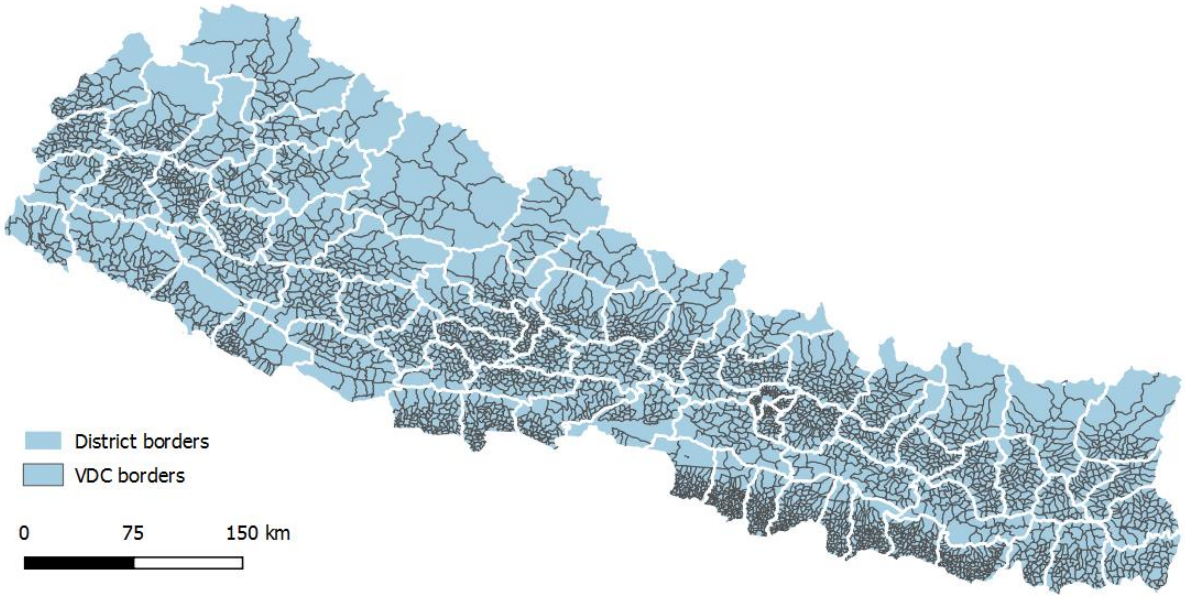


Figure 1.2: Nepal administrative level 3 and 4 borders (data source: HDX, 2016).

Challenge #4: Raster generalization

All input data for the model should be defined at the selected administrative level. This means that input data in raster format will have to be generalized. Generalization of raster data is the retrieval of single values for larger cells or entities, in this case districts or VDCs, from a continuous raster layer. Depending on the application aimed for, simply calculating zonal statistics such as the mean or median of all cell values within a polygon is not necessarily the most desirable method. In this study, often not the whole geographical area covered by a zone is of interest, but only those areas that are populated. For this reason alternative methods to generating single zone values for administrative areas based on a continuous raster overlay are explored.

Challenge #5: Enable rapid execution

While developing the model it should be kept in mind that aid coordinators should be able to run it rapidly after an earthquake has struck in a development region. An experienced field worker mentioned that on the first day there is hardly ever an informative map available at all. “A map with impact estimates in the affected area would be very helpful in the first day, even if it is just to be able to visualise the project area where you will work” (Becks, 2016). Therefore it is important that the model runs on event-specific data that is quickly available. Also a limited amount of tasks and data processing and transformations should be required to run the model on newly available event data.

Challenge #5: Global scope

Although a single-country and single-event model is developed in this study, the methodology of the model aims to form a base for the development of a model targeting all earthquake prone areas on earth. This will be taken into account for example during data selection, by favouring datasets that are available for multiple countries in development regions over those that are uncommon. Also when comparing general usability aspects of the different statistical models this scope will be kept in mind.

2 Research Objectives

In this chapter the general research objectives are defined and more specific research questions towards achieving these goals are presented.

2.1 Research Objectives

The main objective of this study is to explore the possibilities and feasibility of using pre- and post-event open data to train a model to rapidly estimates post-earthquake aid neediness for any earthquake-prone area on earth. Such a model aims to identify aid priority areas and thereby support decisions about the spatial distribution of humanitarian aid resources. It enables rapid and empirically-based humanitarian decision making. Potentially it can be part of IASC's Preliminary Scenario Definition carried out in the first 72 hours after an event. Scientifically, it provides both a theoretical and an empirical base to existing (GIS) methodologies for constructing spatial priority indices. By comparing two different statistical training methods more insight is gained into what models could best be used for aid priority indices. By incorporating post-event damage assessments into the construction of the model it will have a strong empirical foundation.

2.2 Research Questions

Based on the research problem, -objectives and their context as laid out above a main research question is formulated. At the same time the main- and according sub-questions help to structure the research implementation and reporting. The main research question relates to the more general objective as mentioned above. The sub research questions apply specifically to the case of the Nepal 2015 earthquake. The results and insights of this case study help to answer the main question:

Based on a case study of the Gorkha 2015 earthquake, what is the usability of pre- and post-event open data of past earthquakes in estimating priority areas for humanitarian aid rapidly after an earthquake at any place on earth?

The applied definition of usability is derived from an ISO standard (ISO 9241-11): "usability is the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (International Organization for Standardization 1998). Translated to his study, usability concerns the extent to which pre- and post-event open data can help PIM admin- and end-users to effectively, efficiently and satisfyingly estimate post-earthquake aid priority areas. The related sub research questions are:

1. What impact assessment data is available for quantifying aid neediness following the Gorkha earthquake and which fits the study's objectives best?
2. What variables, derived from openly available data, are candidate predictors of the defined response variable?
3. Based on a multivariate linear regression model and a random forest regression model, which candidate predictors can together make the best possible prediction of the defined response

variable, and what is their (relative) importance?

4. For predictor variables that are derived from continuous raster data; how is their predictive value influenced by adjusting their spatial extent to populated areas rather than complete zonal coverage?
5. How do the models perform regarding the prediction of an independent dataset?
6. How do the models compare to each other in terms of predictive performance and general usability?

2.3 Scope

In this study a PIM will be developed for seismic hazards in Nepal. Though, the development of a globally operating model will be kept in mind while making decisions. With regard to the methodology, this research focusses on comparing different statistical models for defining empirical relations and making optimal use of the available open data. Apart from landslides, secondary hazards of earthquakes such as tsunamis and fires are excluded from the analysis. Landslide risks will be included since they caused a lot of damage in the Gorkha earthquake (Government of Nepal National Planning Commission, 2015). Priority is given to shaking intensity because 88% of all earthquake damage is caused by primary effects, which is ground shaking (Erdik et al., 2011). Since earthquakes are usually not single moment events, the temporal extent of the analysis will depend on the defined response variable. If any significant aftershocks took place before the collection date of the response data, these aftershocks should also be included in the analysis. A major aftershock (7.3 M_W) occurred on May 12th in the Middle-Eastern part of Nepal. Of the 75 districts in total the government declared 31 as affected of which 14 as heavily affected (Government of Nepal National Planning Commission, 2015). Nevertheless, the spatial extent of the study area depends on the areas included in the assessment data used for the response variable and the area covered by essential predictor data.

3 Scientific Context

This chapter aims to place the research within its scientific context. What relevant or similar researches have been performed and what knowledge gained from these can be useful for this research? Relevant concepts are the use of open data in the humanitarian sector, earthquake impact modelling, priority indices and the use of past event impact assessments for model training.

3.1 Disaster Response in an Open Data Environment

The contemporary data environment is increasingly characterized by big data and open data. Data collectors and owners are encouraged to openly disseminate it. Open data are data that "can be freely used, modified, and shared by anyone for any purpose" (Open Definition, 2016), though license restrictions can be in place. Many open data initiatives are supposed to ultimately foster collaboration, creativity and innovation (Hofmokl, 2010). The public sector is one of the major producers and holders of information, which ranges, e.g., from maps to companies registers (Aichholzer & Burkert, 2004). But also the diffusion of open government data kept a fast pace in recent years (Vetrò et al., 2016). As Ortmann et al. (2011) state: "disaster management has seen a revolution in data collection. Local victims as well as people all over the world collect observations and make them available on the web." The open availability and usage of data also create new possibilities for finding solutions to aid prioritisation problems in specific. As more governments disseminate national datasets freely accessible and modifiable on the web, humanitarian aid organization gain more possibilities to use them for identifying vulnerable areas and communities. The use of open data by the humanitarian sector is highlighted by UN OCHA's initiative to establish the Humanitarian Data Exchange (HDX) online platform. The goal of this data sharing platform is to make humanitarian data easy to find and use for analysis. Humanitarian data is defined as 1) data about the context in which a humanitarian crisis is occurring, 2) data about the people affected by a crisis and their needs and 3) data about the response by agencies and people seeking to help those who need assistance (HDX, 2016). Since credible information is often lacking in the first days after a sudden onset disaster, initial assessments can instead make use of preliminary available open data and minimize the amount of post-event data needed. However, users should be cautious since especially in development regions governmental open data sets can come with their own limitations, such as being outdated or lacking metadata.

3.2 Empirical and Analytical Approaches

In earthquake damage modelling the functions between shaking intensity and damage are generally constructed either based on an empirical or an analytical approach (King & Rojahn 1996; Jaiswal et al. 2009; Lang 2012; Calvi et al. 2006). Though the analytical approach is more upcoming and becoming more advanced, it is argued that it is less suitable for development regions. Building inventories or systematic analysis of their vulnerabilities are typically lacking in such regions, making analytical tools inadequate (Jaiswal & Wald 2008). Regarding an empirical approach on the other hand, Jaiswal et al. (2009) argue that for regions which have experienced numerous earthquakes with high fatalities historically, typically developing countries with dense populations living in vulnerable structures,

enough data exists to calibrate from the historical record alone. They explain that hybrid and analytical analysis require a series of parameters (for example, knowledge of regional building inventory, structural vulnerability of each building type, occupancy at the time of earthquake, fatality rate given structural damage) which are often unavailable in certain countries or difficult to obtain in cases where it is available, due to inconsistent and poorly characterized historical earthquake casualty data. The empirical approach, on the other hand, is generally regression based, can effectively utilize the available quality and quantity of historical earthquake casualty data and depends on very few free parameters of loss models (Jaiswal et al., 2009). Also, currently most priority indices combine indicators using weights and aggregations decided by analysts. Often the rationales for these are weak. In such situations, a data-driven methodology may be preferable (Benini, 2015). An example of a system using an empirical approach is PAGER. "PAGER rapidly assesses earthquake impacts by comparing the population exposed to each level of shaking intensity with models of economic and fatality losses based on past earthquakes in each country or region of the world (US Geological Survey n.d.)". Within 30 minutes after impact it openly distributes a ShakeMap including the predicted number of people and houses exposed and a range of possible fatalities and economic losses.

3.3 Earthquake Rapid Response Systems

The rapid assessment of spatial distribution and severity of human and structural losses (damage to buildings) after an earthquake can help improve the reduction of human suffering (Erdik et al., 2011). This information comes from rapid response systems. An increasing amount of research has been done about mathematical models and systems which help in the decision aid processes developed when trying to respond to the consequences of a disaster (Ortuño et al., 2013). Earthquake impact models come in all forms and sizes. Some models are very detailed simulations that try to predict damage on the building level. Others are very generic, like the USGS's PAGER predicting for each significant earthquake on the globe the total amount of economic damage and the human losses. With regard to the scope of rapid spatial prioritization of aid resources an analysis on building level is not desired. Currently operating near-real-time loss estimation tools can be classified under two main categories: global and local systems. Methodologies of global rapid loss estimations are relevant for this study since post-earthquake humanitarian needs logically correlate with structural damages and human losses. These systems generally include several features. For example, Benini and Chataigner (2014) formulate post-disaster needs in a study about typhoon priority indices as

$$Needs = k * Magnitude * Intensity * f(Pre\text{-}existing\ conditions)$$

Where *k* is an unknown constant expressing proportionality, *Magnitude* expresses the number of affected people, *Intensity* is the fraction of totally destroyed houses and *Pre-existing conditions* indicate the poverty rate. Another example comes from the Office for the Coordination of Humanitarian Affairs' (OCHA) INFORM Severity Index for Nepal. It used a similar equation to quantify the disaster's impact on the population for prioritisation:

$$Severity = Hazard * Exposure * Vulnerability$$

Hazard is measured by earthquake intensity as derived from USGS ShakeMaps, *Exposure* indicates the total population in a VDC and *Vulnerability* resembles a normalized weight of a housing quality measure (wall and roof type) and a poverty measure (Human Poverty Index) (OCHA 2015a). Both equations include earthquake intensity and social and physical vulnerability related factors. Nearly all earthquake

impact or severity models combine the three factors of hazard, exposure and vulnerability. The predictor categories for this study are defined as:

Hazard

Bird and Bommer (2004) have explained that 88% of damage due to earthquakes is caused by ground shaking rather than secondary hazard. Therefore, one very important part of loss estimation methodologies is the quantification of ground motions. Many rapid response systems use USGS's ShakeMaps for post-earthquake response, public and scientific information and loss assessments (Erdik et al., 2011). ShakeMap uses instrumental recordings of ground motions, kriging techniques, and empirical ground motion functions to generate an approximately continuous representation of shaking intensity shortly (minutes) after the occurrence of an earthquake (Wald et al., 2008). The ground motion distributions that are generated via ShakeMap can be used as input for casualty and damage assessment routines for rapid earthquake loss estimation (Erdik et al., 2011).

Exposure

Exposure refers to the inventory of elements in an area in which hazard events may occur (Cardona et al., 2012). Hence, if population and economic resources were not exposed to potentially dangerous settings, no problem of disaster risk would exist. While the literature and common usage often mistakenly conflate exposure and vulnerability, they are distinct. Exposure is a necessary, but not sufficient, determinant of risk. It is possible to be exposed but not vulnerable (for example by living in a floodplain but having sufficient means to modify building structure and behaviour to mitigate potential loss). However, to be vulnerable to an extreme event, it is necessary to also be exposed.

Physical vulnerability

Physical vulnerability to earthquakes can be induced either by hazards caused directly by the building environment or indirectly by secondary hazards. Erdik (2011) argues that for the assessment of direct physical damages, general building stock inventory data and the related vulnerability relationships are needed. However, not only for building loss estimations, but also for human loss and aid needs estimations the spatial distribution and location of buildings is an important factor to take into account. Not only is building destruction the main reason for people to get injured or parish during an earthquake, but also is loss of shelter an important reason to be in need of aid in the event's aftermath. Unfortunately, only a limited number of countries and cities have well developed building inventories (Erdik, 2011). The ELER (Earthquake Loss Estimation Routine) software system solves this issue by using a proxy procedure that relies land use cover and population distributions to create an aggregated full country covering raster with continuous values indicating the amount of buildings per cell (Hancilar et al. 2010). Also, initiatives like OpenStreetMap distribute openly available crowd-produced datasets of buildings and roads.

Secondary hazards related to earthquakes are landslides, tsunamis, seiches, floods and fires. Sometimes landslides are seen as a primary hazard, together with surface rupture, ground motion and liquefaction, but in this study the will be labelled as secondary hazards. Modelling susceptibility to secondary hazards is a complicated task in itself, as these natural phenomena are generally unpredictable and depend on many different environmental factors. A solution can be to use a proxy indicator that can be considered as a main inducer of secondary hazard.

Social Economical Vulnerability

Socio-economic status plays an important role in increase of social vulnerability related to hazards. It is difficult for the people with low socio-economic status to restore their living order, which was disrupted due to the disaster (Yucel & Arun, 2012). Especially in the humanitarian sector it is important to also take social and economic vulnerability into account, since these influence the personal recovery capacities after initial impact of a natural disaster. For example, in the severity index that OCHA produced after the Gorkha earthquake socioeconomic vulnerability was weighted at 20%, against 40% assigned to earthquake impact (human and structural damages) and 40% to physical vulnerability (landslide hazard and road accessibility). Socioeconomic vulnerability was constructed of: poverty (30%); caste, ethnicity and gender inequality (30%); youth, elderly and disabled people (20%); and labour capacity represented by international migration (20%) (OCHA 2015a).

Learning Models

Multiple existing systems are based on learning from past events. The PAGER system produces information on earthquake location, magnitude, depth, number of people exposed to varying levels of shaking intensity and a region's fragility. A PAGER feature of special interest is their estimate of the total number of fatalities based on empirical correlations between casualties and intensity. Another example is the Japanese HERAS (Hazards Estimation and Restoration Aid System) system. This system estimates damage to railways based on damage experiences of past earthquakes (Yamazaki & Meguro, 1998). Another example is a severity index distributed by the INFORM working group two days after the event in Nepal. Their model used equal weighing of hazard (USGS ShakeMap), exposure (a population overlay) and vulnerability. The latter was composed of housing vulnerability (50%) and poverty (50%). The Human Poverty Index (HPI) was used to represent poverty. Data aggregated to different administrative levels (levels 3 and 4) were combined in this analysis (INFORM, 2015)

Uncertainties

In most modelling, uncertainties are present since approximations and simplifications of the "real world" are necessary in order to perform a comprehensive analysis. In earthquake loss estimation models uncertainties can be derived either from the seismic hazard analysis (is the shaking accurately represented) or from the vulnerability relationships (do wooden structures indeed have less seismic capacity). Input datasets can be the subject to noise, outliers and errors (Hammer & Villmann, 2007). For example, needs assessments can contain documentation errors and national census datasets could be incomplete or outdated. Uncertainties can also be inherent in pre-processed datasets. For example, there exits considerable amount of epistemic uncertainty and aleatory variability in ShakeMaps, depending on the proximity of a ground motion observation location and the estimation of ground motions from the GMPE (Wald et al., 2008). It is possible to examine the effect of cumulative uncertainties in loss estimates using discrete event simulation (or Monte-Carlo) techniques if the hazard and probability distribution of each of the constituent relationships are known. The general finding of studies on uncertainties in earthquake loss estimation is that uncertainties are large and at least equal to uncertainties in hazard analyses (Stafford et al. 2007).

3.4 Needs Assessments and the Value of Priority Index Models

Often the field of humanitarian logistics has been approached in a similar way as business supply-demand chains. Key differences are an unpredictable demand, a short lead time and suddenness of demand for large amounts of different products and services, a lack of initial resources and multiple decision makers who can be difficult to identify (Ortuño et al., 2013). In case of a natural disaster these decision makers are informed by needs assessments, relating to four main questions: 1) whether to intervene, 2) the nature and scale of an intervention, 3) prioritisation and allocation of resources, and 4) programme design and planning (Darcy & Hofmann 2003). The answer to all of these questions starts with an assessment of the spatial distribution of the disaster's impact. Since assessments are often not available the first days, an estimation of the disasters' impact, such as produced by a PIM, is thus the starting point for planning intervention. Ebener et al. (2014) identify that the "need for accurate and up-to-date data to support disaster risk reduction and emergency management has long been recognized". Additionally, various authors specify that the information needs to be accurate, appropriate, timely and valid (ACAPS 2016a; Comes et al. 2015; Homberg van den et al. 2016).

Priority Index Modelling

Such a rapid needs assessment that helps aid distributers to prioritize can be performed by means of a Priority Index Model (PIM). Priority indices have grown popular for identifying communities most affected by disasters (Benini, 2015). A PIM geographically disaggregates the affected area and indicates for each entity the degree to which it is in need of aid. This aid can be any type of humanitarian aid (in total IASC distinguishes between eleven aid-clusters ranging from shelter to nutrition to health (IASC 2012a). Such an index has a position to act as a stopgap before the more detailed assessments are available (Johnson, 2015). The amount of information in an output should be both limited and credible.

The contemporary data environment is characterized by overwhelming amounts of openly available data on a wide range of topics. The number of different ways and applications to make sense out of them are increasing. This creates new possibilities for finding solutions to accelerate aid prioritisation problems. As explained before, since credible information is often lacking in the first days after a sudden onset disaster, initial assessments can instead make use of preliminary available open data and minimize the amount of post-event data needed. This is where PIM's can be part of the solution, since they require only a very limited amount of post-event data. For almost any type of sudden onset natural disasters there are institutions active that rapidly produce datasets on geophysical characteristics of the hazard itself after its occurrence. By collecting and organizing the right set of pre-crisis datasets a PIM can provide decision support for emergency response in an information poor situation.

Within the Red Cross societies some work concerning priority indices has been performed in the past. In a blogpost by Andrej Verity (2014) he suggested the combination of pre-crisis datasets and post-event contextual data to show the impact of the disaster. In response to this, Simon Johnson, GIS expert at the British Red Cross, created try-outs of such a model for Cyclone Pam and Typhoon Maysak (Johnson, 2015). Model input included data on population, wind speed and poverty levels. Poverty was included since he observed in the field himself that "there was a big focus on not just affected population totals, but also on the areas with high poverty as they were the least likely to self-recover" (Johnson, 2015). However, he explains that although the generated output can be useful, the models were made on intuition and that improvements can be made by comparing output against formal ground assessments and by fitting the models mathematically to determine which parameters matter "to see if it is possible

to develop consistency of parameters between countries and if priority indices should be pursued at all" (Johnson, 2015). Also within the Netherlands Red Cross a priority index has been developed for typhoons, trained on data from five past typhoons in the Philippines. The most important conclusions drawn from the development of this model were that the importance of poverty data seems to be overestimated in many other severity indices and that it is essential to use features that are proportional to the population. Otherwise population is by far the most important feature in any model (see: http://510.global, 2016).

3.5 Machine Learning for Damage Assessments

The application of machine learning techniques for (natural) disaster damage assessments is relatively new. Not many studies on the development of such applications have been published.

As Breiman (2001) explains, there are two cultures in the use of statistical modelling to reach conclusions from data: "One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown". The latter is also known as machine learning. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. Such algorithms overcome the approach of strictly static program instructions by making data-driven predictions or decisions. Breiman (2001) argues that statisticians rely too heavily on data modelling, and that machine learning techniques are making progress by instead relying on the predictive accuracy of models. Also for complex prediction problems data models are often not sufficient.

A distinction can be made between supervised and unsupervised learning. For this study supervised learning algorithms will be applied. In supervised machine learning the input data is called training data and has a known label or result (in this case the quantified aid neediness measure). Examples of algorithms that are suitable for multivariate regression problems as in this study are Linear Regression, Decision Forest, Neural Network and Random Forest (Caruana & Niculescu-Mizil, 2006). Accuracy, training time and linearity are often the three considerations when choosing an algorithm (Rohrer, 2016). Training time will most likely not be an issue since the dataset will be relatively small. Whether or not to use an algorithm that uses linearity depends on the data trends observed from scatterplots. Multiple suitable algorithms will be applied to learn from the data and the one with the highest predictive accuracy will be selected for the model.

Since algorithmic modelling is a relatively fast way to make sense out of data, multiple models can be iterated easily based on different parameter settings. This enables the methodology to be extended and applied for other countries and other types of disasters more easily. When new event-specific data and damage assessment are collected and structured, the algorithm can learn from these and improve itself. The model maker will not have to deal with studying and defining the optimal relationships. On the downside, he or she will not gain as much insights in how variables relate to each other or whether they influence the output positively or negatively.

The typhoon priority index that was developed within the Red Cross (see http://510.global/philippines-typhoon-haima-priority-index/) made use of a random forest regression. 13 explaining variables are part of the model. Four of them are event-specific (distance to typhoon path, typhoon position, wind speed and rainfall). Geographical variables were ruggedness, average slope gradient, coastline-inland line

ratio, elevation and area. Finally, wall material, roof material population and poverty were part of it. The response variable was a governmental count of the amount of partially and fully destroyed houses per municipality. Overall the model explains 88% of all variance in housing damage. The R-squared is 0.58 and the mean damage error is 1,290 houses per municipality.

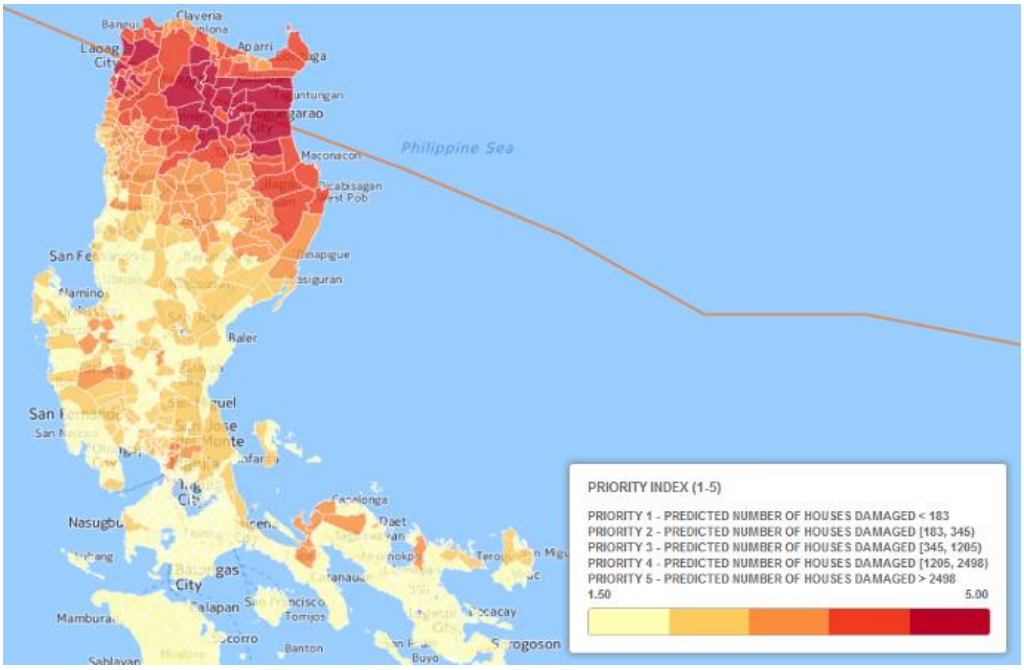


Fig 3.1 – Output Priority Index Typhoon Haima (510, 2016).

It should be noted that the model made much better predictions for fully damaged houses than for partially damaged houses. This might be caused by the fact that threshold for labelling a building as partially damaged differs between municipalities, as each of them did an individual count (510, 2016).

Besides the activities within the Red Cross regarding priority indices, an example of the application of machine learning techniques for disaster damage forecasting is a study by Kohara & Hasegawa (2009). They applied Self-Organizing Maps, multiple regression and decision trees to forecast typhoon damage in Japan. 111 data records of typhoons from 1981 up to 1995 were used to train the model and 86 for testing. Damage data included fatalities, injured, destroyed houses and flooded houses. Event-specific data included month of occurrence, latitude and longitude, atmospheric pressure, maximum wind speed and precipitation. In the data records there was relatively much small scale damage (>80%), therefore the model made very accurate predictions for small scale damage but was less accurate for large scale damage. Therefore they applied the selective-learning-rate approach: the learning rate for training data corresponding to small changes is reduced. Their study focussed on quantitative damage predictions and thus did not include any social vulnerability parameters related to aid needs or recovery.

4 Methodology

This chapter opens by presenting an overview of the methodology by visualizing all research steps in a methodological model. The subsequent paragraphs discuss the approaches for providing answers to each individual sub research question.

4.1 Methodological Model

Broadly speaking the research is completed by performing several consecutive tasks, providing answers to the sub research questions and ultimately making it possible to answer the main research question. The conceptual model below shows all steps. The blue circles indicate the according research questions. First of all, data for the predicted and candidate predictor variables are collected (Q1 and Q2). Where necessary data processing and transformation takes place to extract the desired features from these data. All variable data will be collected in a structured feature matrix. Hereafter the LM and RF models are trained on the data in the matrix, resulting in several models that can be evaluated by comparing the predicted output to the actually measured output (Q3). Based on these evaluations the process can loop back to both model training (adjusting parameter settings) and feature extraction (changing input variables). During this process the comparison of the two different raster generalization techniques will also take place (Q4). After going through the training process again two improved models are created. These will be validated by comparing them to two different independent datasets (Q5). These validation results will in turn form the main input to compare the drawback and advantages of both models (Q6).

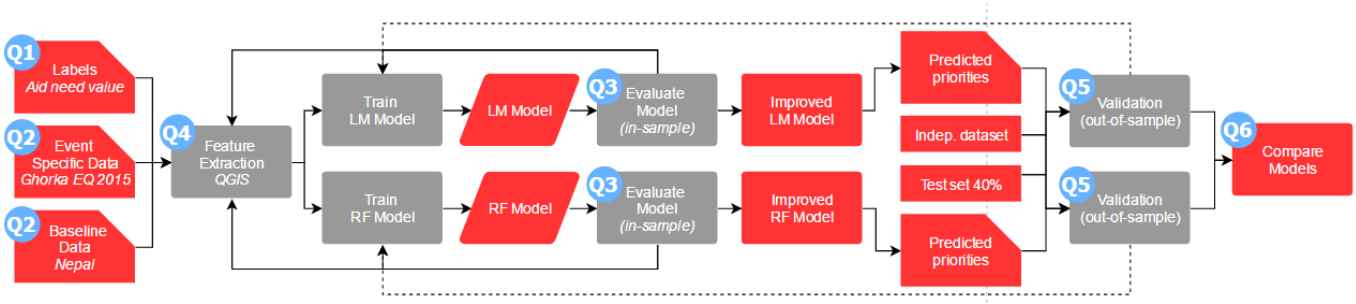


Figure 4.1: Visualization of applied research methodology.

4.2 Quantifying Aid Neediness

Q1: What impact assessment data is available for quantifying aid neediness following the Gorkha earthquake and which fits the study's objectives best?

The first task is to define the response variable for the model. As explained in the Introduction a proxy indicator has to be used due to the absence of an objective and quantitative measure for aid neediness (see Challenge #1).

An extensive data search will be performed, mainly using information channels of the Red Cross Societies and data portals for online sharing of data within the humanitarian sector. After assessing the

metadata and suitability of multiple datasets one of them will be selected as the final predicted variable. Some of the most important selection criteria are the level of aggregation, number of observations, commonness of the measure used, trustworthiness, temporal coverage, spatial coverage and completeness.

After these possible datasets are collected, their suitability to represent multisector aid neediness will be discussed with several aid workers experienced with the coordination of emergency relief. The main goals of these discussions is to find out what indicator(s), in case of an earthquake, would help aid workers in the field best to identify priority areas in the first few days after initial impact.

4.3 Defining Candidate Variables

Q2: What variables, derived from openly available data, are candidate predictors of the defined response variable?

The second research question relates to the definition of candidate predictor variables. They are being referred to as candidates since they can be eliminated throughout the process of model training. The assumption is made that all predictors are available for inclusion or exclusion from the model. From the review of existing earthquake impact models as presented in Section 3.3 Earthquake Rapid Response Systems it was concluded that candidate variables of four different categories should be included: hazard related variables, exposure related variables and physical and social vulnerability related variables. Sub question 2 is answered by determining requirements for variables in each category, collecting data and defining relevant variables in each of the four categories.

However, the collected set of explaining variables does not have to be perfect, since automated variable selection procedures offer the opportunity to eliminate variables with insufficient influence on the predicted variable. This way no fixed assumptions about what social or physical vulnerability exactly entails have to be made. Apart from the previously found predictive value of certain indicators, the data collected maybe depends even more on availability, attainability, accuracy and level of geographical aggregation. All relevant data will be p-coded (p-codes are unique geographic identification codes assigned to administrative areas worldwide by OCHA) and organized by the same level of aggregation as that of the predicted variable.

4.4 Model Fitting

Q3: Based on a multivariate linear regression model and a random forest regression model, which candidate predictors can together make the best possible prediction of the defined response variable, and what is their (relative) importance?

The aim of the third research questions is to find the best fitting empirical relationships for both models between the previously defined response variable and candidate predictor variables. Before the actual model training a few preparations take place.

Data Exploration

First an exploration of all the data takes place. Data exploration consists of three subsequent tasks. First the complete data matrix is checked for missing values. In case values are missing first an explanation is sought, after which will be decided either to leave the corresponding observation out of analysis or to fill in the (assumedly) correct value. Secondly, it is checked whether the data are normally distributed. If the frequency distributions of the variables indicate skewness appropriate data transformation such as a logarithmic transformation will be considered. The third issue to look at is collinearity among covariates. By studying pairwise scatterplots and interpreting absolute correlation coefficients multicollinearity between predictor variables is studied. Appropriate ways to deal with covariance will be discussed and reported in the according Results section.

Cross Validation

After all assumptions for the model training are sufficiently met the next step is to prepare for cross validation by splitting up the data in a training set and a test set. This enables the possibility to check afterwards whether the model created is not over-fitted to the training data and can also make an accurate prediction for a set of observations that was not known to it during training. The size of these sets depends on the size of the complete dataset. The sets will be assembled randomly and stay the same throughout the whole modelling process.

As explained before, for the detection of empirical relationships in the data two different algorithms will be modelled: a multivariate linear regression model and a random forest regression model. To find the best fitting models both algorithms have different aspects to pay attention to.

Linear Regression Model

The first model to be made is a multivariate linear regression model (LM), which is a logical decision given the multivariate regression problem. To find the best fitting LM a stepwise variable selection is applied, meaning that the choice of predictive variables is carried out by an automatic procedure, in this case the 'regression subsets selection' function. This function performs an exhaustive search for the best subsets of predictor variables for predicting the response variable. It produces sets of candidate models of different sizes, leaving room for own interpretation. This method prevents for unnecessarily holding on to no longer significant variables or permanently eliminating variables which later on can become relevant again, both of which can occur when applying forward selection or backward elimination techniques. Stepwise regression techniques are sometimes criticized for creating over-simplifications of the data, but this will be taken into account by performing cross validation on the independent test data set. The regsubsets function returns a table of models showing which variables are in each model, ordered by a specified selection statistic (Lumley, 2017). The selection criteria for the final LM is the adjusted R-squared (R^2_{adj}), which indicates how much of the variance in the response variable is explained by the model and is adjusted to the number of predictors in the model. During the model training, usage of this measures prevents for overfitting to the training data. After a final model has been selected, the normal R-squared will be used.

Several rounds of training are carried out. This is mostly to see how the model's performance is influenced by defining different types of candidate variables, which cannot be included simultaneously due to collinearity. This provides insights into how best to extract features from geographical input data in order to make the best possible predictions of post-earthquake aid neediness. The same manner also enables to independently compare different spatial aggregation techniques, as is the subject of research question 4, explained in the next section. Once the model with the best R^2_{adj} is selected the coefficients

can be interpreted, which will provide insights into the influence of individual variables on the aid neediness variable.

After the final LM is selected, validity of the model will be assessed by testing for multiple model assumptions. These assumptions are linearity of residuals, heteroscedasticity of residuals and normality of residuals.

Random Forest Model

As the name suggests, the random forest model (RF) consists of multiple (decision) trees. Decision trees predict the outcome of the response variable by splitting up the input variables at relevant points. A RF classifier uses a number of decision trees in order to improve the prediction rate. It makes corrections with each model iteration during the training and averages the outputs of multiple decision trees. In standard trees, each node is split using the best split among all variables. In a RF, each node is split using the best among a subset of predictors randomly chosen at that node (Liaw & Wiener, 2002). Therefore, the RF model can be viewed as a machine learning version of a linear regression model. In RStudio the RF function model returns the number of trees, the number of variables tried at each split, the mean of squared residuals and the percentage of variance explained. Besides this it does not give much insight into the relationships defined. As with the fitting of the linear regression model, also here several training rounds take place, allowing to independently compare the model's performance with a different selection of predictor variables to choose from and with different parameter settings. The selection criteria for the best fitting model is R^2.

During the fitting of both models only in-sample validation takes place, meaning that the models fitting best to the training data are being selected, leaving validation with the test data set out of consideration for now. The in-sample validation is done by means of comparing and interpreting the R^2 scores of the models and checking for randomly distributed residuals by means of scatterplots.

4.5 Comparing Aggregation Techniques

Q4: For predictor variables that are derived from continuous raster data; how is their predictive value influenced by adjusting their spatial extent to populated areas rather than complete zonal coverage?

As explained under Challenge #4 in the Introduction, generalizing a continuous raster layer to a single value for a larger entity per definition causes a loss of information, because the resolution of the layer increases. By finding an answer to this research question the ultimate aim is to gain insight and form recommendations about the comparative advantage of using one raster generalization technique over another. As explained in Section 4.1 Conceptual Model the process takes place during the fitting of both models to the training data. Two techniques for aggregation to VDC or district level will be compared, both performed with QGIS software. This is performed only for continuous raster layers for which it can be argued that taking into account only built-up areas could be more sensible, such as the earthquake's shaking intensity. Also, because the model strives to produce an output within 6 hours this will not be applied to event-specific data which will be available only after the event, as it takes quite some computation time.

Zonal Statistics

This generalization method is relatively simple, in the sense that is requires only few computation steps. Zonal Statistics as a basic GIS computation whereby the target layer is a continuous raster layer and a vector layer with lines or polygons defines the zones to which statistical values should be assigned. To calculate a mean elevation value for one geographical entity for example, the slope values of all cells intersecting with this entity are added up and divided by the total number of intersecting cells. Depending on the interest this can give a false perception, as the average shaking intensity in an entity can be relatively high, while the average intensity at the most populated areas in this entity can be low.

Clipping to built-up areas

By adding several computation steps the information loss can be reduced. The extent of the continuous raster data should cover only populated areas, as these are the target of emergency relief. Spatial population distribution is represented by building locations. As no official data on built-up areas of Nepal is openly available, an alternative method for defining these areas is to derive a building dataset from OpenStreetMap (OSM). Fortunately, from April 25 to June 8 2015, a significant crowdsourcing effort of the OSM community with more than 7,500 contributors from around the world supported the logistic of the Nepal government, UN Agencies and International organizations responding the Nepal Earthquake human relief (wiki.openstreetmap.org, 2015). These efforts were coordinated by the Humanitarian OSM Team (HOT). Currently, more than 1.1 million building outlines are mapped in the 16 most affected districts. First of all, a distinction can be made between higher and lower densities of populations. Via a Kernel Density operation a raster layer representing the density of points around each cell is produced. The next step is to clip the raster layer to built-up areas. By drawing 50 meter buffers around these buildings built-up areas can be defined (buffers are drawn around building centroids instead of polygons due to processing limitations) (Figure 4.3).

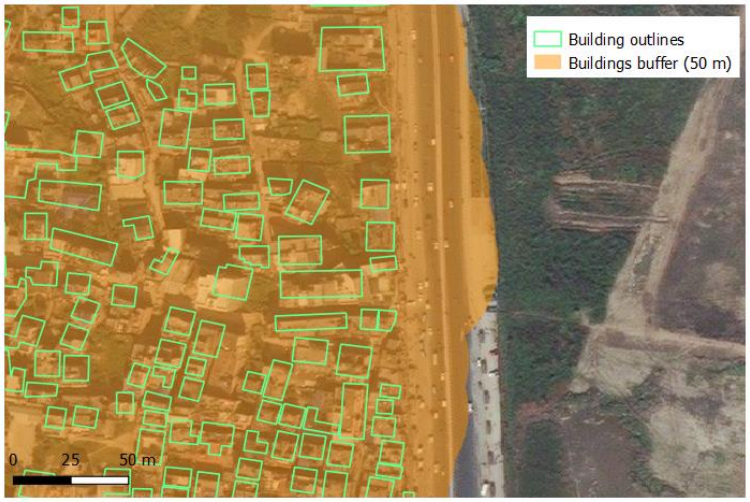


Figure 4.2: 50 meter buffer drawn around OSM building outlines.

Hereafter, the cell values of the input raster (e.g. shaking intensity) are multiplied by the intersecting cell values of the building density raster. The resulting output is clipped to the extent of the buffer area (Figure 4.3). To retrieve single values for each geographic entity from this layer the same steps as described above for Zonal Statistics are repeated. Seven VDCs did not have a building mapped within them. For these VDCs the slope values based on the first generalization method were used.

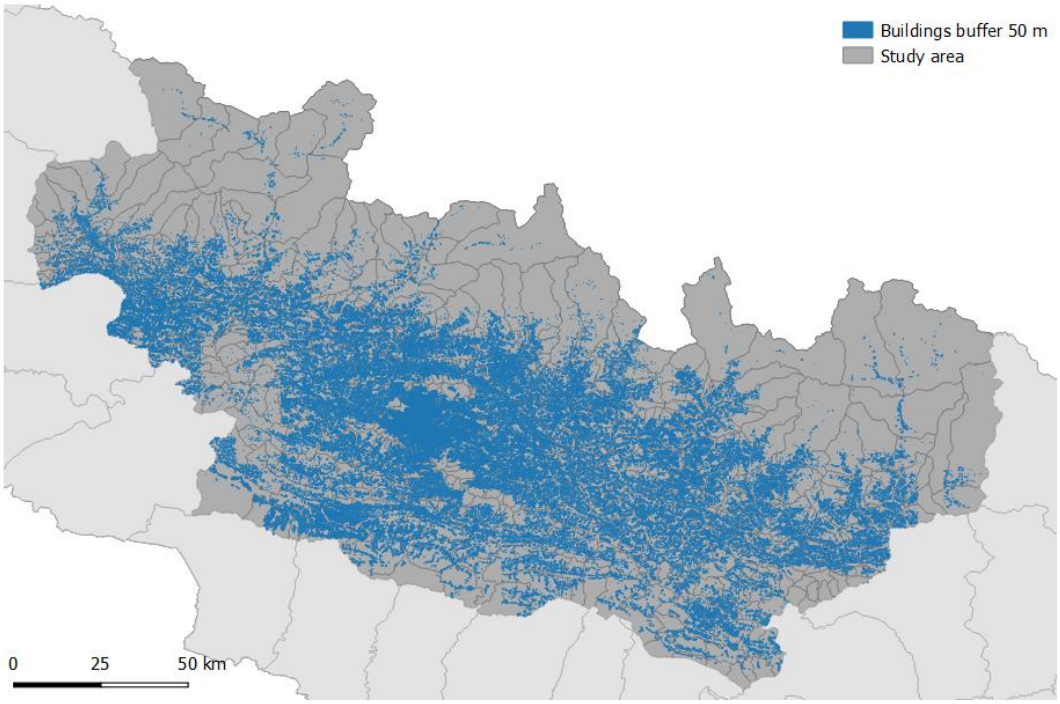


Figure 4.3: 50 meter buffer around OSM building outlines for the complete study area (source buildings data: OSM, 2017).

The aggregated values will be included in the data matrix. To compare them independently they will be included as candidate predictor separately. Both models are trained once on a matrix including the first variable, and once on a matrix including the second. The relative importance of both variables in the resulting models are compared, allowing to draw conclusions on which would be better to use.

4.6 Model Validation

Q5: How do the models perform regarding the prediction of an independent dataset?

So far performance of the models was judged only by interpreting (adjusted) R^2 values to assess how well they "fit" the data they were trained on. However, it is also important to see how well the model can make predictions for data other than the data it was trained on. This is referred to as out-of-sample validation.

Out-of-sample validation

The selected models will make predictions for the test dataset that was set apart before the model training. Both the selected LM and RF model will be run on the test cases. Their performance in this validation test is again judged by the R^2 values, a scatterplot and the root mean squared error (RMSE) as a measure for the prediction error. If the models perform much lower on the test set this indicates overfitting. This can be overcome by simplifying the model by eliminating the predictor variables with the lowest significance. Also, maps of both the measured and predicted values are compared in order to draw conclusions about the spatial characteristics of the predictive accuracy.

4.7 Model Comparison

Q6: How do both models compare to each other in terms of predictive performance and general usability?

To make sense out of the insights gained during previous research steps and to be able to answer the sixth research question the LM and RF model are compared to each other. First of all, their predictive performance is compared by summarizing the coefficients of determination and average prediction errors. Additionally, maps displaying the residuals of each model output are compared. Focus is on the correct identification of the highest priority areas. Secondly, the comparison of the models general usability is mostly descriptive and focusses on their drawbacks and opportunities when the PIM is scaled up to a global level. A distinction is made here between admin-users, those users that will work on further development of the model and produce model output, and end-users, being humanitarian aid field workers with a role in decision making about aid distribution. Especially for this last group of users, intuitiveness of model output is an important aspects of usability.

5 Results

In this chapter the results and insights of the previously presented methodological steps are reported. All sub research questions will be answered consecutively in order to draw conclusions and provide recommendations in the final chapters.

5.1 Defining a Response Variable

Q1: What impact assessment data is available for quantifying aid neediness following the Gorkha earthquake and which fits the study's objectives best?

In the weeks after the Gorkha earthquake various organisations and institutions have performed multiple ways to assess the impact on citizens and their environment. Most assessments are based on surveys, but other examples of assessment methods are satellite imagery interpretation (see UNITAR/UNOSAT, 2015) and large scale mobile phone tracking (see Wilson et al., 2016). Open data availability, suitability and usability for model training are the main criteria for the selection of a suitable dataset to indicate the level of aid neediness.

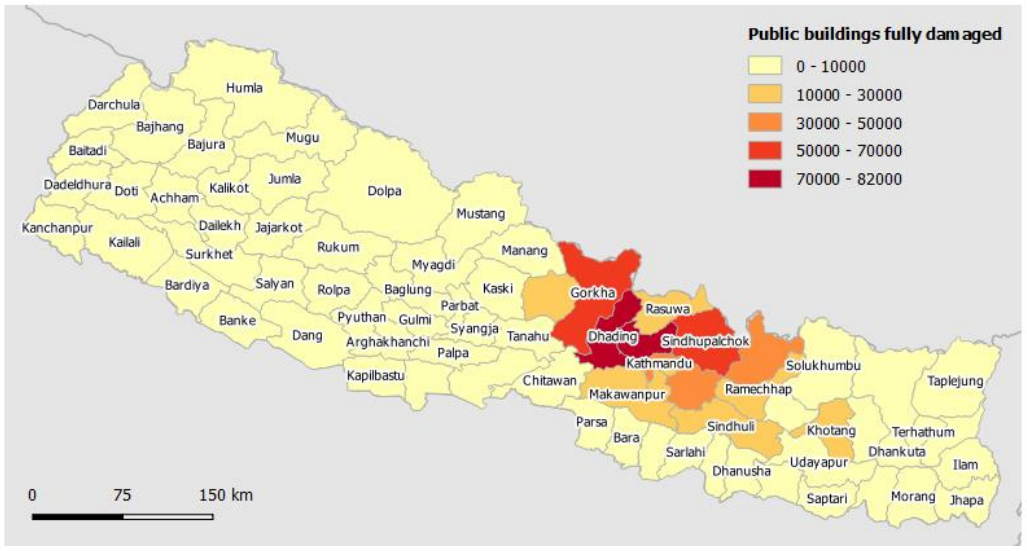
As explained, an extensive search via mainly Red Cross information channels and online humanitarian data sharing portals (the online Humanitarian Data Exchange portal (HDX) in particular) was performed. All datasets that could possibly indicate aid neediness were kept track of in a structured table including metadata on each dataset. These included the number of 'people in need' (OSOCC Assessment Cell, 2015), the origin of people in shelter camps (IOM, 2015), structural damage identified via satellite imagery (Yun et al., 2015), number of people displaced and affected (Nepalese Red Cross Society, 2015), the number of houses fully and partially damaged (Nepalese Red Cross Society, 2015a; OCHA Nepal, 2015) activity logs of all humanitarian organizations working on housing reconstruction (Housing Recovery and Reconstruction Platform, 2016), above normal population inflow derived from phone tracking data (Wilson et al., 2016) and relief items distributed (Ministry of Home Affairs Nepal, 2015).

From discussions with experienced humanitarian aid field workers the main conclusion was that from all collected datasets the ones representing structural damages to residential buildings would be the best aid neediness proxy indicators. The main reason for this was mentioned in multiple conversations, for example by M. Becks, Head Resilience Advisory Unit at The Netherlands Red Cross (2015): "Damage to houses is important for many aid sectors. Of course for the Shelter cluster, but therefore also for the WASH (Water, Sanitation and Hygiene) cluster, and therefore the Health cluster. And depending on how and where food is stored also for the Food Security cluster". Logically, structural damages should then preferably concern residential buildings only. Two of the collected datasets reported structural damaged to residential buildings. These datasets and their characteristics are discussed in the two Sections below, based on which a final selection can be made.

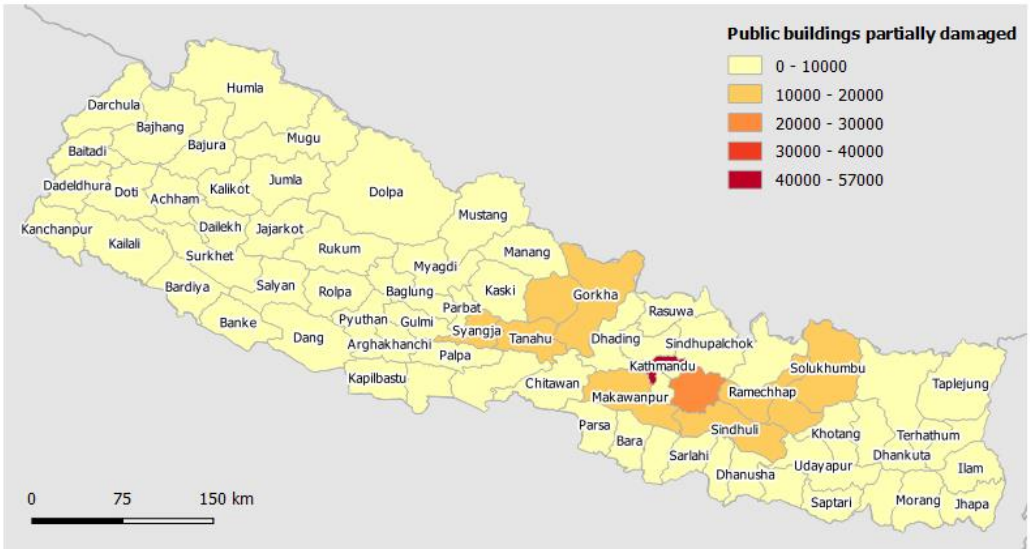
Other datasets representing aid neediness rather well (such as the number of people in need), were discarded mostly due to a limited number of observations, subjectivity of the measure(s) used and uncommonness of such an assessment. Training a model based on measurements that are often reported after an earthquake in any place on earth keeps more options open for future training and improvement of the model.

5.1.1. Dataset A: Structural Damage on District Level

The first dataset was compiled by OCHA from reports of the Nepalese Ministry of Home Affairs and the Nepali Police. It was derived from the online HDX platform (see: OCHA Nepal, 2015). It distinguishes between four different types of impact indicators: 'deaths', 'injuries', 'governmental buildings damaged', 'governmental buildings partially damaged', 'public buildings damaged' and 'public buildings partially damaged' ('public buildings' are residential buildings and 'public/governmental buildings damaged' means fully damaged). The indicator of interest here is 'public buildings damaged'. This is expected to be a good proxy indicator for multisector aid neediness, as it indicates how many people have lost their shelter and thereby possibly their sanitary facilities and food supplies.



a)



b)

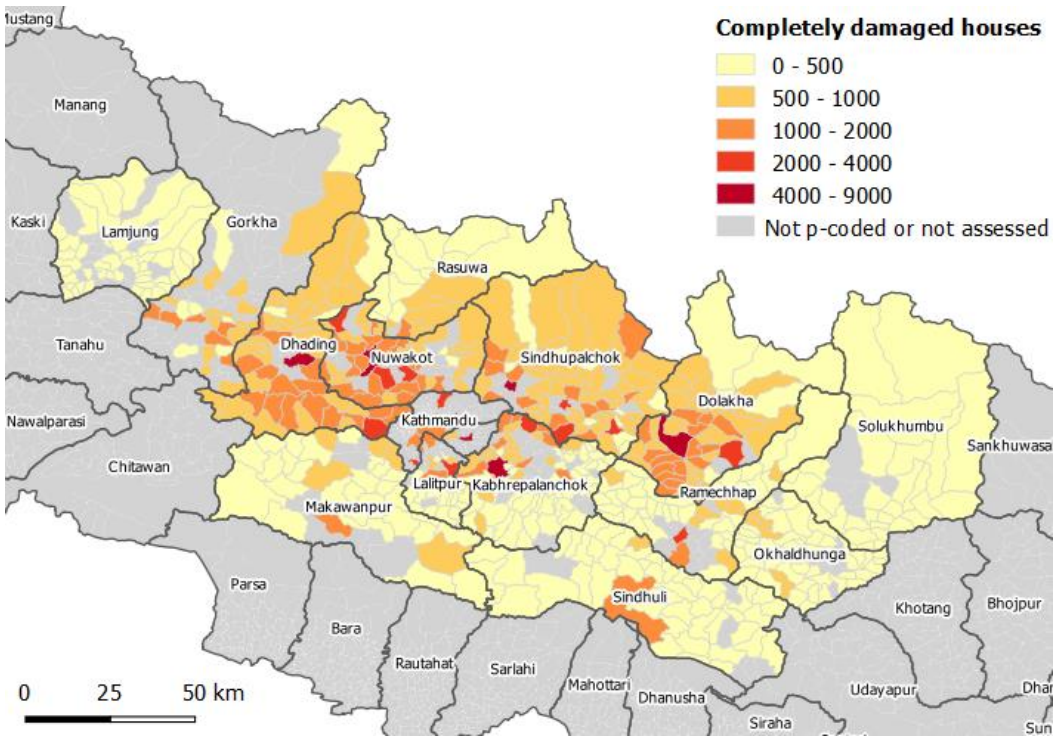
Figure 5.1a-b: Number of fully (a) and partially (b) damaged public buildings in Nepal 2015 (data source: OCHA Nepal, 2015).

The numbers are defined for each of the 75 districts in Nepal. The dataset includes 22 separate measurement dates between the 29th of April and the 5th of June 2015. When divided by the total number of households as defined in the National Population Census of 2011 there is one district (Dolakha) exceeding 100%, with 107% of all houses fully damaged.

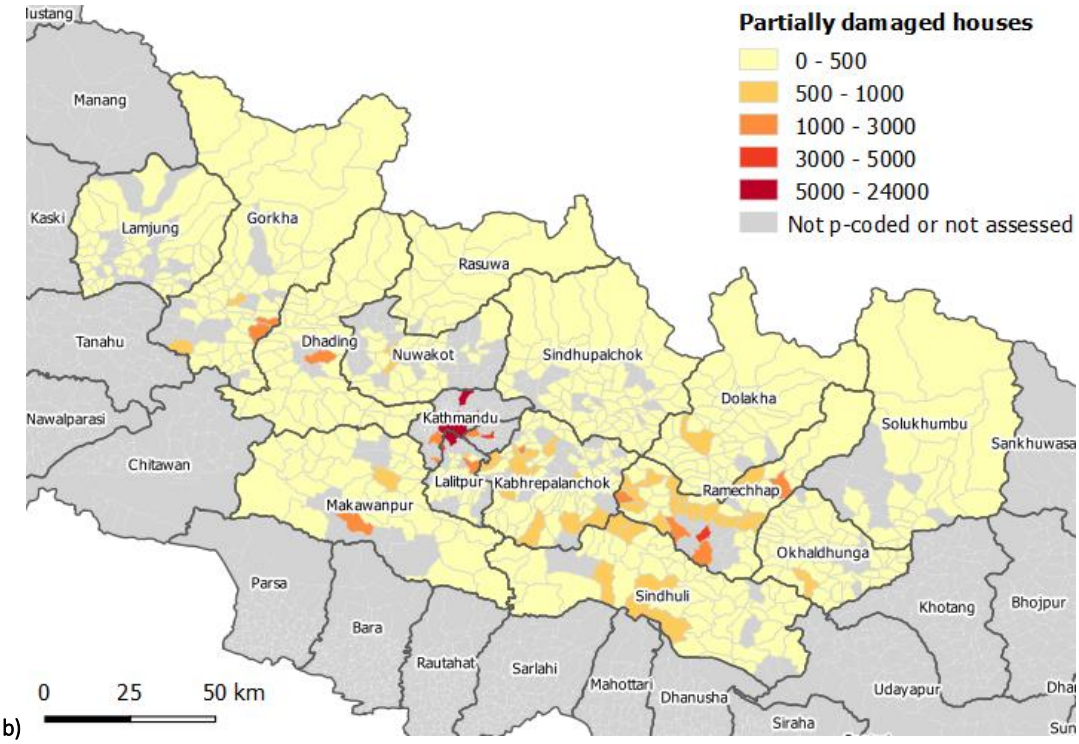
5.1.2 Dataset B: Structural Damage on VDC Level

The second likely suitable dataset is composed by the Nepalese Red Cross (see Nepalese Red Cross Society, 2015a). This data was gathered during an Initial Rapid Assessment (IRA) that defines the number of deaths, injured, affected households, affected males and females, displaced households, displaced people, completely damaged houses and partially damaged houses. All were reported in the first week after the main shock. Every chapter office had sent out local volunteers to fill in templates (see Appendix I – IRA Assessment Template), which were sent back to the Red Cross headquarters in Kathmandu. Volunteers derived the numbers either from estimating, counting or talking to local people (Knight, 2016). IRA's are a common way of assessing needs in the humanitarian sector.

In total there are numbers reported for 1,017 VDCs located in the 25 most affected districts. Unfortunately, 400 observations in the dataset do not include an identification code like a p-code (see Section 4.3) or governmental code, and therefore cannot be located automatically. The p-coding of the document was done automatically by the British Red Cross by means of an algorithm searching for matching letters. Most likely these missing labels are caused by the fact that administrative borders in Nepal are rather dynamic (as explained in Challenge #3). Also some local volunteers had reported municipality names instead of VDC names (Knight, 2016). The VDCs for which no p-code was assigned are left out of the analysis. The remaining 617 observations are within 16 (heavily) affected districts (Figure 5.2a-b).



a)



b)

Figure 5.2a-b: Number of completely damaged houses (a) and partially damaged houses (b) (data source: Nepalese Red Cross Society, 2015a).

To get a better understanding of the relative impact all absolute numbers in the dataset can be converted to relative numbers (Figure 5.3) by dividing them by the total number of households as reported in the 2011 National Population Census (data on the number of houses is not available).

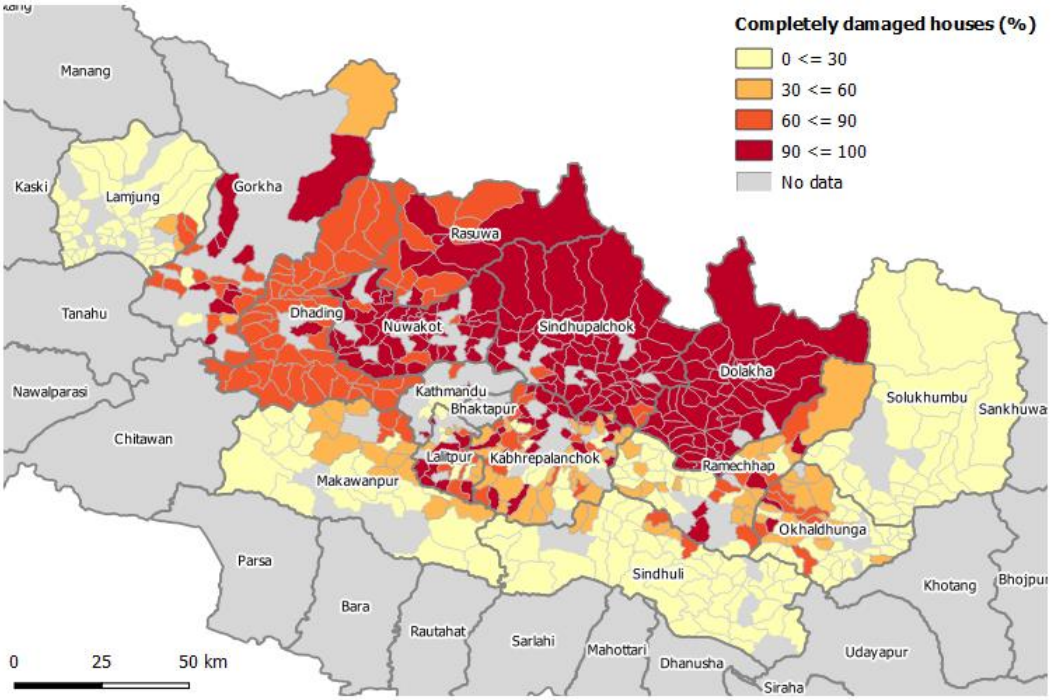


Figure 5.3: Percentage of completely damaged houses (data sources: Nepalese Red Cross Society, 2015a; National Population Census Nepal, 2011).

It is important to mention that while calculating these relative numbers some abnormalities appeared:

- For several districts the number of displaced households divided by total number of households

gave exactly the same output as the number of fully destroyed houses divided by the total number of households, indicating that somewhere in the data gathering process one measure was copied from the other. Nevertheless, it is likely that indeed the number of displaced households closely resembled the number of fully destroyed houses in reality.

- In the district of Dhading all calculations resulted in equal numbers for nearly all VDCs under a category: 100% of households affected, 80% of households displaced, 65% of houses completely damaged and 15% of houses partially damaged. Therefore it is likely that the 47 VDCs in this district were not individually assessed.
- In the district of Sindupalchwok nearly all VDCs (25 out of 28) give an outcome of exactly 100% when the number of affected households was divided by the total number of households, again indicating that no individual assessment of this measure actually took place.
- 114 of the 612 calculated numbers for the percentage of completely damaged houses reached above 100%. The highest resulting number was 720%. However, errors do not necessarily stem from the IRA dataset. Other possible sources of uncertainties are that the number of households as reported in the census is not a good representation of the number of households in reality, since only officially registered households are included in the census. Also the census data are four years older than the assessment data. Between 2011 and 2015 the total population of Nepal increased with 4.9% (The World Bank, 2017).

Comparing Datasets

An advantage of the first dataset, reporting damaged houses on a district level, is that the multiple measurements over time create the opportunity to validate the model on post-aftershock situation. A downside of this dataset is that it reports damages on the relatively coarse district level. As a result there are only 14 to 30 observations of real interest. In the end, the model output is more valuable if it can correctly estimate variations between the most affected entities.

With regard to the IRA dataset, the main drawback is the abnormalities such as the many duplicate values which could negatively influence the predictive power of a model. Nevertheless, because of the higher number of observations and the lower level of aggregation, training of the model on these data can give a more information-rich output and better distinguish between different amounts of damage in the highly affected areas. The downside of the low administrative level of the data is that it could be harder to collect data for predictor variables on the same level, since not much data is reported on this level in Nepal. Finally, the fact that commonness of this type of assessment increases possibilities for further model training. For these reasons the IRA dataset is judged to suit the study objectives best. It is recognized that the selected dataset is far from perfect, nor completely objective, which stresses the importance of careful data exploration. However, the challenge in this study is to see how good of a prediction can be made based on the best available open data.

Final Response Variable

Different response variables concerning damage to residential buildings can be derived from the selected dataset. The assessment distinguishes between completely and partially damaged houses. The former is expected to be more suitable for model training, as it is generally less subjective as there is no agreed notion of when a building is partially damaged. Therefore the number defined under this category is very dependent on the observer's interpretation and hence not that objective. During development of the typhoon PIM, mentioned in Section 3.5, it was also observed that partially damaged buildings are harder to predict (510,2016). Also in a Post Disaster Needs Assessment by the Nepalese National Planning Commission (2015) it is mentioned that there is indeed no uniform criterion for

partially damaged. Also, the number of partially damaged houses in itself does not say a lot. Two different entities might both have a relatively low number of partially damaged houses. Despite this, one of them may be heavily affected as nearly all houses got completely destroyed (leaving hardly any houses as only partially damaged), while in the other entity not a single house got completely destroyed. Nonetheless, 'partially damaged' houses are indeed also of interest to emergency relief suppliers. Since the number of partially damage does hold information a solution could be to enrich the number of completely damaged houses with the number of partially damaged houses to create a House Damage Factor, hereafter referred to as HDF (Figure 5.4). The applied equation of this HDF is:

$$HDF = (0.75 \cdot CDH) + (0.25 \cdot PDH)$$

HDF: house damage factor
CDH: number of completely damaged houses in one administrative area
PDH: number of partially damaged houses in one administrative area

Because the number of partially damaged houses is an indistinctive measure which does not say much about relative impact in itself, as explained above, it is assigned a relatively low weight of 25%. This division remains rather arbitrary, but gave better model prediction results in comparison to compositions of '*0.66 · CDH + 0.33 · PDH*' or '*1.00 · CDH + 0.25 · PDH*'.

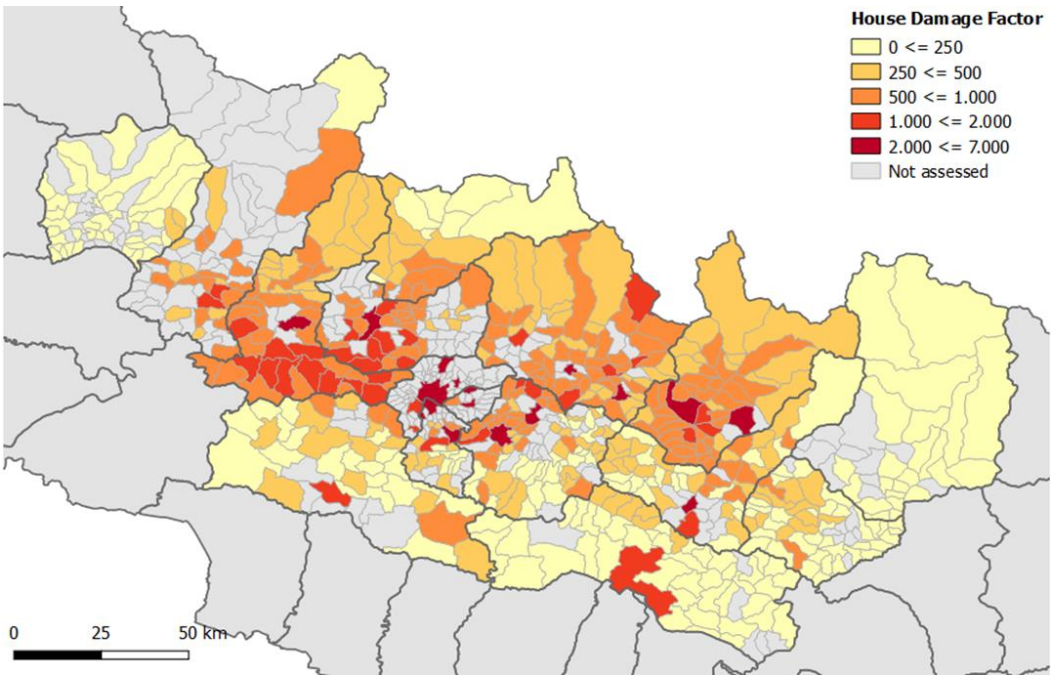


Figure 5.4: House Damage Factor (own calculation based on data from: Nepalese Red Cross Society, 2015a).

To summarize, three possible response variables for the model have been selected: the absolute number of completely damaged houses per VDC, the percentage of completely damaged houses per VDC and the HDF per VDC. Both predictive accuracy and usability will determine which of these three variables will be selected for the advised for an Earthquake PIM predicting for any place on earth.

5.2 Candidate Predictor Variables

Q2: What variables, derived from openly available data, are candidate predictors of the defined response variable?

In order to answer the second research question another online search was performed, focussing on openly available datasets that could serve to predict damage to residential buildings caused by the Gorkha earthquake. As was derived from the analysis of existing models (Section 3.3.), these datasets should be part of one of four categories: hazard related variables, exposure related variables, physical vulnerability related variables or socio-economical vulnerability related variables.

5.2.1 Hazard Predictors

The hazard related predictor variables are meant to represent the intensity of ground motions during an earthquake. An important requirement for this variable is that its data becomes openly available soon after an initial shock. This enables dissemination of the first PIM output within 24 hours. Other requirements are that updates on the data should be available in case of aftershocks in order for the model to improve its output and that the resolution is low enough to distinguish between different ground motion intensity levels for different VDCs.

Mostly inspired by other earthquake damage prediction models the search for a suitable hazard related predictor quickly turned to the US Geological Survey (USGS) ShakeMaps (see: http://earthquake.usgs.gov/earthquakes/shakemap/). These maps provide near-real-time maps of ground motion and shaking intensity following significant earthquakes (USGS, 2016b). They are especially valuable for this application since they are rapidly available (within +/- 10 minutes) and are constantly updated. Shaking intensity is an indicator of impact of ground motion on built environment and is expressed as the Macroseismic Intensity (MI), which is usually the Mean Mercalli Intensity (USGS, 2017). The MI measures different characteristics of an earthquake than magnitude. Using earthquake magnitude to estimate intensity can be misleading because magnitude only measures the energy released at the source of the earthquake, which does not say much about the intensity of ground shaking at the surface without additional information about environmental factors (Jaiswal et al., 2009). In contrast, the MI expresses the effect of the earthquake on the earth's surface. Rather than focusing on the magnitude and epicenter of an earthquake, it displays a range of ground shaking levels at sites throughout the region depending on distance from the earthquake, the rock and soil conditions at sites, and variations in the propagation of seismic waves from the earthquake due to complexities in the structure of the earth's crust (USGS, 2016a). All complex processes including parameters such as peak ground acceleration and peak ground velocity are included in an automated assessment producing one composite shaking intensity measure, which is widely used across the globe by federal, state, and local organizations for post-earthquake response and recovery (USGS, 2016b). ShakeMaps' in- and output datasets are also immediately open and free of charge available in a format ranging from JSON to GeoTIFF to SHP. Figure 5.5 shows the ShakeMap of the initial shock of the Gorkha earthquake. As visible, the map extent is not adapted to country borders and does not cover the full spatial extent of Nepal.

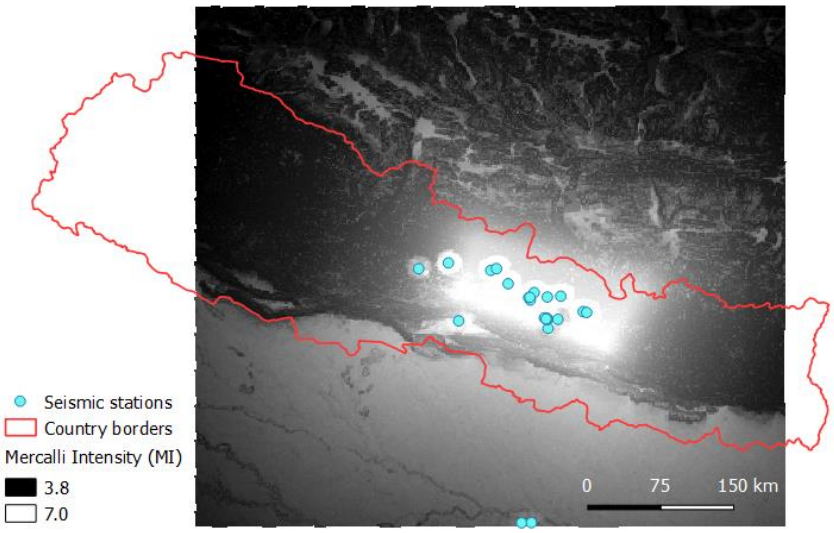


Figure 5.5: ShakeMap Gorkha earthquake 25 April 2015, WGS84 (data sources: (USGS, 2015)).

Another feature that stands out from the map is the concentration of higher MI values around seismic measurement stations. Station locations are the best indicator of where the map is most accurate: near seismic stations the shaking is well constrained by data; far from such stations, the shaking is estimated using standard seismological inferences and interpolation (USGS, 2017). The accompanying uncertainty map (Figure 5.6) shows a similar pattern, with higher certainty near measurement stations. It appears as if higher certainty coincides with higher MI values. Indicating that in areas with lower uncertainty the MI values are likelier underestimated than overestimated. Where the ratio is 1.0 (meaning the ShakeMap is purely predictive), the map is coloured light grey. Where the ratio is greater than 1.0 (meaning that the ShakeMap uncertainty is high because of unknown fault geometry), the map shades toward dark red, and where the uncertainty is less than 1.0 (because the presence of data decreases the uncertainty), the map shades toward dark blue (USGS, 2017). However, the challenge in this study is to explore the possibilities of developing a well estimating algorithm based on the best possible event-specific data that is rapidly available. As with many rapidly available descriptive data on sudden onset natural hazards this means dealing with uncertainties stemming from estimations (in this case interpolation).

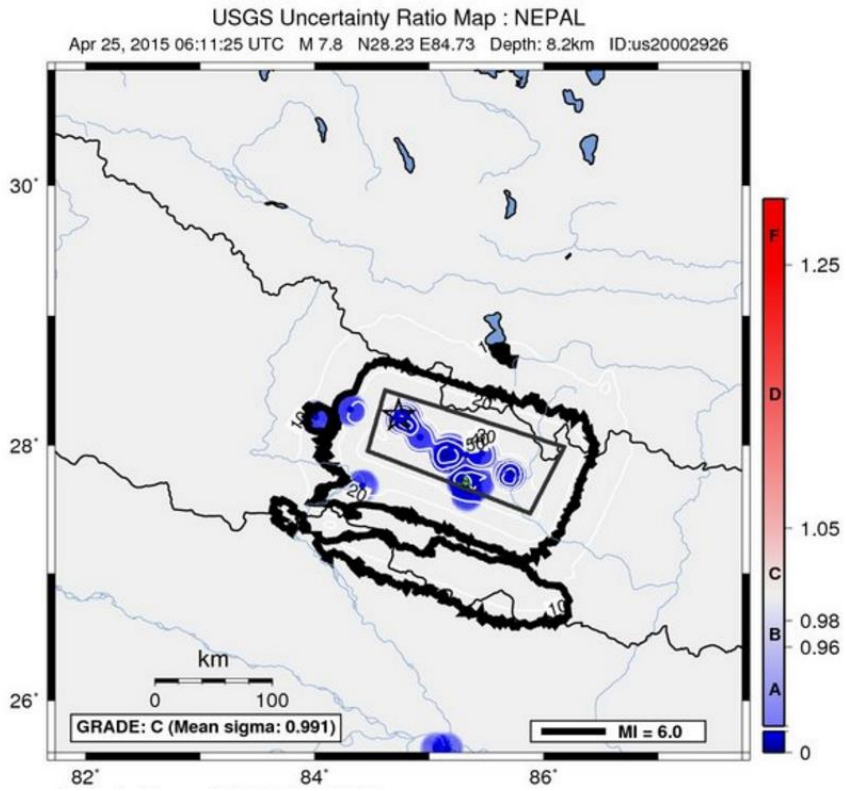


Figure 5.6: Uncertainty map Gorkha earthquake 25 April 2015 (source: USGS, 2015).

All USGS ShakeMap products are provided in WGS84 format. Therefore all other predictor variable data will be converted to the same Coordinate Reference System. The USGS ShakeMap is the only hazard related candidate predictor that will be included in the model. Based on the raster ShakeMap mean MI values can be calculated for each VDC, resulting in the values displayed in Figure 5.7.

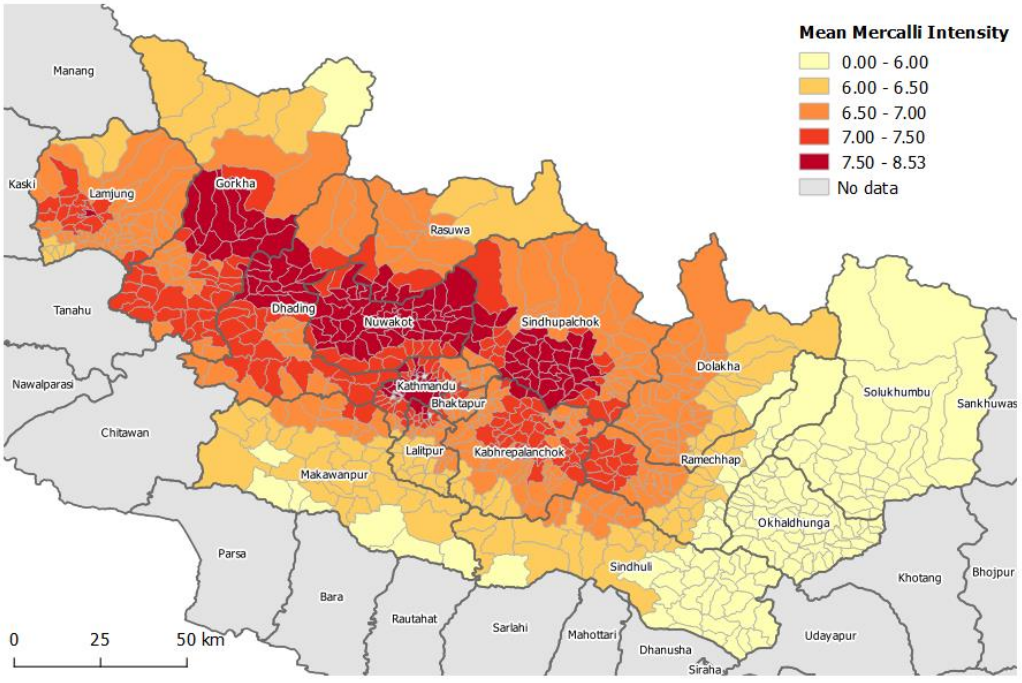


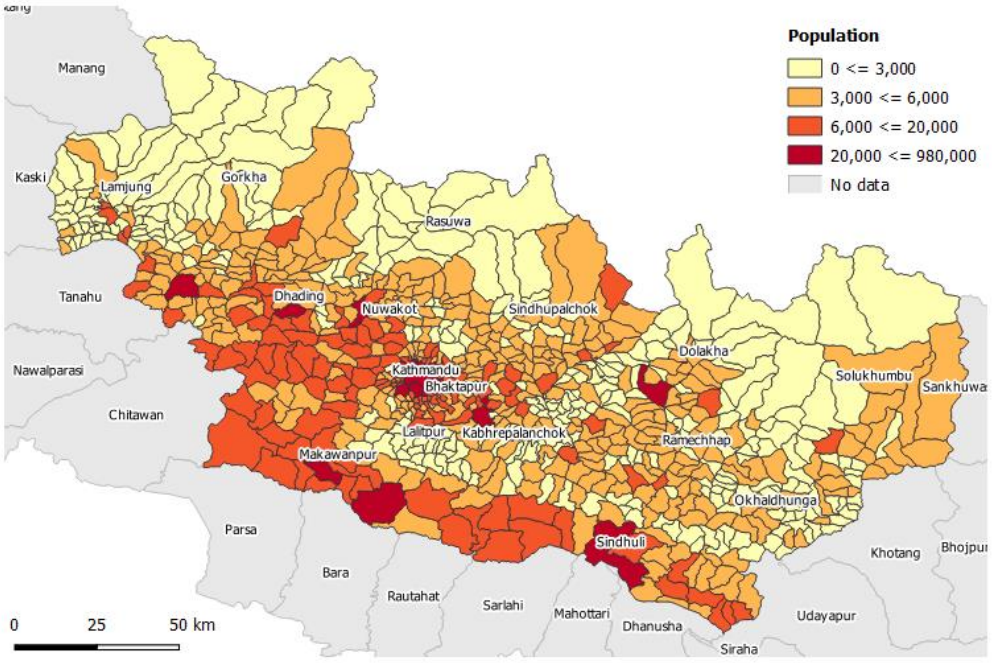
Figure 5.7: Mean Macroseismic Intensity per VDC in the study area (data source: USGS, 2015).

5.2.2 Exposure Predictors

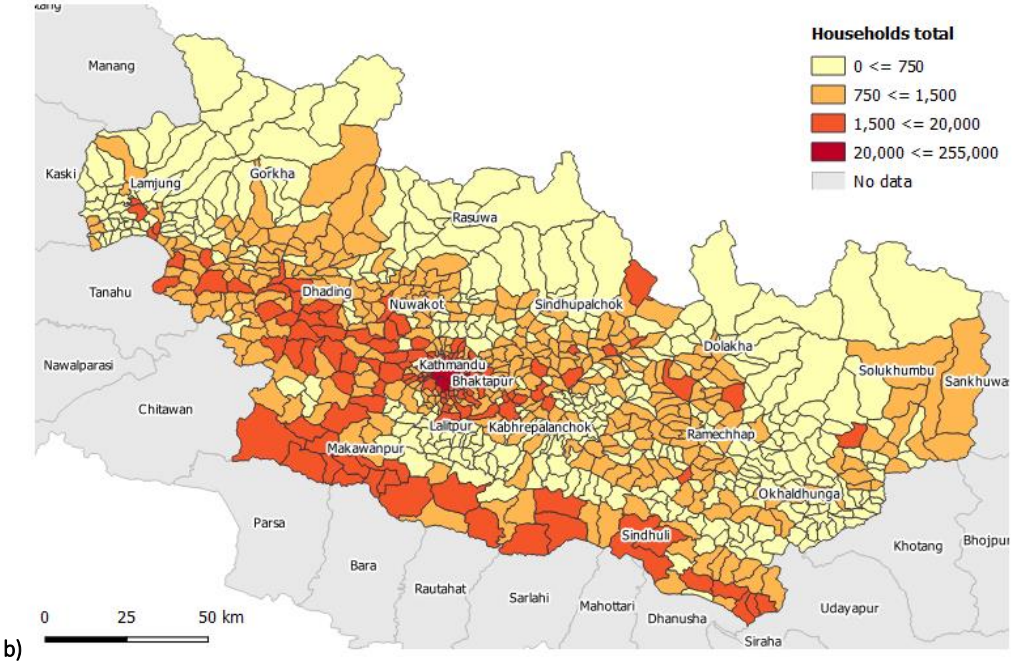
Exposure refers to the inventory of elements in an area in which hazard events may occur (Cardona et al.). The exposure variable should thus be a quantification of the elements at risk. Since the selected response variable is the absolute or relative number of completely damaged houses or factor of this, the elements at risk, and thus subject of exposure, are residential buildings. The number of residential buildings per VDC would be the preferred exposure variable. However, there are no openly available data records of the number of residential buildings, or buildings in general for that matter, on VDC level in Nepal. Alternatives that likely have a similar ratio to the amount of residential buildings per VDC are: the total population of VDCs derived from the 2011 National Population Census, the number of households per VDC from the 2011 census or the number of buildings per VDC as derived from OpenStreetMap (OSM).

Due to the earlier mentioned post-event mapping efforts the OSM building data for the most heavily affected areas is very accurate. However, for some of the sixteen districts of the IRA many buildings are likely not mapped. A calculation of a buildings-to-people ratio (*population / number of OSM buildings*) for each district showed that while eleven of the districts returned a ratio below 1:5 (which seems reasonable), the other five districts had ratios between 1:7 and 1:11, while the average household size ranges between 4.0 and 5.1 (Nepal Central Bureau of Statistics, 2012). The OSM data is thus not complete enough to form a suitable exposure variable. Another exposure indicator has to be defined for now.

This leaves either the population or number of households as possible exposure variables (Figure 5.8a-b). Both maps show that middle and Southern parts of the study area are most populated and that the Northern VDCs, higher in the Himalaya Mountains, are less populated. Both variables will be included in the final predictor variable selection process. The one with the best influence on the predictive accuracy of the model will be selected.



a)



b)

Figure 5.8a-b: absolute number of people (a) and households (b) in IRA districts Nepal (data source: Nepal Central Bureau of Statistics, 2012).

5.2.3 Physical Vulnerability Predictors

The physical vulnerability of the elements at risk, residential buildings, is viewed in two perspectives. On the one hand it is influenced by the building quality of the structures, and on the other hand it is influenced by the occurrence of secondary hazards.

Building Quality

Regarding the former, the National Population and Housing Census 2011 of Nepal provides information on building material by making a distinction between five different types of foundation materials, six different types of wall materials and seven types of roof materials. It reports for every VDC the amount of households within each category. Because it is likely that separate building material variables will correlate, it might be necessary for the LM to construct composite variables. In a Post Disaster Needs Assessment by the Nepalese National Planning Commission (2015) a distinction is made between four main building types in the affected area based on their vertical and lateral load bearing systems, these are:

1. Low-strength masonry buildings
2. Cement-mortared masonry buildings
3. Reinforced concrete frame with refill
4. Wood and bamboo buildings

For each category a description of construction materials of foundation, walls and roof is provided (see Appendix II). In the same report an overview of the earthquake caused damage per building type is presented (Figure 5.9).

Typology of Buildings	Fully Collapsed or Beyond Repairs	Partially Damaged (can be repaired/ retrofitted)
Low strength masonry	474,025 (95%)	173,867 (67.7%)
Cement based masonry	18,214 (3.7%)	65,859 (25.6%)
Reinforced concrete frame	6,613 (1.7%)	16,971 (6.7%)
Total	**498,852**	**256,697**

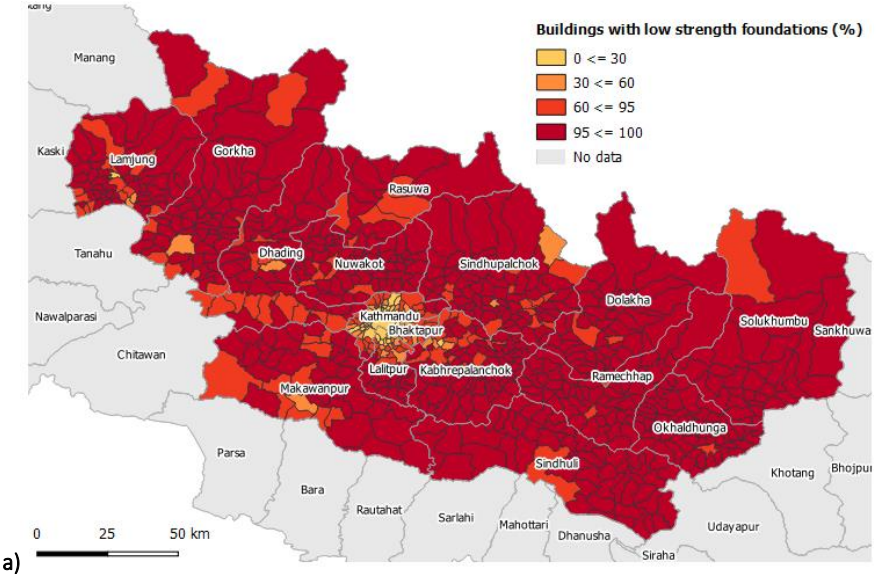
Figure 5.9: Building types and damage caused by Gorkha earthquake (source: Government of Nepal National Planning Commission, 2015).

As the high percentage of damage to low strength masonry buildings already shows, the writers confirm that "the seismic capacity of these buildings is very low, limited by the integrity of structural components and strength of walls and lack of elements tying the structure together (ring beams at wall or roof level). Vertical and horizontal wooden elements are sometimes embedded in walls, providing some level of earthquake resistance, but this is very uncommon" (Government of Nepal National Planning Commission, 2015).

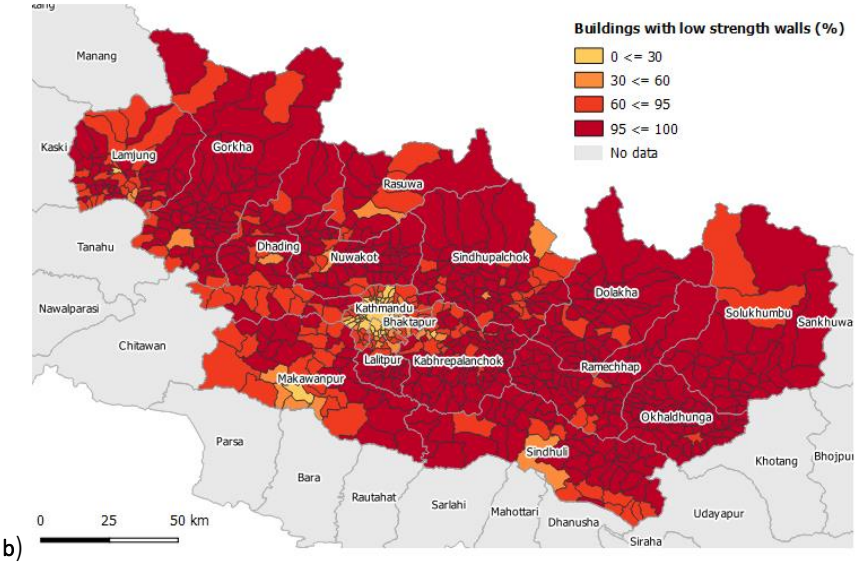
Based on the description of construction materials used in low strength masonry buildings four groups were created from the 18 separate material groups:

1. Low strength foundations; including mud bonded and wooden pillar foundations, excluding cement bonded and RCC pillar foundations
2. Low strength walls; including mud bonded, wooden, bamboo and unbaked brick walls, excluding cement bonded walls
3. Low strength roofs; including thatch, tile, wooden and mud roofs, excluding galvanized iron and RCC roofs.

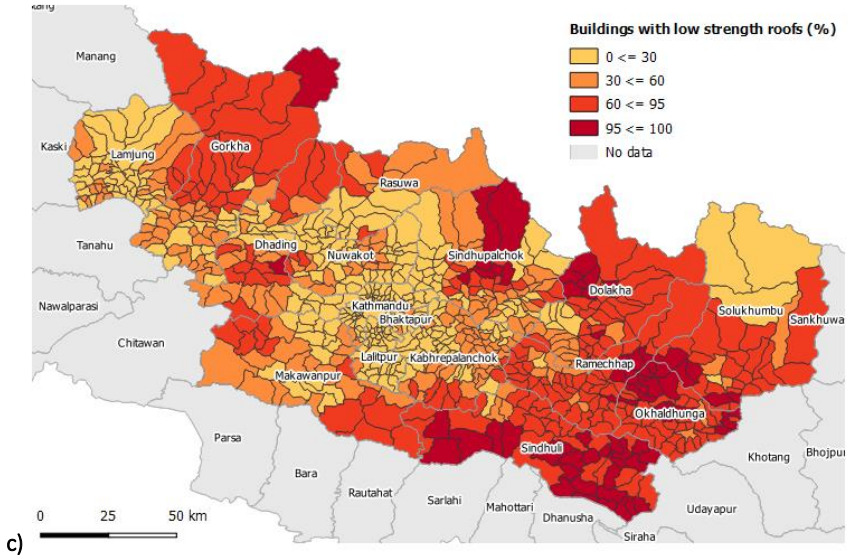
These three new features are expressed as the percentage of households with low strength foundations/walls/roofs of the total amount of households for which the materials are defined. The category 'other' was included in the total, while the categories 'not stated' and 'unknown' were excluded from the total, as they cannot be said with certainty to be not one of the material in the low strength categories. It is expected that these three categories can give an insight into what part of a building (foundation, walls or roof) influence mostly whether or not buildings were completely destroyed in the Gorkha event.



a)



b)



c)

Figure 5.10a-c: Percentage of households living in buildings with low strength foundations (a), walls (b) and roofs (c) (Data source: Nepal Central Bureau of Statistics, 2012).

Secondary hazard susceptibility

To represent people's vulnerability to secondary hazard, landslide susceptibility is defined as a candidate predictor, mainly because of the relatively high occurrence of landslides in this region and because it was reported that landslides caused a lot of damage in the Gorkha earthquake (Government of Nepal National Planning Commission, 2015). Landslide susceptibility mapping is a complicated task. Most methods are applicable to one specific area only. No methodology for the assessment of worldwide earthquake induced landslide susceptibility exists. Therefore a proxy indicator will be included in the model training to represent seismic induced landslide susceptibility instead. There are many aspects that influence the occurrence and size of a landslide in case of seismic activity (soil type, vegetation mass, vegetation root strength, moisture, debris stiffness, see Walker & Shiels, 2013). However, one aspect that has a relative high influence and is a prerequisite for a landslide to occur in the first place is slope inclination. Slope maps can be derived from Digital Elevation Models (DEMs) which are openly available for nearly all regions on earth (for example at srtm.csi.cgiar.org). This availability is important having the global scope in mind.

To do a simple verification of using the slope map as a proxy for landslide susceptibility it is compared to Shapefiles of landslides that occurred as a result of the Gorkha earthquake (Figure 5.11). These landslide locations were mapped based on satellite imagery by staff at Durham University and the British Geological Survey (HDX, 2015). 5,578 landslides where mapped in total. The map shows that landslides mostly occurred in the middle and Northern part of the district where slope values are higher (often more than 30%) than in the Southern part.

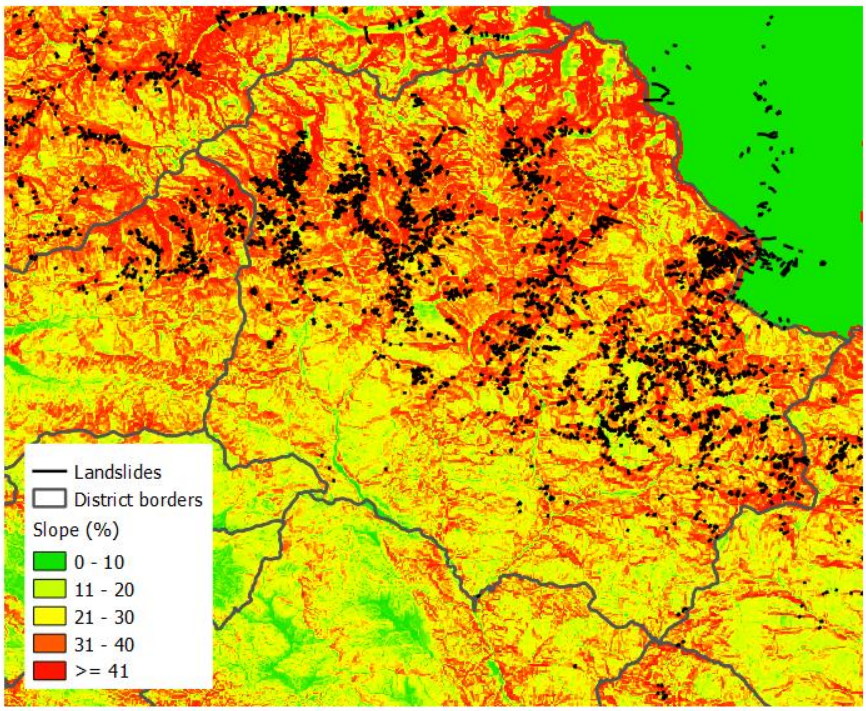


Fig 5.11: Slope and landslides triggered by the Gorkha earthquake in Sindupalchwok district
(DEM data source: (CGIAR Consortium for Spatial Information, 2004), Landslides data source: HDX, 2015).

A 90 meter resolution DEM raster file is derived from the SRTM (Shuttle Radar Topography Mission) Digital Elevation GeoPortal. A pre-defined algorithm is applied to calculate a slope value for each cell by assigning it the average rate of change in value from the cell itself to its eight neighbouring cells (ESRI, 2017).

Since the slope map in first instance is a continuous raster file the issue of raster generalization (see Challenge #4) arises. As these considerations are part of research questions four they are dealt with in Section 5.4 Raster Generalization.

5.2.4 Socio-economic Vulnerability Predictors

As explained in Chapter 3, socio-economical vulnerability of those affected by the earthquake can influence their need of humanitarian aid. However, as the selection of a suitable response variable resulted in a variable that quantifies structural damages the situation changes. This direct physical impact cannot be explained by the more social related variables such as house ownership or female/child headed households which describe characteristics of people rather than buildings. While such aspects do not in itself influence damage to a house, they do influence the capacity of a household or family to find or rebuild new shelter or to retrieve food and water supplies. In particular aspects that resemble poverty or (economic) development, can be related to the physical quality of the buildings people live in, and thus also structural damages. Communities living in economically more developed areas might enjoy better (monitoring of) building standards.

No direct or uniform measure of poverty on administrative level 4 is openly available for Nepal. Indirectly however, several variables can serve as poverty proxies. From the National Housing and Population Census 2011 five variables resembling developmental-economic vulnerability are derived:

1. Toilet type: the census defines for each household the type of toilet, distinguishing between 'no toilet facility', 'flush toilet' and ordinary toilet'. One variable is derived from these numbers by calculating the percentage of households in each VDC without a toilet facility.
2. Drinking water source: for each household the main source of drinking water is reported, distinguishing between 'tap/piped water', 'tubewell/handpump', 'covered well', 'uncovered well', 'spout water' and 'river/stream'. The variable derived from this included the percentage of households deriving their drinking water from tap/piped water, since people with access to this water source are expected to live in a higher level of economic development. Also tap/piped water is often present in houses with a higher building quality. Also, the World Health Organization classifies tap or piped water as an improved drinking water source (World Health Organization, 2011).
3. Literacy: in the census literacy rates for the population aged above 5 years in each VDC are presented. Literacy is included as a candidate predictor as a possible indicator of both poverty and development. Also literacy could be related to people's abilities to improve seismic capacity of their houses.
4. School attendance: the percentage of the population between 5 to 25 years who are currently going to school. School attendance is generally accepted as a poverty indicator.
5. Household size: the average household size in a VDC is considered to be a proxy indicator for poverty. The relationship between household size and poverty is challenged, but not proven wrong (see: Lanjouw & Ravallion, 1994). Variable selection methods will show whether or not it should be included in the model.

The maps in Figure 5.12 visualize the spatial distribution of the five socio-economic vulnerability variables. To a certain extent all maps show similar patterns, with higher concentrations in Northern Gorkha and Rasuwa, the Mid-Southern area and the North-Western part. This stresses the importance of checking for multicollinearity. In the end, only the best damage predictor(s) have to be included in the model.

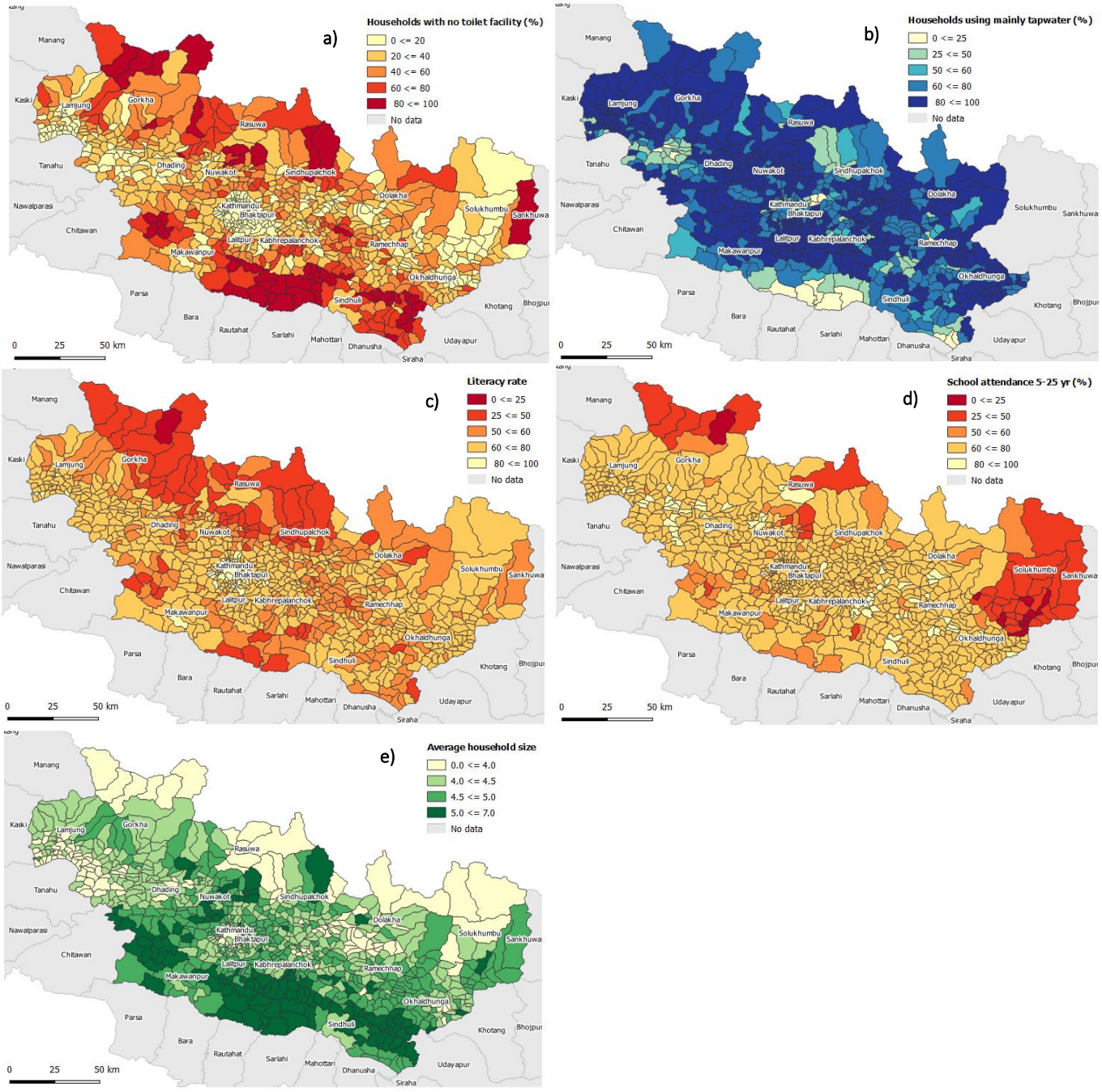


Figure 5.12a-e: Households without a toilet (a), households using mainly tap water (b), average literacy rate (c), school attendance (d), and average household size (e) (Data source: Nepal National Population and Housing Census 2011).

5.3 Model Fitting

Q3: Based on a multivariate linear regression and a random forest regression model, which candidate predictors can together make the best possible prediction of the defined response variable, and what is their (relative) importance?

This research question is answered in several sequential steps. Section 5.3.1 concerns data exploration, checking for and handling missing values, skewness in frequency distributions and multicollinearity. Section 5.3.2 concerns variable selection for the LM by means of automated selection procedures.

Hereafter, Section 5.3.3 discusses the fitting of the RF model, which is a different procedure. The fit of different RF models will be discussed by interpreting error measures. All best performing models are selected. The results of this section concern only the fit of the models to the training data (in-sample validation). No statements about the predictive accuracy of the models can be made until after the out-of-sample validation (Section 5.5).

5.3.1 Data Exploration

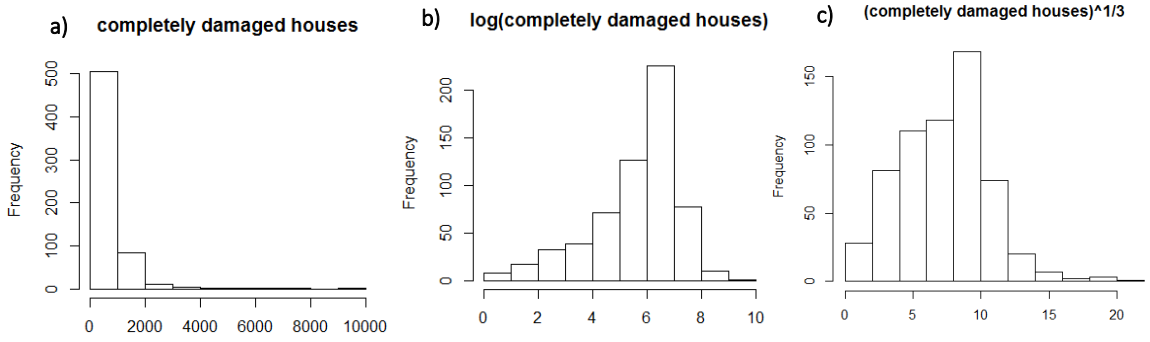
The complete dataset, including training and testing set, contains 612 observations, each representing a separate VDC (level 4 administrative geographical entity). There are three possible dependent variables, which are all continuous, and 13 candidate predictor variables. Descriptive statistics of all variables such as maximum, minimum and mean can be found in Appendix III. This data exploration section applies to the complete dataset. From Section 5.3.2 to Section 5.3.3 only the training dataset is part of the analysis.

Missing Values

One requirement for the LM is that there are no missing values in either dependent or independent variables. The possible response variables are 1) the absolute number of completely damaged houses per VDC, 2) the relative number of completely damaged houses per VDC and 3) the House Damage Factor (HDF). As already mentioned in Section 5.1 of the 1,017 VDCs in this dataset 670 have a p-code assigned to them. Of these 670 VDCs the number of completely damaged houses is defined for 612 VDCs. The number of both completely and partially damaged houses is defined for 517 VDCs. Since most of the candidate predictor variables were derived from national census data (or calculations based on census data) no values were missing here. The slope and Macroseismic Intensity variables were derived from raster data files that overlapped all 16 districts of interest. Also here, no values are missing.

Frequency Distributions and Outliers

Normally distributed dependent and independent variables in a LM or RF model can improve model fit and predictive accuracy. Also outliers (assumed to be observations outside 1.5 times the interquartile range above the upper quartile or below the lower quartile) are important to identify, since they can have implications for both the error measure and the model fit. The histogram in Figure 5.13a shows that the absolute number of completely damaged houses is very right skewed, with outliers on the right side.



Figures 5.13a-c – Frequency distributions showing the number of VDCs on the vertical axis and the value for the according variable on the horizontal axis. The according variables are a) the absolute number of completely damaged houses per VDC, b) the number of completely damaged houses per VDC logartihmic transformed and c) the number of completely damaged houses per VDC cube root transformed.

In 503 of the 612 VDCs less than 1,000 houses got completely damaged. In case of such negative skewness a common logarithmic ($log_{10}(x)$) or cube root transformation ($x^{1/3}$) is appropriate. Prior to

these transformations the seven VDCs for which a zero was reported are changed into a one. After the logarithmic transformation the distribution is slightly left skewed (Figure 5.13b). This transformation caused an increase in the number of outliers from 24 (Figure 5.14a) to 28 (Figure 5.14b, due to duplicate values this is not visible in the boxplot). The cube root transformation resulted in a slightly right skewed distribution (Figure 5.13c), but had only seven outliers (Figure 5.14c) and thus seemed most appropriate. There is no reason to assume that these outliers originate from incorrectly measured or repoted values, hence they will be part of the analysis as they are.

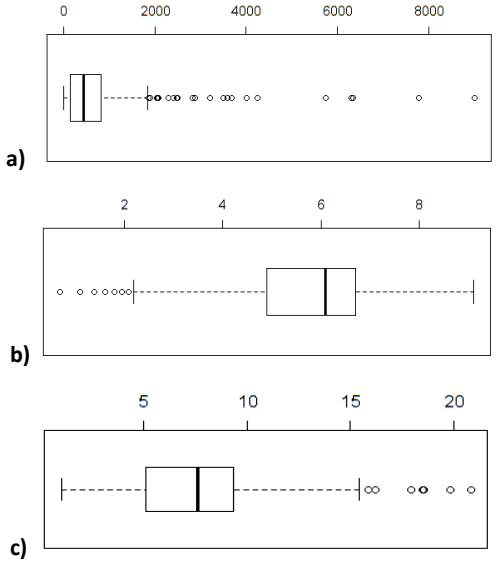


Figure 5.14a-c – Boxplot and outliers of a) completely damaged houses b) log of completely damaged houses and c) cube root of completely damaged houses.

The second possible dependent variable is the relative amount of completely damaged houses per VDC (Figure 5.15a). The distribution is right skew with seven high outliers. These are VDCs with a relative damage amount between 227% and 710%. After a logarithmic transformation the distribution is somewhat more normally distributed but does show left skewness (Figure 5.15b) and contains 38 outliers in the lower values. The milder cube root transformation resulted in a more normal distribution (Figure 5.15c) with only three outliers in the higher values.

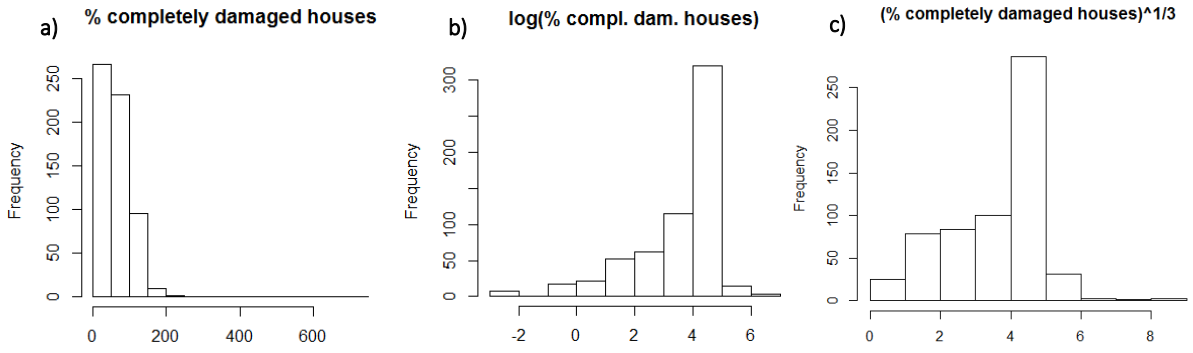


Figure 5.15a-c: Frequency distributions showing the number of VDCs on the vertical axis and the value for the according variable on the horizontal axis. The according variables are a) the percentage of completely damaged houses per VDC, b) the percentage of completely damaged houses logarithmic transformed and c) the percentage of completely damaged houses cube root transformed.

The third possible dependent variable is the composed Housing Damage Factor (HDF), combining data from the amount of both completely and partially damaged houses. Again the data is right skew distributed (Figure 5.16a), with 30 outliers (all above +/- 2,000). A logarithmic transformation gives a rather left skew distribution with 29 outliers (all below +/- 20) (Figure 5.16b). The cube root transformation of this variable (Figure 5.16c) returns a slightly right skew distribution with 13 outliers of observations above 2,780. For all possible response variables a cube root transformation gave the best result in terms of normality, hence these will be part of the further analysis (Section 5.3.2 - 5.3.3).

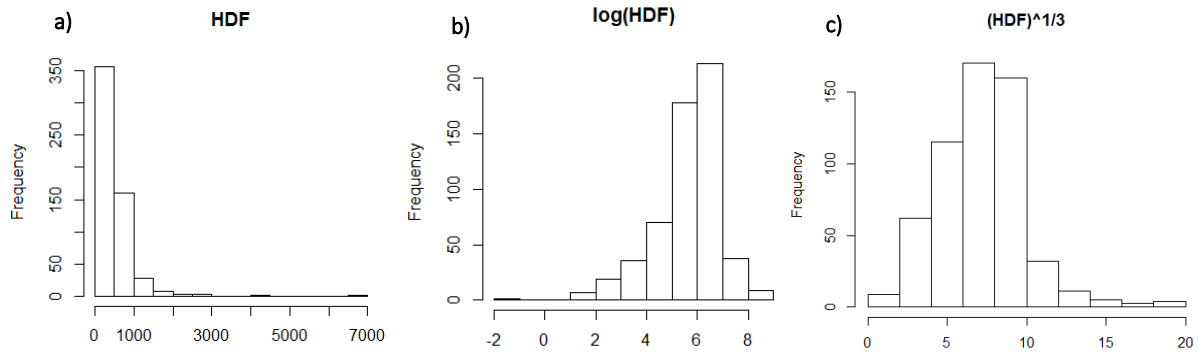


Figure 5.16a-c: Frequency distributions showing the number of VDCs on the vertical axis and the value for the according variable on the horizontal axis. The according variables are a) the house damage factor per VDC, b) the house damage factor per VDC logarithmic transformed and c) the house damage factor per VDC cube root transformed.

The frequency distributions of all candidate predictor variables are presented in Appendix IV. Many variables show right skewness, with a few high outliers. The variables that are quite normally distributed are household size, slope, Macroseismic intensity and literacy rate. Both population and the number of households are right skew, caused by a few metropolitan cities. Nearly all building material variables are left or right skew. This is again mostly caused by high outliers in VDCs with a relatively high population. The population variable has 38 outliers, which are VDCs with a population above 8,800. Transformations to more normal distribution are only applied in case this significantly improves the model's predictive accuracy. The distributions of the variables after applied transformation are presented in Appendix V.

Multicollinearity

Figure 5.17 presents a correlation plot containing all candidate predictor variables. The colour scale on the right side displays the Pearson's correlation coefficient (PCC). A value between -1 and +1 where -1 indicates a perfect negative correlation and +1 a perfect positive correlation. There is quite some positive correlation visible between population and building material variables and between the different building material variables. For the LM no strongly correlating variables can be part of the model simultaneously. Population and household correlate perfectly positive (PCC = 1.00). The total number of households will be eliminated from the model as population figures are more common.

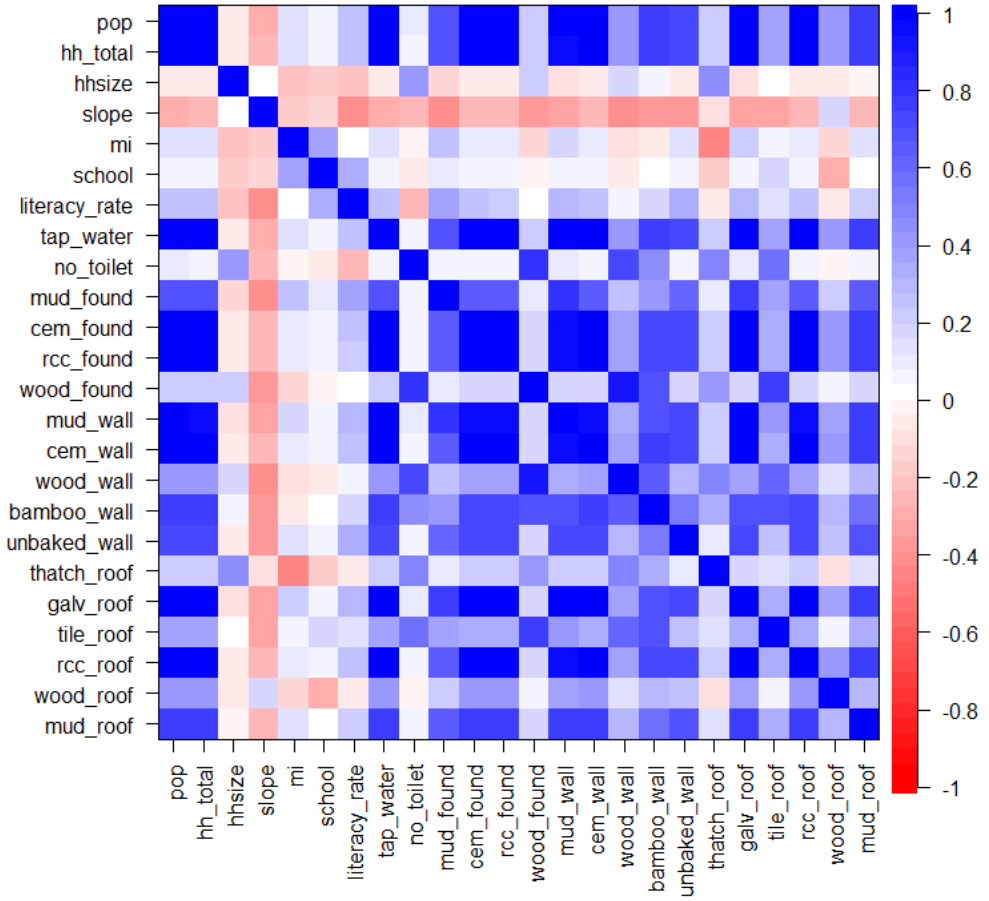


Figure 5.17 – Correlation plot with Pearson's coefficient for all candidate predictor variables. (explanation of variable codes: pop = population, hh_total = total number of households, hhsize = average household size, slope = mean slope value per VDC, mi = Macroseismic intensity, school = relative school attendance, literacy_rate = literacy rate, tap_water = percentage of households with tap water as their main source for drinking water, no_toilet = percentage of households without a toilet facility, mud_found = number of households with mud bonded bricks/stone foundations, cem_found = " cement bonded bricks/stone foundation, rcc_found = " RCC with pillar foundations, wood_found = " wooden pillar foundations, mud_wall = " mud bonded bricks/stone outer walls, cem_wall = " cement bonded bricks/stone outer walls, wood_wall = " wood/planks outer walls, bamboo_wall = " bamboo outer walls, unbaked_wall = " unbaked brick outer walls, thatch_roof = " thatch/straw roofs, galv_roof = " galvanized iron roofs, tile_roof = " tile/slate roofs, rcc_roof = " RCC roofs, wood_roof = " wood/planks roofs, mud_roof = " mud roofs)

Many building material variables correlate with each other. Therefore, they will be combined to three composite variables. These variables are composed as explained in Section 5.2.3 Physical Vulnerability Predictors. The new correlation plot is displayed in Figure 5.18. A strong correlation between low strength walls and foundations (PCC = 0.98) is still present. The low strength walls variable will not be part of the model. There is a medium strong correlation between the literacy rate and the number of households without a toilet (PCC = 0.67). Both variables will be included for automated variable selection. For the LM there will thus be ten candidate predictor variables included in the model.

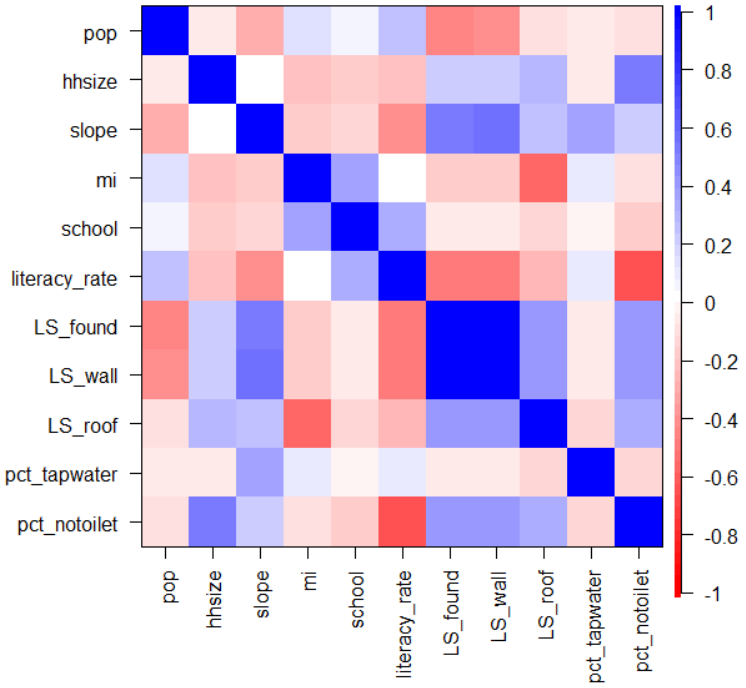


Figure 5.18: Correlation plot with Pearon's Coefficient for all candidate predictor variables, including composite building material variables. (explanation of variable codes: LS_found = total numer of households with a low strenght foundation, LS_wall = " low strength walls, LS_roof = " low strenght roof, pct_tapwater = percentage of households using tap water as their main source for drinking water, pct_notoilet = total numer of households without a toilet facility)

5.3.2 Linear Model Training

For the LM a pre-selection of the best predicting set of predictor variables took place. If the final LM performance estimates are to be unbiased, it must be tested on data that has not been used to tune any aspect of the model, including variable selection. Therefore, from this point onwards only the training data are part of the analysis. As explained in Section 4.4 Model Fitting the variables are selected using the regression subsets selection function in R, with the exhaustive search method. The function was applied to all three possible response variables and their transformed versions. Variables that correlated were not included simultaneously. All regression subset selection plots are presented in Appendix VI. Figure 5.19 presents the plot of that version of each response variable (not transformed, logarithmic transformed or cube root transformed) that achieved the highest $R^2_{adj.}$, which was the cube root transformed versions for all variables. The displayed models are not necessarily statistically significant.

For all three possible response variables the highest R^2_{adj} was achieved with the cube root transformation of the values. The horizontal axis of the plots shows the candidate predictor variables. The vertical axis shows the different R^2_{adj} values of the selected model. Each row represents one model, where a black box indicates that this variable is included in the model and a white box indicates that it is not. The greyscale corresponds to the R^2_{adj} square value.

All three displayed plots show a similar pattern to some extent. The percentage of houses with a low strength roof and the percentage of households using tap water appeared not to contribute anything to the prediction of any of the response variables. Also the household size is hardly ever included, and when it is included the prediction of the model does not decrease by leaving it out. The MI is part of nearly all models. The transformed population variable is also included in most models. Only when predicting the cube root transformation of the relative amount of completely damaged houses this variables is less often included. This is logical since the response variable was already adjusted to population size by dividing by the number of

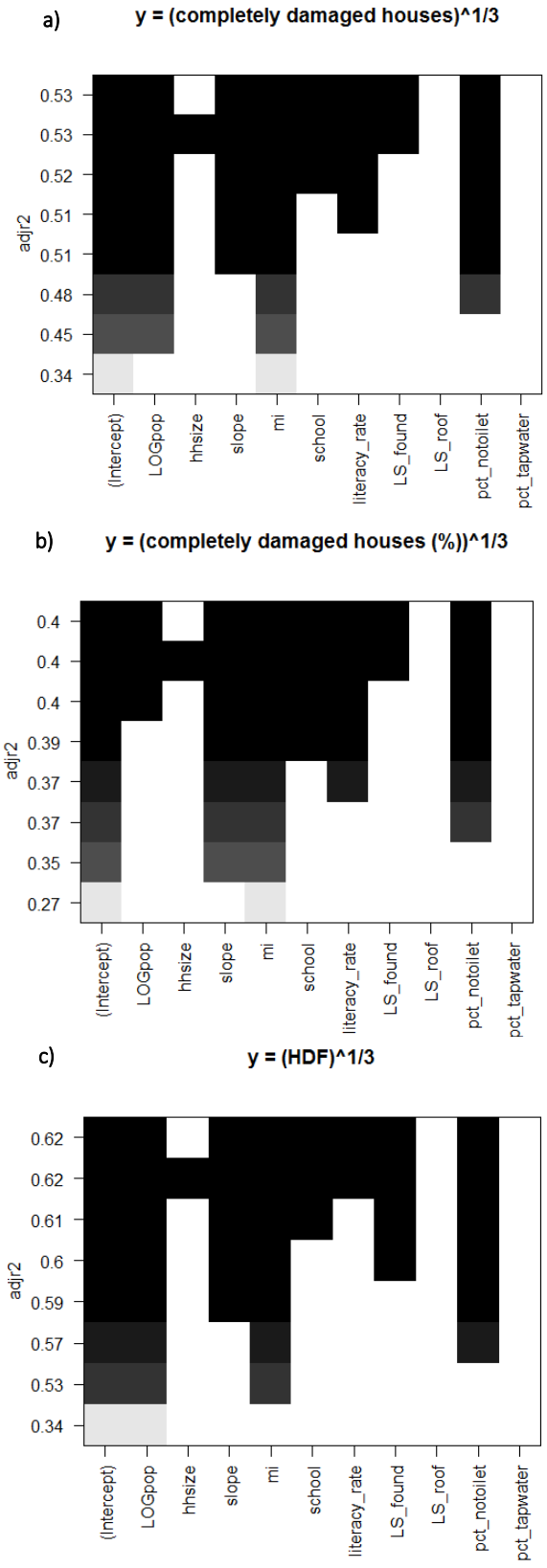


Figure 19a-c: Regression Subset Selection plots for each possible response variable.

damaged houses by total number of households. The third plot shows that with only the population and MI variable likely 53% of the variance in the response variable can be explained.

The highest R^2_{adj} is achieved by predicting the cube root transformation of the HDF by seven or eight predictor variables. Variables of all categories (exposure, hazard, physical vulnerability (both secondary hazard susceptibility and building quality) and socio-economic vulnerability) are selected as contributing factors. This model will be referred to as LM1.

```
Call:
lm(formula = cuberoot_hdf ~ LOGpop + slope + mi + school + literacy_rate +
    LS_found + pct_notoilet, data = train_v5)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9267 -0.9203  0.0594  1.0387  6.0502

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.52629    2.95073  -3.228 0.001366 **
LOGpop         1.73644    0.17743   9.786  < 2e-16 ***
slope          0.07662    0.02183   3.510 0.000508 ***
mi             1.13368    0.13225   8.572 3.59e-16 ***
school         4.01567    0.93347   4.302 2.21e-05 ***
literacy_rate -0.05986    0.01706  -3.509 0.000511 ***
LS_found      -4.99979    1.05523  -4.738 3.17e-06 ***
pct_notoilet  -2.91156    0.54183  -5.374 1.43e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.68 on 341 degrees of freedom
Multiple R-squared:  0.6315,    Adjusted R-squared:  0.624
F-statistic:  83.5 on 7 and 341 DF,  p-value: < 2.2e-16
```

Figure 5.20: LM1 summary. (explanation of codes: LOGpop = logarithmic transformation of total population, slope = mean slope value, mi = mean Macroseismic intensity, school = relative school attendance, literacy_rate = literacy rate (0/100), LS_found = percentage of households with a low strength foundation, pct_notoilet = percentage of households without a toilet facility)

The according R^2_{adj} value indicates that 62.4% of the variance in the response variable can be explained by the selected predictors (which were all significant) (Figure 5.20). The normal R^2 (not adjusted) is 0.63. In an equation this LM takes the following form.

$$HDF = -9.53 + 1.13\,X_1 + 1.74\,X_2 + 0.07\,X_3 - 5.00\,_{X4} + 4.01\,X_5 - 0.06\,X_6 - 2.91\,X_7$$

HDF	*house damage factor = ((compl. dam. houses 0.75) + (part. dam. houses 0.25))$^{1/3}$*
X_1	*mi, mean Macroseismic Intensity*
X_2	*LOGpop, population^{log}*
X_3	*slope, mean slope (%)*
X_4	*LS_found, buildings with a mud or wooden foundation (%)*
X_5	*school, school attendance 5-25 year old's (%)*
X_6	*literacy_rate, literacy rate (0/100)*
X_7	*pct_notoilet, households without a toilet facility (%)*

The coefficient assigned to each predictor represents the mean change in the response variable for one unit of change in the predictor variable. Their signs (being positive or negative) can be interpreted accordingly. As expected, an increased MI (X_1) relates to an increased number of houses damaged. Secondly, a higher population (X_2) associates with more absolute damage. Thirdly, VDCs with higher average slope values (X_3) experienced more damage to houses.

For the second physical vulnerability variable, the relative amount of buildings with a low strength foundation, the coefficient is negative. Meaning that for the Gorkha case VDCs with relatively more good quality foundations experienced more damage. The maps displaying the spatial distribution of low strength foundations (Figure 5.10a) and the HDF spread (Figure 5.4) confirm this relationship. A possible explanation for this pattern could be that big cities are more populated, thus having a higher absolute

amount of houses damaged. At the same time these urban areas also have a relatively higher amount of good quality buildings than rural areas. Also, it is possible that the relative amount of houses with a low strength foundation is higher in reality, but that these are not reported in the census since they are often unregistered.

Concerning the socio-economic vulnerability variables, a lower school attendance rate and a higher literacy rate correlated with heavier damage as expected. The percentage of households without a toilet also showed a coefficient sign opposite to what was expected. The damage was higher in VDCs where more households had a toilet facility. Possibly this is due to similar reasons of more developed building quality of registered households in urban areas.

The absolute value of the t–statistic for each model parameter (Figure 5.21) indicates the relative importance of predictor variables in the model. With a t-test the null hypothesis that the coefficient associated with a predictor variable is not significant. The further away the t-value is from zero, the likelier it is that the null hypothesis should be rejected.

	Overall
LOGpop	9.786446
slope	3.510388
mi	8.572127
school	4.301883
literacy_rate	3.508689
LS_found	4.738123
pct_notoilet	5.373559

Figure 5.21: absolute t-statistics predictors LM1.

The values show that in LM1 the population and MI are the most important explaining variables, indicating that for the modelled case exposure and hazard related variables are most important in explaining damage. Hereafter follow consecutively toilet absence, low strength foundations, school attendance, slope and literacy rate.

For the multivariate LM to be valid it should meet the assumptions of: linearity of residuals, heteroscedasticity of residuals and normality of residuals. Figure 5.22a plots the residuals against the predicted values. These displayed data points concern only the training dataset.

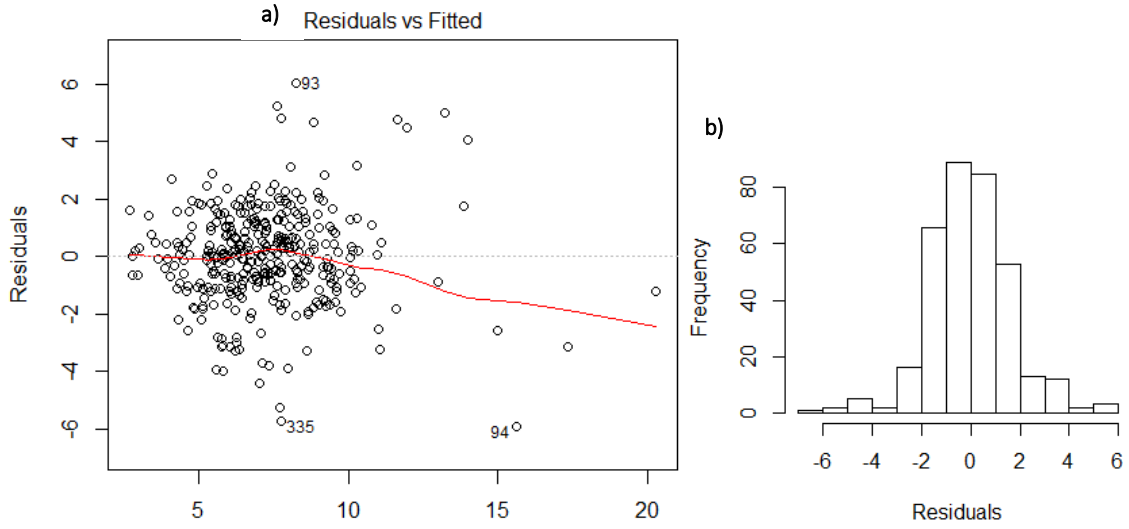


Figure 5.22a-b: Residuals to predicted values (x-axis) of LM1 (a) and frequency distribution of the residuals of LM1 (b).

Concerning the linearity assumption, the residuals are spread more or less randomly around the 0-line. This suggests that the assumption that the relationship is linear is reasonable. The residuals do get a bit larger as the damage increases, but the assumption of heteroscedasticity is sufficiently met. Concerning normality of the residuals, it can be noted that as the damage gets higher ($HDF^{1/3} > 10$) the model somewhat overestimates the damage, as the real values are below the regression line. The plot also shows that for this part of the LM the prediction relies on very few observations. When the damage is lower ($HDF^{1/3} < 10$) the fitted residuals line is very close to zero.

5.3.3 Random Forest Model Training

The RF model relies on less assumptions than the LM. The only assumption is, as with any other model that the sampling is representative for reality. The algorithm can handle covariance among independent variables and non-linearity. Several RF models are run, testing with the different possible response variables. Composite variables and the variables they are composed of were not included simultaneously. The RF algorithm was run for each possible response variable and its transformed versions. Below the results of the two RF models with the best fit are presented.

Random Forest Model 1: House Damage Factor

The model with the highest explained variance and R^2 score was the one predicting the cube root transformation of the HDF with 24 predictors (hereafter referred to as RF1). The 24 included predicted variables can be reviewed in Figure 5.23a. Due to randomness in the model fitted values vary a little for every model run. The highest R^2 score over 40 model runs is 0.72. The RF algorithm created 500 decision trees, trying 8 variables at each split. After +/- 50 trees the error rate stabilized (Figure 5.23b). No information on significance of variables or the nature of their influence on the response variable is returned. However, the relative importance of predictors can be interpreted by means of the variable importance plot (Figure 5.23a). The x-axis indicates the percentage increase in Mean Squared Error, reported on a 0% to 100% scale, in case the predictor variable of interest is permuted (randomly shuffled).

The Macroseismic Intensity (mi) was most important in predicting the HDF, followed by seven building material variables. Surprisingly, four out of these seven variables relate to the roof material, while the percentage of low strength roofs did not come out as significant in any of the LMs. Population ranks as the tenth most important predictor. This is another difference with the LM. For LM1 the population variable was slightly more important than the MI variable, while in the RF model the MI is almost twice as important as the population. Also, it stands out that the number of buildings with a mud-bonded foundation is an important predictor. The eight most important predictors include variables from each of the four categories.

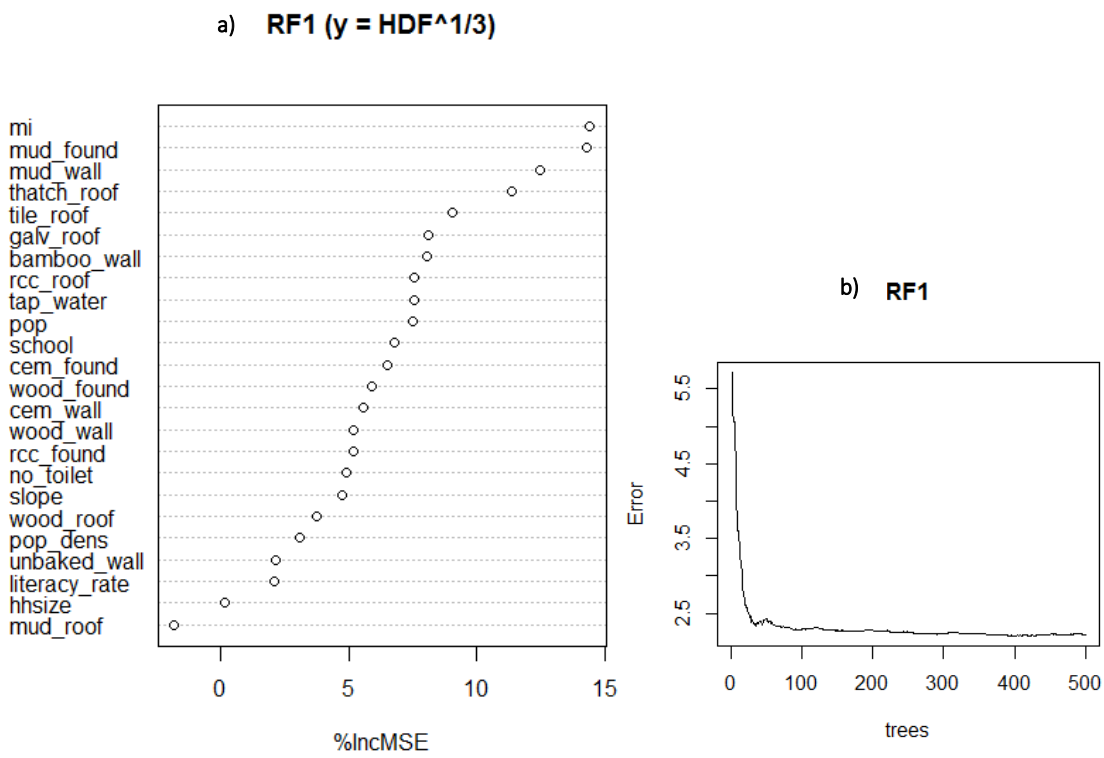


Figure 5.23a-b: Variable importance plot of RF1 a) and error versus RF trees (b). (explanation of variable codes: pop = population, hh_total = total number of households, hhsize = average household size, slope = mean slope value per VDC, mi = Macroseismic intensity, school = relative school attendance, literacy_rate = literacy rate, tap_water = percentage of households with tap water as their main source for drinking water, no_toilet = percentage of households without a toilet facility, mud_found = number of households with mud bonded bricks/stone foundations, cem_found = " cement bonded bricks/stone foundation, rcc_found = " RCC with pillar foundations, wood_found = " wooden pillar foundations, mud_wall = " mud bonded bricks/stone outer walls, cem_wall = " cement bonded bricks/stone outer walls, wood_wall = " wood/planks outer walls, bamboo_wall = " bamboo outer walls, unbaked_wall = " unbaked brick outer walls, thatch_roof = " thatch/straw roofs, galv_roof = " galvanized iron roofs, tile_roof = " tile/slate roofs, rcc_roof = " RCC roofs, wood_roof = " wood/planks roofs, mud_roof = '' mud roofs)

Plotting the residuals of the prediction (*predicted HDF – measured HDF*) to the predicted values results in a rather linear pattern (Figure 5.24).

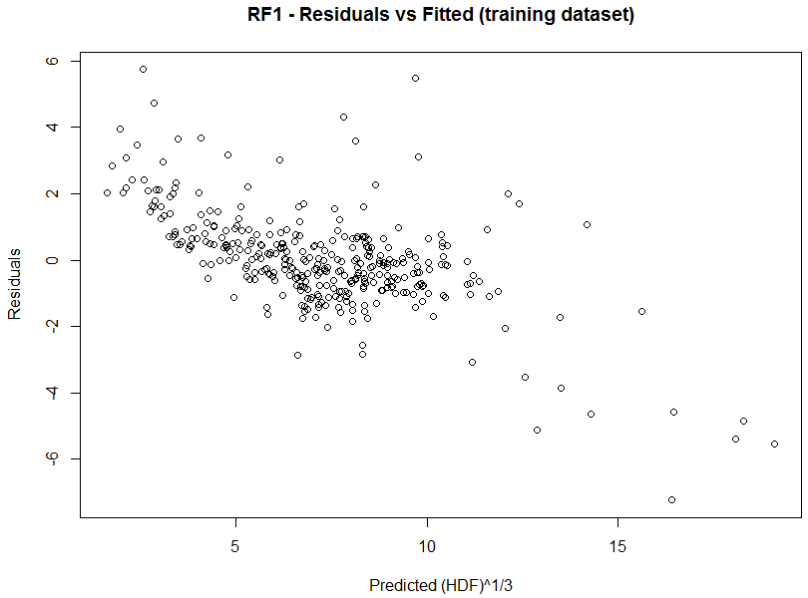


Figure 5.24: Absolute residuals (vertical axis) to predicted values (horizontal axis) of RF1.

The damage in less affected VDCs is overestimated and the damage in more affected VDCs is underestimated. In general, the prediction thus 'flattens' the reality, resulting in smaller differences in damage between VDCs. The model thus fits less good for relatively low and high values.

Random Forest Model 2: Completely Damaged Houses

The second highest R^2 score of the RF models was achieved by the prediction of the cube root transformation of the number of completely damaged houses by the same 24 predictor variables as in RF1. Over 40 model runs the highest R^2 score reached by this model (referred to as RF2) was 0.70. Additionally, 69.2% of the variance in completely damaged houses in the training dataset was explained by the model. Figure 5.25 shows the decrease in Mean Squared Error in case of variable permutation. In general, the variable importance plot resembles the one of RF1. The five most important variables in this model are several building materials (thatch roofs, mud bonded foundations and mud walls, MI and population). Again, the roof material variables appear to be important. Especially the thatch roof variable stands out in its relative importance. In comparison to the RF1 model, the slope variable ranks quite high, and thus has more value in predicting the number of completely damaged houses than in predicting the HDF.

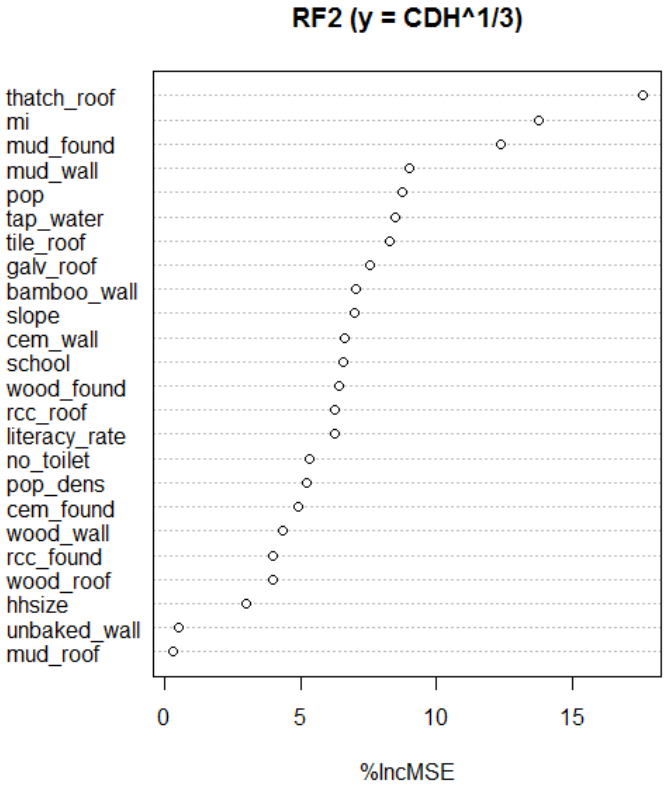


Figure 5.25: Variable importance plot for RF2.

Figure 5.26 displays the distribution of the residuals in relation to the predicted values. Again, the observed pattern is linear. In VDCs with more destroyed houses the damage is underestimated and in VDCs with less destroyed houses the damage is overestimated. For the VDCs in which the measured damage was neither very low nor high, more residuals are closer to zero.

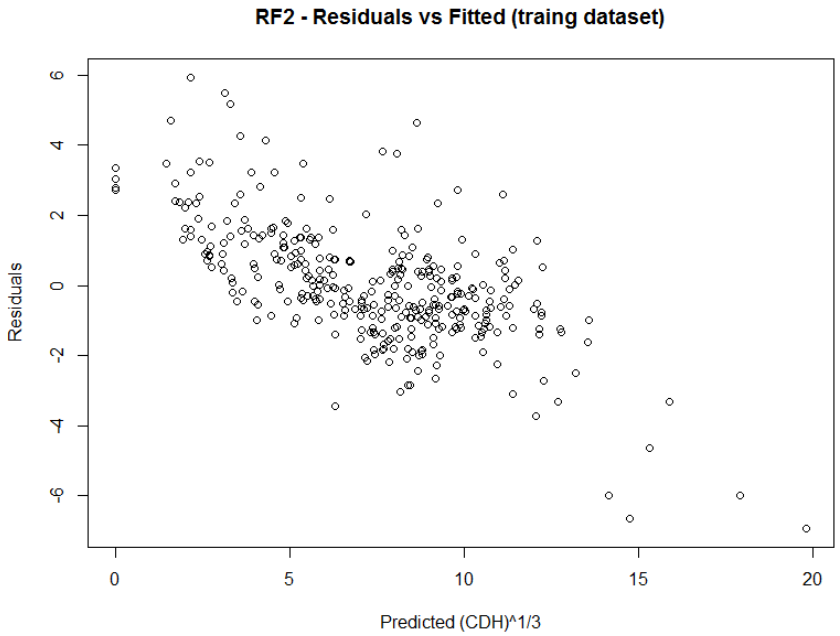


Figure 5.26: Absolute residuals (verical axis) to predicted values (horizontal axis) of RF2

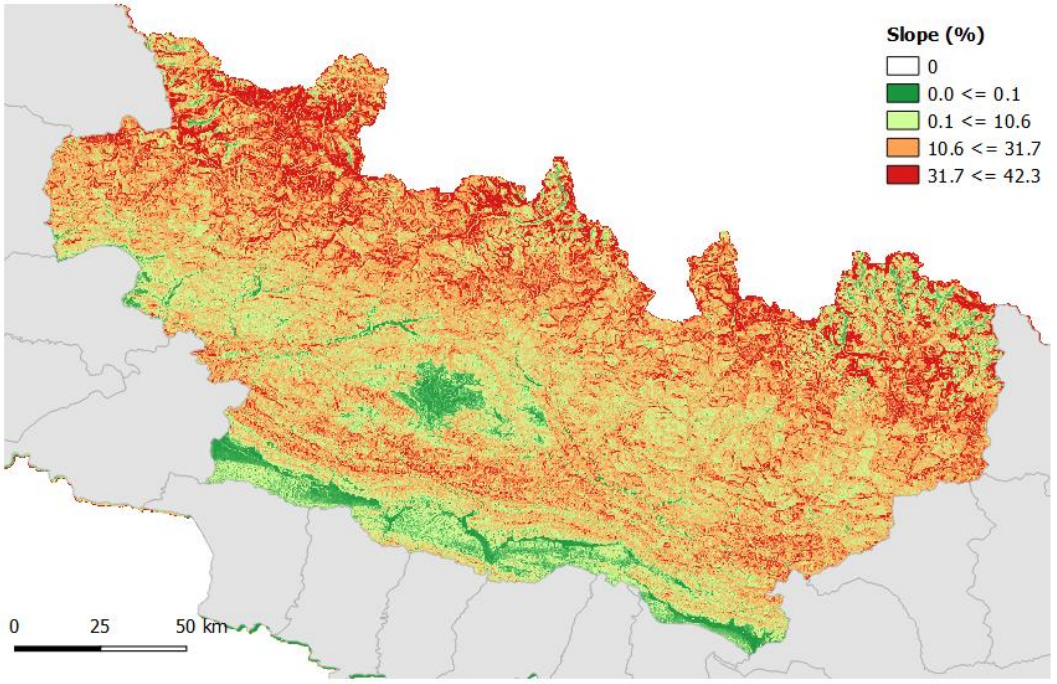
5.4 Optimal Raster Generalization

Q4: For predictor variables that are derived from continuous raster data; how is their predictive value influenced by adjusting their spatial extent to populated areas rather than complete zonal coverage?

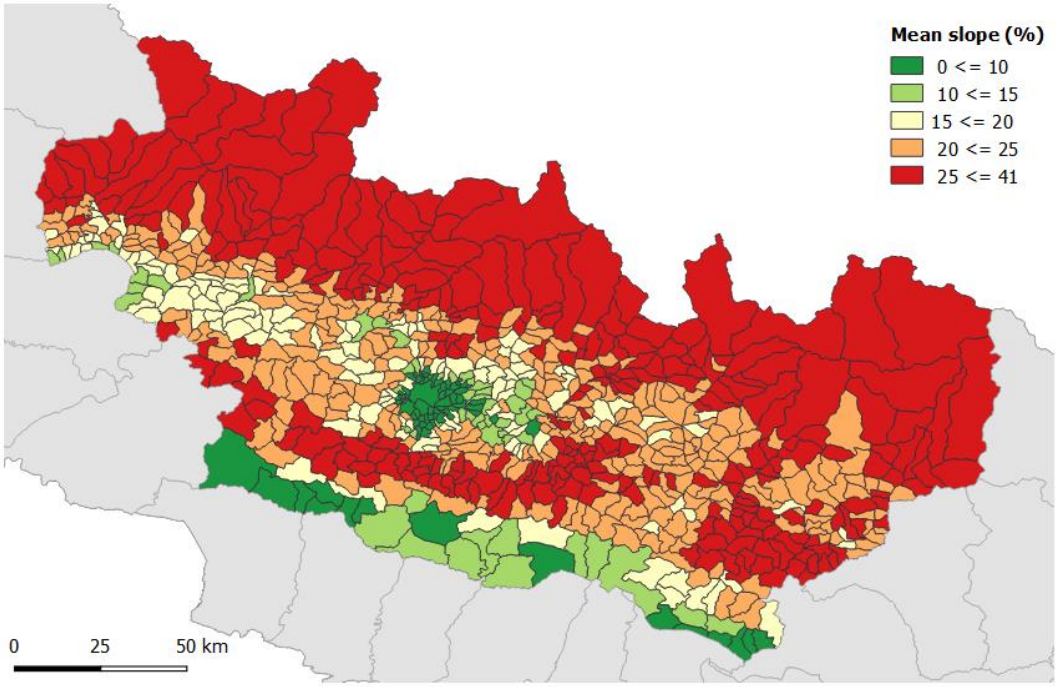
As explained, two different methods of generalizing the slope raster data to single VDC values are tested. There are two raster predictor variables: slope and MI. However, this analysis does not apply to MI since this data is not available until after an event occurred and thus needs to be included in the model with limited pre-processing. Section 5.4.1 explains how simple zonal statistics were derived and what the relative importance of the resulting predicting variable was in the models. Section 5.4.2 explains how the VDCs values were limited to building locations and what influence this had on the model fit and relative variable importance.

5.4.1 Non-adjusted Slope

The first generalization method was to derive a mean value based on all cells intersecting a zone. Figures 5.27a presents the raster slope map derived from the DEM. Since the whole study area is located in the Himalayas, steep slopes are present all over the area. In the Northern part of the study area steeper slopes are more prevalent than in the southern part, although even in the most Northern parts there are still some relatively large flatter areas visible. Figure 5.27b shows the calculated mean slope values for each VDC based on this first generalization technique.



a)



b)

Figure 5.27a-b: Slope raster in study area (a) and mean slope values per VDC based on total area coverage (b).

The map shows that high mean slope values were calculated for all of the Northern VDCs. The variable based on these slope values significantly contributed to the LM prediction of house damages, as presented in Section 5.3.2. However, its relative importance in the model is generally low. It ranked as the sixth most important variable out of seven variables. R^2_{adj} increased from 0.6115 to 0.624 by adding the slope variable. In the RF1 and RF2 models it ranked as the 16th and 10th most important variable out of 24 variables. Permutation of the variable resulted in a 4.7% decrease in MSE.

5.4.2 Slope Adjusted to Built-up Areas

The second generalization method was to extract slope values only from built-up areas, based on building outlines derived from OSM (see Section 4.5). The resulting VDC values are displayed in Figure 5.28.

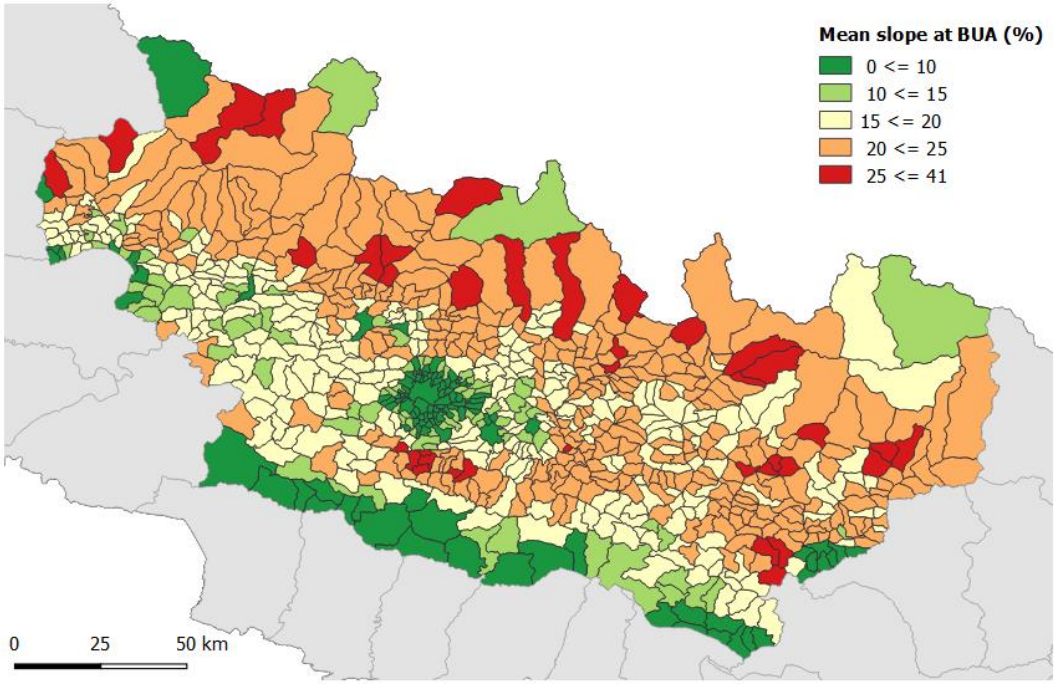


Figure 5.28: Slope values derived from built-up areas only.

The pattern resulting from this method looks distinct from the previous one. On average, the slope values in the built-up area are lower compared to values based on the first generalization technique. The highest average slope value of a VDC is now 33% instead of 41%. Only for a few VDCs a mean slope value above 25% was calculated. It is clearly visible, that for the most Northern VDCs which had relatively large flatter areas the mean slope value is now lower in comparison to the first method. This indicates that most buildings in these VDCs are located in relatively flat areas. Therefore, these buildings will be less susceptible to landslides and possibly experienced less damage.

Nevertheless, in comparison to the previous method, the variable derived from this generalization method resulted in an R^2_{adj} increase of 0.6115 to 0.6232 by adding the built-up area slope variable. The R^2_{adj} thus decreased with 0.13% in comparison to the slope variable not adjusted to built-up areas. In LM1 it ranked as the 7th, thus least, important variable. In the RF1 and RF2 model it ranked as the 18th and 11th most important variable out of 24 variables.

In comparison, for both the LM and RF models a better model fit is reached with the slope variable that is not adjusted to built-up areas. However, the change is nihil. The initial slope variable was of relatively low importance in both models. An adjustment to built-up areas did not change this.

5.5 Model Validation

Q5: How do the models perform regarding the prediction of an independent dataset?

After the best fitting LM and RF models have been defined their predictive values can be assessed by performing out-of-sample validation.

5.5.1 Out-of-sample validation

Validation is performed by running the selected models on the test dataset containing 40% of the observations. Different measures of predictive accuracy result from this.

Linear Model

The R^2_{adj} of the LM when applied on the test data is 0.52. Indicating that 52% of the variance in the HDF variable of the test dataset is explained by the model. The Root Mean Squared Error (RMSE) of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

+

where X_{obs} is observed values and X_{model} is modelled values at time/place i. The lower the RMSE, the better the model fit. The RMSE of the LM1 model on the testing data is 1.89 for the predicted $HDF^{1/3}$ (564 for HDF). This means that the standard deviation of the unexplained variance in this model is a $HDF^{1/3}$ of 1.89. RMSE values can also be used to distinguish model performance on the training and on the testing data. The RMSE on the training data was 1.72 for $HDF^{1/3}$ (517 for HDF). The model thus fitted better on the training data, but the difference in RMSE is limited and does not indicate heavy overfitting.

Figure 5.29a plots the predicted against the measured $HDF^{1/3}$ values for the test dataset. Especially in the VDCs with a higher $HDF^{1/3}$, the model predictions deviate more from the real values. Figure 5.29b shows that again the residuals are linearly distributed, overestimating low values and underestimating higher values.

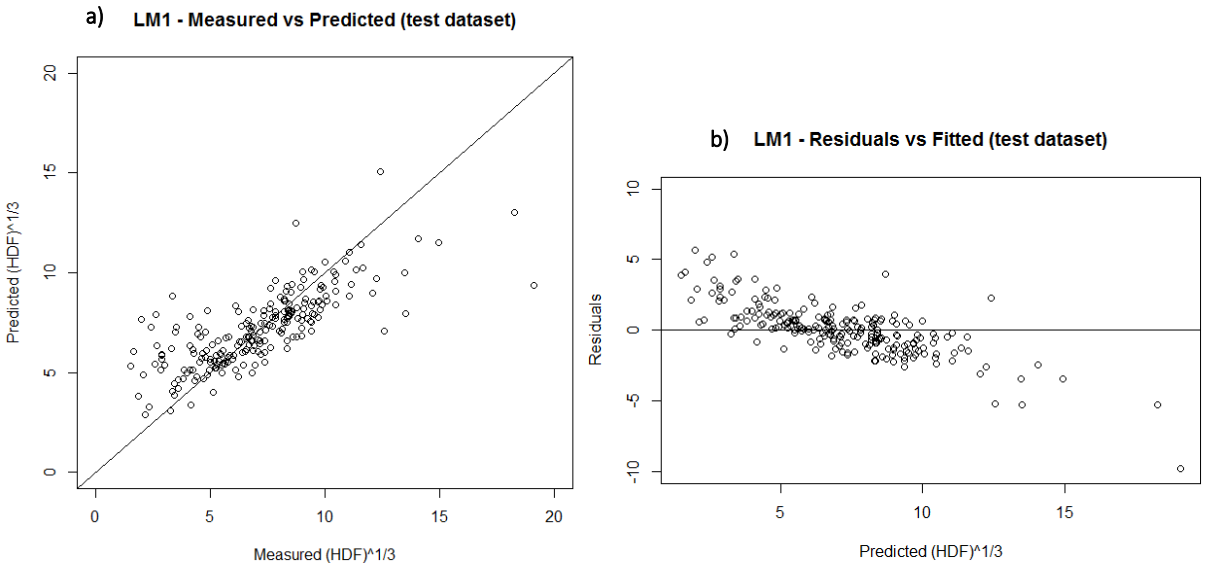
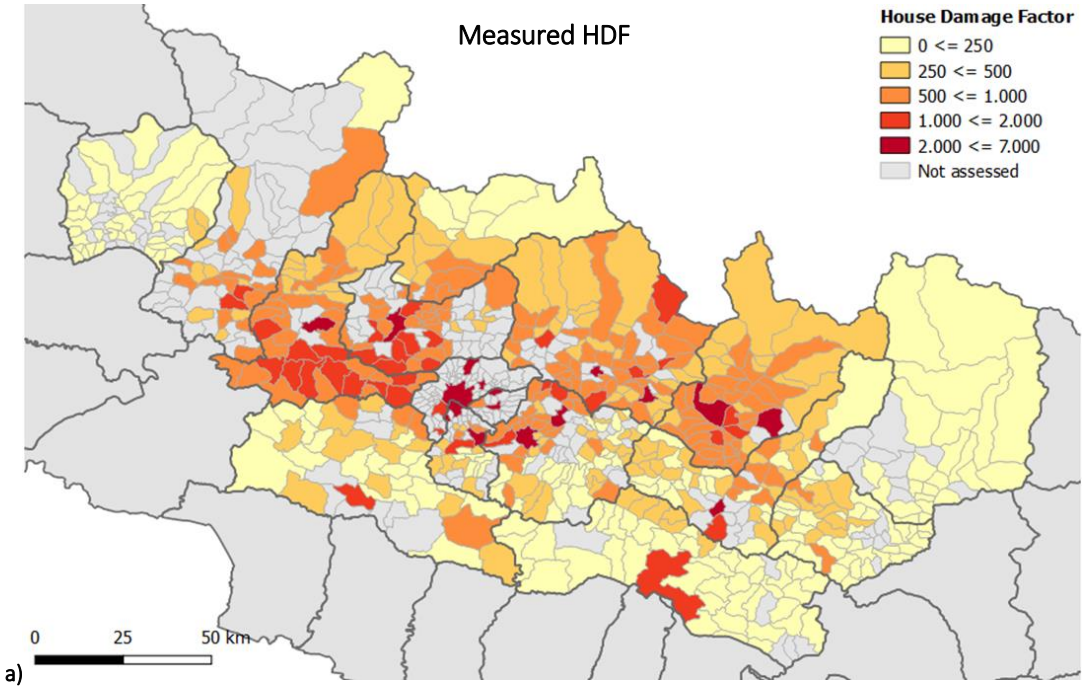


Figure 5.29a-b: Predictive accuracy of LM1 for $HDF^{1/3}$. Plots of measured against predicted values (a) and predicted values against residuals (b).

Figures 5.30 a and b display the reported and the by the LM predicted HDF for the VDCs where both the number of partially and completely damaged houses was reported. The correct identification of the VDCs with the highest damage are important with regard to prioritisation decision making by aid workers. Quite a number of the VDCs in the highest HDF category got predicted correctly, though the number of VDCs with a HDF above 2,000 is higher in reality than in the prediction. The number of VDCs for which the model predicted a HDF above 2,000 while it was lower in reality is limited to four. Overall the patterns are similar. The displayed predictions do include training data also.



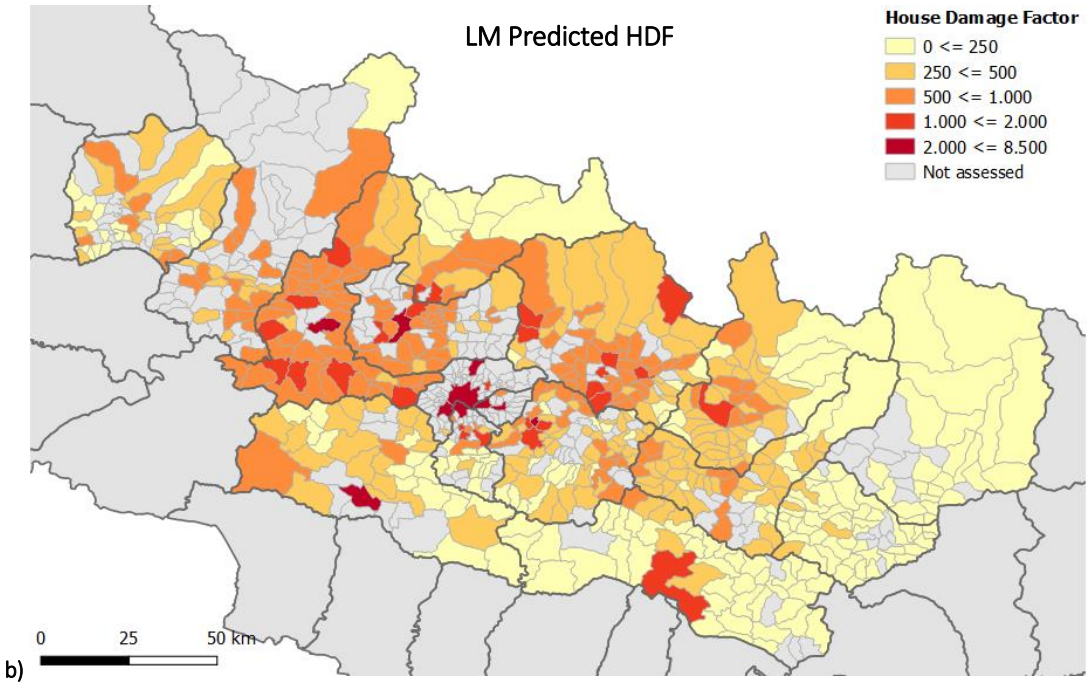


Figure 5.30a-b: Measured HDF (a) and LM predicted HDF values (b).

Random Forest Model 1

For the RF model a prediction on the test dataset resulted in an R^2 of 0.63. The RMSE is 1.68 for $HDF^{1/3}$ (503 for HDF), indicating that the standard deviation of the unexplained variance in this model is a $HDF^{1/3}$ of 1.68. On the training dataset the RMSE was 1.56 for $HDF^{1/3}$ (466 for HDF), so again there is no heavy overfitting of the model to the training data. Figure 5.31a displays the predicted to the measured $HDF^{1/3}$ values. Similar as with the LM prediction, especially in the VDCs with a higher HDF the model predictions deviate more from the real values. The residuals distribution (Figure 5.31b) confirm this.

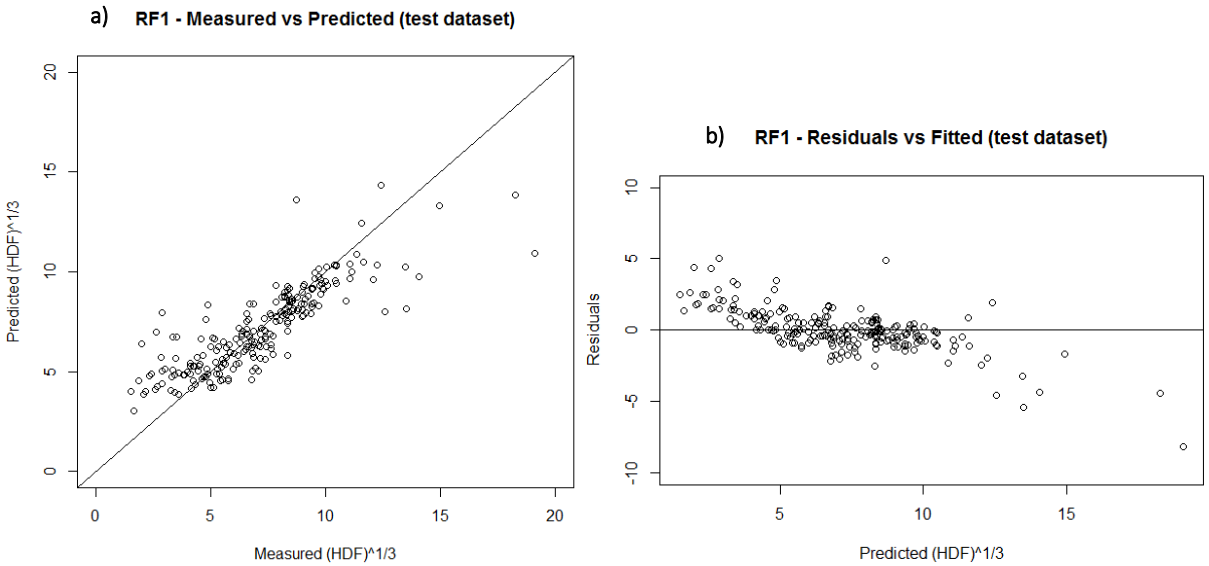


Figure 5.31a-b: Predictive accuracy of RF1 for $HDF^{1/3}$. Plots of measured against predicted values (a) and predicted values against residuals (b).

Figure 5.32 displays the HDF values predicted by the RF1 model. The RF1 model identifies more highest damaged VDCs than the LM model (16 in stead of 9).

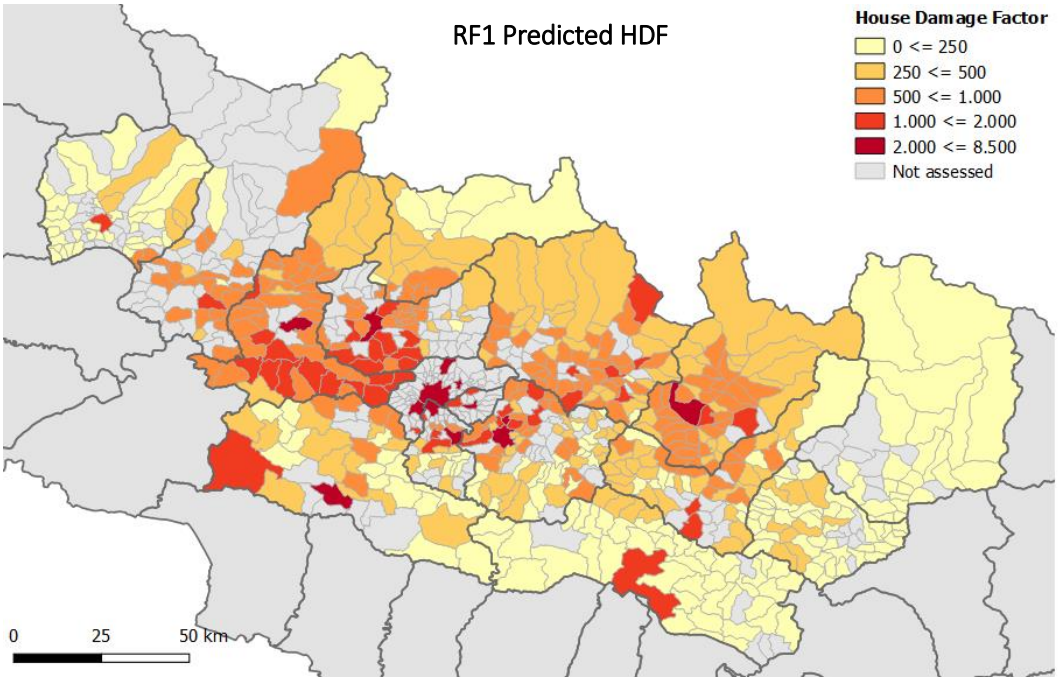


Figure 5.32: RF1 predicted HDF values.

Random Forest Model 2

For the second RF model (predicting the number of completely damaged houses rather than the HDF) a prediction on the test dataset resulted in an R^2 of 0.60. The RMSE is 1.17 for 'completely damage houses$^{1/3}$'(626 for completely damaged houses), indicating that the standard deviation of the unexplained variance in this model is 1.17. The RMSE on the training data for RF2 was (555 for completely damaged houses). An increase of 12.8% in RMSE indicates that the model might be slightly overfitted to the training data. Figure 5.33a displays the predicted to the measured values. The plot shows that the model tends to overestimate less damaged VDCs and underestimate higher damaged VDCs. Resulting in a more equal spread of damage categories, as is confirmed by the maps in Figures 5.34 a and b.

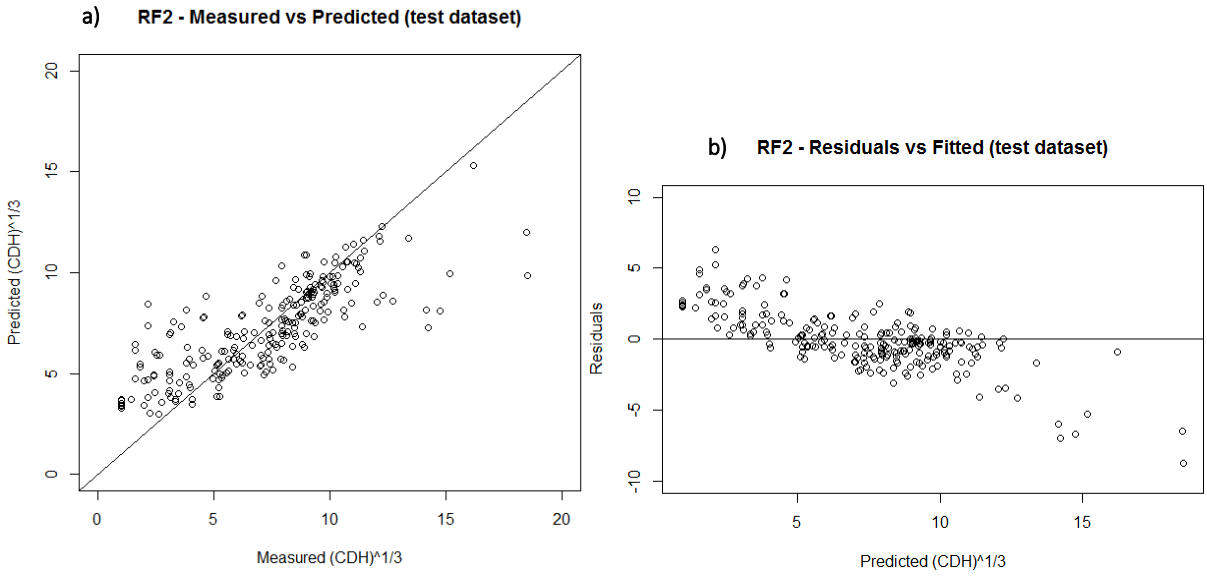
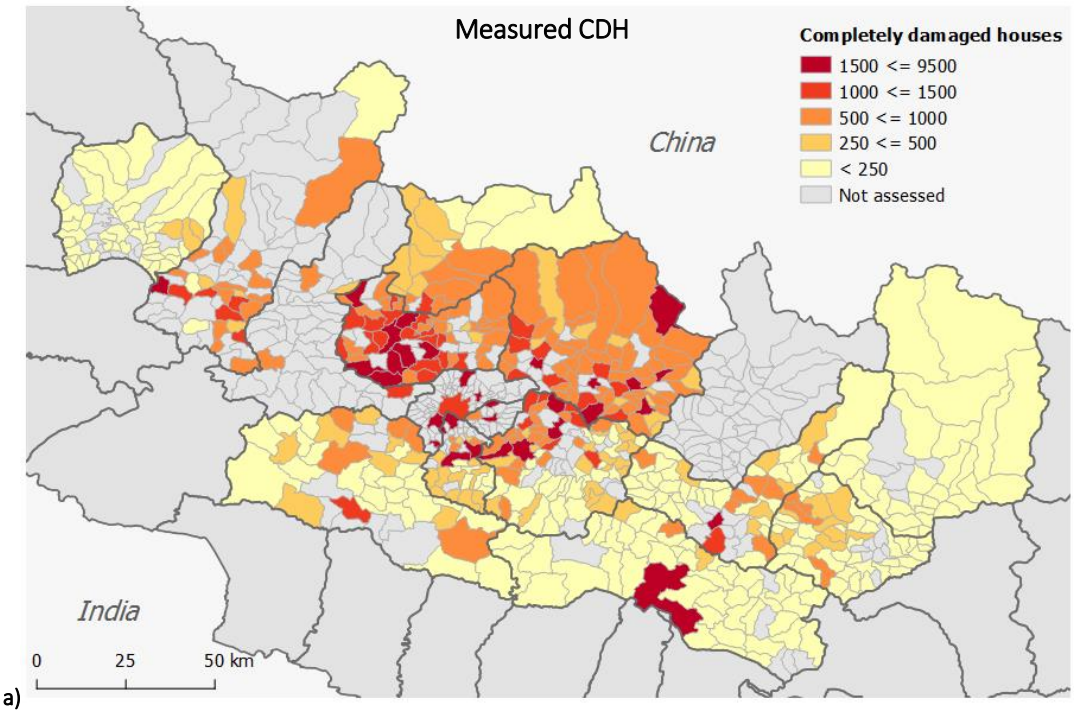
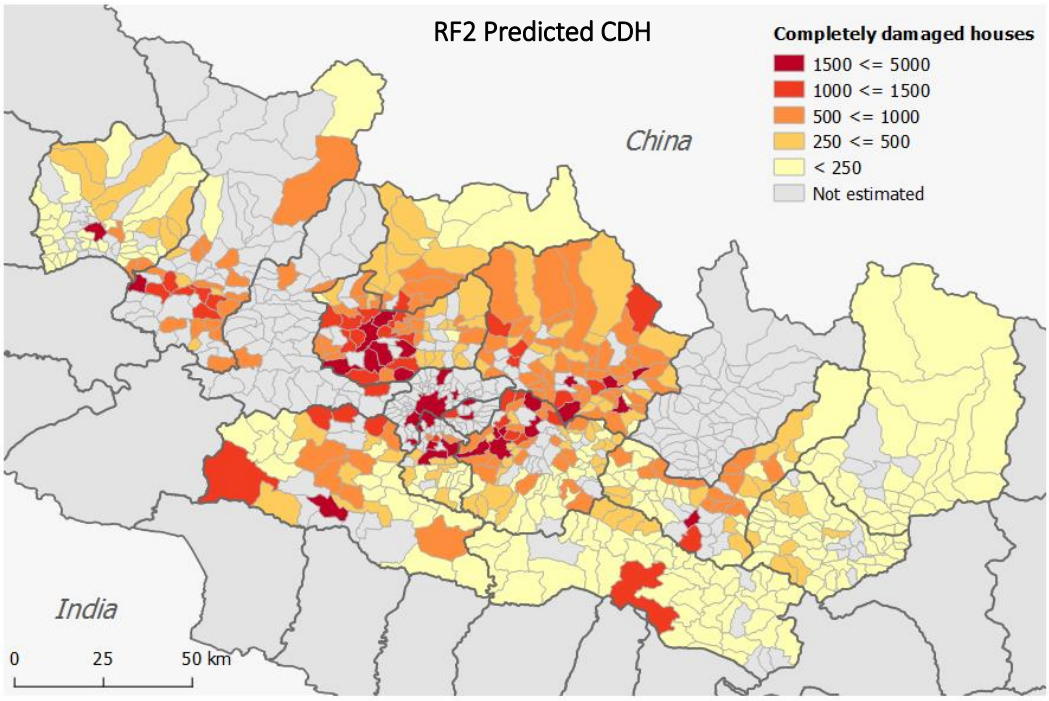


Figure 5.33a-b: Predictive accuracy of RF2 for (completely damaged houses)$^{1/3}$. Plots of measured against predicted values (a) and predicted values against residuals (b).

Figure 5.33 displays the reported (a) and the by RF2 predicted number of completely damaged houses per VDC (b). Quite some VDCs are predicted in a damage category lower than reported.



a)



b)

Figure 5.34a-b: Measured completely damaged houses (a) and RF2 predicted completely damaged houses (b)

5.6 Model Comparison

Q6: How do the models compare to each other in terms of predictive performance and general usability?

In this section the three models are compared in terms of their predictive performance (as presented in the previous section) and their general usability.

5.6.1 Predictions Compared

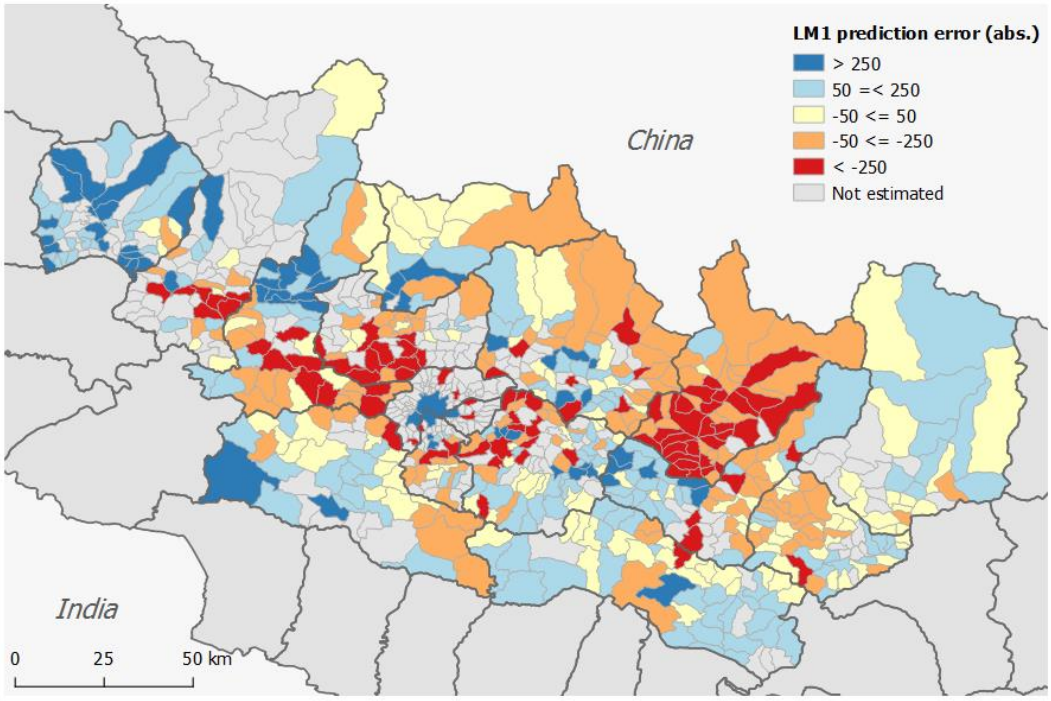
Table 5.1 summarizes the different measures of predictive accuracy and some general model characteristics as presented in previous sections. The most accurate prediction is made by the first random forest model predicting the House Damage Factor. This model can explain more variance in response variables than the other two models. Also the average prediction error per VDC is lowest and the model seems to be least sensitive to overfitting, judging from the relatively low decrease in R^2 between training and testing data. Despite the fact that the response variable used for the RF2 model has the lowest number of outliers, the Root Mean Squared Error (RMSE) is highest, with an average error of 626 houses per VDC.

Model	Response variable	Nr. of predictors	R^2_{test}	$(R^2_{train}) - (R^2_{test})$	RMSE	Nr. of outliers
LM1	House Damage Factor	7	0.52	-0.10	564	13
RF1	House Damage Factor	24	0.63	-0.08	502	13
RF2	Completely damaged houses	24	0.60	-0.10	626	7

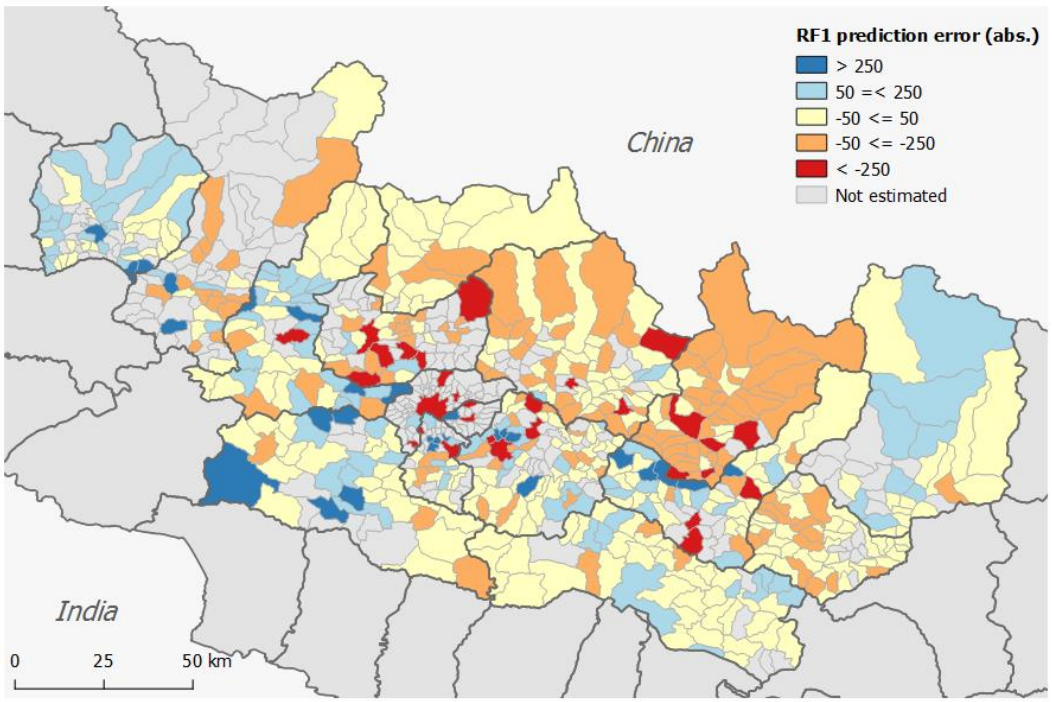
Table 5.1: Models predictive accuracy measures and characteristics

Comparing Prediction Errors

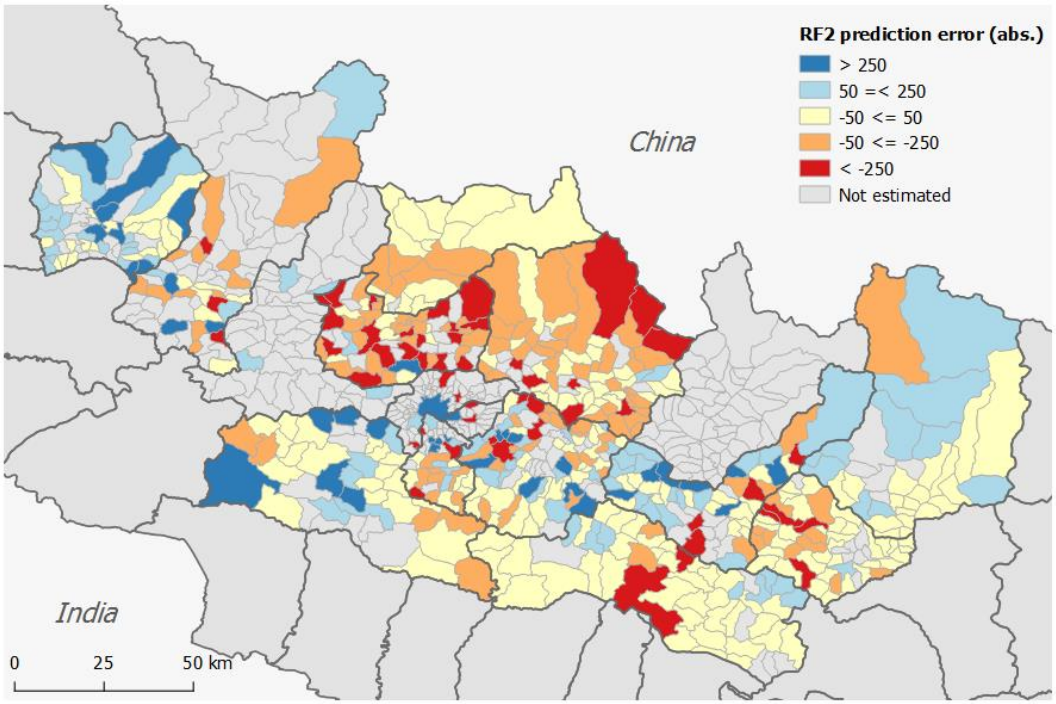
Figures 5.35 presents the spatial distribution of prediction errors for consecutively LM1, RF1 and RF2. Prediction errors were calculated by subtracting the measured values from the predicted values. When comparing the residual distributions of the different models to each other, first of all it stands out that for all three models the Eastern, Southern and Western border areas of the study area are subject to mostly overestimations. The damage in the middle and more Northern areas tends to be more underestimated. Another observation is that the LM1 prediction has the most VDCs with an error of more than 250 or less than -250. Especially in the district of Dolakha the damage in many VDCs is heavily underestimated. Looking back at the original data, the number of fully damaged houses divided by the total number of households results in a value above 100% for 46 out of the 50 assessed VDCs in Dolakha (see also Figure 5.3). Even though the true total number of households might be higher than is reported in the 2011 Census (due to unregistered households) it is unlikely that this results in a percentage above 100 for nearly all VDCs in the district. Therefore, this leads to suspect that in reality the damage was lower than what was reported, and thus an underestimation by the model is not very surprising.



a)



b)



c)

Figure 5.35a-c: Residuals of LM (a), residuals of RF1 (b) and residuals of RF2 (c) (*residual = predicted - measured*)

Also, nearly all models overestimated damage for a lot of VDCs in the most Western district Lamjung. In comparison to other Eastern, Southern and Western bordering districts the reported numbers of houses damaged were very low in Lamjung. For 33 out of the 44 assessed VDCs the number of houses completely damaged was below 50.

Comparing Prediction of Priority Areas

Since an important aim of PIMs is to identify aid priority areas in the initial phase after a disaster it is important to compare how accurately the models predicted the highest priority areas, being the VDCs

with the highest measured damage. Tables 5.2a, b and c show the fifteen VDCs with the highest predicted damage. The second column indicates where the same VDC was ranked in the original measured data. The VDCs displayed are all part of the testing dataset consisting of 222 observations in total.

For the LM1 model the table shows that out of the 15 VDCs with the highest predicted HDF, 7 were also within the top 15 VDCs with the highest measured HDF. For the RF1 model, 10 out of the 15 VDCs with the highest predicted HDF were also within the top 15 highest measured HDF VDCs. For RF2, out of the 15 VDCs with the highest predicted number of completely damaged houses, 8 were also within the top 15 of VDCs with the highest measured number of completely damaged houses. The RF1 model thus identified most, two-third, of the fifteen highest priority areas correctly. The LM1 model identified a little less than half of the priority areas correctly and for the RF2 model this was a little more than half.

a) Linear Model 1

Priority rank Predicted	Priority rank Measured	VDC name
1	2	Nilkanth
2	54	Banepa Municipality
3	4	Chautara
4	3	Panauti Municipality
5	45	Chapagaun
6	17	Thulo Sirubari
7	10	Tatopani
8	21	Ichok
9	48	Ugratara Janagal
10	219	Gaunshahar
11	7	Mahadevsthan Mandan
12	12	Pida
13	44	Dhursa
14	40	Katunje
15	1	Panchkhal

b) Random Forest 1

Priority rank Predicted	Priority rank Measured	VDC name
1	3	Panauti Municipality
2	2	Nilkanth
3	54	Banepa Municipality
4	48	Ugratara Janagal
5	45	Chapaguan
6	1	Panchkhal
7	4	Chautara
8	6	Manthali
9	10	Tatopani
10	11	Barhabise
11	7	Mahadevsthan Mandan
12	219	Gaunshahar
13	23	Salyan Tar
14	8	Madanpur
15	15	Jiwanpur

c) Random Forest 2

Priority rank predicted	Priority rank measured	VDC name
1	3	Panauti Municipality
2	1	Panchkhal
3	2	Nilkanth
4	214	Gaunshahar
5	122	Banepa Municipality
6	6	Mahadevsthan Mandan
7	51	Chapaguan
8	18	Jiwanpur
9	7	Madanpur
10	11	Barhabise
11	72	Ugratara Janagal
12	28	Chhatre Deurali
13	23	Salyan Tar
14	12	Pida
15	8	Manthali

Table 5.2a-c: Highest damage ranking LM1, b) highest damage ranking RF1 and c) highest damage ranking RF2.

Comparing the RF1 and RF2 model predictions can give additional insights since both models are the same, except for the response variable they predict (HDF for RF1 and number of completely damaged houses for RF2). Out of the 15 priority VDCs 13 overlap between both models. This indicates that the

addition of partially damaged houses in the HDF variable hardly influences the priority areas identified by the models.

In predictions by both models some VDCs stand out because of their high predicted rank and their low priority rank in reality. The VDC Gaunshahar, for example, is ranked as 9th and 4th most damaged VDC, while in the original dataset it was among the 20 VDCs with the least damage. The RF models predicted 1,690 completely damaged houses and an HDF of 1,377, while in reality only 4 houses were reported to be completely damaged and an HDF of 4.8 was calculated. Since the VDC was subject to quite some damage predictors (MI=7.4, population=6,611, ranking high in mud bonded foundations and walls), a reported number of 4 completely damaged houses is rather surprising. Especially since in surrounding VDCs with similar population numbers around 1,000 to 2,000 damaged houses were reported. It appeared that at the time of assessment formally the VDC Gaunshahar did not exist. In May 2014 this VDC together with 2 other VDCs (Udipur and Chandisthan) were merged into the existing Besishahar Municipality VDC (District Development Committee Lamjung, 2014). For Udipur no numbers were reported at all in the original assessment file and for Chandisthan 14 completely damaged houses were reported. Essentially, because of outdated administrative border files the model made predictions for VDCs that are no longer existent. It is likely that more 'non-existing' VDCs are present in the original assessment file and the model output, because the total number of VDCs in the administrative borders file used is 3,754, while the most recently reported number of VDCs is 3,157 (Techsansar, 2016).

Another way to compare the model's accuracies in predicting priority areas is by dividing all VDCs of the test dataset in five equally sized priority classes based on the reported damage, and to compare this to the predicted priority classification. Table 5.3 shows for each of the three models the percentage of VDCs for which the correct priority level was predicted and the percentage of predictions that was in a category too high or too low. Each class consisted of either 44 or 45 observations. VDCs with the most damage got assigned a priority level 1.

	LM1			RF1			RF2		
Priority level	Correct level predicted	Too high	Too low	Correct level predicted	Too high	Too low	Correct level predicted	Too high	Too low
1	62%	38%	-	69%	31%	-	66%	34%	-
2	41%	29%	30%	45%	28%	27%	48%	29%	23%
3	32%	29%	39%	39%	29%	32%	30%	25%	45%
4	36%	30%	34%	43%	21%	36%	43%	19%	38%
5	51%	-	49%	69%	-	31%	69%	-	31%
Total:	45%	25%	30%	53%	22%	25%	51%	21%	28%

Table 5.3: Share of predictions in correct priority categories for LM1, RF1 and RF2. For each of the five priority levels the numbers define how much VDCs got predicted the same priority level as was measured (green), how much VDCs got predicted one or more priority level(s) higher than measured (blue) and how much VDCs got predicted one or more priority level(s) lower than measured (red).

The accuracies for the different models are very similar. All models assign most VDCs correctly in priority levels 1 and 5, and they all assign less than 40% of the observations correctly for the level 3 category. From the wrong predictions in this category, all models assigned more VDCs to a level too low than to a level too high. In general, the models thus underestimated damage in the middle priority class. Moreover, for the three middle categories (levels 2, 3 and 4) all models assign a wrong priority level to more than half of the observations. The total predictive accuracy of the priority levels resembles the

compared performance of the models in terms of R^2 scores. The RF1 model predicts best, followed by the RF2 and the LM1 model. All models predict more VDCs in priority categories too low than in categories too high. In that sense all models tend to underestimate rather than overestimate.

5.6.2 General Usability of Models

Apart from the predictive power of the models also their usability plays an important role in determining which model is most suitable for implementation in humanitarian disaster response. This usability can be viewed from two different perspectives. On the one hand, there is the usability of model output for end-users (being aid workers abroad in the affected area). On the other hand, there are admin-users (being those in local aid offices responsible for the tasks from data preparation to producing a visual PIM output).

Admin-user Perspective

For those who are responsible for producing a useful model output for their colleagues in the field it is important that they can run the model on an affected area without having to take many time intensive steps. Once the final models are in use, both the LM and RF models require very little computation time because of the relatively small datasets. However, since the models will need a lot of further training and fine-tuning before they can be implemented, it is also important to consider their time intensiveness for constructing and training the models. In this sense, the LM takes more time to construct beforehand, since the admin-user needs to check for normal distributions of variables, covariance among the variables, preselect the best predictor variable subset and after fitting check for the model assumptions concerning residuals. The RF models do not require any of these steps, since the only assumption is reliable input data. Also, he or she does not need to perform a separate analysis in order to select the best variable subset. In terms of further model training and improvement the RF models are thus preferred over the LM model.

However, also the LM model has some advantages over the RF models. One of these advantages is that admin-users can gain insight by analysing the nature of the individual coefficients of the variables. This can show if the relation between a variable in the model and the damage is positive or negative. For this case study this gave the insight that the relation between damage on the one hand and foundation type, school attendance and toilet presence on the other hand was opposite to what was expected. Such findings can be an incentive for the admin-user to find explanations and possibly to redefine the predictor variables used.

Comparing the two RF models, despite the decreased R^2 and the relatively high prediction error, the RF2 model might be preferable because its response variable is composed of a single measured value (completely damaged houses) instead of two measured values (partially and completely damaged houses). Because of that, less post-event data is required for future model training. Additionally, as mentioned in Section 5.1.2 partially damaged measures are often questionable. In that sense, to train on only completely damaged houses could give a more reliable output of priority areas.

End-user Perspective

One important model requirement from the end-user's perspective is the output's intuitiveness, meaning that the model output can be easily grasped without much thought. From this perspective, it can be stated that the RF2 output, representing the absolute number of completely damaged houses, is easier to interpret for end-users than the predicted House Damage Factor represented by LM1 and RF1. The RF2 output makes it easier for end-users to form an image in their minds of the disaster's impact. This also makes the output easier and faster to communicate between end-users. In a post-disaster environment decision makers have to be able to communicate clearly. With the HDF output an end-user would first have to invest time and effort to understand the meaning of the measure and to

figure out how to interpret this in the field. With the number of completely damaged houses this is not the case.

Figure 5.36 presents a possible priority index output map based on the RF2 model prediction for the complete study area. The map shows two isolines of the earthquake's Macroseismic intensity. When disseminated, a map like this should either be accompanied by a spreadsheet containing all values and VDC names, or be presented online enabling users to zoom and click for exact prediction values. Districts bordering the study area are not included in this analysis, though they could be included if also country specific data were prepared for these districts. Additionally, the models uncertainty should be included in the map, for example by reporting the expected average prediction error of 626 houses per VDC.

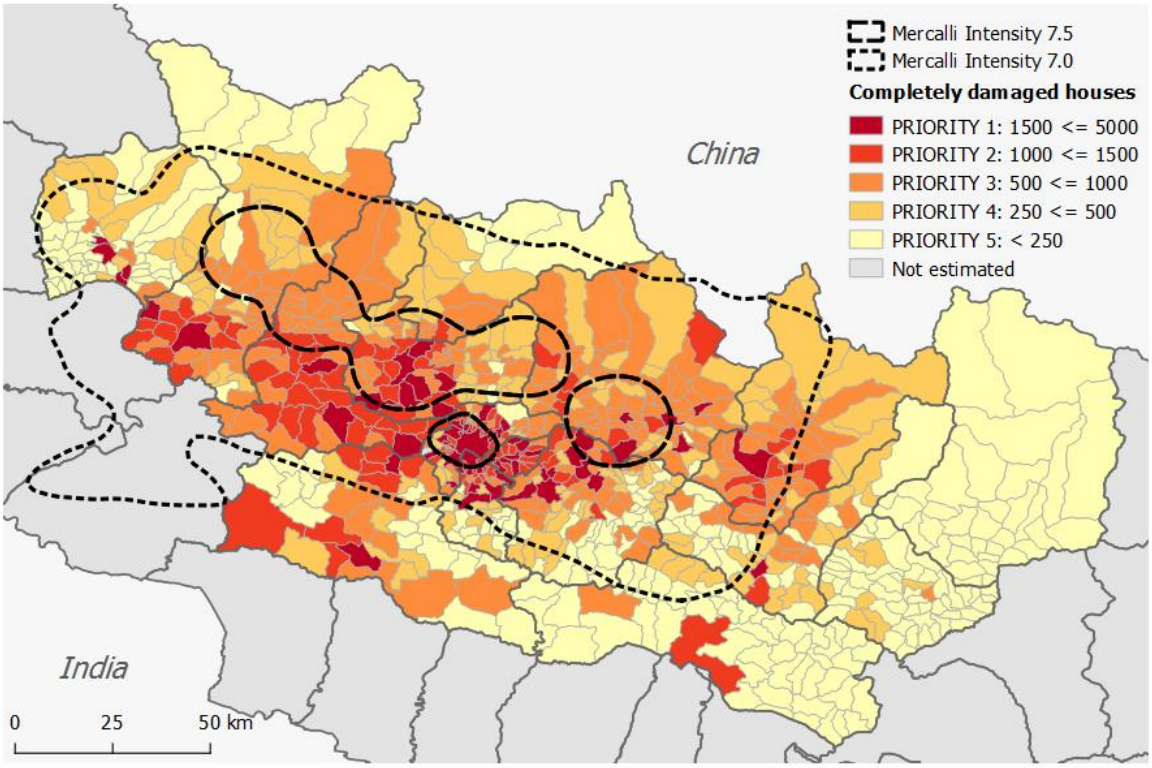


Figure 5.36: Possible visual PIM output, displaying the RF2 predicted damage values for the study area.

Another important requirement from the end-user's perspective concerns the information that he or she can derive from the model output. The absolute measure of the RF2 model output offers the possibility to derive other measures from it. For example, by dividing the predicted number of completely damaged houses by the total number of households a relative measure of housing damage can be created, presenting the percentage of houses completely damaged per VDC. The models predicting the HDF are not suitable for such a calculation.

Also, the output produced by the RF2 model can be more easily corrected and/or supplemented in the field. As soon as the first damage assessments have taken place they can be included in the model, either as new, corrected or confirmed values. For correction or addition of HDF values both fully and partially damaged houses will have to be assessed.

6 Discussion

This chapter first presents a short overview of the applied research methodology. Hereafter, the most important findings and their implications are discussed. At the same time, limitations of the research approach are addressed. An additional section of recommendations for follow-up research is presented at the end of the Chapter.

6.1 Research Summary

In the first days following a natural disaster, humanitarian decision makers often deal with a scarcity of information on the spatial distribution of the event's impact, and thus the need for humanitarian aid of the affected population. By learning from data of past events Priority Index Models can rapidly produce an estimate of a disaster's impact, which can help decision makers to identify aid priority areas. The main objective set in this research was to explore the possibilities for a model that rapidly estimates post-earthquake aid neediness for any earthquake-prone area on earth, learning from data of past events. To achieve this, the case of the Gorkha earthquake of 2015 in Nepal was used as a test case. Pre- and post-event open data related to the earthquake were collected to construct a training dataset. In order to find relationships in the dataset a multivariate linear regression model and a random forest regression model were fitted to the data. Based on a comparison of on the one hand the predictive accuracy and on the other hand the general usability of the best fitted models, conclusions can be drawn about which model is most suitable to be successfully extended to other earthquake-prone areas in the future.

6.2 Main Findings, Limitations and Recommendations

The first step towards creating a predictive model for aid priority areas after the Gorkha event was to select a suitable impact assessment dataset to reflect different levels of aid neediness after the earthquake. The Initial Rapid Assessment dataset by the Nepalese Red Cross was considered the most suitable. The main reasons to select this dataset were that damage to residential buildings is an incentive for many aid clusters; the dataset has many data points on a low administrative level which enables to make distinctions within the most affected area and; the use of a common measure creates more possibilities for future model training. The two possible aid priority indicators that were derived from the dataset were the 'number of completely damaged houses per VDC' and the 'House Damage Factor'.

A critical remark here is that the quantification of aid neediness by structural damages remains questionable. Damages to residential buildings can give a good idea about how the inhabitants are affected from multiple aid-cluster perspectives. However, it does not resemble their capacity to cope with this impact. Family structures for example, can be very important in explaining aid neediness. It is often observed in a post-disaster scenario that female-headed households have more difficulties in rebuilding their house. Nevertheless, it seems almost impossible to define an aid neediness indicator that includes aspects of coping capacity and at the same time can be defined similarly for multiple past events. In this perspective, the strength of structural damages as a proxy measure is that it is a single-feature and commonly assessed measure, which makes it easier to reproduce for future model training. Also, it can be objectively measured and is straightforward in its interpretation, which increases the

reliability of reported numbers. Possibly, Priority Index Models should not strive to produce a quantitative 'aid neediness' output that includes coping capacity. Instead, during decision making processes, damage estimations can be combined with information about the coping capacities of affected communities. This way, aid workers that are familiar with the local context can decide themselves what a destroyed house means in terms of aid neediness for different parts of the population.

Another remark concerns the quality and reliability of the Initial Rapid Assessment dataset. It was observed that for quite some VDCs rough estimations were made based on total household figures, sometimes providing the same relative number of houses damaged for all VDCs in a district. Also, the list of assessed VDCs contained VDCs that were formally non-existent at the time of data collection. Data quality issues caused by rough estimations, gap-filling or outdated base-information are likely an inherent aspect of rapid assessments done (by volunteers) in a post-disaster environment. This means that the model is trained on an estimated situation which does not necessarily represent the real post-disaster situation. Therefore, detailed damage assessments, done after the first post-disaster phase, should be preferred over initial rapid assessments for future training of the model.

The second research question aimed at defining candidate predictor variables from open data to predict the previously defined aid neediness indicator(s). A total of 27 candidate predictors were defined, distributed over four different categories: hazard, exposure, building quality, susceptibility to secondary hazards, and socio-economic vulnerability. The data behind most variables was derived from the 2011 National Population and Housing Census Nepal. The hazard variable was expressed as the mean Macroseismic Intensity per VDC as derived from USGS ShakeMaps.

One main drawback observed at this part of the case study is the absence of a good overview of the total number of houses per VDC at the time of the earthquake (the exposure variable). Also the population figures derived from the 2011 Census are likely quite different from the actual numbers in 2015. The census reports a total population of 26,494,504 in 2011, while the World Bank (2017b) estimated the total number in 2015 at 28,513,700. It is likely that census data will be outdated for other earthquake-prone countries also, since most governments conduct a population (and housing) census once every ten years. On top of that, censuses can undercount vulnerable people (low-incomes, children and minorities) and informal settlements. To overcome these limitations, future models could make use of WorldPop data (see: http://worldpop.org.uk). They provide high spatial resolution data on estimated human population distribution in raster format. This way, a possible error marge will be more uniform for different countries. Another important limitation are the uncertainties in the USGS ShakeMap and its correlation with the location of seismic measurement stations. Especially since Macroseismic intensity is one of the most important predictors in all models, the final model output is likely influenced by these uncertainties. This could explain for example why the most Eastern district Okhaldunga has quite some heavily damaged VDCs while the MI was reported below 6.0 for the whole district. The uncertainties in this area are relatively high (see Figure 5.6) and thus the MI might have been higher in reality. Another limitation is that in this case study aftershocks were not taken into account because the date of the damage assessment was not clearly defined. If assessment dates are defined for future training cases, it is advisable to include aftershock intensity data from ShakeMap in the hazard variable. A final drawback regarding the model's predictor variables is that for Nepal no openly available composite measures of the population's socio-economical vulnerability such as poverty and development indexes were collected/found. This is discussed in more detail in Section 6.3.

To answer the third research question, a multivariate linear regression model and a random forest regression model were fitted to a training dataset containing 60% of all observations. For the linear model (LM1), the best fitting model consisted of seven predictor variables ($population^{log}$, slope,

Macroseismic intensity, school attendance, literacy rate, foundation type and toilet presence) predicting the House Damage Factor$^{1/3}$. The population and the Macroseismic Intensity variables evidently were the most important in this model. From the random forest algorithm the two best performing models were selected. The first included 24 predictor variables predicting the 'House Damage Factor'$^{1/3}$ (RF1). The second best fitting model, with the same input variables, predicted the absolute number of 'completely damaged houses'$^{1/3}$ (RF2). In both models the 'Macroseismic Intensity', 'mud bonded bricks/stone walls', 'mud-bonded bricks/stone foundations' and 'thatch/straw roofs' ranked as the four most important variables. Low strength building quality variables were very important in the RF models.

In all models the Hazard related variable was most or second-most important. But where the LM1 model assigned second-most importance to population, the RF models favoured multiple building material variables. One explanation for this could be that in the LM1 model the building material variables were included as composite variables because of multicollinearity. Also, when comparing models RF1 and RF2 it is evident that the population variable was more important in the model predicting completely damaged houses (RF2). When using composite building material variables the predictive accuracy of the models decreases, but when extending the model to other countries individual building material classes might differ, making a generalization to 'low-strength' and 'high-strength' composites inevitable.

In the LM1 model, some variable relationships turned out to be opposite to what was expected. VDCs with more damaged houses corresponded to VDCs with higher school attendance, higher toilet presence and better foundation qualities. The significance of these variables indicates that they are suitable predictors for the Nepal case. However, because of the unexpected coefficients these relationships are likely case-specific and not generalizable to other countries. If the nature of these relationships indeed differs for other countries, it can be analysed if a more general relationship exists between damage and a composite measure of socio-economic vulnerability. Alternatively, a single (proxy-) measure for socio-economic vulnerability such as literacy rate could be used. The relative importance of both socio-economic and physical vulnerability is lower than what could be expected based on existing models. Possibly these factors are overestimated in existing models.

To answer the fourth question, it was analysed if and how the predictive value of the slope variable was influenced by adjusting its spatial extent to populated areas only, in comparison to complete area coverage. A visual inspection of landslide locations and slope values led to the conclusion that indeed slope could be a proxy indicator for earthquake induced landslide susceptibility. However, the initial slope variable was of relatively low importance in all three models. An adjustment to built-up areas did not change that. For all models a better model-fit was achieved with the slope variable that was not adjusted to built-up areas. However, the change was nihil. Based on the Gorkha case there seems to be no motive to adjust slope data to populated areas in Earthquake PIMs.

A remark has to be made here about the completeness of the OSM buildings layer. Of the sixteen most affected districts, five had a 'buildings-to-population-ratio' between 1:7 and 1:11, which seems to be unrealistic. Additionally, since most of the buildings were mapped after the event it is likely that, depending on the area, for future affected areas the buildings layer is far more incomplete. The case study results give no information about whether slope values are related to structural damages in other countries or during other events also, since susceptibility to earthquake-induced landslides depends on more factors than slope only, which will differ per region. Future model training on events in other countries should indicate whether slope values are related to structural damages, not only in Nepal but in other countries also.

Hereafter, the models were validated 'out-of-sample' by running them on a test dataset consisting of 40% of the original dataset. The RF1 model predicted 63% of the variance in the spatial distribution of

damage in the test dataset (R^2=0.63). Similarly, the RF2 model explained 60% (R^2=0.60) and the LM1 model 53% (R^2=0.53). All models overestimated the damage in less affected VDCs and underestimated the damage in more affected VDCs.

The fact that extreme values were not covered sufficiently by the models could indicate that some predictor variables are missing in the model. This missing variable(s) should be able to explain higher differences. Since the model underestimated damage for a lot of highly populated VDCs it was analysed if a distinction between rural and urban areas would improve the predictions. In the Nepalese Census of 2011 in total 58 VDCs are labelled as urban areas (Nepal Central Bureau of Statistics, 2012). Of these 58 VDCs only 12 are within the study area. These are not enough data points to test the model on urban areas only. However, by leaving these 12 VDCs out the models can be applied to rural areas only. As a result, the LM1 R^2 increased from 0.52 to 0.53, the RF1 stayed the same with R^2=0.63 and the RF2 increases a little from 0.60 to R^2=0.61 (all on testing data). Future model training can show if another distinction between urban and rural entities can further improve the predictive accuracies of the models. Another critical remark here is that due to the fact that the model is trained on an uncertain quantification of the impact, the model's predictions could be closer or further away from the real situation than the error measures indicate.

During the out-of-sample validation, the importance of the hazard and exposure variables was confirmed again. The RF models were able to explain 49% of the variance in damage with MI and population as the only input variables. Similarly, the LM explained 48%. By adding the other 22 variables the predictive accuracy gradually increases. Based on this it can be stated that depending on data availability the exact set of included variables can differ per event or country, but that the MI and population should always be included.

The final methodological step of the research was to compare the selected models in terms of their predictive accuracy and general usability for its target users. For all measures of predictive accuracy (R^2 and RMSE) the RF1 model performed best. Also for the prediction of correct priority levels the RF1 performed best, predicting the right level for 53% of the 222 VDCs in the test dataset and 69% of the VDCs in the highest priority category. Apart from the quantitative predictive accuracy of the models, their performance can be assessed by their usefulness for targeted users. From an admin-user perspective the RF2 model, despite the relatively high prediction error, will be most suitable for continuation of model training because there are less model assumptions in comparison to LM1 and the response variable requires less data and is likely more reliable in comparison to RF1. Also from an end-user perspective the RF2 model has the advantage that its output is more intuitive and can be more easily enriched in information.

From this case study, it could be concluded that in general the RF2 model is most favourable, since the higher usability for both admin- and end-users outweighs the small decrease in predictive accuracy in comparison to RF1. Final conclusions on this matter can only be drawn after model verification for other countries has taken place. Nevertheless, the linear model approach should not be disregarded completely in future model training cases because of the insights it provides about whether or not a certain relationship between a predictor and response variable is case- of country-specific.

6.3.1 Model Extrapolation

MQ: Based on a case study of the Gorkha 2015 earthquake, what is the usability of pre- and post-event open data of past earthquakes in estimating priority areas for humanitarian aid rapidly after an earthquake at any place on earth?

Finally, what remains is to discuss the possibilities to extrapolate the developed Priority Index Model to other earthquakes, both within Nepal and outside of Nepal. In the end, this determines the usability of pre- and post-event open data for estimating aid priority areas at any place on earth, which forms the main research question. Based on the Gorkha case study, several statements can be made about the expectations and preconditions for usability of the applied modelling approach for future events.

For the model in its current state, it is likely that an estimation of similar accuracy can be produced for a future event in Nepal. The main reason for this expectation is that for the whole of Nepal data on the 24 exact same input variables is available. A similar estimation means that around two-third of the highest priority areas will be identified correctly. Similar to the Gorkha model output, it is likely that the estimation will not cover extreme values very well. Damage in heavily damaged VDCs will thus likely be underestimated and in the least damaged VDCs it will be overestimated. No impact assessment or validation data is required to produce a useful model output. However, when these data become available after several days the prediction can be improved. Damage assessment numbers from individual VDCs can be added as single observations to the training dataset. Additionally, a future model estimation could be improved by decreasing the training dataset to include only VDCs that resemble the now affected VDCs. For example, in case a new earthquake in Nepal occurs in the Eastern part of the country, not affecting any dense urban districts like Kathmandu and Dhading, the training dataset could be limited to the less populated VDCs. This would require the definition of a good threshold for 'more populated' and 'less populated' areas that can be applied to any country. As described above, applying the distinction made by the Nepalese government hardly improved the models, though some positive change was observed.

Next, something can be said about the usability of the current model for an event outside of Nepal. Even if data on all of the input variables is available and prepared, based on the current model, which is trained on only one case, the model output will likely not be accurate. This is expected mostly due to the expected presence of case- or country-specific relations in the current model. These relations were partially highlighted by the linear model applied in this study, but only additional model training on other events can confirm this, after which they can be eliminated and possibly replaced.

However, after training the model on multiple cases across different countries and if the same input data is collected for at least one variable in each category (hazard, exposure, physical vulnerability and socio-economic vulnerability) the current model can be expected to produce an estimation that can be useful to support relief distribution decision-making. The extent of this usability depends on several factors. First of all, the usability stands or falls with data availability and -preparedness. Data of some variables most certainly can be obtained for other countries also, but other variables will be more difficult to obtain in a similar manner.

The hazard variable data obtained through USGS ShakeMaps will be available for all significant earthquakes at any place on earth (USGS, 2017a). Data on total population numbers for each

geographical entity within a country can usually be obtained through openly available census data. In case these are not available previously mentioned WorldPop data could be used. When combined with a Shapefile of administrative entities population density can be calculated from the population numbers. Harmonization of input data sources can also contribute to more similar model performances between estimations for different countries. The slope variable, as one of the physical vulnerability variables, can also be defined uniformly for any country through DEMs obtained through SRTM Digital Elevation GeoPortal. Building material data on the other hand, is less widely available. As explained above, differing materials per country will make the use of composite building quality variables inevitable. For the socio-economic vulnerability variables the availability differs. Literacy rates and household size data are usually obtainable through national population censuses or national bureaus of statistics. Drinking water sources, toilet presence and school attendance are less prevalent. Such information is often provided by NGOs and can be found through portals such as the HDX platform. Not only because these data are not so prevalent, but also because their relationship to damage is likely not universal (except for drinking water source) in the future they could be replaced by an alternative variable. This can be either an existing socio-economic vulnerability index or a suitable indicator. The precondition for this alternative variable is that it should be standardized and available for any country on a low administrative level (preferably level 4).

Despite the fact that future model estimations are expected to be valuable for aid prioritisation decision making, some critical remarks should be made. One remark is that there are many factors influencing the severity of damage to houses that differ between countries. For each seismic event the mechanisms coming with it will be different. Concerning secondary hazards for example, earthquakes can lead to fires, liquefaction or tsunamis, which can all cause severe damage to houses. Such factors are unique to the environment that an event takes place within. They could be included for specific cases only, just like slope was included as a landslide susceptibility variable in this study's model. However, because this requires training on many additional predictor variables it increases model complexity. This study aims at model simplicity to make it more universally applicable. From that perspective, also slope could be excluded from the model. Another remark is that besides the 24 defined predictor variables there are many more factors that play a role. For example, houses can be damaged because of their height or because high neighbouring buildings collapse. No model could ever cover all damage contributing factors. Neither should a PIM aim to do so.

In short, the extent to which the model can be applied across different countries and events can be improved by: excluding secondary hazard susceptibility variables, finding an alternative uniform socio-economic vulnerability variable and using composite building quality variables. These are all actions that, while improving the general applicability of the model, will likely also worsen the performance for individual cases. But especially the need for a rapidly produced model output is an argument to take these steps.

Finally, it should be mentioned that usability of the model output also relates to the availability of alternative information source that can help aid prioritisation decision making. Generally, there is an information scarcity in the immediate post-disaster phase. A PIM output could in this situation fill a gap by providing a visual and spatial overview of the impact. In any situation, PIM output should be field verified and used as a supplementary information source.

6.4 Suggestions for Follow-up Research

Due to the explorative nature of this research these suggestions for follow-up research mainly concern the next logical steps to be taken towards the development of an Earthquake PIM. First of all, it will be necessary to extent the model's reliability by training it on more cases. When adding new training cases to the model, it is advisable to select a suitable case based on the availability of a reliable and complete damage assessment dataset. This data should take the same form as the one used in this study: absolute number of completely damaged residential buildings per geographical entity (preferably on a low administrative level). Apart from impact data, total population numbers and a ShakeMap covering the impacted area are the minimum requirements for new training cases. The structured training dataset and statistical analysis script (see Appendix VII) of the Gorkha case can speed up the process of future training. Depending on data availability, several predictor variables might be eliminated from the model. It can be expected that a limited number of predictor variables and a clearly measurable output variable will result in a model producing more reliable predictions for other events in other countries. Model simplicity can thus contribute to generalizability of the output. However, further research should verify this hypothesis.

Another interesting research could be to further investigate the suitability of damage to residential buildings as a proxy indicator for humanitarian aid neediness. One of the interviewees stated: "The next thing is to know, after knowing what houses are damaged, is where people went. Since people are the target of relief, you want to know where they are, rather than their damaged homes" (Becks, 2016). Such a topic might be investigated by statistically analysing the relationship between damage assessment data and aid distribution data.

Another suggestion is to further investigate a crucial part for the successful usage of PIMs in general. This concerns the way in which PIM output can best be implemented and presented. Regarding presentation, a PIM output should not be presented in isolation, but complemented with (visual) information on the communities' vulnerability and pre-event situation. This additional information could be classified based on the related aid cluster, since these are determinant for the structure of aid coordination. Becks suggested that leading organizations of the separate cluster could be the entry point for implementing PIM: "PIMs could be implemented in the pre-deployment training. At the Red Cross everyone takes a two-week training before being deployed to an emergency area, called FACT training. Here they could learn how to make sense of such a model" (Becks, 2016).

A final suggestion concerns the absence of a single composite variable indicating socio-economic vulnerability on a low administrative level. The objective of this research could be to identify a uniform socio-economic vulnerability variable that covers many disaster prone countries on a low administrative level. One aspect to look into here could be the option to create an index of the different socio-economic variables included in the current model. Than for other countries a similar index can be constructed if not all the same variables are available. A proposed methodology for this is Principle Component Analysis.

7 Conclusion

This chapter provides a brief summary of the overall conclusions drawn from the discussed results based on the research objectives. It answers the main research question.

MQ: Based on a case study of the Gorkha 2015 earthquake, what is the usability of pre- and post-event open data of past earthquakes in estimating priority areas for humanitarian aid rapidly after an earthquake at any place on earth?

To summarize, for the Gorkha case study the collected pre- and post-event open data proved to be substantially useful for the prediction of multi-cluster aid neediness, and thereby the identification of priority areas. A random forest regression model predicting the absolute number of completely damaged houses per VDC explained 60% of the variance in damage. A better prediction could be made by adding information on the number of partially damaged houses to the response variable. However, partial damage inherently is an inconsistent measure and the prediction of only completely damaged houses has advantages for both admin and end-users.

The main objective of this study was to explore the possibilities and feasibility of using pre- and post-event open data to train a model to rapidly estimate post-earthquake aid neediness for any earthquake prone area on earth. This objective is reflected in the main research question. Based on the above discussion of the research findings, conclusions can be drawn about the general usability of both pre- and post-event open data for Earthquake PIMs.

First of all, it is important to mention here that this study was explorative of nature and comprised one single-event case study. Nevertheless, the findings obtained during the study led to expectations about the performance of an Earthquake PIM covering any place on earth. A random forest model trained on the Gorkha case correctly predicted the highest aid priority level for two third of the observations. Therefore, for the case study itself both pre- and post-event open data have proven to be sufficiently usable to estimate aid priority areas up to the extent were they could provide useful information for post-event aid distribution. It was observed that extreme values were not covered adequately by the model, indicating that some explanatory variable is still missing.

For a future event within Nepal a model output of similar accuracy is expected, though the training dataset might have to be adjusted to represent the newly affected area more in terms of population numbers. For a future event outside of Nepal, the current model, trained on only the Gorkha case, will not be useful due to the presence of case- and country-specific relationships between predictor and response variables. This concerns socio-economic vulnerability and single building material variables.

However, after training the model on multiple other events across different countries it is expected that if the same input data is collected for at least one variable in each category (hazard, exposure, physical vulnerability and socio-economic vulnerability) for a future event, the same model can produce an estimation that is useful to support relief distribution decision-making. Data availability and preparedness are key factors in the actual usability of the model output. Fortunately, data on the most important predictor (hazard and exposure) is widely and uniformly available. For physical and socio-economic vulnerability predictors alternative methods of including them in a generalized manner might have to be sought. Additionally, the usability of model output can also be assigned to the overall

information-scarcity (usually) existing in the immediate post-disaster phase. Model output will always have to be field verified and used supplementary.

The extent to which the model can be successfully applied to different countries and cases can be improved by excluding secondary hazard susceptibility variables, finding an alternative uniform socio-economic vulnerability variable and using composite building quality variables. Additionally, by experimenting with different combinations of predictor variables and adding more training cases, country- or case-specific predictors can be eliminated with time. In general, overfitting a PIM to specific cases or countries is a pitfall, therefore a next important step in developing these models is to combine multiple cases in one training dataset.

Apart from this general usability, the research has also provided insights that are valuable for further model development. First of all, despite the limitations in data quality discussed earlier, especially hazard and exposure related open data have proven to be very useful for the prediction of structural damages. As was expected based on existing earthquake impact models, both had a strong and positive influence on structural damages. These could be considered as indispensable model components and are widely available.

Furthermore, during the initial research stage there turned out to be many initiatives that support and facilitate the collection and dissemination of open data that can be used for earthquake PIMs, for example USGS ShakeMap, OpenStreetMap (OSM) and Humanitarian Data Exchange (HDX). A valuable aspect of these data is that they are often uniform in standards and measures across different countries. Besides these initiatives, national population and housing censuses remain a valuable resource for both country-specific exposure and vulnerability related data.

Finally, based on the Gorkha case it can be stated that the reliability of model output greatly depends on data availability. As a result of this, data preparedness will likely be a major factor in adapting to the unexpectedness that comes with seismic hazards.

Concerning the usability of post-event open data specifically, it was found that residential building damage assessment data are a relatively objective and common measure that relates to multiple country specific factors. As a response variable it produces a model output that can be relevant for multiple humanitarian aid clusters. Nevertheless, a great drawback is that training a PIM on voluntary collected rapid assessment data means training on an estimation rather than on a real situation. Another important post-event dataset concerned the hazard data. Despite its uncertainties, in each model the USGS ShakeMap was an important predictor of structural damages. In combination with its spatial coverage and rapid and open dissemination it has proved to be a valuable source for the further development of Earthquake PIM's.

Also conclusions can be drawn about the usability of certain pre-event open data. It was already mentioned that population had an important function in the model. The physical and socio-economic vulnerability predictors on the other hand were harder to capture in the models and appeared to be less obvious related to damaged. This provides another argument to advise the use of standardized or composite measures of communities' vulnerabilities. Additionally, if data on building materials is available for other countries it is like that the materials differ. Therefore, a generalization of building materials into composite building quality variables is likely inevitable.

Based on the comparison of three different models, a maximum usability of the collected pre- and post-event data can be reached by applying a random forest regression algorithm predicting only the number of completely damaged houses. For targeted end-users this model has the advantage of an intuitive

output which can be easily enriched. Also for admin-users, the limited model assumptions, reliable output and less data required make it favourable over the other statistical approaches. A machine learning approach brings advantages mainly in terms of predictive accuracy and time spent on constructing. Nevertheless, the multivariate linear regression model has proven to be valuable especially in this explorative phase, because of the insights it gave in individual variable relationships.

References

510, 2016. A Priority Index for Humanitarian Aid after a Typhoon. Available at: http://510.global/philippines-typhoon-haima-priority-index/ [Accessed June 5, 2017].

ACAPS, 2016a. *Meeting Information Needs? A review of ten year of multisector coordinated needs assesment reports*, Available at: http://www.alnap.org/resource/21659 [Accessed May 15, 2017].

ACAPS, 2016b. Who we are | ACAPS. Available at: https://www.acaps.org/who-we-are [Accessed September 28, 2016].

Aichholzer, G. & Burkert, H., 2004. *Public Sector Information in the Digital Age: Between Markets, Public Management and Citizens' Rights*, Edward Elgar Publishing. Available at: https://books.google.nl/books?id=a0AbDHMb5rAC&source=gbs_navlinks_s [Accessed May 14, 2017].

Anon, 2016. The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge. Available at: http://opendefinition.org/ [Accessed November 5, 2016].

Becks, Michel. The Netherlands Red Cross. Interviewed by: Bulte, Evelien. (16 December 2016).

Benini, A., 2015. *The Use of Data Envelopment Analysis*, A note for ACAPS. Available at: http://aldo-benini.org/Level2/HumanitData/Acaps_150708_Using_DEA_for_prioritization.pdf [Accessed December 10, 2016].

Benini, A. & Chataigner, P., 2014. *Composite measures of local disaster impact - Lessons from Typhoon Yolanda, Philippines*. Available at: http://www.alnap.org/resource/13008 [Accessed October 20, 2016].

Bird, J.F. & Bommer, J.J., 2004. Earthquake losses due to ground failure. *Engineering Geology*, 75(2), pp.147–179.

Breiman, L., 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), pp.199–231.

Calvi, G.M. et al., 2006. Development of Seismic Vulnerability Assessment Methodologies over the Past 30 Years. *ISET Journal of Earthquake Technology*, 43(3), pp.75–104.

Cardona, O.D. et al., 2012. Determinants of Risk: Exposure and Vulnerability. In *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University Press, pp. 65–108. Available at: https://www.ipcc.ch/pdf/special-reports/srex/SREX-Chap2_FINAL.pdf [Accessed May 14, 2017].

Caruana, R. & Niculescu-Mizil, A., 2006. An Empirical Comparison of Supervised Learning Algorithms. Available at: https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf [Accessed January 13, 2017].

CGIAR Consortium for Spatial Information, 2004. SRTM Data Search. Available at: http://srtm.csi.cgiar.org/SELECTION/inputCoord.asp [Accessed May 17, 2017].

Comes, T., Vybornova, O. & Van de Walle, B., 2015. Bringing Structure to the Disaster Data Typhoon:

An Analysis of Decision-Makers' Information Needs in the Response to Haiyan. Available at: https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10288 [Accessed May 15, 2017].

CRED, 2015. *The Human Cost of Natural Disasters, A Global Perspective*. Available at: http://emdat.be/human_cost_natdis [Accessed September 26, 2016].

Darcy, J. & Hofmann, C.-A., 2003. *According to need? Britain's leading independent think-tank on international development and humanitarian issues*, Londen. Available at: www.odi.org.uk [Accessed May 14, 2017].

District Development Committee Lamjung, 2014. Lamjung District. Available at: http://www.ddclamjung.gov.np [Accessed May 19, 2017].

Ebener, S., Castro, F. & Dimailig, L.A., 2014. *Increasing Availability, Quality, and Accessibility of Common and Fundamental Operational Datasets to Support Disaster Risk Reduction and Emergency Management in the Philippines*, Available at: http://www.gaia-geosystems.org/PROJECTS/SIIEM/PHL/Green_Paper_DSWD-SIIEM_305014.pdf [Accessed May 15, 2017].

Erdik, M. et al., 2011. Rapid earthquake loss assessment after damaging earthquakes. *Soil Dynamics and Earthquake Engineering*, 31(2), pp.247–266.

ESRI, 2017. How Slope works—Help | ArcGIS for Desktop. Available at: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-slope-works.htm [Accessed June 2, 2017].

Fiedrich, F., Gehbauer, F. & Rickers, U., 2000. Optimized resource allocation for emergency response after earthquake disasters. *Safety Science*, 35(1), pp.41–57.

Government of Nepal National Planning Commission, 2015. *Post Disaster Needs Assessment*, Vol. A: Key Findings, Kathmandu. Available at: http://www.recoveryplatform.org/jp/pdf/Nepal_PDNA%20Volume%20A%20Final.pdf [Accessed November 17, 2016].

Hammer, B. & Villmann, T., 2007. How to process uncertainty in machine learning? European Symposium on Artificial Neural Networks Bruges (Belgium). Available at: https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2007-7.pdf [Accessed December 17, 2016].

Hancilar, U. et al., 2010. ELER software – a new tool for urban earthquake loss assessment. *Nat. Hazards Earth Syst. Sci*, 10, pp.2677–2696. Available at: www.nat-hazards-earth-syst-sci.net/10/2677/2010/ [Accessed November 28, 2016].

HDX, 2016. About - Humanitarian Data Exchange. Available at: https://data.humdata.org/about/terms [Accessed November 24, 2016].

HDX, 2015. Nepal earthquake landslide locations, 30 June 2015 - Humanitarian Data Exchange. Available at: https://data.humdata.org/dataset/nepal-earthquake-landslide-locations-30-june-2015 [Accessed December 7, 2016].

Hofmokl, J., 2010. The Internet commons: towards an eclectic theoretical framework. *International Journal of the Commons*, 4(1), pp.226–250.

Homberg van den, M., Monne, R. & Spruit, M., 2016. Bridging the Information Gap: Mapping Data Sets on Information Needs in the Preparedness and Response Phase. Available at: https://www.cordaid.org/en/publications/bridging-information-gap/ [Accessed May 15, 2017].

Housing Recovery and Reconstruction Platform, 2016. Nepal - Who's Doing What Where. Available at: https://data.humdata.org/dataset/160625-hrrp-4w-national [Accessed May 16, 2017].

IASC, 2012a. *Multi-Cluster/Sector Initial Rapid Assessment*, Geneva. Available at: https://docs.unocha.org/sites/dms/documents/mira_final_version2012.pdf [Accessed November 22, 2016].

IASC, 2012b. *Operational Guidance for Coordinated Assessments in Humanitarian Crises perational Guidance for Coordinated Assessments in Humanitarian Crises*, Geneva. Available at: https://docs.unocha.org/sites/dms/CAP/ops_guidance_finalversion2012.pdf [Accessed January 5, 2016].

INFORM, 2015. Nepal Earthquake Severity Index - Version 4, 30 April Severity Index. Available at: http://reliefweb.int/sites/reliefweb.int/files/resources/nepal_earthquake_severity_index_version_4_30_april.pdf [Accessed May 14, 2017].

International Organization for Standardization, 1998. ISO 9241-11:1998(en), Ergonomic requirements for office work with visual display terminals (VDTs) — Part 11: Guidance on usability. Available at: https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-1:v1:en [Accessed May 14, 2017].

IOM, 2015. Nepal - IOM DTM Earthquake Dataset - Site assessment data. Available at: https://data.humdata.org/dataset/io [Accessed May 16, 2017].

Jaiswal, K., Wald, D.J. & Hearne, M., 2009. *Estimating Casualties for Large Earthquakes Worldwide Using an Empirical Approach*, Available at: http://earthquake.usgs.gov/ [Accessed September 29, 2016].

Jaiswal, K.S. & Wald, D.J., 2008. *Development of a Semi-Empirical Loss Model Within the USGS Prompt Assessment of Global Earthquakes for Response (PAGER) System.* Available at: ftp://hazards.cr.usgs.gov/web/data/pager/Jaiswal_Wald_2010_Semi.pdf [Accessed October 10, 2016].

Johnson, S., 2015. Disaster Impact Before Ground Assessments – Medium. Available at: https://medium.com/@Simon_B_Johnson/disaster-impact-before-ground-assessments-f0e128f1af19#.8rhw3csv2 [Accessed October 5, 2016].

Johnson, S.B., 2015. Disaster Impact Before Ground Assessments – Medium. Available at: https://medium.com/@Simon_B_Johnson/disaster-impact-before-ground-assessments-f0e128f1af19#.g5h7mc3u2 [Accessed October 18, 2016].

King, S.A. & Rojahn, C., 1996. *A Comparison of Earthquake Damage and Loss Estimation Methodologies.* Eleventh World Conference on Earthquake Engineering. Paper No. 1482. Available at: http://www.iitk.ac.in/nicee/wcee/article/11_1482.PDF [Accessed October 25,

2016].

Knight, Paul. British Red Cross. Interviewed by: Bulte, Evelien. (12 December 2016).

Kohara, K. & Hasegawa, R., 2009. Typhoon Damage Forecasting with Self-Organizing Maps, Multiple Regression and Decision Trees. Available at: http://vigir.missouri.edu/~gdesouza/Research/Conference_CDs/IFAC_ICINCO_2009/ICINCO/Workshops/Workshop ANNIIP/ANN Applications in Prediction & Forecasting/Workshop ANNIIP_2009_23_CR.pdf [Accessed May 15, 2017].

Lang, D.H., 2012. *Earthquake Damage and Loss Assessment – Predicting the Unpredictable*. Doctoral Thesis, University of Bergen. Available at: http://bora.uib.no/handle/1956/6753 [Accessed November 1, 2016].

Lanjouw, P. & Ravallion, M., 1994. *Poverty and household size*, Available at: http://documents.worldbank.org/curated/en/641891468741345861/Poverty-and-household-size [Accessed May 17, 2017].

Liaw, A. & Wiener, M., 2002. Classification and Regression by randomForest. *R News*, 2(3), pp.18–22. Available at: http://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf [Accessed May 15, 2017].

Lumley, T., 2017. *Regression Subset Selection*, Available at: https://cran.r-project.org/web/packages/leaps/leaps.pdf [Accessed May 16, 2017].

Ministry of Home Affairs Nepal, 2015. Nepal Disaster Risk Reduction Portal. Available at: http://drrportal.gov.np/ [Accessed May 16, 2017].

Nepal Central Bureau of Statistics, 2012. *National Population and Housing Census 2011 (National Report)*, Kathmandu. Available at: https://unstats.un.org/unsd/demographic/sources/census/wphc/Nepal/Nepal-Census-2011-Vol1.pdf [Accessed May 16, 2017].

Nepalese Red Cross Society, 2015a. *Initial Rapid Assessment*. Obtained through personal contacts at Brithish Red Cross.

Nepalese Red Cross Society, 2015b. *NRCS SitRep*, Available at: https://trello.com/c/ocRnF7O2/169-nrcs-sitrep-leaflet-damages-distributions-responders [Accessed May 16, 2017].

OCHA, 2015a. *Concept Note Severity Index Nepal*. Available at: http://un.info.np/Net/NeoDocs/View/4893 [Accessed December 14, 2016].

OCHA, 2015b. *Nepal Flash Appeal Revision Earthquake*, Available at: http://reliefweb.int/sites/reliefweb.int/files/resources/nepal_earthquake_2015_revised_flash_appeal_june.pdf [Accessed May 14, 2017].

OCHA Nepal, 2015. Nepal: Official figures for casualties and damage. Available at: https://data.humdata.org/dataset/official-figures-for-casualties-and-damage [Accessed May 16, 2017].

Ortmann, J. et al., 2011. Crowdsourcing Linked Open Data for Disaster Management. Available at:

http://ceur-ws.org/Vol-798/proceedings.pdf. [Accessed January 23, 2017].

Ortuño, M.T. et al., 2013. Decision Aid Models and Systems for Humanitarian Logistics. A Survey. In Atlantis Press, pp. 17–44. Available at: http://link.springer.com/10.2991/978-94-91216-74-9_2 [Accessed September 8, 2016].

OSOCC Assessment Cell, 2015. *Nepal Earthquake District Profiles*, Available at: https://www.humanitarianresponse.info/sites/www.humanitarianresponse.info/files/assessments/090515_gorkha_district_profile_osocc_assessment_cell_0.pdf [Accessed May 16, 2017].

Pedraza-Martinez, A.J., 2013. On the Use of Information in Humanitarian Operations. In *Decision Aid Models for Disaster Management and Emergencies*. pp. 1–16. Available at: http://link.springer.com/10.2991/978-94-91216-74-9_6 [Accessed September 30, 2016].

Rohrer, B., 2016. How to choose machine learning algorithms | Microsoft Azure. Available at: https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/ [Accessed November 7, 2016].

Smolka, A. et al., 2004. *The Principle of Risk Partnership and the Role of Insurancein Risk Mitigation*, Vancouver. Available at: http://www.iitk.ac.in/nicee/wcee/article/13_2020.pdf [Accessed October 21, 2016].

Techsansar, 2016. Village Development Committees (VDCs) in Nepal. Available at: http://techsansar.com/vdc-nepal-list/ [Accessed October 13, 2016].

The Sphere Project, 2011. *Humanitarian Charter and Minimum Standards in Humanitarian Response*, Available at: http://www.ifrc.org/PageFiles/95530/The-Sphere-Project-Handbook-20111.pdf [Accessed May 14, 2017].

The World Bank, 2017. Population Total Nepal. Available at: http://data.worldbank.org/indicator/SP.POP.TOTL [Accessed May 20, 2017].

USGS, 2017. PAGER Scientific Background. Available at: https://earthquake.usgs.gov/data/pager/background.php [Accessed May 14, 2017].

USGS, 2017a. ShakeMap. Available at: https://earthquake.usgs.gov/data/shakemap/ [Accessed June 5, 2017].

USGS, 2017b. ShakeMap Documentation: Products and Formats. Available at: https://usgs.github.io/shakemap/products.html [Accessed May 16, 2017].

USGS, 2015. ShakeMap M 7.8 - 36km E of Khudi, Nepal. Available at: https://earthquake.usgs.gov/earthquakes/eventpage/us20002926#shakemap [Accessed May 16, 2017].

USGS, 2016a. ShakeMap Scientific Background. Available at: http://earthquake.usgs.gov/earthquakes/shakemap/background.php [Accessed December 7, 2016].

USGS, 2016b. ShakeMaps. Available at: http://earthquake.usgs.gov/earthquakes/shakemap/ [Accessed October 3, 2016].

Verity, A., 2014. Estimating Disaster Impact with contextual Pre and... Available at: http://blog.veritythink.com/post/104948097769/estimating-disaster-impact-with-contextual-pre-and [Accessed October 5, 2016].

Vetrò, A. et al., 2016. Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly*, 33(2), pp.325–337.

Wald, D.J. et al., 2008. Development of the U.S. Geological Survey's PAGER system (Prompt Assessment of Global Earthquakes for Response).

Wilson, R. et al., 2016. Rapid and near Real-time Assessments of Population Displacement Using Mobile Phone Data Following Disasters: The 2015 Nepal Earthquake. *PLoS Currents*. Available at: http://currents.plos.org/disasters/?p=27109 [Accessed May 16, 2017].

World Health Organization, 2011. *Guidelines for drinking-water quality*, World Health Organization. Available at: http://www.who.int/water_sanitation_health/publications/dwq-guidelines-4/en/ [Accessed May 17, 2017].

Yamazaki, Fumio & Meguro, Kimiro, 1998. Developments of Early Earthquake Damage Assessment Systems in Japan. Proceeding of ICOSSAR '97, pp. 1573-1580.

Yucel, G. & Arun, G., 2012. Earthquake and Physical and Social Vulnerability Assessment for Settlements: Case Study Avcilar District. Available at: http://www.iitk.ac.in/nicee/wcee/article/WCEE2012_3676.pdf [Accessed May 14, 2017].

Yun, S.-H. et al., 2015. Rapid Damage Mapping for the 2015 M w 7.8 Gorkha Earthquake Using Synthetic Aperture Radar Data from COSMO–SkyMed and ALOS-2 Satellites. *Seismological Research Letters*, 86(6), pp.1549–1556. Available at: http://authors.library.caltech.edu/62605/1/1549.full.pdf [Accessed May 16, 2017].

Appendices

Appendix I – IRA Assessment Template

Nepal Red Cross Society
…….Sub Chapter……..
VDC level rapid assessment templat-2066

VDC/Municipality:
Date of Disaster

S. No.	Ward no	Population			Affected families			Displaced families		Damages household		Needs identified
		Deaths	Missing	Injured	Households	Population		Household	Population	Completely	Partially	
						Male	Female					

1. Damaged community and individual property

S. No.	Descriptions of damages	Number	Total estimated loss (in amount)	Approach to collect information	Remarks
1	Completely destroyed houses				
2	Partially destroyed houses				
3	School building				
4	Bridges/Roads				
5	Community resource centre(VDC, Library, evacuation shelter)				
6	Health post				
7	Cultivated crops (in Bigaha/ Ropani)				
8	Fertile lands (in Bigaha/ Ropani)				
9	Livelihood scopes • Livestock • IG scheme • fish ponds • Small shops • etc				
10	Others				

2. Partners in the field

S. No.	Descriptions of response sector	Name of the organization				
		VDC				
1	For example, Support to house maintenance/ reconstruction					
2	NFRI distribution					
3	Water and sanitation service					
4						

Prepared by:
Name:
Position:
Office stamp:

Approved by:
Name:
Position:
Office stamp:

Appendix II – Description of Building Materials

Based on the predominant building types in the affected areas, housing can be categorized into four main types based on their vertical and lateral load-bearing systems, in line with the 2011 Census:

• Low-strength masonry buildings are constructed locally (in available stone, fired brick and sun-dried brick) in mud mortar. They are typically two-storey buildings excluding the attic, with timber or bamboo floors overlaid with mud. The roofs are mostly of timber or bamboo covered with tiles, slate, shingles or corrugated galvanized iron (CGI) sheets. The walls tend to be very thick, depending upon the type of walling units. The seismic capacity of these buildings is very low, limited by the integrity of structural components, strength of walls, and lack of elements tying the structure together (ring beams at wall or roof level). Vertical and horizontal wooden elements are sometimes embedded in walls, providing some level of earthquake resistance, but this is very uncommon.

• Cement-mortared masonry buildings have walls of fired brick, concrete block or stone in cement-sand mortar and are usually constructed up to three storeys. The floors and roofs are made of reinforced concrete or reinforced brick concrete. Despite the use of high-quality materials, these buildings suffer from deficient construction practices. Provision of earthquake-resistant features is not commonly found in these buildings.

• Reinforced concrete frames with masonry infill consist of cast-in-situ concrete frames with masonry partition and infill walls (brick, block or stone masonry) that are not tied to the frame. With floors and roofs of reinforced concrete slabs, these buildings are usually constructed up to four storeys, but buildings up to even 20 storeys have been observed. Despite the use of high quality materials and the fact that seismic detailing has become more common in recent years, the vast majority of these buildings suffer from deficient construction practices.

• Wood and bamboo buildings are constructed with wooden planks, thatch or bamboo strip walling materials, with flexible floors and roof. These suffered less damage from the earthquake due to their light weight

Source: Government of Nepal National Planning Commission, 2015

Appendix III – Data Exploration

```
      id              vdc_name          hlcit_code        dist_name          pcode
Min.   :   1.0   Length:612         Length:612         Length:612         Length:612
1st Qu.: 192.0   Class :character   Class :character   Class :character   Class :character
Median : 719.5   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mean   : 936.4
3rd Qu.: 896.2
Max.   :3403.0
```

```
   hh_total           hhsize          slope       compl_damg_houses pct_compl_damg_houses part_damg_houses
Min.   :   118.0   Min.   :2.620   Min.   : 2.24   Min.   :   1.0    Min.   : 0.00          Min.   :    0.0
1st Qu.:   510.8   1st Qu.:4.180   1st Qu.:20.36   1st Qu.: 134.5    1st Qu.: 19.01         1st Qu.:   47.0
Median :   761.0   Median :4.520   Median :23.75   Median : 438.5    Median : 65.00         Median :  123.3
Mean   :  1531.9   Mean   :4.549   Mean   :23.06   Mean   : 618.4    Mean   : 64.96         Mean   :  322.1
3rd Qu.:  1063.8   3rd Qu.:4.872   3rd Qu.:26.65   3rd Qu.: 816.0    3rd Qu.: 97.99         3rd Qu.:  269.8
Max.   :254292.0   Max.   :6.700   Max.   :36.11   Max.   :9012.0    Max.   :720.34         Max.   :23705.0
                                                                                             NA's   :41
school_attendance literacy_rate   pct_lowstrength_found pct_lowstrength_walls pct_lowstrength_roof   mud_found
Min.   :0.1700    Min.   :28.91   Min.   :0.1589       Min.   :0.1375        Min.   :0.006873      Min.   :    3.0
1st Qu.:0.6641    1st Qu.:57.64   1st Qu.:0.9794       1st Qu.:0.9604        1st Qu.:0.211361      1st Qu.:  452.8
Median :0.7227    Median :62.34   Median :0.9936       Median :0.9865        Median :0.473018      Median :  666.0
Mean   :0.6938    Mean   :61.99   Mean   :0.9569       Mean   :0.9431        Mean   :0.507923      Mean   :  862.1
3rd Qu.:0.7620    3rd Qu.:67.27   3rd Qu.:0.9981       3rd Qu.:0.9957        3rd Qu.:0.815544      3rd Qu.:  934.8
Max.   :0.8649    Max.   :88.52   Max.   :1.0000       Max.   :1.0000        Max.   :0.996169      Max.   :37941.0
```

```
  cem_found          rcc_found          wood_found        thatch_roof        tap_water          tube_water
Min.   :     0.0   Min.   :     0   Min.   :   0.00   Min.   :   0.0   Min.   :    0.59   Min.   :    0.0
1st Qu.:     0.0   1st Qu.:     0   1st Qu.:   1.00   1st Qu.:  19.0   1st Qu.:  395.50   1st Qu.:    0.0
Median :     3.0   Median :     0   Median :   5.00   Median :  66.0   Median :  593.50   Median :    0.0
Mean   :   276.8   Mean   :   278   Mean   :  87.73   Mean   : 146.7   Mean   : 1133.37   Mean   :   44.0
3rd Qu.:    13.0   3rd Qu.:     3   3rd Qu.:  17.00   3rd Qu.: 215.2   3rd Qu.:  837.50   3rd Qu.:    1.0
Max.   :100716.0   Max.   :106607   Max.   :5925.00   Max.   :1066.0   Max.   :163339.00   Max.   :18574.0
```

```
 bamboo_wall       unbaked_wall        mud_wall          cem_wall         wood_wall
Min.   :   0.00   Min.   :   0.000   Min.   :   15.0   Min.   :     0.0   Min.   :   0.00
1st Qu.:   0.00   1st Qu.:   0.000   1st Qu.:  464.8   1st Qu.:     2.0   1st Qu.:   1.00
Median :   3.00   Median :   0.000   Median :  660.0   Median :     8.0   Median :   3.00
Mean   :  27.17   Mean   :   3.786   Mean   :  831.3   Mean   :   594.2   Mean   :  45.27
3rd Qu.:  10.00   3rd Qu.:   0.000   3rd Qu.:  926.0   3rd Qu.:    34.0   3rd Qu.:  11.00
Max.   :2089.00   Max.   :307.000   Max.   :30040.0   Max.   :212587.0   Max.   :2546.00
```

```
   mud_roof        pct_part_damg_houses      hdf              mi          pct_tapwater       pct_notoilet
Min.   : 0.0000   Min.   :  0.000      Min.   :  0.25   Min.   :4.863   Min.   :0.006928   Min.   :0.0000
1st Qu.: 0.0000   1st Qu.:  6.454      1st Qu.:156.75   1st Qu.:6.242   1st Qu.:0.736750   1st Qu.:0.2133
Median : 0.0000   Median : 15.000      Median :363.75   Median :6.788   Median :0.869015   Median :0.3494
Mean   : 0.3644   Mean   : 26.683      Mean   :537.67   Mean   :6.700   Mean   :0.813632   Mean   :0.3865
3rd Qu.: 0.0000   3rd Qu.: 34.292      3rd Qu.:650.91   3rd Qu.:7.087   3rd Qu.:0.942959   3rd Qu.:0.5373
Max.   :33.0000   Max.   :454.581      Max.   :6954.50  Max.   :8.533   Max.   :1.000000   Max.   :0.9802
                  NA's   :41           NA's   :41
```

```
  galv_roof          tile_roof          rcc_roof          wood_roof                cov_water          uncov_water
Min.   :    0.0   Min.   :   0.0   Min.   :    0.0   Min.   :  0.00  0.00   Min.   :    0.00   Min.   :   0.00
1st Qu.:  100.8   1st Qu.:  40.0   1st Qu.:    0.0   1st Qu.:  0.00  0.00   1st Qu.:    0.00   1st Qu.:   4.00
Median :  321.5   Median : 142.0   Median :    2.0   Median :  0.00  0.00   Median :    2.00   Median :  19.00
Mean   :  515.2   Mean   : 262.9   Mean   :  548.9   Mean   : 26.68  4.06   Mean   :   50.66   Mean   :  65.16
3rd Qu.:  596.5   3rd Qu.: 362.0   3rd Qu.:    9.0   3rd Qu.:  4.00  1.00   3rd Qu.:   11.00   3rd Qu.:  58.00
Max.   :37317.0   Max.   :6466.0   Max.   :204824.0   Max.   :662.00  4.00   Max.   :10890.00   Max.   :1348.00
```

```
   area_sqm             pop            pop_dens          spout_water      cuberoot_cdh       LOGpop
Min.   :  3125822   Min.   :   415   Min.   :   0.828   Min.   :   0.00   Min.   : 1.000   Min.   : 6.028
1st Qu.: 13129893   1st Qu.:  2297   1st Qu.: 112.210   1st Qu.:   5.00   1st Qu.: 5.124   1st Qu.: 7.739
Median : 20264379   Median :  3400   Median : 170.947   Median :  31.00   Median : 7.597   Median : 8.132
Mean   : 36578754   Mean   :  6560   Mean   : 314.309   Mean   :  97.19   Mean   : 7.287   Mean   : 8.172
3rd Qu.: 34216454   3rd Qu.:  4852   3rd Qu.: 251.500   3rd Qu.: 101.25   3rd Qu.: 9.345   3rd Qu.: 8.487
Max.   :702155430   Max.   :975453   Max.   :19724.494   Max.   :4830.00   Max.   :20.810   Max.   :13.791
```

```
  no_toilet        flush_toilet       ord_toilet        river_water
Min.   :   0.0   Min.   :     0.0   Min.   :   0.00   Min.   :   0.00
1st Qu.: 144.0   1st Qu.:   104.0   1st Qu.:  76.75   1st Qu.:   1.00
Median : 258.5   Median :   234.0   Median : 179.00   Median :   4.00
Mean   : 325.1   Mean   :   939.5   Mean   : 257.38   Mean   :  12.46
3rd Qu.: 423.0   3rd Qu.:   435.2   3rd Qu.: 313.25   3rd Qu.:  13.00
Max.   :2538.0   Max.   :234299.0   Max.   :17274.00   Max.   :315.00
```

Explanation of codes: id = identification code, vdc_name = name of VDC, hlcit_code = government code, dist_name = district name, pcode = OCHA p-code, hh_total = total number of households, hhsize = average household size, slope = mean slope (%), compl_damg_houses = completely damaged houses, pct_compl_damg_houses = completely damaged houses (%), part_damg_houses = partially damaged houses, school_attendance = school attendance 5 – 25 year olds (%), literacy_rate = literacy rate (%), pct_lowstrength_found = low strength foundations (%), pct_lowstrength_walls = low strength walls (%), pct_lowstrength_roof = low strength roofs (%), mud_found = mud bonded bricks/stone foundations, cem_found = cement bonded bricks/stone foundation, rcc_found = RCC

with pillar foundations, wood_found = wooden pillar foundations, thatch_roof = thatch/straw roofs, tap_water = tap water as main drinking water source, tube water = tube water as main drinking water source, bamboo_wall = bamboo outer walls, unbaked_wall = unbaked brick outer walls, mud_wall = " mud bonded bricks/stone outer walls, cem_wall = cement bonded bricks/stone outer walls, wood_wall = wood/planks outer walls, mud_roof = mud roofs, pct_part_damg_houses = partially damaged houses (%), hdf = house damage factor, mi = macroseismic intensity, pct_tapwater = tap water as main drinking water source (%), pct_notoilet = households without a toilet facility (%), galv_roof = galvanized iron roofs, tile_roof = tile/slate roofs, rcc_roof = RCC roofs, wood_roof = wood/planks roofs, cov_water = covered well as main drinking water source, uncov_water = uncovered well as main drinking water source, area_sqm = area in m^2, pop = total population, pop_dens = population density, spout_water = spout as main drinking water source, cuberoot_hdf = $hdf^{1/3}$, LOGpop = total $population^{log}$, no_toilet = households without a toilet, flush_toilet = flush toilet, ord_toilet = ordinary toilet, river_water = river water as main drinking water source.

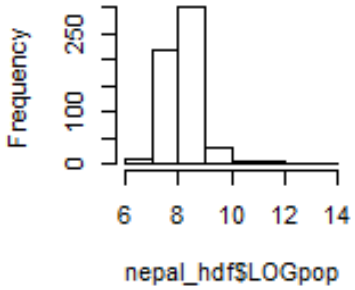
Appendix IV – Frequency Distributions Candidate Predictor Variables

Explanation of variable codes: pop = population, pop_dens = population density, hhsize = average household size, slope = mean slope value per VDC, mi = Macroseismic intensity, school_attendance = relative school attendance, literacy_rate = literacy rate, pct_lowstrength_found = low strength foundations (%), pct_lowstrength_walls = low strength walls (%), pct_lowstrength_roof = low strength roofs (%), pct_tapwater = percentage of households with tap water as their main source for drinking water, pct_notoilet = percentage of households without a toilet facility, mud_found = number of households with mud bonded bricks/stone foundations, cem_found = " cement bonded bricks/stone foundation, rcc_found = " RCC with pillar foundations, wood_found = " wooden pillar foundations,

mud_wall = “ mud bonded bricks/stone outer walls, cem_wall = “ cement bonded bricks/stone outer walls, wood_wall = “ wood/planks outer walls, bamboo_wall = “ bamboo outer walls, unbaked_wall = “ unbaked brick outer walls, thatch_roof = “ thatch/straw roofs, galv_roof = “ galvanized iron roofs, tile_roof = “ tile/slate roofs, rcc_roof = “ RCC roofs, wood_roof = “ wood/planks roofs, mud_roof = ‘’ mud roofs)

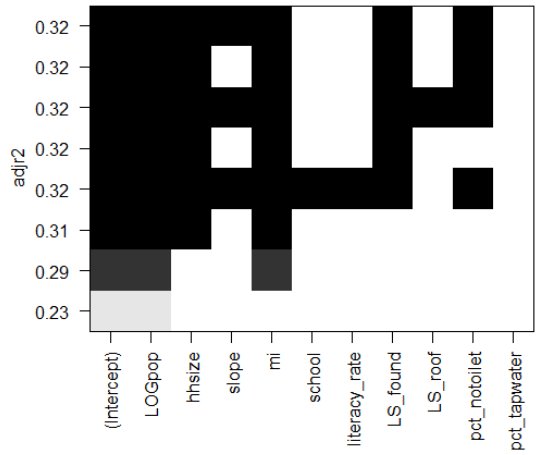
Appendix V – Frequency Distributions after Transformation

Logarithmic transformation of population variable:

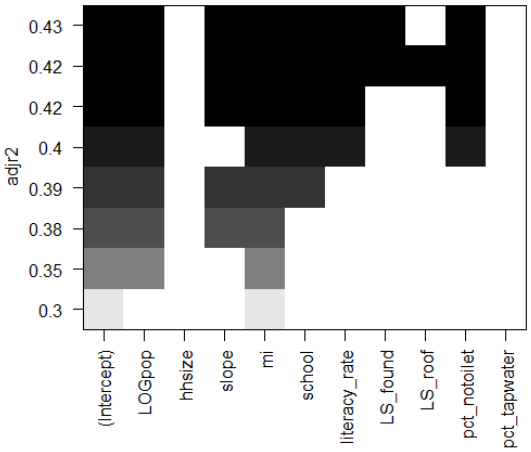


Appendix VI – Regression Subset Selection Plots

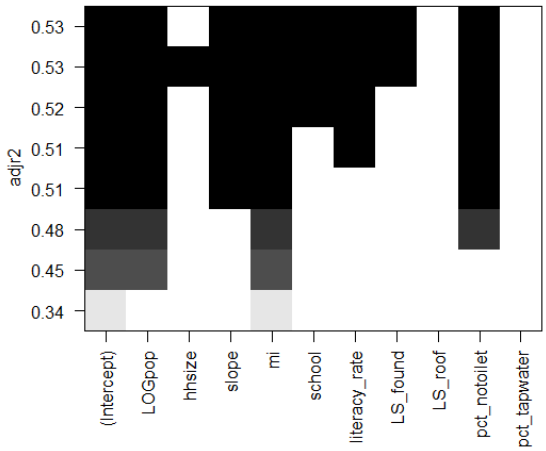
y = completely damaged houses



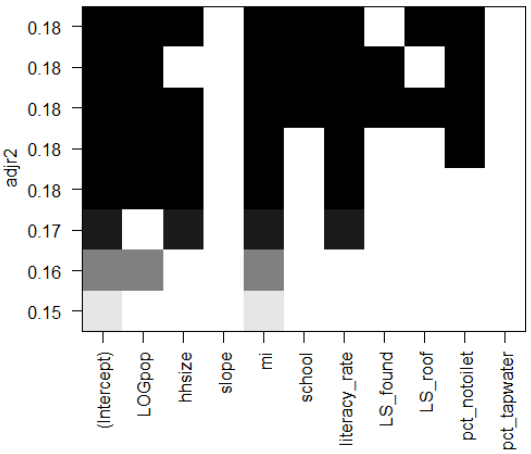
y = log(completely damaged houses)



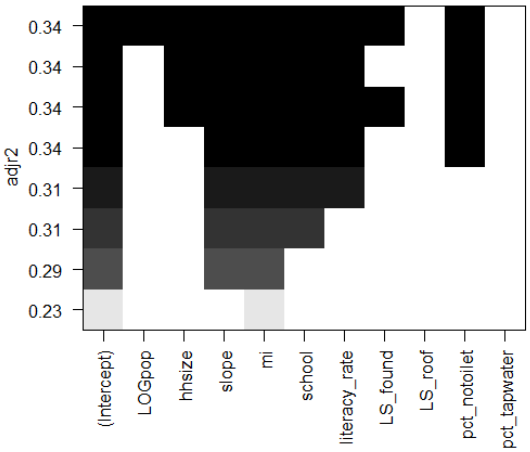
y = (completely damaged houses)^1/3



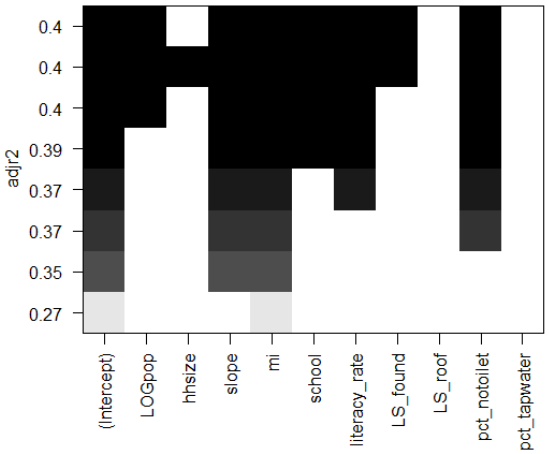
y = completely damaged houses (%)

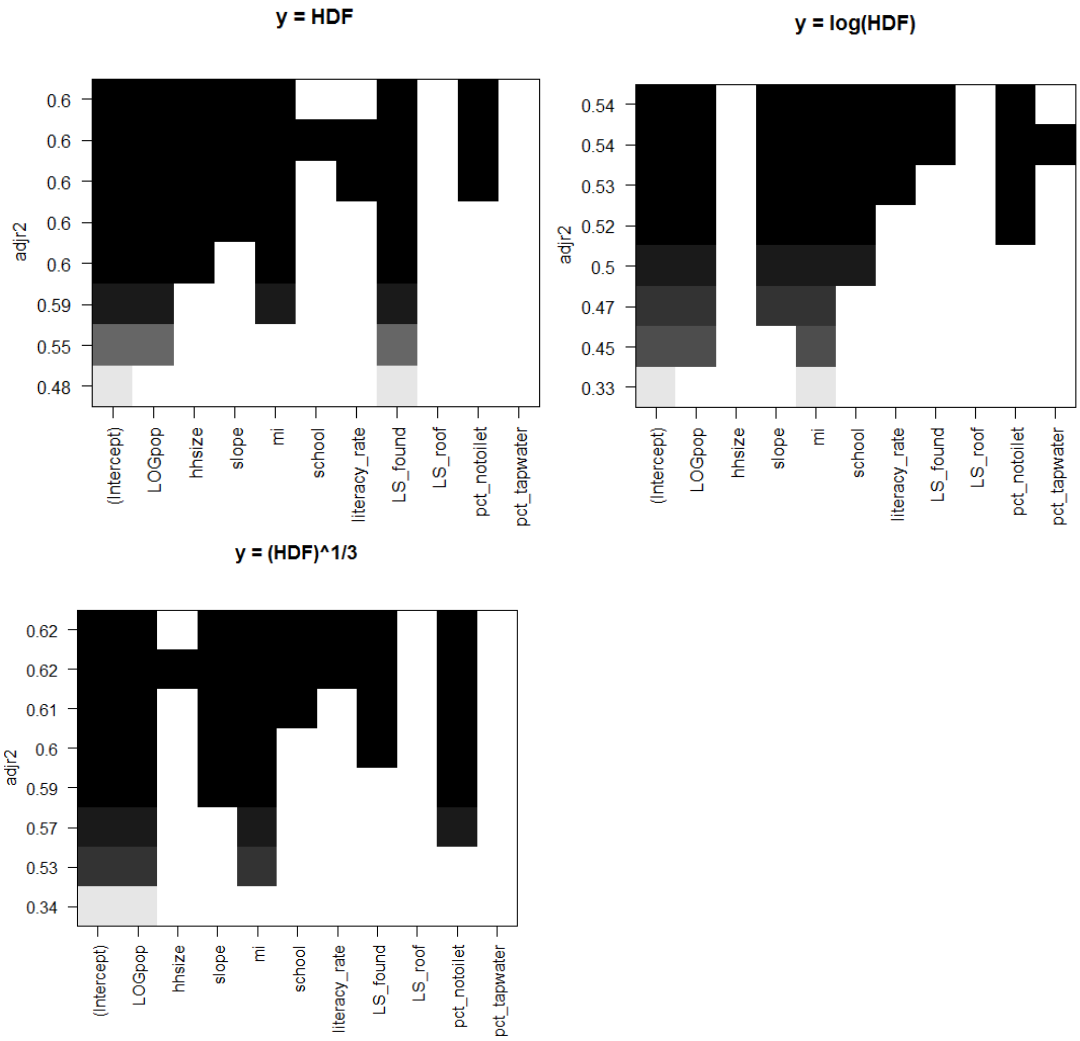


y = log(completely damaged houses (%))



y = (completely damaged houses (%))^1/3





Explanation of codes: adjr2 = adjusted R-squared, LOGpop = population^{log}, hhsize = average household size, slope = slope (%), mi = macroseismic intensity, school = school attendance among 5 – 25 year olds, literacy_rate = literacy rate (%), LS_found = low strength foundations, LS_roof = low strength roofs, pct_notoilet = households without a toilet (%), pct_tapwater = households using tap water as their main drinking water source.

Appendix VII – PIM Training Gorkha Case R Script

```
#This script trains and predicts the nepal priority index model on
the gorkha case of nepal 2015. It is meant to show how the model was
constructed and is not suitable for applying to new events.

#load packages
library(caret)
library(readxl)
library(C50)
library(leaps)
library(psych)
library(randomForest)

# import file (nepal_hdf has all the observations for which a hdf
was caluclated (571) and nepal_cdh has all the observations for
which a number of completely damaged houses was reported (612))
nepal_hdf <- read_excel("C:/Users/Evelien/Dropbox/Priority Index -
Earthquake Nepal/Model Training/nepal_hdf.xlsx")
nepal_cdh <- read_excel("C:/Users/Evelien/Dropbox/Priority Index -
Earthquake Nepal/Model Training/nepal_cdh.xlsx")
# dataset for the whole study area
nepal_SA <- read_excel("C:/Users/Evelien/Dropbox/Priority Index -
Earthquake Nepal/Model Training/nepal_studyarea.xlsx")

# DATA EXPLORATION
# explore frequency distribution of possible y-variables (some
examples)
hist((nepal_cdh$compl_damg_houses),main="y = completely damaged
houses")
hist(log(nepal_hdf$hdf),main="log(HDF)")
summary(boxplot(nepal_hdf$hdf))

# add LOG and CUBEROOT transformation of Y variables to the table
# remove 0's before transformations
nepal_cdh$compl_damg_houses[nepal_cdh$compl_damg_houses==0]=1
nepal_hdf$hdf[nepal_hdf$hdf==0.250]=1.0
# creating cube root transformed variables
nepal_hdf$cuberoot_hdf = ((nepal_hdf$hdf)^(1/3))
nepal_cdh$cuberoot_cdh = ((nepal_cdh$compl_damg_houses)^(1/3))
# visualizing the final two possible response variables
hist((nepal_hdf$cuberoot_hdf),main="(HDF)^1/3")
hist((nepal_cdh$cuberoot_cdh),main="(completely damaged
houses)^1/3")

# check for skewness among candidate predictor variables and
transform if necessary (log for right-skwew, reflect + log for left-
skew)
multi.hist(nepal_hdf[,c(5,7,8,9,16,17,18,42,49,24:27,29:33,35:40)],d
ensity=TRUE,freq=TRUE,bcol="cyan",main=" ")
nepal_hdf$LOGpop=log(nepal_hdf$pop)
nepal_cdh$LOGpop=log(nepal_cdh$pop)
nepal_SA$LOGpop=log(nepal_SA$pop)
nepal_hdf$found_reflect=(1.001-(nepal_hdf$pct_lowstrength_found))
```

```
nepal_hdf$LOGfound=(log(nepal_hdf$found_reflect))

names(nepal_SA)[names(nepal_SA) == 'mercalli_intensity'] <- 'mi'

# split up test and training data (hdf dataset)
set.seed(123)
train_ind=(runif(nrow(nepal_hdf))<=0.60)
train_hdf <- nepal_hdf[train_ind, ]
test_hdf <- nepal_hdf[!train_ind, ]
# same for the cdh dataset
set.seed(123)
train_ind=(runif(nrow(nepal_cdh))<=0.60)
train_cdh <- nepal_cdh[train_ind, ]
test_cdh <- nepal_cdh[!train_ind, ]

# check for multicollinearity
pairs.panels(train_hdf[,
c(56,7,8,10,14:19)],cor=TRUE,lm=TRUE,hist.col="cyan",method="pearson
",scale=FALSE,pch = 20, cex = 1)
cor.plot(train_hdf[,
c(56,7,8,10,14:19)],colors=TRUE,main="Correlation plot (abs)")

  # LINEAR REGRESSION
# automated variable selection
regss=regsubsets(cuberoot_hdf ~ LOGpop + hhsize + slope + mi +
school_attendance + literacy_rate + pct_notoilet + pct_tapwater +
pct_lowstrength_found + pct_lowstrength_roof,data=train_hdf)
plot(regss,scale="adjr2",main="y = (HDF)^1/3")
summary(regss)
# train the linear regression model with the selected variables
lm1=lm(cuberoot_hdf ~ LOGpop + mi + hhsize + slope +
school_attendance + literacy_rate + pct_lowstrength_found +
pct_notoilet,data=train_hdf)
summary(lm1)
# relative importance of the predictor variables
varImp(lm1, scale = FALSE)

# checking performance of the model on the training dataset
plot(lm1,main="LM1")
# assign the prediction to the matrix, still cube root transformed
train_hdf$predLM_cr=lm1$fitted.values
# calculate root mean squared error and rquared on training data
postResample(train_hdf$predLM_cr,train_hdf$cuberoot_hdf)
#plot measured to predicted
plot(train_hdf$predLM_cr,train_hdf$cuberoot_hdf,main="LM1 - Measured
vs Predicted (training dataset)",xlab="Predicted
(HDF)^1/3",ylab="Measured (HDF)^1/3")

# run the LM model on the test dataset
predLM_cr=predict(lm1,test_hdf)
# assign the prediction to the matrix
test_hdf$predLM_cr=predLM_cr
# calculate the RMSE and rsquared
postResample(test_hdf$predLM_cr,test_hdf$cuberoot_hdf)
# plot the residuals
test_hdf$resLM_cr=((test_hdf$predLM_cr)-(test_hdf$cuberoot_hdf))
```

```
plot(test_hdf$cuberoot_hdf,test_hdf$resLM_cr,main="LM1 - Residuals
vs Fitted (test dataset)",xlab="Predicted
(HDF)^1/3",ylab="Residuals")
# undo the cuberoot transformation, to get realistic RMSE
test_hdf$predLM=((test_hdf$predLM_cr)^3)
postResample(test_hdf$predLM,test_hdf$hdf)
# plot measured to predicted
plot(test_hdf$predLM_cr,test_hdf$cuberoot_hdf,main="LM1 - Measured
vs Predicted (test dataset)",xlab="Predicted
(HDF)^1/3",ylab="Measured (HDF)^1/3")
# apply the LM model on the complete study area
predLM_cr=predict(lm1,nepal_SA)
nepal_SA$predLM_cr=predLM_cr
nepal_SA$predLM=((nepal_SA$predLM_cr)^3)


  # RANDOM FOREST
#rf1 model predicting (house damage factor)^1/3
rf1=randomForest(cuberoot_hdf ~ pop + mi + slope + pop_dens + hhsize
+ literacy_rate + school_attendance + thatch_roof + mud_roof +
rcc_roof + wood_roof + tile_roof + galv_roof + wood_wall +
bamboo_wall + mud_wall + unbaked_wall + cem_wall + mud_found +
wood_found + rcc_found + cem_found + tap_water +
no_toilet,data=train_hdf,mtry=8,importance=TRUE,ntree=200)
print(rf1)
#rf2 model predicting (completely damaged houses)^1/3
rf2=randomForest(cuberoot_cdh ~ pop + mi + slope + pop_dens + hhsize
+ literacy_rate + school_attendance + thatch_roof + mud_roof +
rcc_roof + wood_roof + tile_roof + galv_roof + wood_wall +
bamboo_wall + mud_wall + unbaked_wall + cem_wall + mud_found +
wood_found + rcc_found + cem_found + tap_water +
no_toilet,data=train_cdh,mtry=8,importance=TRUE,ntree=200)
print(rf2)

# checking performance of the model on trainig data (only RF2)
summary(rf2)
#calculate RMSE and rsquared on training data
train_cdh$predRF2_cr=rf2$predicted
postResample(train_cdh$predRF2_cr,train_cdh$cuberoot_cdh)
# check relative importance of predictors
importance(rf2,type=1)
varImpPlot(rf1,type=1,main="RF1 (y = HDF^1/3)")
varImpPlot(rf2,type=1,main="RF2 (y = CDH^1/3)")
# add predicted values and retransformation of them to table (and
plot against eachother)
train_cdh$predRF2=((train_cdh$predRF2_cr)^3)
plot(train_cdh$predRF2,train_cdh$compl_damg_houses,main="Measured VS
predicted (training data)")

# run the RF1 model on the test dataset
predRF1_cr=predict(rf1,test_hdf)
# assign the prediction to the matrix
test_hdf$predRF1_cr=predRF1_cr
# calculate the RMSE and rsquared
postResample(test_hdf$predRF1_cr,test_hdf$cuberoot_hdf)
```

```
plot(test_hdf$predRF1_cr,test_hdf$cuberoot_hdf, main="RF1 - Measured
vs Predicted (test dataset)",xlab="Predicted
(HDF)^1/3",ylab="Measured (HDF)^1/3")
# plot the residuals
test_hdf$resRF1_cr=((test_hdf$predRF1_cr)-(test_hdf$cuberoot_hdf))
plot(test_hdf$cuberoot_hdf,test_hdf$resRF1_cr,main="RF1 - Residuals
vs Fitted (test dataset)",xlab="Predicted
(HDF)^1/3",ylab="Residuals")
# undo cr transformation to get realistic RMSE
test_hdf$predRF1=((test_hdf$predRF1_cr)^3)
postResample(test_hdf$predRF1,test_hdf$hdf)
# plot measured to predicted
plot(test_hdf$predRF1_cr,test_hdf$cuberoot_hdf,main="RF1 - Measured
vs Predicted (test dataset)",xlab="Predicted
(HDF)^1/3",ylab="Measured (HDF)^1/3")
# apply the RF model on the complete study area (SA)
predRF1_cr=predict(rf1,nepal_SA)
nepal_SA$predRF1_cr=predRF1_cr
nepal_SA$predRF1=((nepal_SA$predRF1_cr)^3)

# run the RF2 model on the test dataset
predRF2_cr=predict(rf2,test_cdh)
# assign the prediction to the matrix
test_cdh$predRF2_cr=predRF2_cr
# calculate the RMSE and rsquared
postResample(test_cdh$predRF2_cr,test_cdh$cuberoot_cdh)
plot(test_cdh$predRF2_cr,test_cdh$cuberoot_cdh, main="RF2 - Measured
vs Predicted (test dataset)",xlab="Predicted
(CDH)^1/3",ylab="Measured (CDH)^1/3")
# plot the residuals
test_cdh$resRF2_cr=((test_cdh$predRF2_cr)-(test_cdh$cuberoot_cdh))
plot(test_cdh$cuberoot_cdh,test_cdh$resRF2_cr,main="RF2 - Residuals
vs Fitted (test dataset)",xlab="Predicted
(CDH)^1/3",ylab="Residuals")
# undo cr transformation to get realistic RMSE
test_cdh$predRF2=((test_cdh$predRF2_cr)^3)
postResample(test_cdh$predRF2,test_cdh$compl_damg_houses)
# plot measured to predicted
plot(test_cdh$predRF2_cr,test_cdh$cuberoot_cdh,main="RF2 - Measured
vs Predicted (test dataset)",xlab="Predicted
(CDH)^1/3",ylab="Measured (CDH)^1/3")
# apply the RF model on the complete study area (SA)
predRF2_cr=predict(rf2,nepal_SA)
nepal_SA$predRF2_cr=predRF2_cr
nepal_SA$predRF2=((nepal_SA$predRF2_cr)^3)

# save dataframes with prediction to local drive
write.csv(nepal_SA, file = "C:/Users/Evelien/Dropbox/Priority Index
- Earthquake Nepal/Model Training/nepal_SApred.csv")

#END
```