POLITECNICO
MILANO 1863

Polo Territoriale di Lecco

Department of Civil and Environmental Engineering

AN EMPIRICAL FLOOD MORTALITY MODEL USING RANDOM FOREST ALGORITHM

Mina Yazdani

914997

A thesis presented for the degree of:
Master of Science in Civil Engineering for Risk Mitigation

Supervisor: Daniela Molinari, PhD
Co-supervisor: Christian Natale Gencarelli, PhD

Lecco, Italy

December-2021

Acknowledgment

First and foremost, I would like to express my utmost gratitude to Dr. Daniela Molinari, who guided me through this work with continuous support, offering advice and encouragement at every stage of the research. I am grateful for the valuable opportunity to work on this project with her, an experience through which I have acquired knowledge and skills that will help me for a better performance in the future. Without her immense knowledge and her commitment to this work, this achievement would not be possible.

My gratitude extends to my co-supervisor, Dr. Christian Gencarelli, for his support and patience, insightful comments, and suggestions, and for the time he dedicated to the preparation of this research project.

I am grateful to the team of researchers at the Italian National Council of Research, especially Dr. Simone Sterlacchini, Paola Salvati, for their collaboration and assistance during this work.

I extend my gratitude to Giovanni Ravazzani for his assistance and suggestions for the development of this project.

I am very deeply thankful to my family for their unconditional love and support and for providing me with the opportunity to explore various possibilities in life and to always believe in me. I love you.

Abstract

With an increase in the occurrence of natural disasters, the management and mitigation of the adverse consequences such as the casualties due to these phenomena are considered as a priority for decision-makers. Floods are one of the most common hazards affecting the lives of millions of people worldwide; thus, studying the dynamics of these events and the most significant conditions leading to fatalities is deemed crucial in the management of flood risk, both in emergency conditions and in time of peace. However, the modeling capabilities in this field are currently limited, and the creation of such tools is believed to be essential. In such a context, the current study aims at the creation of an empirical flood mortality model for the Italian context, with the use of the Random Forest (RF) algorithm, based on an initial dataset of flood mortalities in the Po river district in 1970-2019 developed and managed by the Italian National Council of Research (CNR-IRPI) of Perugia. By considering the information available in the literature, the main explanatory variables related to the occurrence of flood mortality have been identified. Next, for each fatality record in the dataset, the information describing these variables has been extracted using the available data on the locations of the mortalities as well as data available in national and regional geodatabases. Moreover, the investigation of the hydrological and hydraulic data relating to the corresponding flood events, made possible their characterization in terms of hazard intensity. This process resulted in the creation of a dataset of 127 mortality records, each characterized by ten explanatory variables. In addition, in order to allow the RF algorithm to identify the role and the importance of the different explanatory variables, a synthetic dataset consisting of records of the individuals who were involved in the event but did not lose their lives was created, using the frequency distributions of the explanatory variables in the flooded areas. The resulted dataset consists of 1270 records of non-fatalities. These two datasets together are used to create the RF model. After training the model, the validation of the RF algorithm on two different datasets (data that was not used for the creation of the model) led to the choice of the final model setup, with a classification accuracy of 89%, characterized by the parameters of “Age”, “Place” of the accident, “Morphological Zone”, “Distance from the river”, “Density of the buildings in the municipality”, “Corine land cover code”, “Hazard scenario code”, “Solid transport” carried by the flood, and the “Return Period” of the flood event and excluding the parameter “Gender” which is identified by the model as the least significant in the final outcome. This study resulted in the identification of the most important explanatory flood mortality parameters, which can be used as an information base for the identification of strategies addressed to mitigate and manage the risk of loss of life due to floods.

Keywords: Flood mortality, flood damage model, random forest, Po River Basin

Table of Contents

List of Figures

List of Tables

Chapter1

Introduction

In recent decades, changes in climate have caused impacts on natural and human systems. The recent detection of increasing trends in extreme precipitation and discharge in some catchments implies greater risks of flooding at a regional scale [1]. The statistics indicate that increased warming may result in a larger fraction of the global population being affected by major river floods [2]. The rate of population growth, more intensive urbanization in flood-prone areas, and the limited development of sustainable flood-control strategies will increase the (potential) impacts of these phenomena [3]. From 2010 to 2019, 46% (1,298) of disasters triggered by natural hazards were floods, with more than 673 million people affected (EM-DAT, 2020), becoming the primary driver of economic and insured losses during the first half of 2019 [4]. Thus, governments and policymakers have taken an interest in this issue, investing in the implementation of flood risk management approaches to reduce flood risk and flood damages. As arranged by "The European Directive 2007/60/CE", suitable flood risk management plans must be defined, focusing on "prevention, protection, and preparedness, aiming at the reduction of adverse consequences for human health, the environment, cultural heritage, and economic activity associated with the floods in the community". The present work focuses on the loss of life linked to flood events by proposing a prediction model to be used in the definition of risk mitigation strategies.

The consequences of floods can be categorized as direct and indirect damages; direct effects caused by floodwaters are such as the damage to infrastructures, roads, buildings, and the effects on humans such as the loss of life and injuries; indirect damages are such as the economic and socio-economic losses [5]. Some of the consequences of flooding, such as the damages to residential buildings, have been more investigated, and more previous work on estimating and modeling them is available. However, among the direct damages, loss of life is one of the most serious and irreversible types of flood effects [3] and is considered as one of the first priorities for the decision-makers when facing these events. Several studies have been conducted on the different drivers of flood mortality and the strategies to adopt in order to cope with them, but due to the variability (in time and space) and the interdependencies among vulnerability and hazard factors, the lack of a univocal definition of the effects floods have on people, and the difficulties in the acquisition of accurate data for the calibration and validation of mortality models, prediction and estimation of flood mortality is still very much desired.

Different approaches exist to model the losses caused by floods; synthetic models are based on expert judgment, following a series of "what-if analyses" to define a relationship between the hazardous event and the resulting damages, and this way, they have a higher level of standardization, thus a higher transferability to different areas or regions. On the other hand, empirical models implement previously observed flood loss data, which makes them less

transferable to other spatial or temporal contexts. Both empirical and synthetic models can be configured as univariable or multivariable. Often multivariable models are better suited for describing the complexity of flood damage processes; however, they require accurate, detailed, and consistent data on both observed damage and its explanatory variables, which are rarely available [6].

Regarding the loss of life due to floods, several studies have been conducted to better investigate these variables, leading to the identification of two main groups of contributing factors: the first group relates to the environment and the second group relates to the victims [7]. The environmental factors both describe the hazard and the specifications of the location of the mortality accident, such as the topographical and the morphological features of the flood plain [8]. The most important features of the flood event that influences the degree of damage to people as reviewed in literature are the water depth, the flood velocity and the flood bedload, which could consist of huge debris, impact of which could cause trauma to the individuals and promote drownings [7], [8]. Moreover, the most significant factors describing the victim's conditions are namely the age and the gender of the victim, vulnerabilities, e.g., physical and/or psychological disabilities, the residency of the fatalities, and whether or not they were involved in a hazardous behavior [7]. However, the degree of importance of each factor could be different, considering the context and the country for which these effects are being investigated.

For Italy, comprehensive information on the human consequences of floods and landslides, including the dead and the injured, missing persons, evacuated and homeless people, the age and gender of the fatalities, and the causes and the circumstances of the fatal events, is available from 68 CE to 2014 [9]. Based on this dataset, in 1965-2014, 771 individuals were killed by 441 flood events at 420 sites [9]. Thus, the availability of models for the prediction of flood mortality in the Italian context is strongly needed, but no context-specific tools are still available.

The current thesis aims at investigating available data on flood fatalities in Italy in order to derive an empirical multivariable prediction tool for the loss of life due to floods in the Italian context. With respect to the whole database previously discussed, the study analyzes only information on 138 fatalities, happened in 1970-2019 in the Po river basin area [9], [10]. The investigated time period is limited in the past as completeness and reliability of data related to deaths that occurred before 1970 is limited. The geographical context has been chosen in coherence with the wider context in which this study is located. Indeed, the work is developed in collaboration with the Italian National Research Centre (CNR), within a research project for the District Authority of the Po River aimed at the updating of flood damage maps, for the next cycle of the Floods Directive. The prediction model is created using the Random Forest Algorithm (RFA). This algorithm has been previously used for studying flood effects, such as the estimation of flood risk and the damage to residential buildings, and also to estimate flood mortality in vehicle-related deaths.

Figure 1 shows the flowchart of the work. First, a review of the literature was performed in order to identify the significant contributing factors (i.e., explanatory variables) for the estimation of flood mortality. The obtained information was compared with information available in the dataset to evaluate for how many explanatory variables data was already available in the record. Then, two actions were required: (i) the cleansing of available data in order to make them usable by the

RFA; (ii) the definition of new variables related to those factors for which no information is included in the original dataset, and the corresponding assignment of values. For example, a study of gray literature on the flood events associated with the data records was carried out in order to identify and define parameters to describe the flood hazard. Next, in order to allow RFA to properly assess the role and the importance of the different explanatory variables, a synthetic dataset consisting of records of the individuals who were involved in the event but did not die, was created, using the statistical distributions of the explanatory variables in the flooded area. The dataset of event mortalities coupled with this synthetic dataset was then used as the inputs of the RFA. The performance of the obtained model was tested in the next step, using two different sets of data. Ultimately, to simplify the process of interpretation of the Random Forest model, an attempt to extract some rules describing the conditions by which mortality is more likely to occur was made. These results aim at creating a more interpretable tool to estimate the loss of life due to floods (Figure 1).

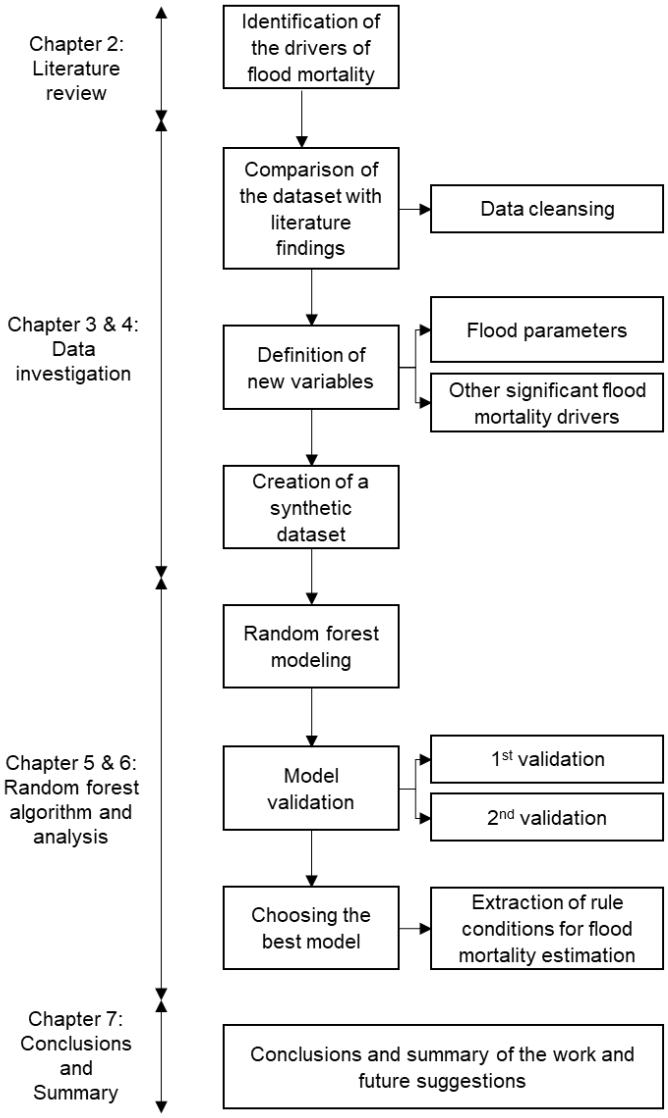

Figure 1- Flowchart showing the organization of the thesis

The present manuscript consists of seven main chapters:

Chapter 1:

A general introduction of the topic, the objectives of the work and the achievements.

Chapter 2:

A review on the literature, investigating the most important explicative variables for flood mortality and available modelling tools.

Chapter 3:

A description of the dataset used for the creation of the model: the original dataset, the definition of the contributing factors (i.e., explanatory variables of flood fatalities), the cleansing process and the definition and evaluation of new variables.

Chapter 4:

A discussion on the hypothesis, considerations, and the framework for creating the “non-fatalities” synthetic dataset.

Chapter 5:

A brief introduction on the Random Forest Algorithm and the analysis performed on the data using this methodology.

Chapter 6:

The discussion of the results of the analysis performed.

Chapter 7:

The summary and the conclusions of the study, and the suggestions for future developments.

Chapter 2

Literature Review

In 2019, 127 floods affected 69 countries, killed 1,586 people, and displaced 10 million more [11]. The fatalities due to floods occur because of a series of circumstances that should be investigated in detail, to identify the drivers of flood mortality. Different flood fatality studies have approached this topic from various points of view. Loss of life due to floods has been studied for different types of flood events indicating that the average mortality is highest for flash floods, while riverine flood affects the greatest number of people [3]. The speed of the flood is one of the basic factors that determines a flood's impact on people, and the literature acknowledges that fast events, defined as "flash floods," are very dangerous to humans, since the rapidity may surprise people, giving them a very short amount of time to act [7]. Moreover, the high velocity of the floodwater can generate a loss of stability [12], resulting in drowning. Thus, the flood velocity has been considered as one of the most important factors describing a flood event, in various studies [8], [13]. Furthermore, the depth of the flow is another significant factor that characterizes the flood hazard and coupled with the velocity, it can lead to the instability of the individuals in the water [8], [12], [14].

Another important factor is the flood bedload, which in extreme conditions might consist of huge debris and solid material, such as cars and trees, and results in trauma, and injuries, [7], [8], which could weaken the strength of victims during the course of a flood event. The nature of the flood plain (e.g., topographical and morphological features, catchment size, presence of obstructions, the state of land cover, etc. [3], [8], [15].) could influence the runoff characteristics and also, the flood depth and velocity. Other sources mention that factors such as the place where the accident occurred (rural or urban areas) can contribute to the impact of the flood, since e.g., in rural areas, the lack of fast responding units for rescues, evacuations, and road closures, or the lack of mitigating structures, such as bridges over low water crossings can affect the ability of the individuals to get to safety, while in urban areas, an increase in the concentration of human activities can amplify risk factors [7]. In addition, the place of the accident can be interpreted as indoors/outdoors conditions, where the individual might be on a bridge, in the proximity of the river, or in a campsite area, where they might not be well informed about the weather conditions or might be surprised by the flood event [7].

The distance of the individuals from the watercourse is also considered an influencing factor in flood mortality, since the depth and velocity of floodwaters will vary with distance from the source of the flooding (breach, river, overtopping, etc.); therefore, individuals are more likely to be swept away by the water flow if they are closer to the flood source [8].

The time of the accident is another important variable that determines the visibility conditions during the flood event, affecting a person's capability to judge the depth and the speed of the

flowing water [7], [15]. In a study on flood fatalities in Texas, USA, among those fatalities with known information on the time of the accident, 52% of fatalities happened at night [16] confirming the importance of this parameter. However, the information on this variable often is obtained with very low confidence, since it is very difficult to estimate accurately when the victim got involved in the accident, and not the time when the body was found. Moreover, the probabilities of individuals being in a location, will vary by the time of day, the time of year, etc. [8].

As variables describing the victims, the sociodemographic parameters of gender and age are two types of information that is often available for flood victims. The gender of the victim acts differently according to the socio-economic characteristics of their community [7]. In the USA, among 1075 fatalities that occurred between 1996- 2014, 65% of the victims with known information on gender were males [17], and the same pattern is observed for Australia (1900–2015) [18], hinting at the fact that males have a higher exposure to floods because of their mobility and outdoor work, in contrast to females. However, female fatalities are usually higher in underprivileged societies as being a female could be considered a sort of vulnerability [7].

The age of the individuals could be considered a significant parameter in the description of flood mortality, especially combined with other variables. By investigating the combination of the age factor and the place of the accident, literature suggest that most of the elderly people were affected inside their homes rather than in outdoor places [7]. They are more vulnerable to floods since they are more expected to have difficulty moving to safe places or might suffer from chronic medical conditions preventing them from finding shelter immediately. Moreover, in European floods, the percentage of fatalities over 60 was much higher than the deaths for the same age group in the US, where non-elderly people might take risks in floods that their European counterparts do not. On the other hand, in Europe, the elderly might be left to fend for themselves during floods more frequently than in the US [19]. Also, the age factor coupled with vehicle-related deaths, which form a high percentage of flood fatalities especially in the USA, shows that, in this country between 1996-2014, almost 50% of the males in vehicle-related fatalities were between 24 and 62 years old, confirming that male individuals in this group age are more likely to be outdoors, driving, since 80% of licensed drivers are between 20 and 64 years old according to the Federal Highway Administration (FHWA 2009) [17], which increases their chance to be involved in vehicle-related flood deaths. In Australia, young adults between the ages of 10 and 29 and those over 70 years were overrepresented among those drowned [20]. Additionally, younger children (under the age of 5) have a higher death rate with respect to the individuals belonging to the age group of 5 -14 years old as, according to Kellar and Schmidlin [21], they depend more on adults for getting to safety during a flood event [7].

As mentioned before, the health conditions of the individuals could act as an influencing factor in flood mortality, since it could increase their vulnerability to the flood event. The elderly, pregnant women, and people with disabilities are more vulnerable to these events as these conditions can lead to impaired responses, reducing the chance of survival [7], [8]. Moreover, poverty and low level of education can pose as influencing factors in flood mortality. In low-income countries, the relatively high rates of poverty, inadequate mitigation measures [7], and a higher number of illegal settlements, results in a higher vulnerability to floods, while in the communities where evacuation

protocols and mitigation measures are implemented and people are familiar with these measures, the death rate is relatively lower.

In addition, some works of literature discuss the importance of the residency of the individuals because the residents are more familiar with the place they live in and are more accustomed to the dangers of their environment and the safety measures needed to be taken during hazardous events [7].

Other than the above-mentioned parameters, the way the individuals behave during a flood event very much affects flood mortality, as e.g., most of the flood fatalities (>90%) in Australia (1997-2008) occur because of the choices made by people; choices to engage in inappropriate risk-taking behaviors, to enter flooded waterways either by foot or in a vehicle, or attempts to retrieve stock or property [20]. During some of the previous European floods, different forms of flood tourism were reported, including large crowds gathering on riverbanks and bridges and people engaging in recreational boating activities on flooded streams [19].

As investigated through various research literature on the topic of flood mortality, different variables can be considered in order to estimate and model the loss of life due to floods. Previous models are present that consider the topic of Flood Fatality. Boyd [22], Druiser (1989) and Waarts (1992) have estimated mortality for coastal floods (storm surge, hurricane), and river floods as a function of the water depth only.

Jonkman (2001) proposes a method for the determination of loss of life for sea and riverine floods in the Netherlands accounting for the effects of water depth, flow velocity and the possibilities for evacuation, contrary to the previous univariable models [13]. Penning- Rowsell (2005) introduced a model based on a deterministic approach to estimate injury and loss of life due to floods [8]. This model is based on determining a ‘hazard rating’ for different zones of the floodplain (based on depth, velocity, and a debris factor associated with each zone), a score for the ‘area vulnerability’ (in terms of flooding lead time, etc.), the population at risk and the population’s vulnerability (in terms of elderly and sick or disabled), to estimate the number of deaths and injuries. Some of the indicative factors proposed for the model are derived based on expert judgment. Jonkman (2008) [13] also proposed a model in which an analysis of flood characteristics, such as water depth, rise rate and flow velocity is performed, followed by the estimation of the number of people exposed (including the effects of warning, evacuation, and shelter). Then the mortality is assessed among those exposed to the flood.

Terti (2019) [23] proposes a methodology towards the integration of physical and social dynamics leading to an estimation of circumstance-specific human losses during a flash flood. After an investigation on the various circumstances leading to flash flood mortalities, the most significant circumstance is identified as vehicle-related deaths. Thus, in the study, a random forest classifier is applied to assess the likelihood of fatality occurrence for vehicle-related flood deaths as a function of representative indicators, such as the age of the workers commuting by vehicles, the road density, the maximum duration of precipitation, etc., as parameters best describing vehicle-related death circumstances. The parameters are extracted from a database of flashfloods from 2001-2011 in the United States, for which using the information on the victims and the people

involved, and a random forest classifier, a probabilistic assessment of vehicle-related human risk is performed.

While there are different models available on the loss of life due to floods, they mostly have been created in local contexts, making it hard to be transferred to the context of Italy. For this purpose, we have decided to use the Random Forest algorithm to create an empirical mortality model specific to Italy, considering multiple significant variables in flood fatality on a dataset of victim records over a time span of 49 years.

Chapter 3

Methodological Approach and Data Investigation

3.1. Description of the data

The data used for this study is partly derived from the dataset presented in the work of Salvati [9] in which the age and gender of the fatalities, and the causes and circumstances of the fatal events occurred in the 50-year period of 1965–2014 in Italy, including information on 771 persons killed by 441 flood events at 420 sites, and 1,292 persons killed by 405 landslides at 390 sites have been investigated. The dataset is considered as the most complete and representative portion of a larger catalog related to the events from 68 AD to 2014 that includes almost all the fatal events, from very low intensity (causing one fatality) to the most destructive (causing more than 30 fatalities). Information from written reports and interviews with eyewitnesses, newspaper articles and historical accounts, survivors' inquiries and petitions, reimbursement requests filed by those who suffered damage, and official government reports and documents have been investigated. Web sites, social networks, and blogs, and collected and analyzed photographs and videos were examined with much care to ensure that the sources, and particularly the non-official sources (e.g., newspaper articles, blogs, and social networks), were accurate and truthful [9]. Additionally to this data, the records related to the events from 2014 – 2019 are obtained from the information available on the POLARIS Website [10], which publishes accurate information on geo-hydrological risk concerning the population of Italy, including periodic reports on landslide and flood fatal events, the exact number of fatalities occurred, the landslide and flood mortality rates, analyses of specific damaging events, and blog posts on landslide and flood events [24]. For the current work, the information from the above-mentioned sources has been extracted for the basin of the Po river, consisting of 138 records of flood fatalities due to 45 flood events from 1970 – 2019.

In addition to this dataset, the location points associated with each mortality record are also provided. These locations can be used to extract other information to better describe the fatalities. A representation of these location points is shown in Figure 3.

3.1.1. Territory

The basin of the Po river is the largest hydrographic basin in Italy, both due to the length of the Po river (650 km), and also as a result of the size of the outflows. The total area of the basin is approximately 74,000 km^2, of which 70,000 km^2 is in the Italian territory, while the remaining parts belong to the French and Swiss territories. In the Flood Risk Management Plan of the Po District, 35 main basins have been defined, divided into 83 sub-basins, the extension of which occupies 54.8% of mountain areas [25]. A satellite image of the Po river basin district is presented in Figure 2, showing the principal cities in the district, the water bodies, and the regional limits of the area.

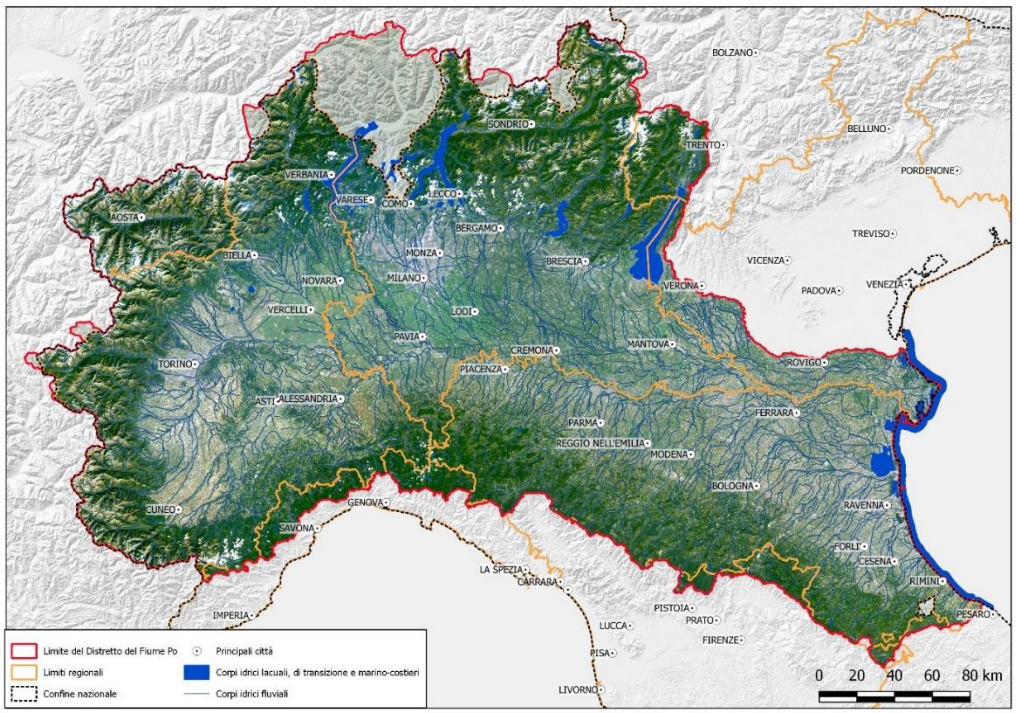

Figure 2- Satellite image of the Po river basin [25]

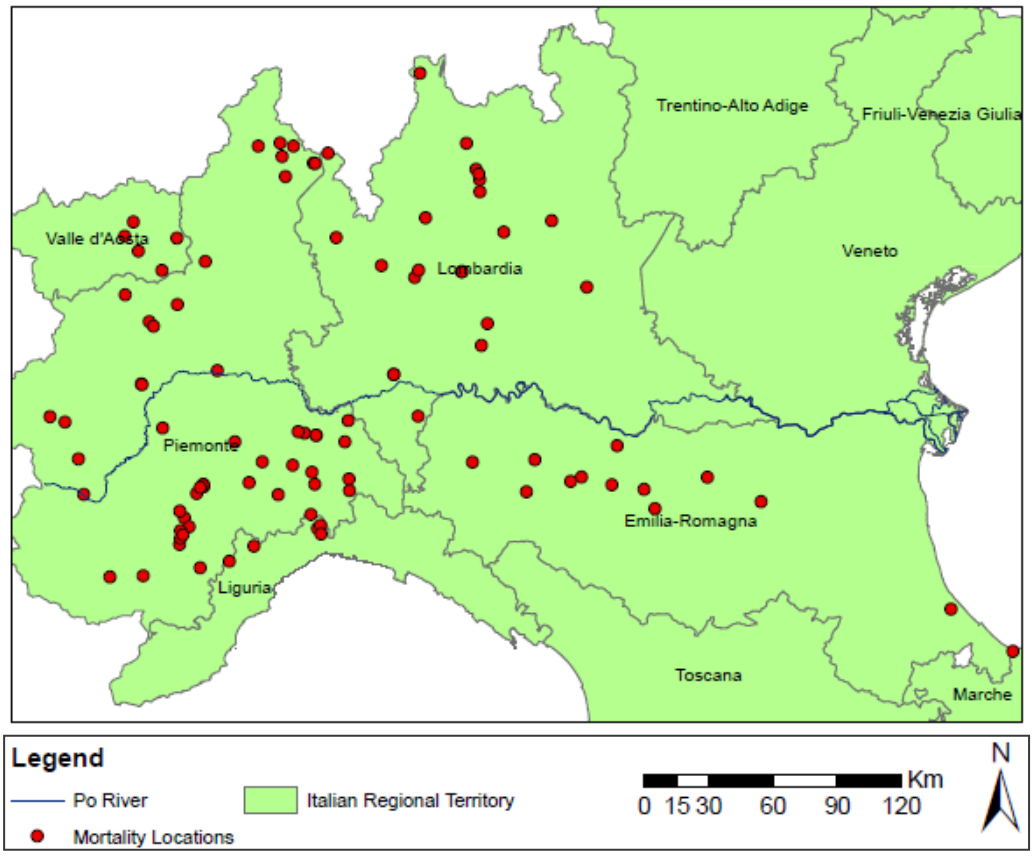

Figure 3- Representation of the mortality locations

3.1.2. Description of the information included in the dataset

The information included in the dataset for each victim record is presented in the following.

- Event identification: The information regarding the date of the event, the region, the province, and the municipality associated with the fatality, and an identification code that relates each record with its associated location point.
- Situational parameters: There are two different parameters of “Time” and “Darkness” describing the situation in which the mortality event occurred. These two variables indicate when the accident happened and also hint at the visibility conditions. The “Time” is described both in a 24-hour clock system, or in descriptive terms such as “early morning”. The “Darkness” parameter has two categories of “Yes” & “No”.
- Victim Identification: The dataset contains information on two of the sociodemographic factors of “Age” and “Gender” of the victims.
- Place of the accident: The information regarding this parameter indicates where the victim was found.
- “Dynamic “and “Manner”: These two variables explain circumstances of the fatality event, if the person was stuck in a flooded room, or was getting out of the car, or had fallen in the watercourse, etc.
- Inappropriate behavior: This parameter explains if a victim was conducting wrong behaviors such as retrieving properties, going to the lower levels of the house, walking in floodwater, etc.
- Cause of death: Explaining if the victim died due to, e.g., drowning, electrocution, or a heart attack.

It should be noted that regarding the data completeness, the information available on the flood victims are partial, and not all records are provided with all of the information on the parameters explained above.

3.2. Frequency analysis of the parameters and comparisons with literature findings

Based on the frequency distribution of the different parameters describing the victim records in the dataset, a comparison between these results and the literature studies in Chapter 2 is performed to see if similar patterns are observed.

Firstly, taking a look at the locations of the fatality records, they are distributed in 6 of the Italian regions as presented in Figure 3 and Figure 4, with Piemonte, having the highest percentage of the fatalities.

Considering the parameters of gender and age, 60.9% of the fatalities were males, while 39.1% were females. Moreover, most of the fatalities were between 15-29 years old, followed by the adults between 65-85 years old. A cross-analysis on these two parameters shows that male individuals are overrepresented, especially between the ages of 15 to 74, while most of the

fatalities younger than 14 years of age and older than 75 belong to females (Figure 5). The same similar pattern is observed in Chapter 2, in the studies done in the context of the USA and Australia [17], [18], [20].

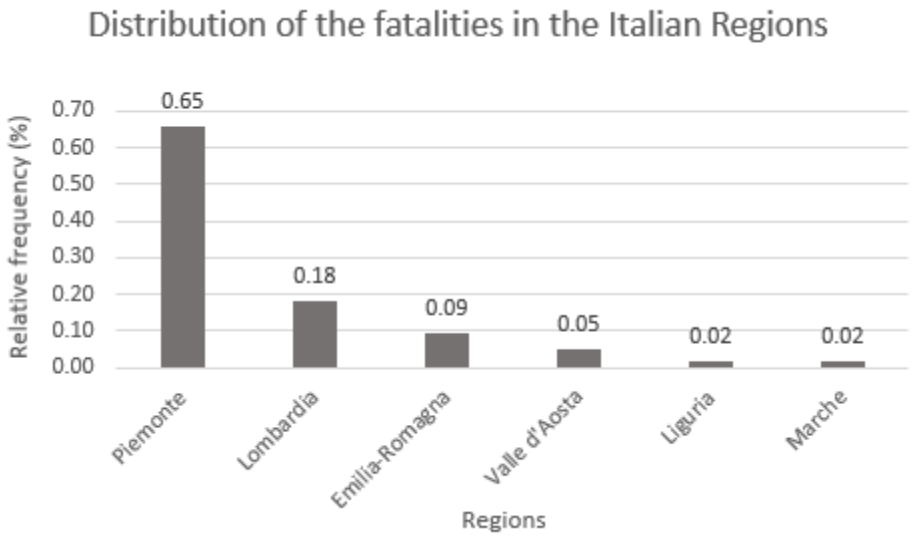

Figure 4- Distribution of the fatalities in the Italian Regions

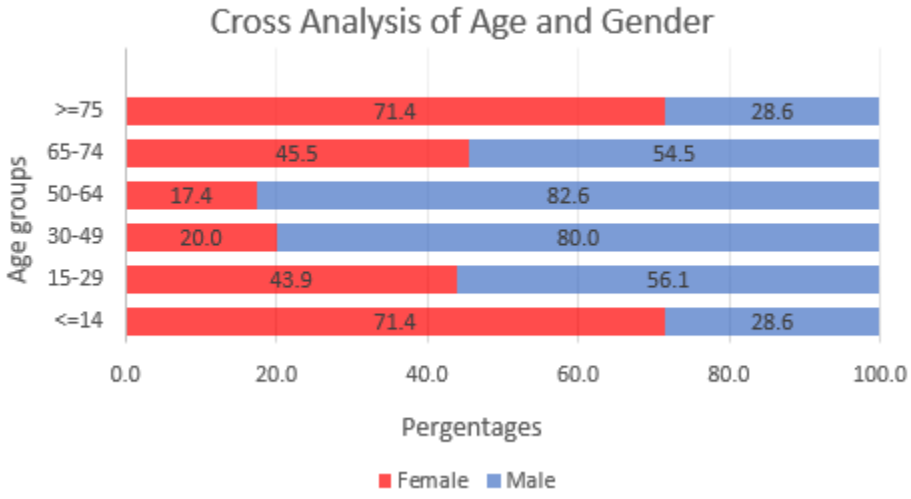

Figure 5- Cross Analysis of Age and Gender of the fatalities

Regarding the place in which the victims were found, 76% of them were outdoors and 24% in indoor spaces. The majority of the victims as shown in Figure 6, were found outdoors on the streets (30.7%), and in private/public buildings (25.2%). These statistics agree with the studies on the circumstances of flood fatalities, as most of the people died outdoors, on the roads, and/or involved in a vehicle- related deaths [21], [20].

Considering the parameter of Inappropriate behavior, among all of the fatalities, around 22% of the victims were acting in an inappropriate way at the time of the flood event, which belonged mostly to males (Figure 7). Similar patterns are observed in other contexts as in the dataset of Australia (1997-2008) in which 26.5% of the victims were swimming or surfing in flooded waterways and 16% were associated with activities like swimming or wading across waterways [20].

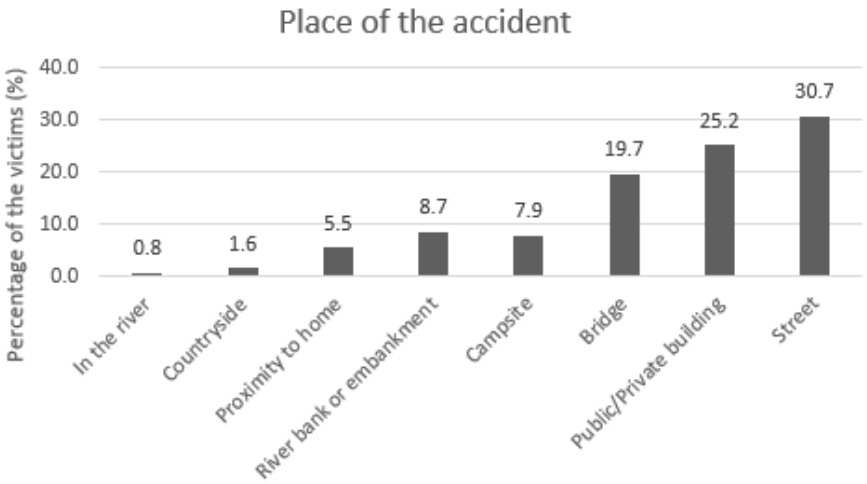

Figure 6- Place of the fatality accidents

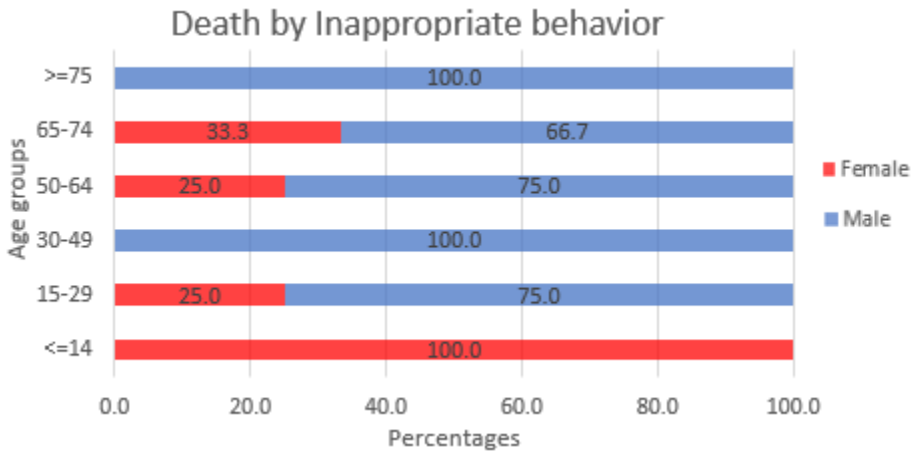

Figure 7- Categorization of Death with respect to Inappropriate behavior among the victims

In the present dataset, the most dangerous dynamic among the mortality events was being “swept up by water”, followed by being “dragged by water while being in the car” as dynamics that were associated with people being outdoors (except for two cases in which the victims were dragged from inside the building to the outer environment). The third most dangerous dynamic in which the accident happened indoors was when the victims were “Stuck in a flooded room” and could not seek shelter (Figure 8). This is in accordance with the study in Australia in which being swept away by water whether on foot or in a car was a frequent dynamic of the accident, while being trapped in a building was not necessarily a frequent case in the circumstances of mortality in this

study [20]. Moreover, in the United States, the most dangerous dynamic was “being involved in vehicle-related circumstances” [23].

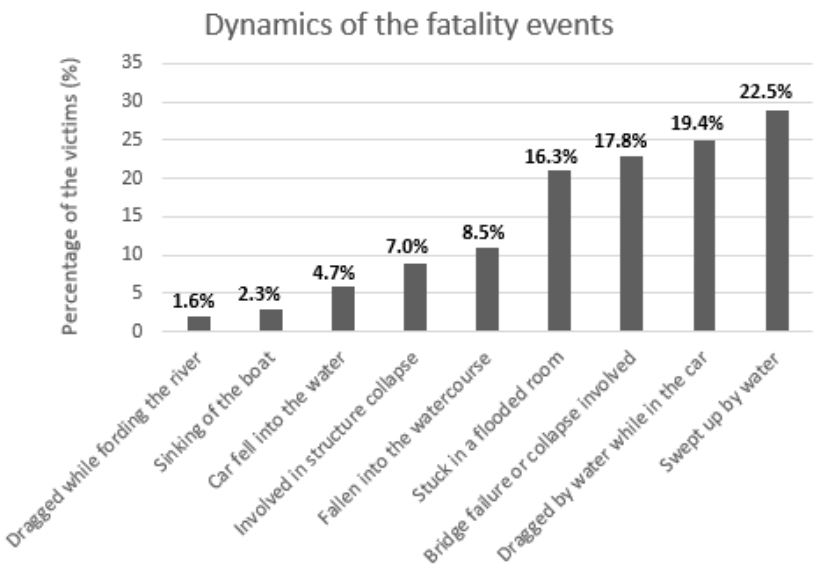

Figure 8- Dynamics in which the victims were involved during the flood events

3.3. Data Cleansing

In order to obtain a more refined set of data records, with well-defined fatality explanatory variables, which can be interpreted and analyzed by the RFA, a number of modifications were performed on the initial dataset.

3.3.1. Removing variables and victim records

A few of the variables and victim records were omitted from the dataset.

- Time and Darkness: These two variables, in spite of providing very affective information on the flood mortality, were decided to be removed from the records, since they contained a lot of missing values (67% missing data on the parameter “Darkness”, and 57% of missing data on the parameters “Time”), which would affect our next analysis adversely.
- Victim records: A few of the victim records which were associated with the cause of death as heart attack (4 records) and sickness (1 record), were removed from the dataset since they were considered unrelated to the analysis.

3.3.2. Modifying variables

Two of the parameters described in the dataset were modified.

- Place: This variable, consists of information describing where the victim was found, such as on the street, in the house, outside near a building, on a bridge, in a riverbank, etc. This variable was re-categorized into 5 general classes to analyze this information better.
- Manner, Dynamic, Inappropriate behavior: The information on these three parameters was combined together to create a new variable, as “Inappropriate behavior”. This new variable

is a binary, “Yes” indicating that the victim was involved in a dangerous activity during the flood, and “No” for when this was not the case.

3.4. Definition of new variables

Based on the suggestions from the literature analysis performed in Chapter 2, aside from the parameters available in the dataset, there are other variables that can be significant in creating the flood mortality model. The information on these variables is obtained based on the location points associated to the mortality records. The variables are explained below. (Most of the operations in this section are performed using the ArcMap 10.7.1 software.)

3.4.1. Morphological Zone

The characteristics of the flood plain, such as the morphology of the area are important environmental drivers of the loss of life due to floods. These features may influence the depth and velocity of floodwaters [8] or the time of concentration. Based on a shapefile created by the Po river district Authority [25] the basin area of the Po river is classified in two categories of Plain and Mountain areas, which can act as a proxy to describe the slope, affecting the flow of water and also the stability of humans. Therefore, each mortality location is associated to a morphological zone of mountain or plain. This classification is observed in Figure 9.

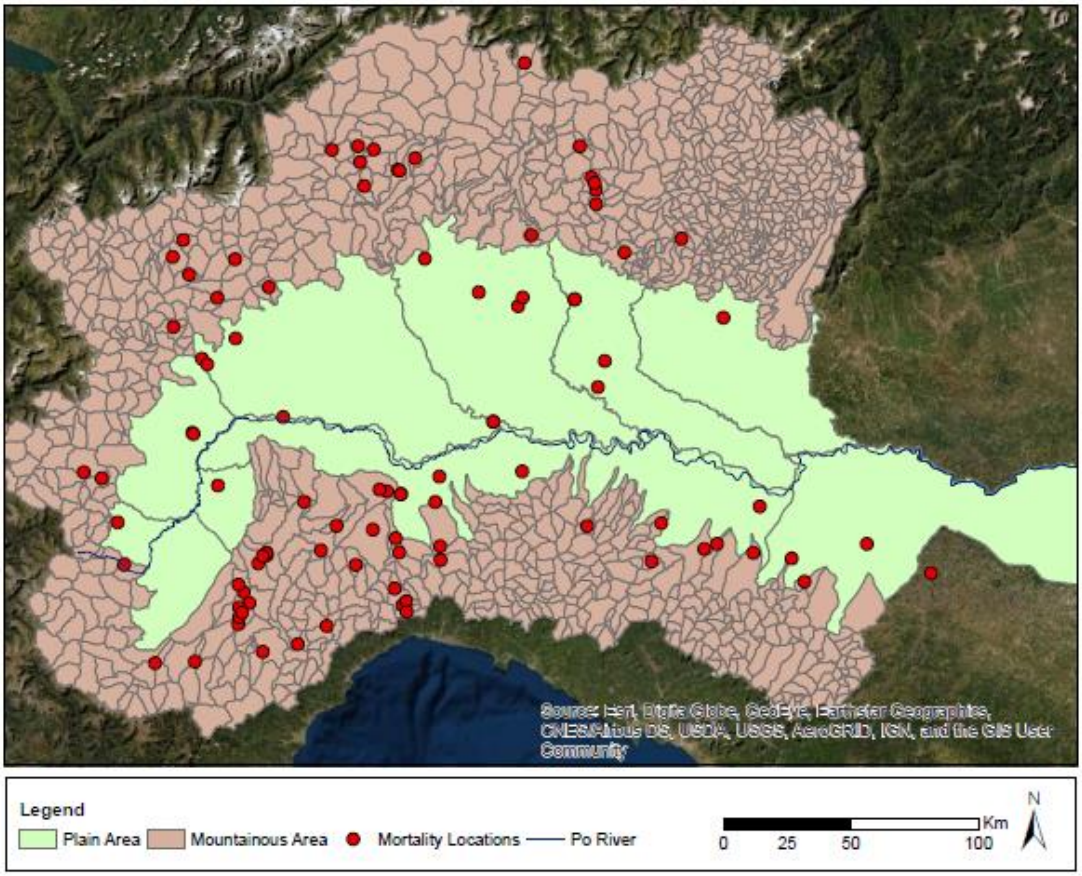

Figure 9- Map of the morphological classification of the Po basin area

3.4.2. Corine Land Cover Code

The land cover of an area influences the hydrological processes such as the runoff and the infiltration rate. In urban areas, most of the land is covered by buildings, streets, and compacted landscapes, increasing the volume and velocity of the runoff, and thus aggravating the conditions in a flood event. Moreover, urban areas offer more possibilities for people to find shelter or get rescued in case of a flood event. For this work, the information regarding the land use data was derived from the National CLC databases produced by "The Eionet network National Reference Centers Land Cover (NRC/LC)", updated in 2018 by Copernicus [26]. The database consists of an inventory of 44 classes, subdivided into 3 levels. The classification up to the second level is presented in Table 1. Figure 10 shows the distribution of CLC codes for the 2nd level classes, for the victim records.

Level 1	Level 2
1. Artificial Surfaces	1.1. Urban fabric
	1.2. Industrial, commercial and transport units
	1.3. Mine, dump, and construction sites
	1.4. Artificial non-agricultural vegetated areas
2. Agricultural Areas	2.1. Arable land
	2.2. Permanent crops
	2.3. Pastures
	2.4. Heterogeneous agricultural areas
3. Forests and Semi-natural Areas	3.1. Forests
	3.2. Shrub and/or herbaceous vegetation associations
	3.3. Open spaces with little or no vegetation
4. Wetlands	4.1. Inland wetlands
	4.2. Coastal wetlands
5. Water Bodies	5.1. Inland waters
	5.2. Marine waters

Table 1- Classification of Corine Land Cover

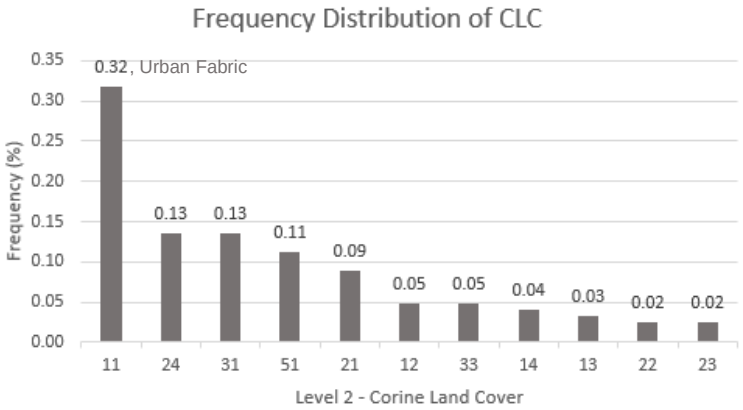

Figure 10- Frequency Distribution of the Corine Land Cover Code associated with the mortality records

3.4.3. Slope

The slope of an area is an indicator of the flow of water and having the information of the DTM (Digital Terrain Model) maps of different areas in Italy, which are available through the ARPA (Agenzia Regionale per la Protezione dell'Ambiente) websites, it is possible to obtain the local slopes related to the locations of the fatalities. Considering the resolution of the DTM maps, slope values were calculated in a 25 km^2 area using the ArcMap tools. Thus, the information of the local slope in the location of the fatalities was added to the dataset.

3.4.4. Distance from the river

The depth and the velocity of floodwater vary with the distance from the source of flooding. Using the satellite images of the area, it was possible to derive the distance of the mortality location point from the river source of the flood event. Thus, based on the satellite images, the river causing the flood event was identified (as the main river course closest to the fatality location, since this information was missing from the initial dataset), and then the distances were obtained. The frequency distribution of the distance values can be observed in Figure 11. Based on Figure 11-b, almost 90% of the values are within a 600-meter distance from the river.

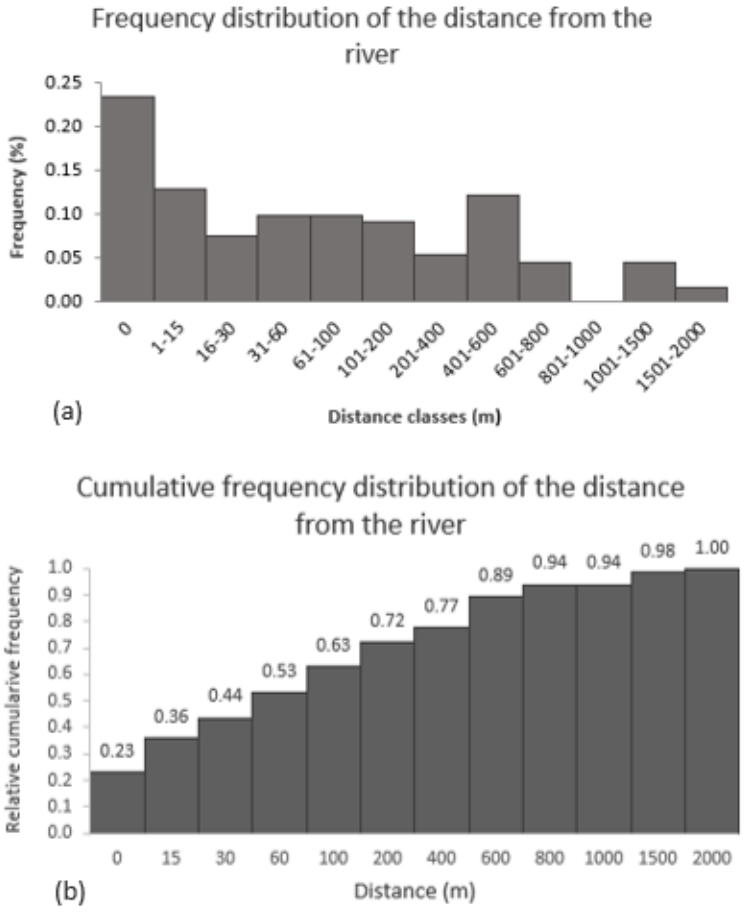

Figure 11- Frequency distribution of the parameter "Distance from the river" in the dataset

3.4.5. Density of the buildings in a municipality

Considering the extent of the buildings in the areas in which the mortalities occurred, it is possible to obtain a proxy to represent the density of the population. By calculating the density of the buildings for the municipality, it is possible to estimate the distribution of the population in the area. This measure is chosen as a proxy of population density because, considering the time span of the data records, building density is a more stable variable in representing the population density since the urban fabric changes less frequently. One does not expect significant differences between the urban fabric of the present day and that of 50 years ago (except in larger urban areas). On the other hand, using the available data on the population (Italian census 2011) [27] may imply significant errors in representing the past population. Moreover, building density provides a proxy for the level of urbanization of the area, better specifying the information on land use. Thus, for each municipality, using the cadastral maps, the ratio between the area of the buildings and the area of the municipality was calculated. The distribution of this parameter is shown in Figure 12.

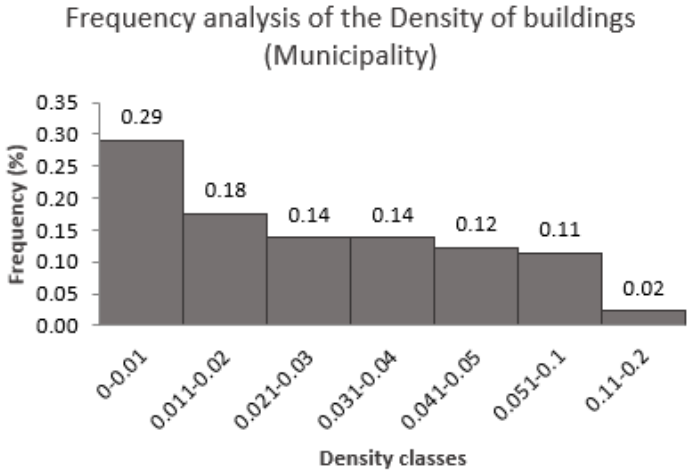

Figure 12- Frequency distribution of the Building density (Municipality) in the dataset

3.4.6. Density of the buildings in a buffer area

With the same reasoning as in 3.4.5, for each mortality point, the density of the buildings in a buffer area equal to the distance of that point from the river (calculated in 3.4.4) was obtained to create a more local understanding of the distribution of individuals in that buffer area.

3.4.7. Risk, Damage and Hazard Scenario classes

The Flood Hazard and Risk maps were developed through the Flood Risk Management Plan (FRMP) of the Po District, with an objective to initiate a new phase of the national policy for the management of flood risk, as indicated by the European Floods Directive. For each river basin district, these maps direct the action on the most significant risk areas [25], for which flood hazard and risk maps were developed. Regarding the Hazard, the maps identify three scenarios as defined below in Table 2. For such scenarios, risk maps identify four different classes of R1, R2,

R3 & R4, classifying the area from the lowest food risk to the highest, for the exposed elements of the population, services, infrastructures, economic activities, etc. A similar classification is defined for the damage maps, with the lowest damage belonging to the D1 category and the highest classified as D4.

Hazard Scenario	Description	Return Period
L	Flood event with a low probability of occurrence	up to 500 years
M	Flood event with an average probability of occurrence	100 – 200 years
H	Flood event with a high probability of occurrence	30 – 50 years

Table 2- Categories of the Hazard scenarios represented in the maps created through FRMP

Therefore, the information for these 3 parameters were derived from the maps and assigned to each mortality record. Figure 13 shows the distribution of the hazard scenario associated with the mortality records in the dataset.

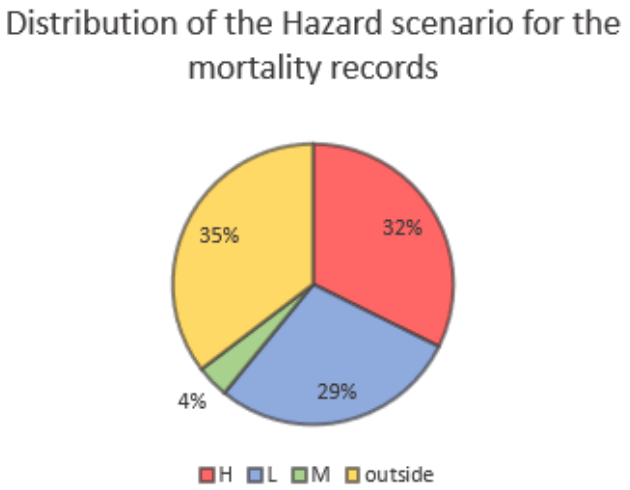

Figure 13- Distribution of the Hazard scenario for the mortality records

3.5. Definition of hazard parameters

Each mortality record belonging to the dataset is associated with a flood event that previously occurred in the Po river basin. However, the dataset does not contain any more information on the intensity of the flood events or their duration. Thus, an attempt was made in order to derive some information on the events and create descriptive parameters for each one of them.

To do so, multiple gray literatures from 1970-2019 were investigated, such as technical reports investigating specific events, yearly reports produced by each region on various natural

phenomena accessible through the ARPA platforms, and documents on the precipitation measurements belonging to various pluviometric stations produced for each region.

The information describing each flood event represented in the documents mentioned above, were not uniformly available for all the events, e.g., for one event the precipitation values and the maximum discharge on a section of the river involved in the event was described. In contrast, for another flood phenomenon, only the 24-hour precipitation values were available.

Considering the 50-year timespan of the flood events in the dataset, the paucity of information for some of the events, especially the ones dating back to the 1970s, is understandable, however limiting the process of finding an appropriate descriptive variable for all of the flood phenomena.

After much consideration and investigation, two parameters were chosen as representatives of the flood events: The return period of the maximum 24-hour precipitation & The solid transport.

3.5.1. Return period of the maximum 24-hour precipitation

One type of information available for almost flood events was the maximum 24-hour precipitation values. Using this data, it was possible to obtain the return period of the maximum 24-hour rainfall causing the flood event, used as a proxy for the return period of the flood. The process of obtaining this information is explained using the example below.

Example.1

Date of the event: June 27, 1997, Bergamo

The flood event occurred due to the triggering of the Oglio river, within the corresponding hydrographic basin of Oglio. At the time of the event, five precipitation stations had been operating. After extracting the precipitation recordings from those stations and then, calculating the maximum 24-hour rainfall values for each station, those values were compared, and the maximum value of the 24-hour precipitation among all the stations was assigned to the event.

As presented in Figure 14, the fatality event occurred in the municipality of Costa Volpino, and the pluviometric station chosen as the representative station is the one in Pantano d’Avio, corresponding to the maximum 24-hour precipitation record in the basin of Oglio for this event.

In the next step, using the information of the DDF curves (Depth-Duration-Frequency) specific to the pluviometric stations, considering that the maximum 24-hr precipitation in Pantano d’Avio station was recorded as 160.2 mm, it was possible to derive the return period associated to the maximum 24-hour rainfall. This return period is used as one of the variables describing the flood event. The DDF curve for the Pantano d’Avio station in this example is shown in Figure 15, in which the corresponding return period is approximately 50 years.

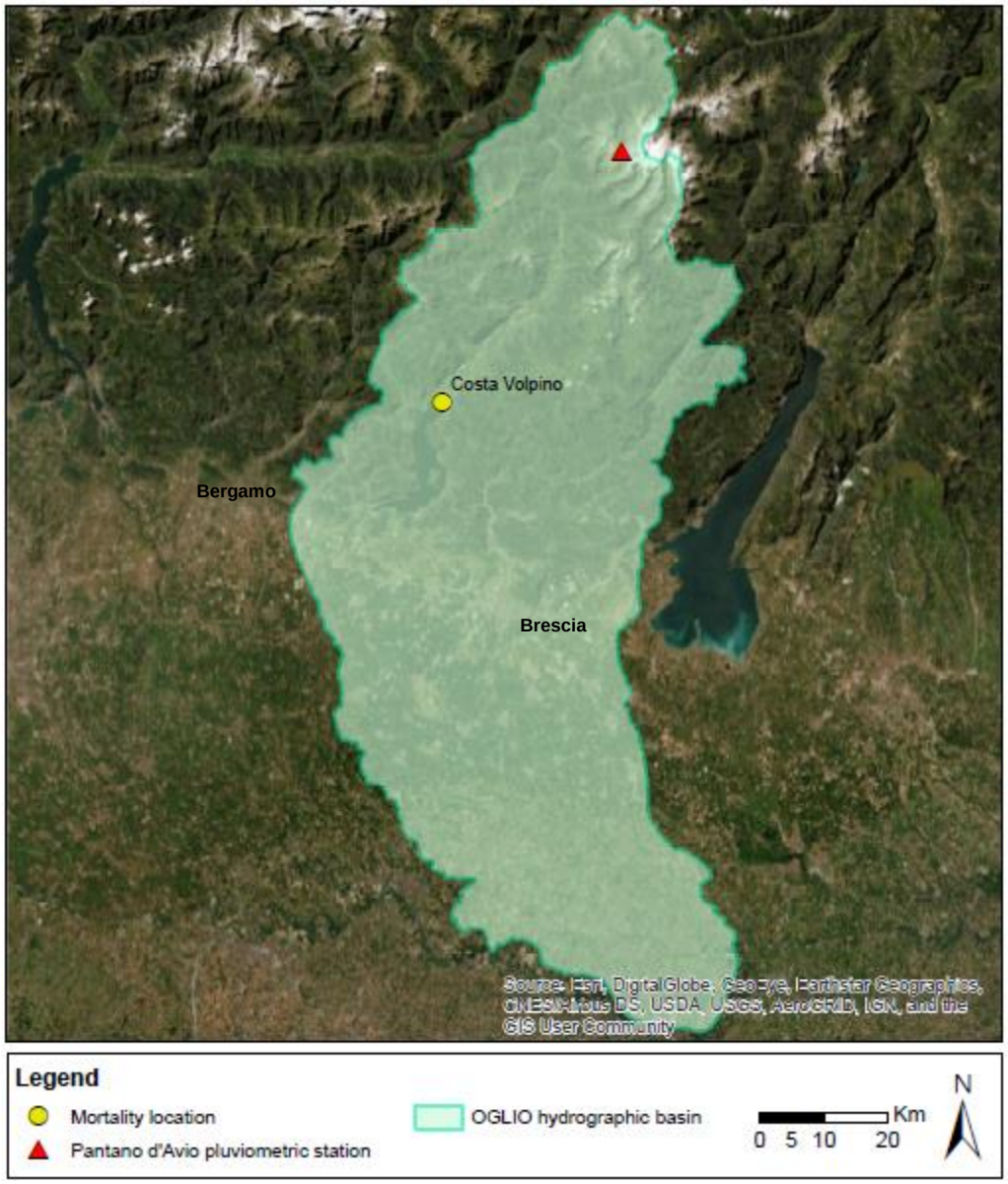

Figure 14- Example 1, the representative pluviometric station for the event of June 27 ,1997, Bergamo

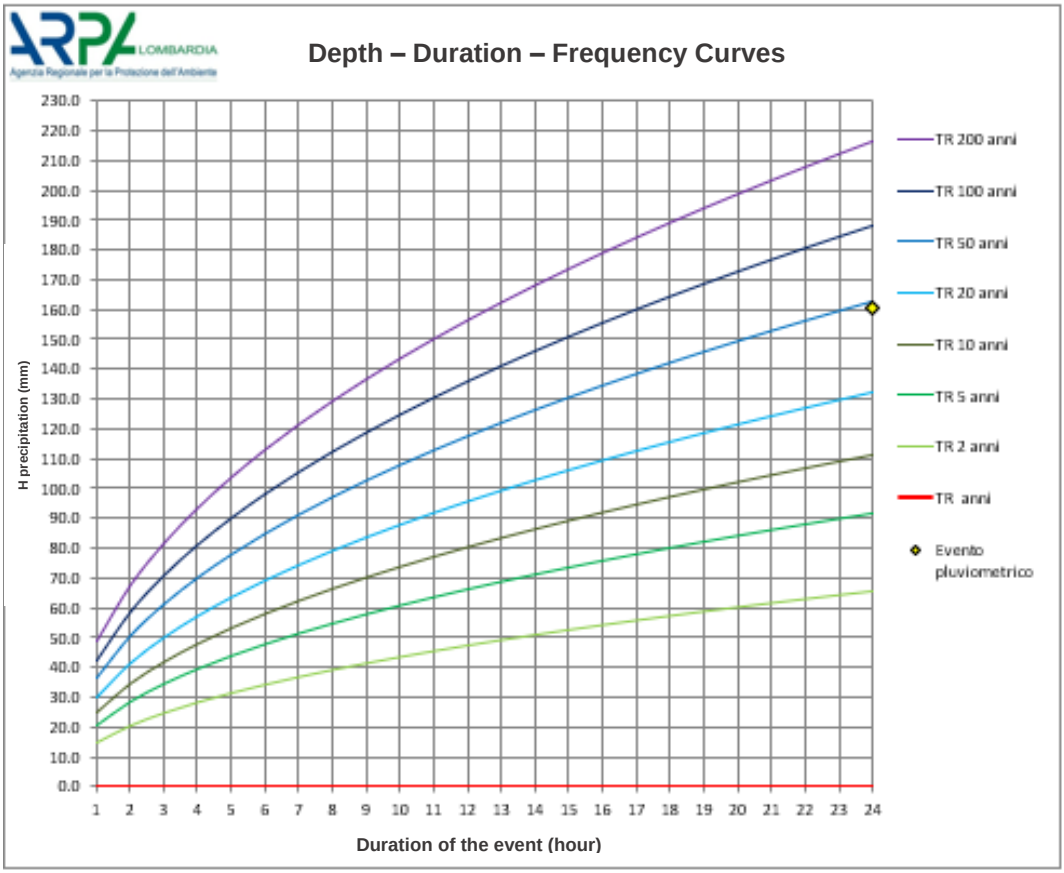

Figure 15- Example 1, DDF curve associated to the Pantano d'Avio rainfall station

Applying the same methodology explained through Example1 to all of the other cases, the return period of the maximum 24-hour precipitation was calculated for the events associated with the mortality records in the dataset.

3.5.2 Solid Transport

As suggested by the literature reviewed in Chapter 2, the bedload and the solid material carried within the flood event can be considered an important driver of flood fatality. An attempt was made to extract this type of information from the different sources investigated for the characterization of the events, in which the availability and the method used to report on this parameter varied from one event to the other. Information on bedload was reported in some cases as the volume of the solid discharge in different sections of the river involved in the event, while in some others only descriptive language was used to take note of the presence of the solid materials. Whereas in some other events, it was hard to access this information as no descriptive report was available on the event. In the end, this type of data was available for 67% of the mortality records, while the information was missing for 33% of the cases. Thus, the parameter defined to take into account the state of the solid transport within the flood event was categorized into three different classes,

described in Table 3. This parameter qualitatively indicates the presence of solid material transported in the flood phenomena.

Solid Transport Categories	Description
Yes	Presence of solid transport material
No	No solid transport material present
NA	No evidence of such indicator

Table 3- Categories of the Solid transport parameter

3.6. Variable uncertainty

3.6.1. Uncertainty regarding the coherence of the mortality locations and the place of the accident

There are some degrees of uncertainty regarding the database of the locations of the mortality events. Considering the place in which the victims were found (recorded in the variable “Place”) and the corresponding locations on the map of the area, some inconsistencies were spotted in a few cases. These inconsistencies could have occurred because the sources from which the information on each victim is obtained (especially the ones that date farther back) are very descriptive; thus, the creation of the location points based on them is a very complex process. For this reason, the other parameters that were derived using these location points might hold some degrees of uncertainty.

- **Slope**

This parameter was defined as the local slope in a pixel of 5×5 m^2 area corresponding to the location of mortality. However, considering the uncertainty mentioned above, it was decided to omit this parameter from the next phases of the analysis, and instead rely on the parameter of Morphological zone as a proxy of the slope variable.

- **Distance from the river**

Another parameter that might be carrying some level of uncertainty is the distance of the mortality location from the river causing the flood event. However, due to the importance of this parameter in relation to fatalities, it was decided to keep it in the next steps of the analysis.

3.6.2. Uncertainty in describing the flood parameters

The different sub-basins within the Po river basin area, as presented in Figure 16, have different sizes. For smaller catchments, the time of concentration (t_c; defined as the time that water needs to flow from the most remote part of the catchment to the downstream section) is shorter in contrast to larger sub-basins, and thus, the choice of the maximum 24-hour rainfall duration might misrepresent the intensity of the floods that occurred in those basins. However, this choice was made because of the limitations encountered during the investigations on the flood events, as

explained in section 3.5, since only the 24-hour rainfall records were accessible for some flood events. To have a consistent measure for all of the floods and to be able to compare the results with each other, the return period of the maximum 24-hr rainfall was chosen as the representative of the hazard. Ideally, having precipitation values with resolutions of 1, 3, 6, and 12 hours for all of the events in the dataset would result in considering the return periods for rainfall durations closer to the concentration-time of each specific catchment, as the representative flood parameter.

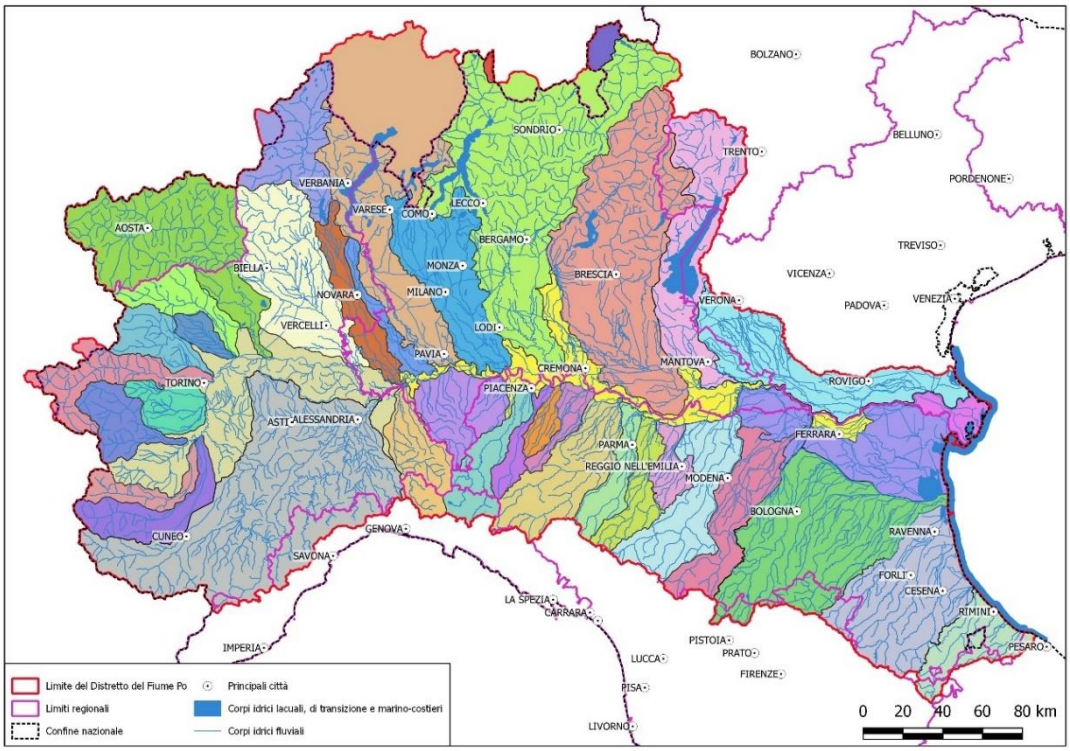

Figure 16- Sub-basins within the Po River Basin [25]

3.7. Descriptive parameters of flood mortality

Considering available information in the dataset and the new variables defined in the previous sections, in the end, 11 indicators were chosen as explanatory variables of flood mortality, to be used for further analysis. As observed in Table 4, the variables of risk and damage categories derived in 3.4.7 were dismissed as they were found misleading for the creation of the model, since they represent the damage and flood risk to various exposed elements such as population, services, infrastructures, economic activities, etc. Moreover, the variable of Inappropriate behavior was disregarded in the current analysis because it can outweigh the influence of other factors, conceptually. As an example, considering the case in which the individual was dwelling along the banks of the river or on a bridge, this behavior overshadows the other parameters, so it does not matter if they were female or male, to which age group they belonged, or in what morphological zone the event occurred, this behavior overshadows the other variables. Thus, to avoid this issue, it was decided not to include this variable in the model parameters.

N	Name of the Parameter	Categories	Description	Note
1	Gender	F M	Gender of the victim	Feminine and Masculine
2	Age	<=14 15-29 30-49 50-64 65-74 >=75	Age of the victim	categorical variable
3	Place	Bridge Building Campsite Outdoor River	The place in which the victim was found	**Bridge:** The victims were found on a bridge that collapsed. **Building:** The victims were inside a generic building (house, cellar, shop, etc.) **Campsite:** The victims were found in a camping area. **Outdoor:** The victims that were found outdoors, whether in a vehicle, or on foot. **River:** The victims were found near the embankments (in a 1m distance from the river) or inside the river.
4	Morphological Zone	Mountain Plain	Indicates the morphological zone corresponding to the location of the fatality event	Categorical variable

Table 4- The list of variables used as indicators of flood fatality

N	Name of the Parameter	Categories	Description	Note
5	Hazard Scenario Code	H M L outside	Describing the category of flood scenario based on the maps produced within PGRA	**H**: Frequent floods, TR 30 - 50 years **M:** floods with medium frequency, TR 100 - 200 years **L:** Rare floods, TR up to 500 years
6	Corine Land Cover Code	15 possible categories	Table 1	Categorical variable
7	Distance from the River	All	Distance of the fatality location from the river causing the flood event	Continuous variable
8	Density of the buildings (Municipality)	All	Density of the buildings with respect to the municipal area	Continuous variable
9	Density of the buildings (Buffer)	All	Density of the buildings with respect to a buffer area equal to the distance from the river	Continuous variable
10	Return Period	15 possible categories	Return period of the max 24-hr precipitation	Categorical variable
11	Solid Transport	Yes No NA	Indicating the presence of solid material	Categorical variable

Table 4- The list of variables used as indicators of flood fatality

Chapter 4

A synthetic dataset of non-fatalities

4.1. Introduction

As discussed in the previous chapter, after performing some modifications on the primary data for flood fatalities, a final dataset of 127 mortality records with 11 descriptive parameters was prepared, in which parameters describe the factors that are significant for causing the loss of life due to floods.

To analyze the role of each parameter and to see what variables are more significant for causing flood mortalities, one should integrate the dataset with data about the individuals who were affected by the flood event but did not lose their lives. However, such a dataset is still missing for the flood phenomena considered in this study; thus, a synthetic dataset of non-fatalities was created to fulfill this purpose. This dataset, which is called the “Non-fatalities”, consists of records of people involved in the same flood events as the fatality dataset (called the “Fatalities”), but who did not lose their lives in the event (they will be mentioned as the people *involved* from this point forward). These two sets of data together enable us to investigate how the descriptive indicators change when a fatality event occurred compared to when an individual involved in the flood survived. The dataset is archived in (https://github.com/Mina-yz/POLARIS-dataset).

4.2. Methodology

The synthetic dataset was created based on a random selection of values from the frequency distribution of the descriptive parameters. For each record in the “Fatalities” dataset, 10 “Non-fatality” records are created. These “Non-fatality” records characterize the individuals involved in the same flood as the victims. At the end of the process, a total dataset of 127 mortalities (“Fatalities”) and 1270 “Non-Fatalities” is then formed, which will be used to determine the role of the most significant parameters, by means of a machine learning algorithm called Random Forest (see chapter 5).

To obtain a reasonable scenario to describe where people were likely to be involved in the flood event, one could consider the extension of the flood in question as the area over which the distribution of the parameters is obtained. However, there is not a lot of accurate information on the extension of the floods in the catalog, thus, as a proxy, the hazard scenario maps [25] introduced in 3.4.7 could be used to estimate the area where the floodwater might have reached, limited to the boundaries of the municipality where the fatality occurred, as well as the upstream and downstream municipalities (this area is called the *selection area* from now on). Then, it is possible to create the dataset of individuals who could have been in those areas at the time of the event but managed to survive.

From the 11 descriptive parameters, 2 of them (“Return Period” of the flood and the “Solid Transport”) were kept the same as the parameters of the flood event causing the fatality, while the remaining parameters were extracted randomly from their frequency distributions.

It is important to note that the parameter of “Density of the buildings (Buffer)” was excluded from the list of the flood mortality indicators because the process to obtain the frequency distribution of this parameter over a large area was very cumbersome.

4.3. Random selection of parameters

As explained before, to create the “Non-fatalities” dataset, each descriptive parameter is randomly selected from the frequency distribution of that parameter over the selection area associated with each record in the “Fatalities” dataset. A more straightforward explanation of the methodology is given through an example in the following (Example 2).

Example. 2

- **Selection area**

This example considers the creation of “Non- fatalities” associated with the mortality records that occurred in the municipality of Alba, involving the Tanaro river. As observed in Figure 17, there are four mortality locations involving the municipality of Alba, and all of them were caused due to the flood of November 4th, 1994. This flood event in the basin of Tanaro is characterized by a Return period of 100-200 years for the 24-hr precipitation. Thus, referring to the hazard scenario maps [24] of this area, as observed in Figure 17, the low-frequency scenario is associated with this event. Therefore, considering the boundaries of the municipality of Alba, the upstream municipality of Roddi, and the downstream municipality of Barbaresco, and the hazard map, the selection area over which the frequency distribution of the parameters is calculated, is obtained (Figure 18).

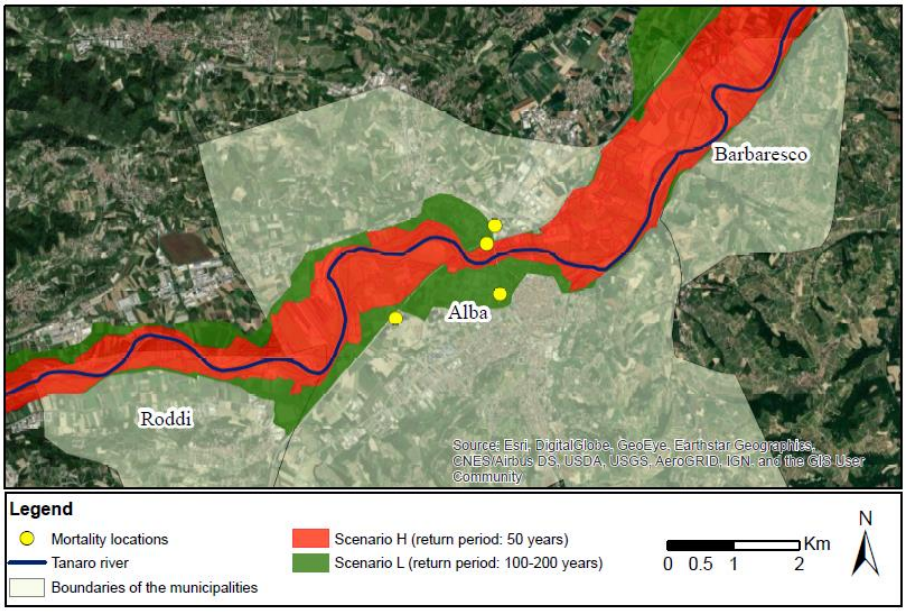

Figure 17- Example. 2- The case of the municipality of Alba

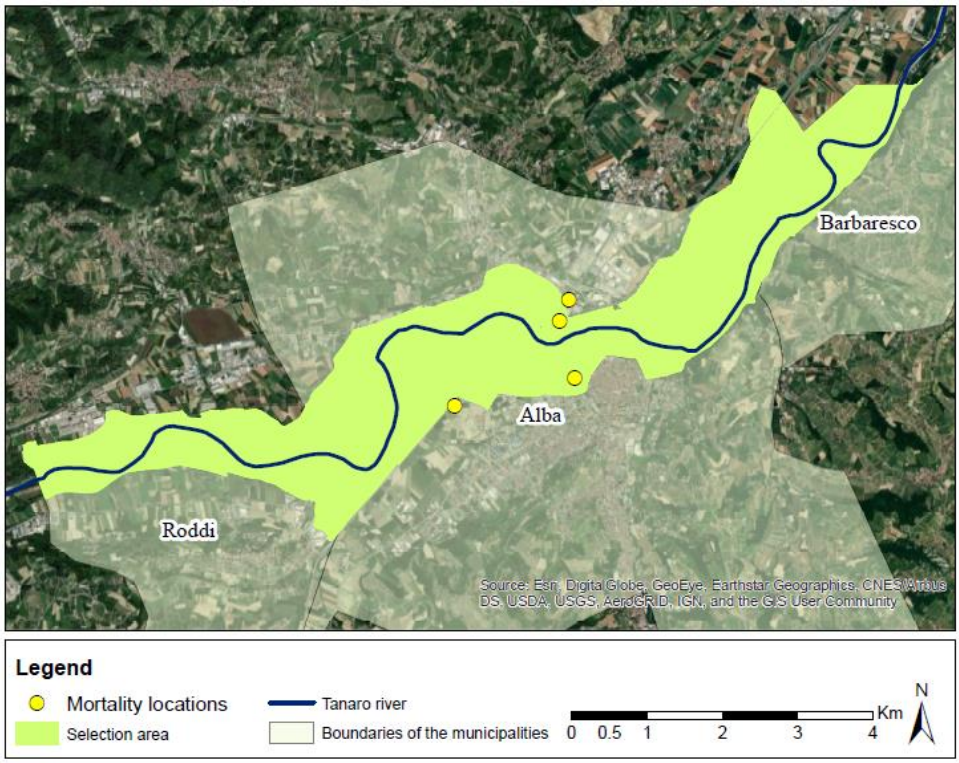

Figure 18- Example.2- Selection area

- **Age and Gender**

The frequency distribution of the two sociodemographic parameters of age and gender were obtained referring to the Italian census data of 2011 [27]. The graphs in Figure 19 show the distribution of these two parameters over the selection area regarding this example. Therefore, for each morality record, ten random values are extracted from the distribution in Figure 19-a for the "Age" parameter, and as for the "Gender", ten random values are extracted from the distribution in Figure 19-b for the creation of ten "Non-fatality" records.

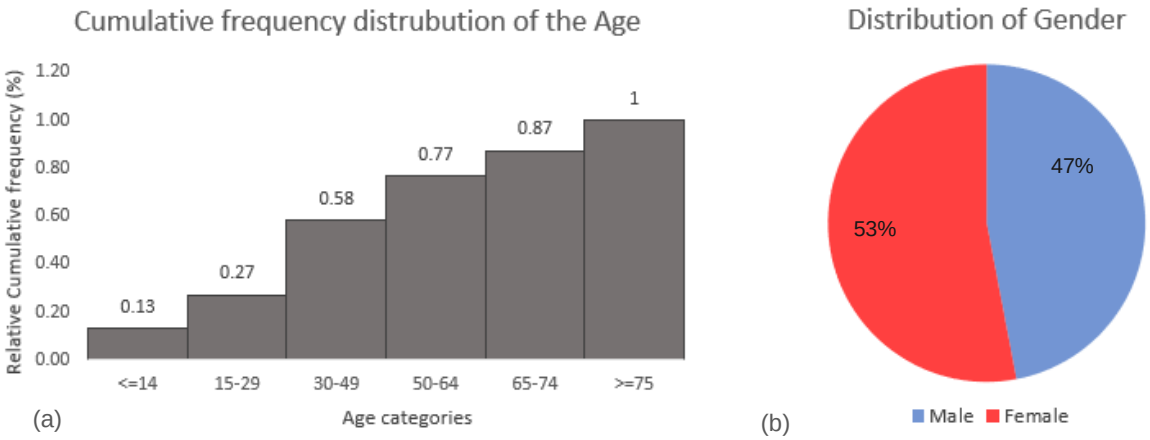

Figure 19- Example.2- Frequency distribution of a- Age, b- Gender

- **Distance from the river**

To obrain the frequency ditributionof this parameter, a point grid of 5m × 5m over the selection area is considered. Then, the distances of those points from the Tanaro river (the river involved in this example) are calculated using ArcMap tools. Next, these values are divided into nine classes for which the relative cumulative frequency distribution is obtained (Figure 20), and the random values are extracted from this distribution to for the “Non- fatalities”.

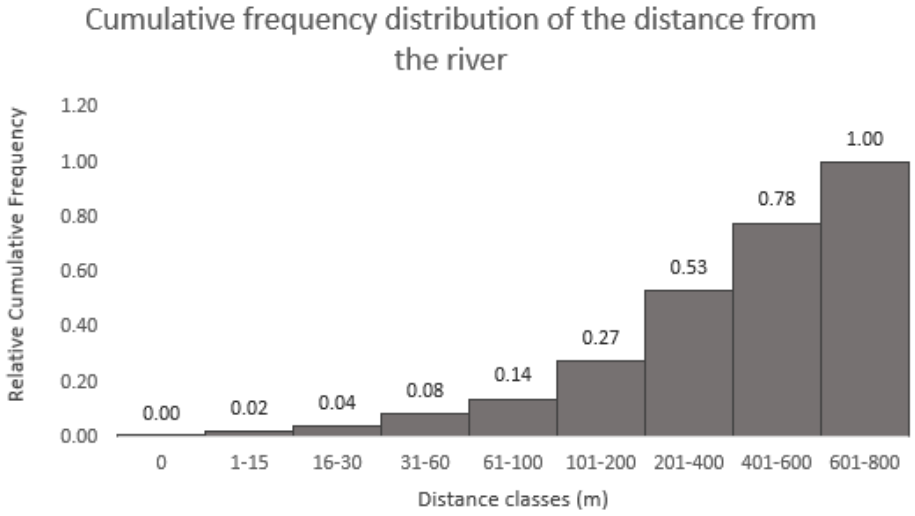

Figure 20- Relative cumulative frequency distribution of "Distance from the river"

- **Place**

For the analysis of this parameter, the first step was to classify the selection area into Indoor (buildings) and Outdoor spaces as shown in Figure 22, and to calculate the frequency distribution of them. Based on the National Human Activity Pattern Survey (NHAPS) [27], a two-year probability-based telephone survey of exposure-related human activities in the United States, the respondents reported spending an average of 87% of their time in enclosed buildings. This statistic is applied to the distribution of the parameter “Place”, and then a final frequency distribution is obtained. This distribution which is presented in Figure 23 shows that the population is 26% likely to be found in indoor spaces, and 74% likely to be in outdoor spaces. Thus, as explained before, ten values are extracted from this distribution with the two categories of indoor and outdoor spaces.

As a next step, the outdoor spaces are divided into 4 sub-categories of bridge, river, campsite, and outdoor area, and then, the frequency distribution of them is obtained (Figure 24). Next, equal to the number of the previously extracted values belonging to the “outdoor spaces”, new values are obtained from the distribution presented in Figure 24. Therefore, 68.08% (0.92× 0.74) of the population in the selection area are likely to be in outdoor areas and only 5.92% in the river, while almost non belong to the category of campsite and bridge because of the very low percentage of the area occupied by them. Thus, the random values for the parameter “Place” needed to create

the “Non-fatalities” records are obtained from these two distributions, declaring the possible place that the people involved during the flood event could have been. A scheme of this process is presented in Figure 21.

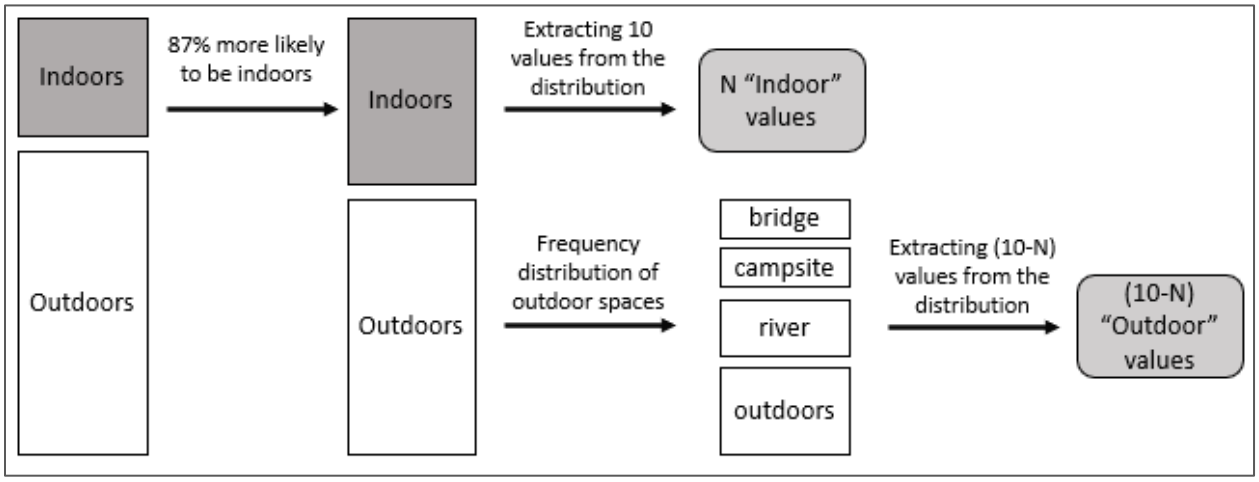

Figure 21- A scheme for the random selection of the parameter “Place”

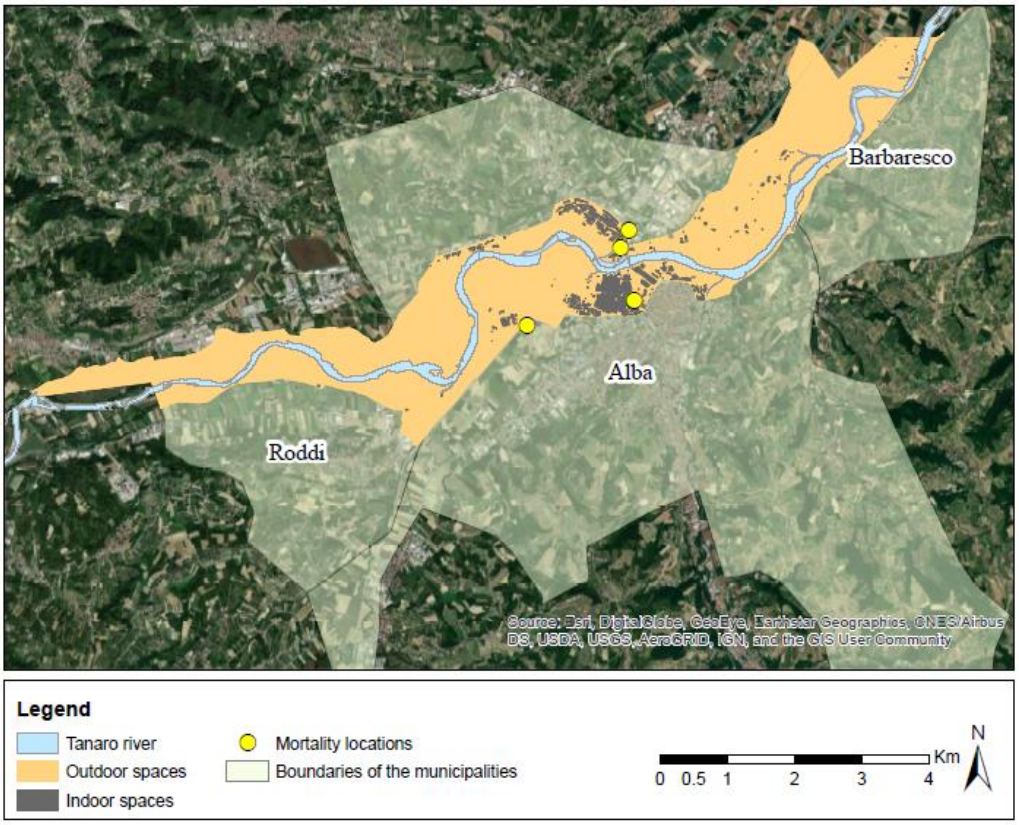

Figure 22- Example.2- Classification of the selection area into Indoor and Outdoor spaces

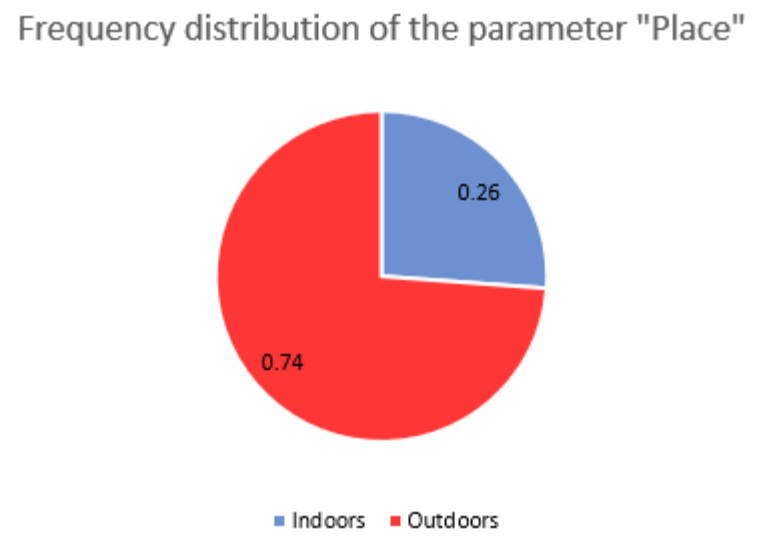

Figure 23- Example.2- The relative frequency distribution of Outdoor/Indoor spaces, applying the NHAPS statistic

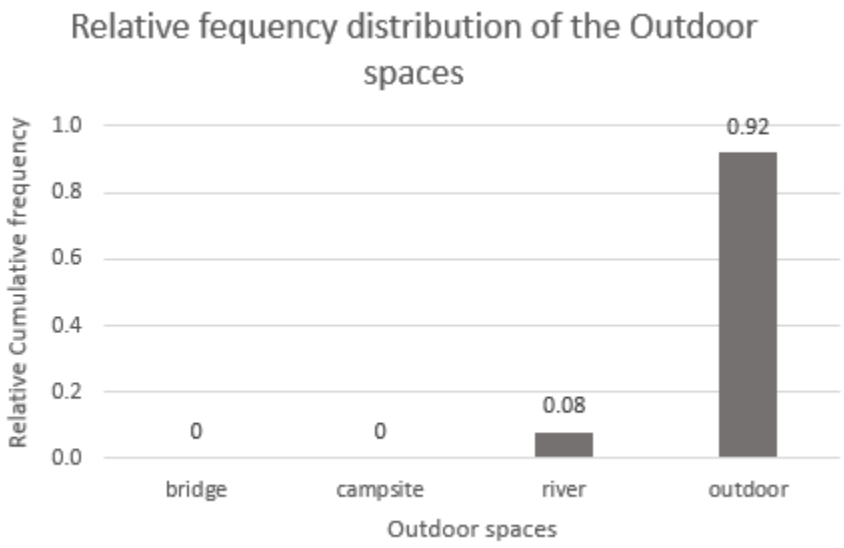

Figure 24- Example.2- The relative frequency distribution of outdoor spaces

- **Density of the buildings (Municipality)**

For this parameter, since it is related to the municipality the person involved in the flood event belongs to, it is calculated for each of the three municipalities of Alba, Roddi, and Barbaresco for this example. Then, for each record of the “Non-fatalities”, this value is randomly chosen from the density values of the three municipalities.

- **Morphological zone**

The selection area is overlapped with the map of the morphological zone discussed in 3.4.1, to obtain the extension of the two categories of “Plain” and “Mountain” in the selection area and calculated the frequency distribution of this parameter. Then, the random values are extracted from that distribution. For this example, the whole selection area belongs to the “Mountain” class, thus, all of the “Non-fatality” records created are from the same morphological zone category.

- **Corine Land Cover**

Similar to the previous methodology to extract the other variables, the map of the Corine Land Cover is overlapped with the selection area of the example and based on the frequency distribution of the different categories (Figure 25), the random values needed to form the “Non-fatalities” are obtained.

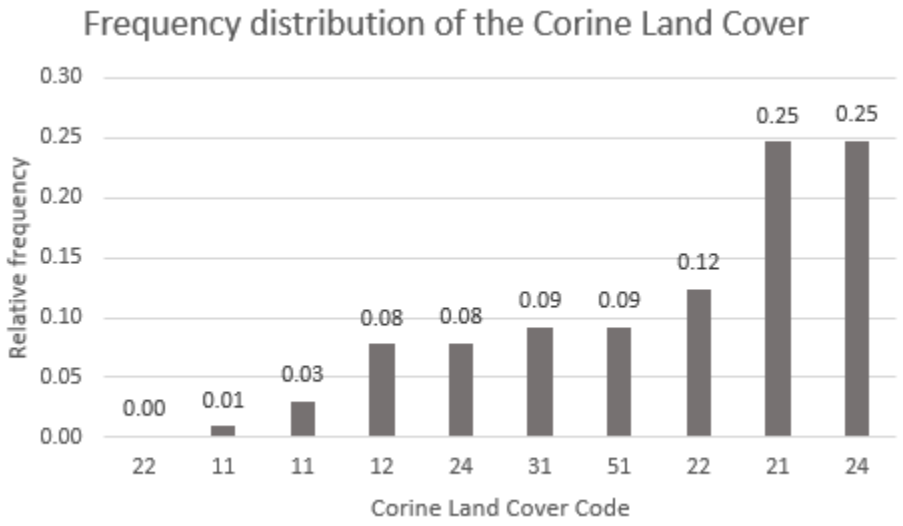

Figure 25- Example.2- Frequency distribution of the Corine Land Cover in the selection area

- **Hazard Scenario**

As indicated in Figure 17, the hazard scenario map for this example classifies the flood scenario into two categories: the high-frequency flood (return period of 50 years) and the low-frequency flood (return period of 50-100 years). Based on the distribution of this parameter over the selection area, the random values needed to form the synthetic records are extracted.

4.4. Specific cases

As explained through Example.2, the frequency distribution of the parameters is obtained over the selection area, which is mainly derived from the hazard scenario maps. However, some cases in which these maps cannot be used to form the selection area exist. These situations include:

1- When the location point of the mortality is located outside the scenario maps
2- When these maps are not available for where the fatality event occurred.

To solve this issue, the distance of the victim from the river comes to play. Thus, to create the "Non-fatalities" records for that victim, it was decided to consider a buffer area equal to the distance of that individual from the river causing the flood.

An example of this is the mortality event that occurred in Ceriano Laghetto municipality, causing the death of a 71-year-old male. For this case, the mortality point was far outside the hazard scenario. The buffer area created is observed in Figure 26. This area is assumed to give an idea of where the floodwater had reached in reality; thus, making up for the inability to use the hazard scenario maps. Therefore, the selection area for these cases is limited to the upstream and downstream municipalities as well as the municipality where the victim was found and the buffer area created, assuming that the flood was extended as far as the distance of the victim from the river.

Figure 26- The hazard buffer area for the mortality event in the municipality of Ceriano Laghetto

Another specific situation could happen when either one of the cases 1 or 2, as explained previously, occurs; however, the distance of the victim from the river is reported as zero. An example of this case is the event that happened in 1994 in the municipality of Dogliani. The fatality event occurred on a bridge causing the death of a 70-year-old female.

For this case, the suggested procedure as explained before was applied, however, since the distance of the victim from the river is zero, referring to Figure 11 in 3.4.4 based on which almost

90% of the distance values of the victims in the “Fatalities” dataset are within 600 meters from the river, the proposed solution was to use this value for the creation of the buffer. This situation is observed in Figure 27.

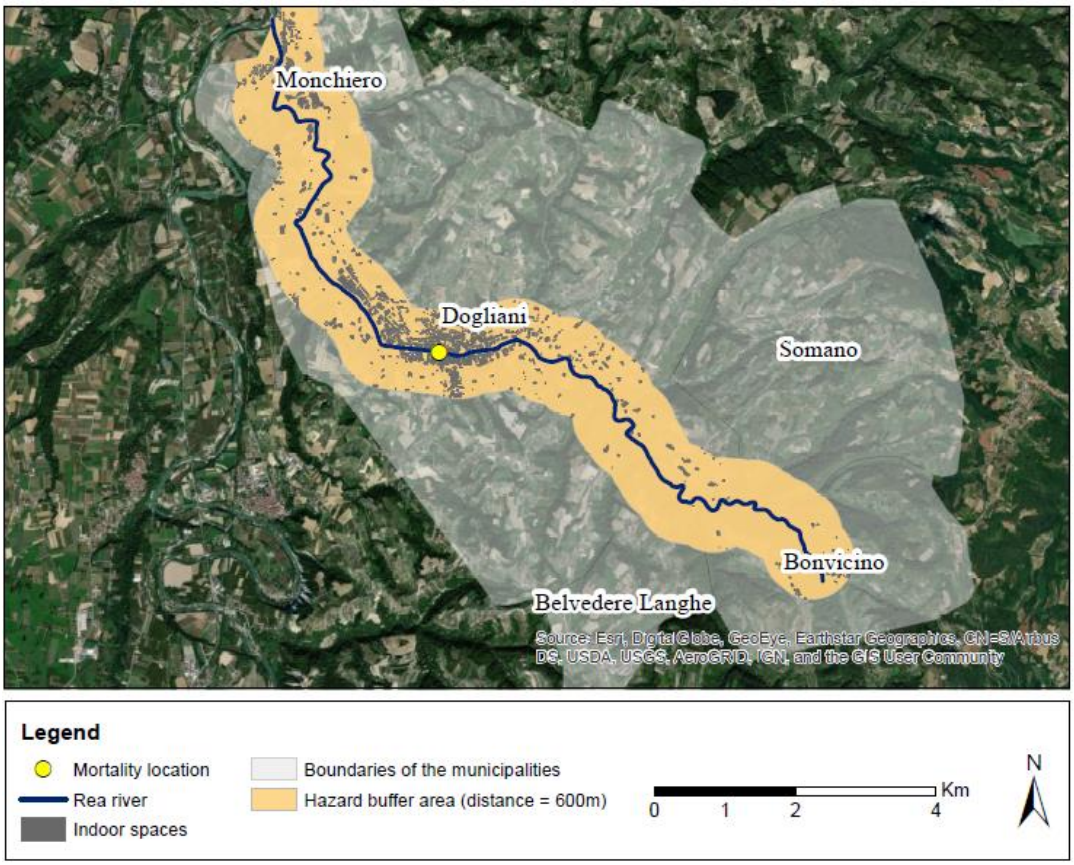

Figure 27- The hazard buffer area for the mortality event in the municipality of Dogliani

The procedure to create the synthetic “Non-fatality” records associated with each flood mortality is described in 4.3 and 4.4. Therefore, for each person who died in a flood event, ten other individuals who were affected by the same flood but did not lose their lives are characterized. This way, a total dataset of 127 mortalities (“Fatalities”) and 1270 “Non-fatalities” are formed, which will determine the role of the most significant parameters that result in death due to floods. This data will be analyzed together in the following chapters, with the use of a machine learning algorithm called Random forest.

Chapter 5

Data Analysis using Random forest

5.1. Introduction

Random forest (RF) [28], a supervised machine learning algorithm based on an ensemble of multiple decision trees, and a powerful tool often used for complex classification and regression problems, is adopted for the analysis of the data in the current study.

The RF algorithm has been previously used for multi-variable flood damage modelling and flood risk assessments, considering the complicated relationship between the model indicators and the target variables in these problems. Compared to other advanced statistical approaches the RF algorithm does not rely on any linear or other relationship between the input predictor variables and the target variable, and it is not sensitive to outliers, being able to handle nonlinear problems [28]. Wang and colleagues [29] used RF for the creation of a flood hazard risk assessment model and the identification of the most important indicators of flood risk, while Wagenaar and colleagues [30] adopted this algorithm as well as other supervised approaches for the modeling of flood damage to residential buildings. Terti and colleagues [23] on the other hand, applied this algorithm to the study of vehicle-related flash flood fatalities in the context of the United States, as a quantitative analysis of human loss of life due to flash floods. In the current study, the RF algorithm is chosen to create an estimation tool for flood mortality as well.

Before the explanation of the methodology, a few basic concepts are reviewed.

5.2. Machine learning algorithms

Machine learning (ML) algorithms are methods which automate data analysis and model building. Through these techniques, the machine is trained on the available data, learning the patterns and the relations present among the predictors (input variables), and then based on those findings, making decisions, and predicting on the unknown observations.

Various types of ML algorithms are present and widely used by data scientists, but in general, most statistical learning problems fall into one of two categories: supervised or unsupervised.

5.2.1. Supervised and Unsupervised learning

A *supervised* learning algorithm uses labeled data (data with both predictors and the associated responses). Thus, for each observation of the predictor measurements x_i, i = 1, …, n there is an associated response measurement y_i, and the algorithm aims at fitting a model that relates the response to the predictors, in order to accurately find the predictions for the future observations (prediction), or better understand the relationship between the response and the predictors (inference) [31]. Some examples of supervised learning algorithms are Linear regression, Logistic regression, or the Random forest algorithm which is use for this study.

In contrast, *unsupervised* learning algorithms use unlabeled data (data with only predictor measurements and no responses), where for every observation i = 1, ..., n, a vector of measurements x_i is present but no associated response y_i. Thus, the model tries to understand the relationships and the patterns between the input variables, to make inferences, or to find specific characteristics between the variables. A very common type of the unsupervised algorithms is clustering in which the observations fall into relatively distinct groups based on their specific characteristics [31].

5.2.2. Regression and Classification problems

Generally, problems with a quantitative response are referred to as *regression* problems, while those involving a qualitative target variable are referred to as *classification* problems. Usually, the selection of a statistical learning method is done on the basis of whether the response variable is quantitative or qualitative. Therefore, whether the predictors (input variables) are qualitative or quantitative is generally considered less important. Moreover, sometimes a statistical learning method can be used for either classification or regression problems.

This thesis addresses a classification problem in which the target variable describes the occurrence or non-occurrence of a fatality event.

5.3. Decision trees and random forests

RF algorithms operate based on an ensemble of multiple decision trees. Each tree produces one final result.

A decision tree is a tree-shaped diagram, which can be seen as a series of binary split points (nodes) leading to an answer in the form of a class (leaf) [30], as shown in Figure 28. Each branch of the tree represents a decision or an occurrence.

Each split (starting from the root node) is built from a binary question based on which the predictor space is divided into a number of simple regions. The splitting usually continues until a certain condition (e.g., the number of remaining observations in each region, or the maximum number of leaf nodes, etc.) is met, resulting in a final prediction.

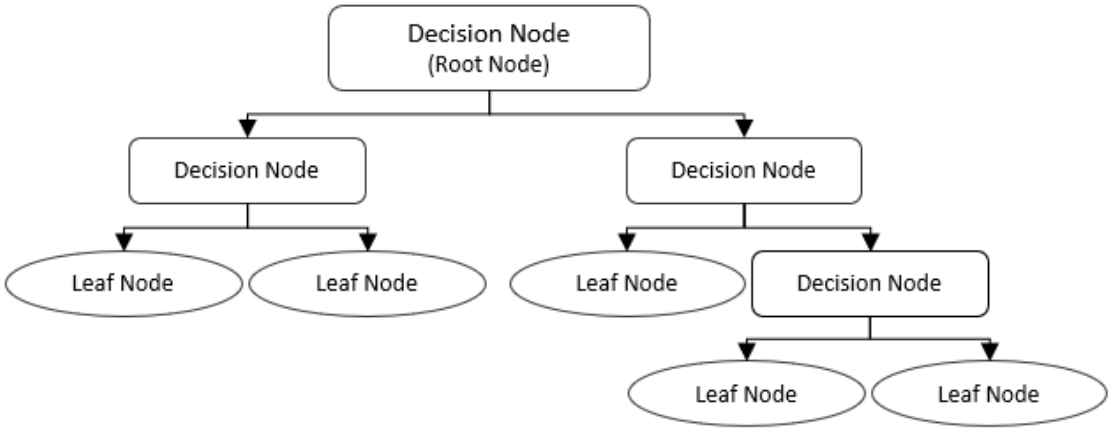

Figure 28- Diagram of a decision tree

For a classification problem using RF, the decision of the majority of the trees (the classification having the most votes among all of the trees in the forest) is considered as the final outcome of the RF. The scheme of a RF algorithm can be observed in Figure 29.

Based on the definition of the RF, each tree in the ensemble forest is built from a new training sample (sample of the data used for model training) drawn randomly with replacement (i.e., a bootstrap sample) from the N cases in the original training set.

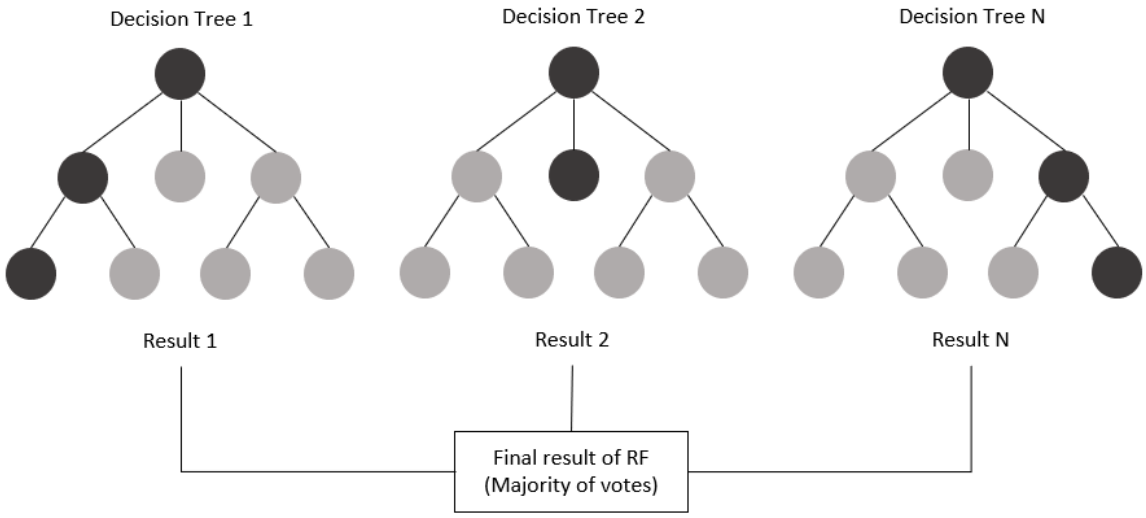

Figure 29- Diagram of a random forest algorithm

5.4. Methodology: classification method

Considering the total dataset of events and non-events, each data record is characterized by ten input variables (mortality indicators), and a target variable ("Death") associating it to a fatality case ("Fatalities") or a case where no fatality was recorded ("Non-fatalities"). These are the final indicators used for the training of the RF model (Table 5).

N	Input variables	Parameter described
1	Gender	Gender
2	Age class	Age
3	Place	Place
4	MorphologicalZone	Morphological Zone
5	HazardScenario	Hazard Scenario Code
6	Corine reclass	Corine Land Cover Code
7	Distance	Distance from the river
8	Density	Density of the buildings (Municipality)
9	ReturnPeriod	Return Period
10	Solid	Solid Transport

Table 5- List of the input variables for the RF algorithm

The processing of the indicators and the further analysis to create the RF model was performed in “RStudio” [32], using the function "randomForest " of the randomForest CRAN package [33]

5.4.1. Data Undersampling

Usually when facing classification problems, the classifying algorithms work well for datasets in which the distribution of the target variable is well balanced. However, in the cases with imbalanced dataset where there is an unequal distribution of classes across the data records, the accuracy of the model will be biased towards the majority class (the class with higher frequency), and the model might face the problem of overfitting (in which it learns the behavior and the patterns of the training data too well, that it even follows the specifications of the training dataset), getting very good at predicting the pattern of the majority class. As proposed by Yap and colleagues [34], undersampling, in which the records in the majority classes are eliminated randomly to achieve equal distribution with the minority class, is found to work well in improving the classification for the imbalanced dataset, similar to the case of this work.

The data available for processing in the current study, consists of 127 “Fatalities” and 1270 “Non-fatalities”, in which an imbalance between the two classes of the target variable is observed (the number of “Non-fatalities” are relatively higher with respect to the mortality data). Thus, to be able to control the number of the data records used for model training and validation, the parameter K (K= 1, 2, ... ,10) is defined as a control measure, specifying the ratio between the number of “Non-fatalities” and the number of “Fatalities”, as in equation 1. Various runs of the model can be performed with different values of K, to better observe the effects it creates on the performance of the model.

$$K = (N°of\ "Non-fatalities")/(N°\ of\ "Fatalities") \quad (1)$$

Therefore, e.g., with K=2, the model uses 127 “Fatalities” and takes a sample of 254 records from the “Non-fatalities” to form the RF input dataset.

5.4.2. Training and test data

As mentioned in 5.2, a machine learning model uses a set of data to learn, and then makes predictions on unknown observations. Usually, the total dataset is divided into two separate subsets for training and testing. The training set is used to create the model, and then, a set of samples can be set aside to evaluate the final model, called the test dataset. The decision to divide the data to training/ test sets may vary for each problem, thus, different cases were tried out for this work, with different ratios of training and test set. For this purpose, the parameter “trainset” is introduced which takes into account the ratio between the training set and the test set.

5.4.3. Number of trees

As explained in 5.3, random forests consist of multiple random decision trees. Based on the function “randomForest” used in the model algorithm for the current work, the user can decide on

the number of the random decision trees that the algorithm grows. This factor, represented by the parameter “ntree” in the model, should not be set to too small a number, to ensure that every input data record gets predicted at least a few times [33], thus increasing the randomness in the tree production. However, the improvement in the results decreases at some point, as the number of the trees increase, i.e., there is a point at which the improvement in the predictions from constructing more trees will be lower than the computational cost. So, considering the increase in the time that the algorithm needs to create a large number of trees, one should find an optimal value to be adopted in the model.

For the development of the RF in this study, the change in the performance of the model while considering the cost in computational time was investigated, and in the end the optimal value of 500 trees were considered for the development of the algorithm.

5.4.4. Model validation

Referring to section 5.4.1, when running the model, from the 1270 data available in the synthetic dataset of “Non-fatalities”, we use (127×K) random samples in addition to the 127 “Fatality” records, to train the model (the portion of data used for this purpose is called “Training set”) and then test the model’s performance (1st validation dataset). The parameter “trainset” describes what percentage of the data is used for the training set and the 1st validation set.

A different validation of the model can be carried out by considering the remaining part of the synthetic data (1270 – 127× K) that was not used in the previous step and evaluating the model’s prediction ability on this new test set consisting of only “Non-fatalities”. This dataset is called the “2nd validation” set. A scheme of this concept is shown in Figure 30.

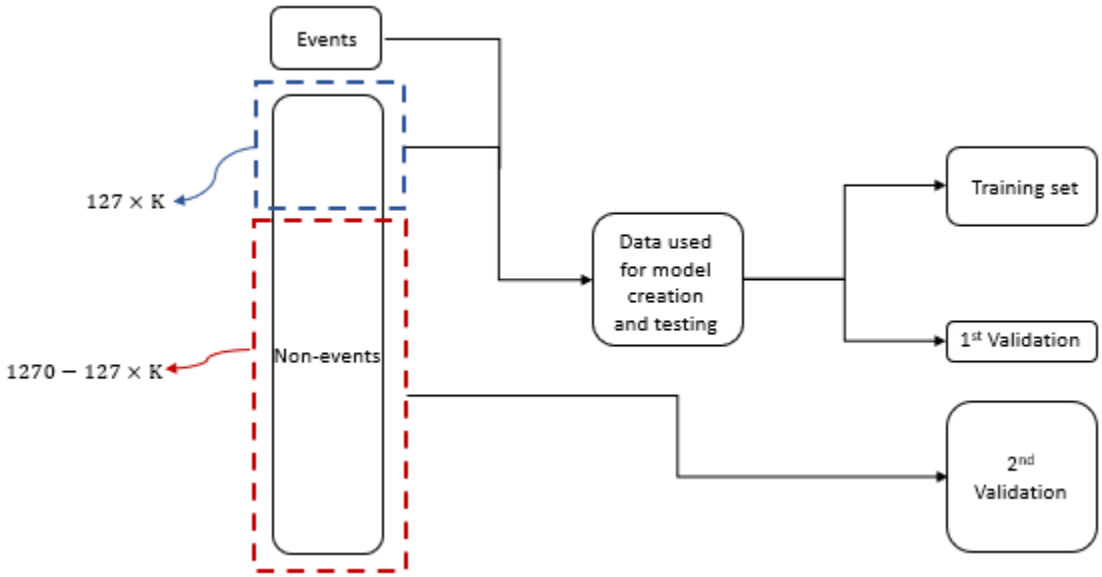

Figure 30- Scheme of the model training and validation datasets

Therefore, the model is created on the training set, and later, to judge the performance of the model, it is applied on the validation datasets (the data belonging to the validation sets were not used for training the model). The result of the model is a classification of the target variable for each record of the validation set, associating each record with an outcome of “Fatality” or “Non-fatality”. Thus, comparing what the model predicts on each record and what the target value is already available in the validation datasets, the prediction ability of the model is investigated.

5.4.5. Measures of model performance

Validation measures:

To measure the performance of the model, the predicted results of the model should be compared with the know values available in the validation dataset. For a more comprehensive investigation of the model’s performance, the classification Accuracy (Acc), the True Positive Rate (TPR), the True Negative Rate (TNR), and the False Negative ratio (FN-ratio) on the test data are investigated. The mentioned concepts are described below (equations 2,3,4,5).

- Classification Accuracy: It is a measure that describes the number of the correct predictions the model makes with respect to all of the predictions.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

- False Negative Ratio: A ratio that defines the number of false negatives (missed alarms) with respect to all of the deaths in the test data. It is a measure that is very important in the context of this study, since a high number of missed fatalities in the results does not represent a good performance by the model.

$$FN - ratio = \frac{FN}{N° \; of \; "Events" \; in \; test \; set} \tag{3}$$

- True Positive Rate: This measure considers the ability of the model to correctly classify the “Fatalities”.

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

- True Negative Rate: It describes the ability of the model to correctly classify the “Non-fatality” cases.

$$TNR = \frac{TN}{TN + FP} \tag{5}$$

To better present the prediction results of the classifier, a confusion matrix was implemented, which summarizes the model performance measures on the first validation dataset. A scheme of the confusion matrix is shown in Table 6.

Target / Prediction	1 ("Fatality")	0 ("Non- fatality ")
1 ("Fatality ")	*TP*	*FP (False Alarm)*
0 ("Non-fatality")	*FN (Missed Alarm)*	*TN*

Table 6- Confusion Matrix

Second validation measures:

Since the dataset used for the second validation is extracted from the "Non-fatalities" data, we can only evaluate the performance of the model on this set based on the two measures of True Negative ratio (TNR_{2ndval}) and the False Positive ratio (FPR_{2ndval}), described in equations 5 and 6.

$$FPR_{2ndval} = \frac{FP}{TN + FP} \quad (6)$$

It is important to mention that the accuracy (Acc) alone typically does not provide enough information based on which one can declare that the model has a good performance, especially, in the cases with imbalanced dataset, as discussed in 5.4.1. In the context of this study, since the classification problem focuses on the predictions of fatalities, the model should not only have a reasonable accuracy, but also it should be able to make predictions with a low number of missed fatalities (missed alarms). Therefore, an attempt was made to find the balance between model accuracy, paying attention to the rate of the true negatives and the true positives, and the false negative ratio (FN-ratio) in order to identify the best model. Also, as a next measure, results of the second validation were taken into consideration.

5.4.6. Sensitivity runs

A total of 88 runs were performed on the RF algorithm, always considering the accuracy measures and the FN-ratio, and the measures of the 2nd validation (TNR_{2ndval} and FPR_{2ndval}). In each of the runs, change in 3 model parameters were kept in check:

1- The change in the K parameter (to control the number of the data records used for model training and validation) and how it affects the model performance
2- The ratio of training & test dataset used in model processing (training/test) and its influence on the results
3- The change in the performance of the model with all the input variables and also, when removing each of the ten input variables one-by-one

The various sensitivity runs help calibrate the model parameters in order to achieve the best result. The structure of the 88 runs performed to obtain the best model is presented in Table 7. The total 88 runs consist of 11 main cases considering point 3 mentioned above, and for each case, 8 sensitivity analyses are performed taking into account points 1 and 2.

5.4.7. Variable Importance

When working on prediction models, not only is it important to have an accurate prediction, but also to be able to interpret the results and to understand the most important features in describing the model response. Knowing the significance of the input variables can result in a better assessment of the logic of the model and to understand if that logic is correct. Moreover, by checking the importance of the model indicators, it is possible to remove those that are not that significant in obtaining the final result and have similar or better performance in a shorter training time. In general, the variable performance is a function that provides the decision maker with an opportunity to estimate an indicator's contribution to the model performance, and decide whether it is more advantageous to keep it as the model input or remove it, given the accuracy it results in.

In this study, the function "varImpPlot" from the randomForest CRAN package [33] was used to plot the variable importance measure for the indicators used in the development of the mortality model. This algorithm uses two types of measures for the description of the significance of the predictor variables, the Gini Index, and the Mean Decrease Accuracy. The latter is used to judge the importance of the input variables in the model development.

The Mean Decrease Accuracy is computed from permuting out-of-bag (OOB) data. For each tree, a subsample with replacement is used to create training samples for the model to learn from. The remaining part of the observations not used to fit a given tree are referred to as the OOB observations. For each tree, the prediction error on the OOB portion of the data is recorded (the error rate for classification problems). Then the same is done after permuting each predictor variable. The difference between the two cases is then averaged over all trees and normalized by the standard deviation of the differences. If the standard deviation of the differences is equal to 0 for a variable, the division is not done. The Mean Decrease Accuracy plot expresses how much accuracy the model loses by excluding each variable. The more the accuracy suffers, the more important the variable is for the classification. In the plot produced by "varImpPlot", the variables are presented in descending importance. Figure 31 shows the variable importance plot for run 2.2.2 listed in Table 7 in which the variable "Gender" is removed from the list of input indicators, with a K value equal to 2 (127 "Fatalities" and 254 "Non-fatalities), and the trainset ratio equal to 0.8 (80% of the data used for model training and 20% used for the 1st validation).

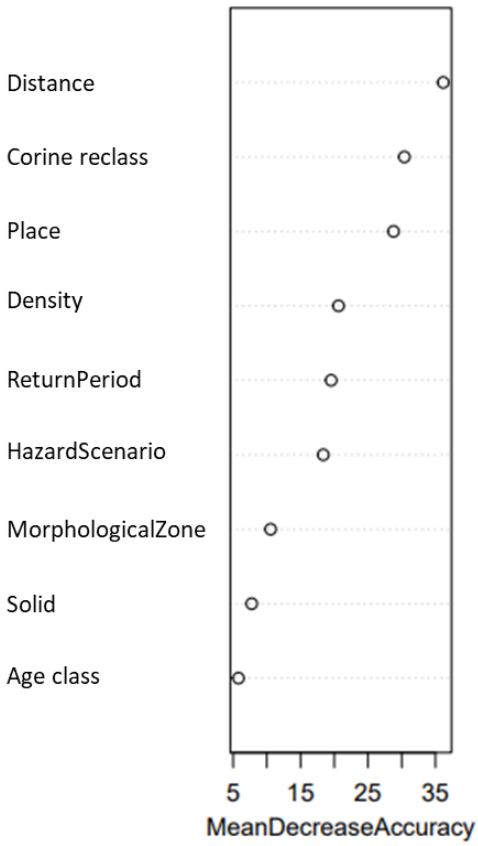

Figure 31- Variable importance plot for run 2.2.2

The results obtained by performing the 88 runs listed in Table 7 are discussed in Chapter 6.

N	Model run	Case description	K	Trainset
1	run 1.1.1	Base Case	1	0.7
2	run 1.1.2			0.8
3	run 1.2.1		2	0.7
4	run 1.2.2			0.8
5	run 1.3.1		3	0.7
6	run 1.3.2			0.8
7	run 1.4.1		8	0.7
8	run 1.4.2			0.8
9	run 2.1.1	Removing "Gender"	1	0.7
10	run 2.1.2			0.8
11	run 2.2.1		2	0.7
12	run 2.2.2			0.8
13	run 2.3.1		3	0.7
14	run 2.3.2			0.8
15	run 2.4.1		8	0.7
16	run 2.4.2			0.8
17	run 3.1.1	Removing "Age"	1	0.7
18	run 3.1.2			0.8
19	run 3.2.1		2	0.7
20	run 3.2.2			0.8
21	run 3.3.1		3	0.7
22	run 3.3.2			0.8
23	run 3.4.1		8	0.7
24	run 3.4.2			0.8
25	run 4.1.1	Removing "Morphological Zone"	1	0.7
26	run 4.1.2			0.8
27	run 4.2.1		2	0.7
28	run 4.2.2			0.8
29	run 4.3.1		3	0.7
30	run 4.3.2.			0.8
31	run 4.4.1		8	0.7
32	run 4.4.2			0.8
33	run 5.1.1	Removing "Solid Transport"	1	0.7
34	run 5.1.2			0.8
35	run 5.2.1		2	0.7
36	run 5.2.2			0.8
37	run 5.3.1		3	0.7
38	run 5.3.2			0.8
39	run 5.4.1		8	0.7
40	run 5.4.2			0.8
41	run 6.1.1	Removing "Density"	1	0.7
42	run 6.1.2			0.8
43	run 6.2.1		2	0.7
44	run 6.2.2			0.8
45	run 6.3.1		3	0.7
46	run 6.3.2			0.8
47	run 6.4.1		8	0.7
48	run 6.4.2			0.8
49	run 7.1.1	Removing "Hazard Scenario"	1	0.7
50	run 7.1.2			0.8
51	run 7.2.1		2	0.7
52	run 7.2.2			0.8
53	run 7.3.1		3	0.7
54	run 7.3.2			0.8
55	run 7.4.1		8	0.7
56	run 7.4.2			0.8
57	run 8.1.1	Removing "Return Period"	1	0.7
58	run 8.1.2			0.8
59	run 8.2.1		2	0.7
60	run 8.2.2			0.8
61	run 8.3.1		3	0.7
62	run 8.3.2			0.8
63	run 8.4.1		8	0.7
64	run 8.4.2			0.8
65	run 9.1.1	Removing "Distance"	1	0.7
66	run 9.1.2			0.8
67	run 9.2.1		2	0.7
68	run 9.2.2			0.8
69	run 9.3.1		3	0.7
70	run 9.3.2			0.8
71	run 9.4.1		8	0.7
72	run 9.4.2			0.8
73	run 10.1.1	Removing "Place"	1	0.7
74	run 10.1.2			0.8
75	run 10.2.1		2	0.7
76	run 10.2.2			0.8
77	run 10.3.1		3	0.7
78	run 10.3.2			0.8
79	run 10.4.1		8	0.7
80	run 10.4.2			0.8
81	run 11.1.1	Removing "Corine Land Cover"	1	0.7
82	run 11.1.2			0.8
83	run 11.2.1		2	0.7
84	run 11.2.2			0.8
85	run 11.3.1		3	0.7
86	run 11.3.2			0.8
87	run 11.4.1		8	0.7
88	run 11.4.2			0.8

Table 7- List of the model runs and the parameters modified

Chapter 6

Discussion of the results

6.1. Introduction

As explained in Chapter 5, the random forest algorithm was used in order to create the flood mortality model for this study. The model is created based on the training data, and its performance is investigated with respect to the results obtained through the 88 runs executed on two validation datasets. These results are compared with respect to the measures defined in section 5.4.5 (FN-ratio, Acc, TPR, TNR, TNR_{2ndval}, and FPR_{2ndval}) in order to calibrate the model parameters, and eventually, the best result is chosen as the final model.

6.2. Results

Table 11 shows the model performance measures for all of the 88 model runs, two of which have been chosen as the best cases of the model's performance, as observed in Table 8 below. These runs of the model represent an acceptable classification accuracy on the first validation set, while keeping a rather low ratio of false negatives (FN-ratio). Moreover, another measure which was considered for choosing these runs is the performance of the model on the second validation set, with high TNR_{2ndval} values.

Runs				1st Validation				2nd Validation	
Label	Description	training/test	K	Acc	FN-ratio	TPR (%)	TNR (%)	TNR_{2ndval}	FPR_{2ndval}
run 1.2.2	Base case	0.8	2	88.3	0.222	77.8	94	93.7	6.3
run 2.2.2	Removing Gender	0.8	2	89.6	0.185	81.5	94	93.2	6.8
run 9.1.2	Removing Distance	0.8	1	68.6	0.25	75	60.9	68.4	31.6

Table 8- Summary of the performance measures for run 1.2.2, run 2.2.2. & run 9.1.2

6.2.1. Run 1.2.2

Looking at the results for run 1.2.2 (one of the base runs in which the RF algorithm is performed with all of the 10 predictor variables) in Table 8 and Figure 32, it is observed that in the 1st validation, with a FN-ratio of 0.222 (6 out of 27 fatalities in the test set), and a classification accuracy of 88.3%, the model is able to correctly classify the fatalities by a ratio of 77.8% indicated by the TPR, and for the "Non-Fatalities" in the test set, the TNR is achieved at 94%. As for the

performance measures of the 2nd validation, , the TNR_{2ndval} is achieved at 93.7%, meaning that the model is able to correctly predict 952 of the “Non-fatality” cases with respect to the total 1016 records, while misclassifying 64 cases, as observed in Figure 32.

Thus, it is possible to conclude that the initial decision on the choice of the 10 descriptive variables for flood mortality was reasonable since an acceptable level of performance was achieved in this run, where all of the predictors were used in model training.

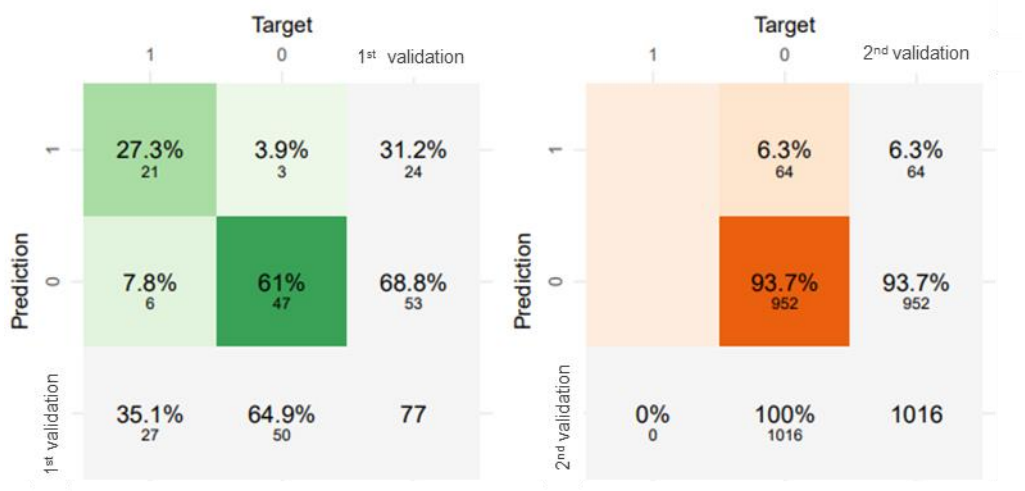

Figure 32- Confusion matrix of the run 1.2.2.- Left: Confusion matrix of the 1st validation, Right: Confusion matrix of the 2nd validation

Moreover, based on Figure 33, which shows the variable importance plot with the conditions of run 1.2.2, in the case of using all of the mortality indicators, the parameter “Distance “is the one with the highest Man Decrease Accuracy, meaning that it holds the highest significance in this run and removing it from the input variables results in the highest loss of accuracy.

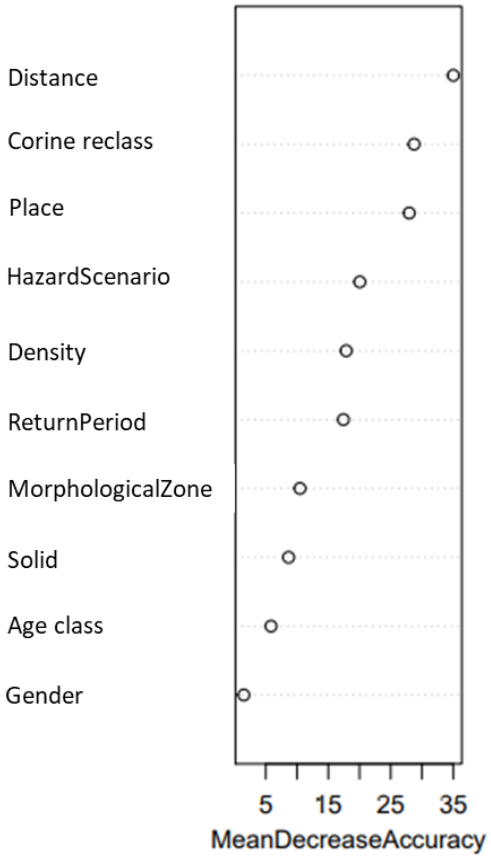

Figure 33- Variable importance for run 1.2.2 (indicated by the Mean Decrease Accuracy)

6.2.2. Run 2.2.2

In this run the model is created based on 9 predictor variables, excluding the parameter of "Gender". As indicated in Figure 33, when running the model with all of the ten predictor variables, the parameter of "Gender" has the least importance for the model.

As shown in Table 8, with a K value equal to 2, 127 "Fatality" records and 254 "Non-fatality" records are used for model training and also for the 1st validation of the model. Therefore, the 1016 remaining "Non-fatalities" are used for the 2nd validation The results of this run are shown in the confusion matrix in Figure 34.

In this run, for which the FN-ratio is equal to 0.185 (having 5 false negatives out of 27 events in the test set), the classification accuracy of the model is at 89.6%, with a true positive rate of 81.5% and a true negative rate of 94%, which is considered one of the best calibrations for the model. As for the results on the 2nd validation, as observed in Figure 34 on the right matrix, from the 1016 records used for this validation, the model classifies the non-fatality cases correctly for 947 records (TNR_{2ndval} : 93.2%), misclassifying 69 of the cases (FPR_{2ndval} : 6.8%).

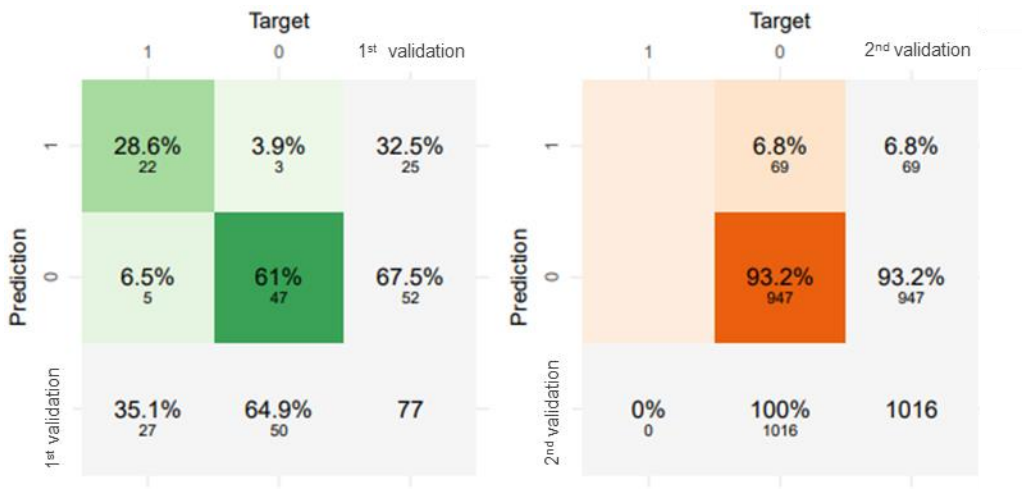

Figure 34- Confusion matrix of the run 2.2.2- Left: Confusion matrix of the 1st validation, Right: Confusion matrix of the 2nd validation

In this run, as shown in Figure 35, the parameter "Distance" is still the highest input variable in the variable importance plot. Moreover, similar to run 1.2.2, the next two most significant variable in describing mortality due to floods are the Corine Land Cover parameter and the Place of the accident. While, removing the parameter "Gender" from the list of input variables puts the parameter of "Age" at the lowest part of the plot.

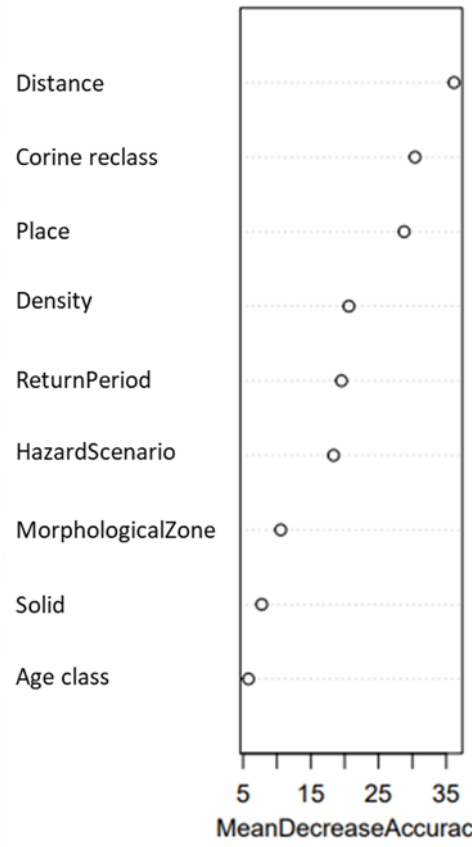

Figure 35- Variable importance for run 2.2.2 (indicated by the Mean Decrease Accuracy)

6.2.3. Discussion on the parameter of "Distance" - Run 9.1.2

As observed in the variable importance plots, for the two best runs obtained from model calibration (Run 1.2.2: Figure 33 and Run 2.2.2: Figure 35), the parameter "Distance' has the highest influence for a successful prediction of flood mortality by the model. However, as explained in 3.6.1, this parameter holds some degrees of uncertainty, thus, among the runs performed on the algorithm (Table 11), it was decided to investigate the results of the model performance for the case where the parameter "Distance" is excluded from the input variables of the model. The best results are obtained for run 9.1.2 (Table 8) in which an equal number of "Fatalities" and "Non-fatalities" are used for the model creation and the 1st validation. With the ratio of 80% and 20% for the training and test set respectively, the test set has 51 records 28 of which result in mortality. As observed in the confusion matrix, in Figure 36, for the 1st validation, the model is able to accurately classify the records on the test set by 68.6%, for which the TPR is 75% and the TNR is 60.9, Also, the model misclassifies 7 cases as false negatives (FN-ratio: 0.25). The results of the 2nd validation indicate a TNR of 68.4 %, 782 correct classifications out of 1143 "Non-fatalities" in the data, while misclassifying 361 records (FPR_{2ndval}: 31.6%). Overall, we observed very lower performance in this run than in runs 2.2.2 and 1.2.2., highlighting the need of having accurate data on the variable distance to increase the accuracy of the model.

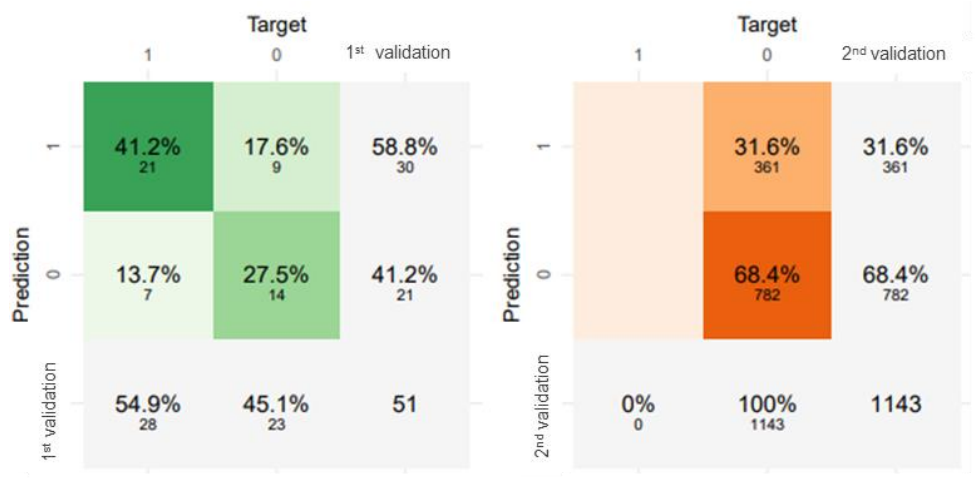

Figure 36- Confusion matrix of the run 9.1.2- Left: Confusion matrix of the 1st validation, Right: Confusion matrix of the 2nd validation

6.2.4. Choice of the final model setup

Based on the runs performed for the calibration of the model parameters, the final choice of the model is with the parameters in runs 2.2.2, which is the model run with the best metrics on the 1st and the 2nd validations. The characteristics of the final model are presented in Table 9.

Final Model Characteristics				1st Validation				2nd Validation	
Input variables		training/test	K	Acc	FN-ratio	TPR (%)	TNR (%)	TNR_{2ndval}	FPR_{2ndval}
Distance Place Corine Land Cover Density Hazard Scenario	Morphological Zone Age Return Period Solid Transport	0.8	2	89.6	0.185	81.5	94	93.2	6.8

Table 9- Final Model Characteristics, run 2.2.2

6.3. Interpreting the results into rules

Tree ensembles such as random forests are accurate machine learning algorithms, but they are usually difficult to understand. Deng and colleagues propose the inTrees package in R (CRAN) [35] a framework that extracts and summarizes rules that govern the splits in each tree in the tree ensemble (in a descriptive way and with the support of logical operators) and calculates frequent variable interactions [36]. This way, the extracted rules are easier to interpret.

Such a tool has been used on the RF model for flood mortality, to create a tool that could be useful for governments and organizations interested in estimating and mitigating flood mortality.

6.3.1. The inTrees framework

The inTrees package consists of algorithms to extract rules, measure (and thus rank) rules, prune irrelevant or redundant variable-value pairs of a rule (prune each rule), select a compact set of relevant and non-redundant rules (select rules), discover frequent variable interactions (extract frequent patterns), and finally, summarize rules into a learner that can be used for predicting new data [36]. This algorithm can be applied to both classification and regression problems. It also can be used to interpret algorithms with an ensemble of decision trees such as random forests.

For the adaptation of the inTrees algorithm for this work, firstly, the rules were extracted from the 500 trees in the random forest characterized in Table 9, then based on the metrics of those rules, they were classified for their qualities. Next, each rule was pruned where it included irrelevant or redundant variable-value pairs. Finally, to summarize the extracted rules, a rule-based learner was implemented.

An overview of the concepts related to this algorithm used in development of the code is presented in the following.

6.3.2. Rules extraction

The extraction of the rules is performed starting from a decision tree's root node to a leaf node (Figure 28). Each rule is expressed as {C ⇒ T}, where C, referred to as the condition of the rule,

is a conjunction of variable-value pairs, and T is the outcome of the rule. The rules extracted from a tree ensemble are a combination of rules extracted from each decision tree in the tree ensemble [36]. For this purpose, the “extractRules” function of the inTrees package [35] was implemented in the algorithm.

6.3.3. Rules metrics

Each rule can be classified by its metrics which describe the rule’s quality (ranking). Three different measures are introduced for this purpose.

- **Frequency:** The frequency measures the popularity of the rule, defined as the proportion of data instances satisfying the rule condition.
- **Error:** For classification problems, it is defined as the number of incorrectly classified instances determined by the rule divided by the number of instances satisfying the rule condition.
- **Length:** The complexity of a rule is measured by the length of the rule condition and is defined as the number of variable-value pairs in the condition.

It should be noted that given two rules with similar frequency and error, the rule with a smaller length may be preferred as it is more interpretable.

The “getRuleMetric” function of the inTrees package [35] was used to obtain the rules’ metrics.

6.3.4. Pruning the rules

The process of pruning is performed to minimize the number of rules since usually, the rules that are extracted using the inTrees package are in a large quantity, so one could find it harder to interpret them.

Each rule extracted from the trees may include irrelevant variable-value pairs. The act of pruning purifies the rule from the irrelevant and redundant variable-value pairs. For each rule, a value defined as “$decay_i$” is introduced that takes into account the increase of error for the rule after removing the *i*th variable-value pair. When “$decay_i$” is smaller than a threshold, the *i*th variable-value pair may be considered unimportant for the rule and thus can be removed [36]. The function “pruneRule” from the inTrees package [35] was used for this purpose.

6.3.5. Summarize the rules

The rules extracted from the tree ensemble can be summarized into a rule-based learner, referred to as a simplified tree ensemble learner (STEL). The goal of this learner is to build a list of rules ordered by priority [36]. For the implementation of this algorithm, the “buildLearner” function of the inTrees package [35] was used.

6.3.6. The rules

After executing the functioned mentioned above as a part of the inTrees package, 34 rules for the flood mortality model were extracted from an initial number of 12917 rules (before pruning). The rules are presented in Table 10. Each rule is characterized by its metrics (frequency, error, and length) with the conditions (variable-value pairs) of that rule leading to a prediction of “Fatality”

('1') and "Non-fatality" ('0'). Extracting these rules can bring the model one step closer to a more straightforward interpretation of the results.

N	length	frequency	error	condition	pred
1	4	0.037	0	HazardScenario ∈ {L, M, outside} & Corine reclass ∈ {21, 22, 23, 24, 31, 32, 33} & Age class ∈ {15-29, 50-64 , ≥75} & ReturnPeriod ∈ {< 2 , 20-50}	1
2	3	0.031	0.167	Corine reclass ∈ {24, 31} & Density > 0.007 & ReturnPeriod ∈ {2-5, >200}	1
3	3	0.024	0	Corine reclass ∈ {11, 12, 13, 14, 23, 31, 41} & Age class ∈ {30-49} & ReturnPeriod ∈ {< 2, 2-5, 5,10-20}	1
4	5	0.021	0	HazardScenario ∈ {L, M, outside} & MorphologicalZone ∈ {Mountain} & Distance >128m & Solid ∈ {No, Yes} & ReturnPeriod ∈ {< 2, 5, 10-20,>200}	1
5	3	0.021	0	HazardScenario ∈ {H} & Corine reclass ∈ {11, 12, 14, 23, 41, 51) & Density > 0.028	1
6	4	0.018	0.143	HazardScenario ∈ {L, M, outside} & Corine reclass ∈ {12, 13, 14, 21, 22, 23, 41, 51} & Place ∈ {Bridge, Building, Campsite} & ReturnPeriod ∈ {5, 20-50, 50-100, 100-200, >200}	1
7	4	0.018	0.143	Corine reclass ∈ {23, 24, 32, 33, 51} & Place ∈ {Bridge, Campsite, Outdoor, River} & Age class ∈ {50-64, ≥75} & ReturnPeriod ∈ {< 2, 2-5, 5, 20-50,100-200}	1
8	4	0.016	0	HazardScenario ∈ {M, outside} & Place ∈ {Outdoor, River} & Distance ≤ 42m & Age class ∈ {15-29, 30-49, 65-74}	1
9	4	0.016	0	Corine reclass ∈ {11, 12, 13, 14, 21, 23, 31, 32, 41, 51} & Place ∈ {Building, Campsite} & Distance > 446m & Age class ∈ {15-29, 50-64}	1
10	5	0.016	0	Corine reclass ∈ {11, 13, 14, 22, 41, 51} & Place ∈ {Outdoor, River} & Density > 0.012 & Age class ∈ {≤14, 30-49, ≥75} & ReturnPeriod ∈ {< 2, 2-5, 20-50, 50-100, 100-200, >200}	1
11	2	0.01	0	Age class ∈ {≤14,15-29, 50-64, 65-74} & ReturnPeriod ∈ {50-100}	1
12	2	0.01	0	Density ≤ 0.004 & Age class ∈ {50-64}	1
13	5	0.01	0	Corine reclass ∈ {12, 23, 24, 31, 51} & Place ∈ {Building, Outdoor, River} & Distance ≤72 & Age class ∈ {15-29, 30-49, 50-64, ≥75} & ReturnPeriod ∈ {5, 2-5}	1
14	2	0.01	0	Corine reclass ∈ {12, 22, 31} & Age class ∈ {65-74}	1
15	4	0.037	0.143	Corine reclass ∈ {14, 24, 33} & 2m <Distance ≤ 193m & Age class ∈ {≤14,15-29, 65-74}	0
16	2	0.031	0	Solid ∈ {No} & Age class ∈ {≤14, 15-29, 30-49, 65-74, ≥75}	0

*The parameters of ReturnPeriod and the Age class are represented in years.

Table 10- The rules extracted from the Random forest model

N	length	frequency	error	condition	pred
17	3	0.029	0	HazardScenario ∈ {L, M, outside} & Corine reclass ∈ {13, 14, 22, 33, 41} & Solid ∈ {No, Yes}	0
18	4	0.029	0	Corine reclass ∈ {13, 14, 22, 33, 51} & Place ∈ {Building, Outdoor, River} & Distance ≤146m & Solid ∈ {NA}	0
19	4	0.024	0	Place ∈ {River} & Distance > 7m & Solid ∈ {No, Yes} & ReturnPeriod ∈ {5, 10-20, 20-50, 50-100, 100-200, >200}	0
20	5	0.024	0	HazardScenario ∈ {L, outside} & Corine reclass ∈ {21, 24, 32} & MorphologicalZone ∈ {Mountain} & Solid ∈ {No, Yes} & Age class ∈ {30-49, ≥75}	0
21	3	0.021	0	HazardScenario ∈ {H, outside} & Place ∈ {Bridge, Building, Campsite} & MorphologicalZone ∈ {Plain}	0
22	3	0.021	0.125	Corine reclass ∈ {21, 23, 32, 33} & Distance > 83m & Density ≤ 0.01	0
23	1	0.021	0.125	Corine reclass ∈ {31}	0
24	3	0.018	0	Corine reclass ∈ {11, 12, 13, 14, 22, 41} & Density ≤ 0.014 & Age class ∈ { 15-29, 65-74, ≥75}	0
25	4	0.018	0	HazardScenario ∈ {H, outside} & MorphologicalZone ∈ {Mountain}& Age class ∈ {15-29, 30-49, 65-74, ≥75} & ReturnPeriod ∈ {50-100, >200}	0
26	2	0.018	0	MorphologicalZone ∈ {Plain} & ReturnPeriod ∈ {< 2, 5, 10-20, 50-100}	0
27	4	0.016	0	Corine reclass ∈ {24, 32, 33} & Distance ≤ 47m & Density ≤ 0.022 & Age class ∈ {≤14, 30-49, 50-64, 65-74, ≥75}	0
28	4	0.016	0.167	2m < Distance ≤ 67m & Solid ∈ {No, Yes} & Age class ∈ {15-29, 30-49, 65-74, ≥75}	0
29	3	0.013	0	HazardScenario ∈ {L, outside} & Corine reclass ∈ {21} & Distance > 402m	0
30	4	0.013	0	HazardScenario ∈ {H, M} & Corine reclass ∈ {11, 12, 13, 41, 51} & Distance > 40m & Age class ∈ {15-29, 30-49}	0
31	3	0.013	0	HazardScenario ∈ {H} & 25m < Distance≤ 43m	0
32	3	0.01	0	HazardScenario ∈ {H, L} & Corine reclass ∈ {12, 13} & Distance >2m	0
33	4	0.01	0	Corine reclass ∈ 11, 13, 14, 21, 23, 32, 33, 41} & Distance > 1m & Age class ∈ {50-64} & ReturnPeriod ∈ {< 2, 2-5 , 20-50, 50-100, 100-200}	0
34	1	0.357	0.199	Else	0

*The parameters of ReturnPeriod and the Age class are represented in years.

Table 10- The rules extracted from the Random forest model

N	Label of the model run	training/test	K	1st Validation								2nd Validation	
				Acc (accuracy)	TN	TP	FN	FP	FN-ratio	TPR (%)	TNR (%)	TNR_{2ndval}	FPR_{2ndval}
1	run 1.1.1	0.7	1	77.9	27	33	9	8	0.214	78.6	77.1	72.1	27.9
2	run 1.1.2	0.8	1	78.4	17	23	5	6	0.179	82.1	73.9	73.1	26.9
3	run 1.2.1	0.7	2	86.1	69	30	9	7	0.231	76.9	90.8	91.5	8.5
4	run 1.2.2	0.8	2	88.3	47	21	6	3	0.222	77.8	94.0	93.7	6.3
5	run 1.3.1	0.7	3	86.3	103	29	14	7	0.326	67.4	93.6	95.7	4.3
6	run 1.3.2	0.8	3	88.2	72	18	7	5	0.280	72.0	93.5	95.5	4.5
7	run 1.4.1	0.7	8	94.2	303	20	17	3	0.460	54.1	99.0	99.2	0.8
8	run 1.4.2	0.8	8	93.9	200	15	12	2	0.444	55.6	99.0	99.2	0.8
9	run 2.1.1	0.7	1	76.6	27	32	10	8	0.238	76.2	77.1	69.1	30.9
10	run 2.1.2	0.8	1	78.4	18	22	6	5	0.214	78.6	78.3	72.0	28.0
11	run 2.2.1	0.7	2	84.3	69	28	11	7	0.282	71.8	90.8	91.4	8.6
12	run 2.2.2	0.8	2	89.6	47	22	5	3	0.185	81.5	94.0	93.2	6.8
13	run 2.3.1	0.7	3	85.6	102	29	14	8	0.326	67.4	92.7	95.2	4.8
14	run 2.3.2	0.8	3	89.2	72	19	6	5	0.240	76.0	93.5	94.8	5.2
15	run 2.4.1	0.7	8	95.0	304	22	15	2	0.405	59.5	99.3	98.8	1.2
16	run 2.4.2	0.8	8	94.3	199	17	10	3	0.370	63.0	98.5	98.8	1.2
17	run 3.1.1	0.7	1	77.9	27	33	9	8	0.214	78.6	77.1	69.5	30.5
18	run 3.1.2	0.8	1	80.4	18	23	5	5	0.179	82.1	78.3	71.7	28.3
19	run 3.2.1	0.7	2	83.5	68	28	11	8	0.282	71.8	89.5	90.5	9.5
20	run 3.2.2	0.8	2	87.0	46	21	6	4	0.222	77.8	92.0	92.7	7.3
21	run 3.3.1	0.7	3	85.0	101	29	14	9	0.326	67.4	91.8	96.0	4.0
22	run 3.3.2	0.8	3	88.2	71	19	6	6	0.240	76.0	92.2	94.8	5.2
23	run 3.4.1	0.7	8	95.0	304	22	15	2	0.405	59.5	99.3	98.4	1.6
24	run 3.4.2	0.8	8	95.2	201	17	10	1	0.370	63.0	99.5	98.8	1.2
25	run 4.1.1	0.7	1	77.9	27	33	9	8	0.214	78.6	77.1	69.3	30.7
26	run 4.1.2	0.8	1	78.4	17	23	5	6	0.179	82.1	73.9	72.1	27.9
27	run 4.2.1	0.7	2	84.3	69	28	11	7	0.282	71.8	90.8	91.1	8.9
28	run 4.2.2	0.8	2	87.0	46	21	6	4	0.222	77.8	92.0	92.8	7.2
29	run 4.3.1	0.7	3	85.0	102	28	15	8	0.349	65.1	92.7	95.8	4.2
30	run 4.3.2	0.8	3	90.2	73	19	6	4	0.240	76.0	94.8	95.3	4.7

Table 11- Performance measures of the runs for the 1st and 2nd validation

N	Label of the model run	training/test	K	1st Validation								2nd Validation	
				Acc (accuracy)	TN	TP	FN	FP	FN-ratio	TPR (%)	TNR (%)	TNR_{2ndval}	FPR_{2ndval}
31	run 4.4.1	0.7	8	94.5	303	21	16	3	0.432	56.8	99.0	99.2	0.8
32	run 4.4.2	0.8	8	94.3	200	16	11	2	0.407	59.3	99.0	99.2	0.8
33	run 5.1.1	0.7	1	79.2	27	34	8	8	0.191	81.0	77.1	71.5	28.5
34	run 5.1.2	0.8	1	76.5	16	23	5	7	0.179	82.1	69.6	73.7	26.3
35	run 5.2.1	0.7	2	85.2	69	29	10	7	0.256	74.4	90.8	91.5	8.5
36	run 5.2.2	0.8	2	87.0	46	21	6	4	0.222	77.8	92.0	93.4	6.6
37	run 5.3.1	0.7	3	85.0	102	28	15	8	0.349	65.1	92.7	96.0	4.0
38	run 5.3.2	0.8	3	88.2	71	19	6	6	0.240	76.0	92.2	95.6	4.4
39	run 5.4.1	0.7	8	94.2	303	20	17	3	0.460	54.1	99.0	99.2	0.8
40	run 5.4.2	0.8	8	94.3	201	15	12	1	0.444	55.6	99.5	99.2	0.8
41	run 6.1.1	0.7	1	81.8	30	33	9	5	0.214	78.6	85.7	73.7	26.3
42	run 6.1.2	0.8	1	82.3	19	23	5	4	0.179	82.1	82.6	75.5	24.5
43	run 6.2.1	0.7	2	86.1	69	30	9	7	0.231	76.9	90.8	90.3	9.7
44	run 6.2.2	0.8	2	87.0	46	21	6	4	0.222	77.8	92.0	92.9	7.1
45	run 6.3.1	0.7	3	85.0	101	29	14	9	0.326	67.4	91.8	94.8	5.2
46	run 6.3.2	0.8	3	90.2	72	20	5	5	0.200	80.0	93.5	94.6	5.4
47	run 6.4.1	0.7	8	93.6	300	21	16	6	0.432	56.8	98.0	99.2	0.8
48	run 6.4.2	0.8	8	93.4	199	15	12	3	0.444	55.6	98.5	99.2	0.8
49	run 7.1.1	0.7	1	77.9	25	35	7	10	0.167	83.3	71.4	69.9	30.1
50	run 7.1.2	0.8	1	76.5	16	23	5	7	0.179	82.1	69.6	73.8	26.2
51	run 7.2.1	0.7	2	87.0	71	29	10	5	0.256	74.4	93.4	90.4	9.6
52	run 7.2.2	0.8	2	89.6	48	21	6	2	0.222	77.8	96.0	92.5	7.5
53	run 7.3.1	0.7	3	84.3	100	29	14	10	0.326	67.4	90.9	94.3	5.7
54	run 7.3.2	0.8	3	90.2	74	18	7	3	0.280	72.0	96.1	94.4	5.6
55	run 7.4.1	0.7	8	94.5	303	21	16	3	0.432	56.8	99.0	99.2	0.8
56	run 7.4.2	0.8	8	95.2	201	17	10	1	0.370	63.0	99.5	99.2	0.8
57	run 8.1.1	0.7	1	75.3	26	32	10	9	0.238	76.2	74.3	70.7	29.3
58	run 8.1.2	0.8	1	74.5	16	22	6	7	0.214	78.6	69.6	72.7	27.3
59	run 8.2.1	0.7	2	87.8	72	29	10	4	0.256	74.4	94.7	90.1	9.9
60	run 8.2.2	0.8	2	90.9	49	21	6	1	0.222	77.8	98.0	91.5	8.5

Table 11- Performance measures of the runs for the 1st and 2nd validation

N	Label of the model run	training/test	K	1st Validation								2nd Validation	
				Acc (accuracy)	TN	TP	FN	FP	FN-ratio	TPR (%)	TNR (%)	TNR_{2ndval}	FPR_{2ndval}
61	run 8.3.1	0.7	3	85.0	102	28	15	8	0.349	65.1	92.7	95.3	4.7
62	run 8.3.2	0.8	3	88.2	72	18	7	5	0.280	72.0	93.5	95.3	4.7
63	run 8.4.1	0.7	8	94.2	302	21	16	4	0.432	56.8	98.7	98.8	1.2
64	run 8.4.2	0.8	8	94.8	201	16	11	1	0.407	59.3	99.5	98.8	1.2
65	run 9.1.1	0.7	1	72.7	25	31	11	10	0.262	73.8	71.4	67.5	32.5
66	run 9.1.2	0.8	1	68.6	14	21	7	9	0.250	75.0	60.9	68.4	31.6
67	run 9.2.1	0.7	2	81.7	69	25	14	7	0.359	64.1	90.8	89.5	10.5
68	run 9.2.2	0.8	2	87.0	49	18	9	1	0.333	66.7	98.0	91.3	8.7
69	run 9.3.1	0.7	3	83.0	100	27	16	10	0.372	62.8	90.9	94.2	5.8
70	run 9.3.2	0.8	3	85.3	70	17	8	7	0.320	68.0	90.9	93.6	6.4
71	run 9.4.1	0.7	8	93.9	301	21	16	5	0.432	56.8	98.4	98.4	1.6
72	run 9.4.2	0.8	8	94.8	202	15	12	0	0.444	55.6	100.0	98.8	1.2
73	run 10.1.1	0.7	1	75.3	28	30	12	7	0.286	71.4	80.0	72.7	27.3
74	run 10.1.2	0.8	1	80.4	18	23	5	5	0.179	82.1	78.3	72.9	27.1
75	run 10.2.1	0.7	2	81.7	67	27	12	9	0.308	69.2	88.2	91.1	8.9
76	run 10.2.2	0.8	2	85.7	46	20	7	4	0.259	74.1	92.0	92.1	7.9
77	run 10.3.1	0.7	3	85.6	105	26	17	5	0.395	60.5	95.5	96.6	3.4
78	run 10.3.2	0.8	3	92.2	75	19	6	2	0.240	76.0	97.4	97.1	2.9
79	run 10.4.1	0.7	8	94.2	304	19	18	2	0.487	51.4	99.3	99.2	0.8
80	run 10.4.2	0.8	8	93.0	201	12	15	1	0.556	44.4	99.5	98.8	1.2
81	run 11.1.1	0.7	1	71.4	26	29	13	9	0.310	69.0	74.3	71.7	28.3
82	run 11.1.2	0.8	1	78.4	18	22	6	5	0.214	78.6	78.3	75.4	24.6
83	run 11.2.1	0.7	2	83.5	70	26	13	6	0.333	66.7	92.1	91.7	8.3
84	run 11.2.2	0.8	2	90.9	49	21	6	1	0.222	77.8	98.0	93.5	6.5
85	run 11.3.1	0.7	3	85.0	104	26	17	6	0.395	60.5	94.5	95.4	4.6
86	run 11.3.2	0.8	3	83.3	69	16	9	8	0.360	64.0	89.6	95.7	4.3
87	run 11.4.1	0.7	8	93.6	302	19	18	4	0.487	51.4	98.7	99.6	0.4
88	run 11.4.2	0.8	8	93.0	199	14	13	3	0.482	51.9	98.5	99.6	0.4

Table 11- Performance measures of the runs for the 1st and 2nd validation

Chapter 7

Summary and conclusions

As one of the most common hazards affecting towns and cities worldwide, floods influence the lives of millions of people every year. The loss of life is one of the most serious types of flood effects and is considered as one of the priorities for the decision-makers before and during such events. Several studies have been conducted on the identification of different drivers of flood mortality; however, the current modeling capabilities in this area are limited. This study aims to create an empirical flood mortality model in the Italian context, developed using the random forest algorithm.

This study is based upon an initial dataset of flood mortalities in the Po river district in 1970-2019 developed and managed by the Italian National Council of Research (CNR-IRPI) of Perugia. The dataset consists of information on the age and gender of the victims, the place of the accident, the circumstance of death, and the information on the location where the flood mortalities occurred. After reviewing the literature, the significant contributing factors (drivers) to estimate flood mortality were investigated, and the information obtained was compared to the patterns observed in the dataset.

Moreover, two other actions were required to complement the information described in the dataset for each victim. Firstly, the cleansing of available data to make them usable by the random forest algorithm; secondly, the definition of new variables related to those factors for which no information is included in the original dataset. By analyzing spatial data on the events in the dataset relating to the areas where the deaths occurred, it was possible to define five new explanatory variables. The parameter of “Morphological zone” characterizes each victim location with a plain or mountainous area. This parameter can act as a proxy to describe the slope, affecting the flow of water and the stability of humans. In addition, the information regarding the land use in the location where the mortality happened is reported in the “Corine land cover code” parameter, which affects the hydrological processes in the area, such as the runoff characteristics. This parameter also describes the urbanization characteristics of the area that the death occurred. Other explanatory variables for flood fatality defined for the records in the dataset are the “Distance” of the victims from the river, the “Density of the buildings (municipality)”, and the “hazard scenario code” obtained from the maps created within the Flood Risk Management Plan of the Po river district. In addition, it was possible to identify two descriptive parameters of the flood hazard by investigating the gray literature on the flood events associated with the dataset as well as the hydrological data: (1) the “Solid transport” parameter and the “Return period of the maximum 24-hr precipitation”. In the end, a dataset of 127 flood mortality records is obtained in which ten explanatory variables characterize each record.

In the next step, in order to allow the random forest algorithm to correctly identify the role and the importance of the different explanatory variables, a synthetic dataset consisting of records of the individuals who were involved in the event but did not lose their lives was created, using the

statistical distributions of the explanatory variables in the flooded area. The dataset of flood mortalities (127 records of “Fatalities”) coupled with this synthetic dataset (1270 records of “Non-fatalities”) was then used as the inputs of the random forest algorithm. Part of the data is used as the “training dataset“ to calibrate the model, and then, the validation of the random forest algorithm on two different datasets (data that was not used for the creation of the model) led to the choice of the final model setup. Furthermore, the analysis made it possible to identify the importance of the different explanatory variables, with the parameter of “Distance from the river” being the most significant and the “Gender” the least significant variables in the estimation of flood mortality.

This final model setup is characterized by the classification accuracy of 89%, while only using nine of the explanatory input variables for the model, excluding the variable of “Gender”, which was already mentioned as the least important variable in the outcome of the model. Finally, to simplify the interpretation of the model, an attempt was made to extract some rules which describe the conditions by which mortality is more likely to occur.

When working on empirical models, the accuracy and degree of confidence in the data used to create the model significantly influence the accuracy of the results produced; however, the availability of accurate information has always been a challenge. In this study, after performing the final setup of the model, it was concluded that the parameter of “Distance from the river” is the most significant input variable for the classification of flood mortality, but this parameter in the dataset used for the creation of the model has some degrees of uncertainty. Therefore, since its importance is proved in this study, it is suggested to repeat the process, trying to improve the confidence in this parameter.

An approximation implemented in this study relates to the synthetic dataset of “Non-fatalities”, in which to characterize the “Age” and the “Gender” of the individuals involved in the flood events, the census data of the year 2011 was used. However, some of the events date back to the 1970s, and the use of the information on the population from 2011 might not well represent that population. A similar approximation is applied to define some of the variables in the datasets of “Fatalities” and “Non-fatalities”. The parameters such as the “Corine land cover code”, and the “Density of the buildings” are also expected to have changed a lot from 1970 through 2019; however, for this study, the Corine land cover is derived from the database of the year 2018, while the Building densities are obtained referring to the cadastral maps of the present day. Thus, it is suggested to implement the information on these parameters produced in the past to better represent the actual conditions of those mortality events.

Moreover, the dataset of “Non-fatalities” created in this study proved necessary in the training of the model to identify the role of the explanatory variables. However, this dataset was created to make up for the lack of an actual database on the individuals involved in the previous flood events who managed to survive. Thus, it would be advantageous to carry out studies and surveys to characterize the individuals who were involved in the events, although due to the large period of the flood events in the dataset, the events might have to be limited to more recent ones. This way, the calculations, and the modeling can be repeated to confirm the model's performance on this new dataset.

In this study, due to the lack of data on the time of the accident, this parameter was removed from the dataset at the beginning of the process. However, this information can act significantly in estimating the fatalities. Other than representing the visibility conditions at the time of the event, this parameter also takes into account the probability of the individuals being exposed to the flood. For example, assuming that a flood event occurs at night when most people are asleep at home, the probability of them being exposed to the flood decreases. Therefore, attempting to implement this parameter in the modeling can be beneficial for a better estimation of the fatalities.

In the event of a flood, the preparedness of the population and the authorities can be crucial for an effective response. Early warnings issued by the authorities can result in the activation of warning and emergency procedures, such as road closures, rescues, and evacuations [7]. Implementing the information on the alerts issued for the events in the dataset can result in an understanding of the importance of these early warnings in the prevention of flood mortalities. Thus, it is suggested to investigate the information on these alerts in future studies for a better estimation of the loss of life due to floods.

Lastly, one of the explanatory variables that was initially decided to be implemented in the model but was then excluded due to the low confidence in the accuracy of the parameter is the “Slope” of the area in which the mortality occurred. Thus, as a proxy, the parameter of “Morphological Zone” was used. One suggestion for future studies is to attempt to use another parameter that might be one step closer to the representation of the slope, which is the slope of the river involved in the flood event. This parameter affects the flow regime and the flow of the sediments carried by the flood and is thus advantageous for modeling flood mortality.

The model developed identifies the most important factors that cause flood mortalities. By interpreting the rules, it is possible to identify the most critical combination of factors that are likely to result in fatalities in the case of an event with similar characteristics identified by the rules. Therefore, these conditions can be considered in the outline of the emergency management plans in order for the authorities to have an effective response when facing flood events. Moreover, it is possible to apply this model to an existing area to estimate the likelihood of the occurrence of fatalities. The results of such analysis can then be turned into a map of the probability of loss of life in a flood scenario. However, since the model works with single records of individuals, it might be a challenge to obtain the information for the model's explanatory variables in such a level of detail.

In addition, due to the dependence of the created model on local data, it might not be easily transferable to other contexts. The characteristics of each hydrographic catchment can be different from the other. However, the parameters proved to be significant in the development of this model can be re-calibrated and used as a first step in developing an estimation tool fitted to a new context.

Bibliography

[1] [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)], "IPCC, 2014: Climate Change 2014: Synthesis Report," Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, IPCC, Geneva, Switzerland, 2014.

[2] [Field, C.B., V.R. Barros, D.J. Dokken, K.J. Mach, M.D. Mastrandrea,T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken,P.R. Mastrandrea, and L.L.White (eds.)], "(IPCC, 2014: Summary for policymakers. In: Climate Change 2014: Impacts,Adaptation, and Vulnerability.Part A: Global and Sectoral Aspects," Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2014.

[3] S. N. JONKMAN, "Global Perspectives on Loss of Human Life," *Natural Hazards,* vol. 34, pp. 151-175, 2005.

[4] "Aon (2019) Global Catastrophe Recap: First Half of 2019.".

[5] "WHO (World Health Organization – Regional office for Europe): 2002, Floods: Climate Change and Adaptation Strategies for Human Health".

[6] Mattia Amadio1, Anna Rita Scorzini2, Francesca Carisi3, Arthur H. Essenfelder1, Alessio Domeneghetti3, Jaroslav Mysiak1, and Attilio Castellarin3, "Testing empirical and synthetic flood damage models: the case of Italy," *Natural Hazards and Earth System Sciences,* 2019.

[7] O. Petrucci, "Factors leading to the occurrence of flood fatalities: a systematic review of research papers published between 2010 and 2020," *Natural Hazards and Earth System Sciences,* 2021.

[8] EDMUND PENNING-ROWSELL, PETER FLOYD, DAVID RAMSBOTTOM and SURESH SURENDRAN, "Estimating Injury and Loss of Life in Floods: A Deterministic Framework," *Natural Hazards,* 2005.

[9] Paola Salvatia, Olga Petrucci, Mauro Rossi, Cinzia Bianchi, Aurora A.Pasqua, FaustoGuzzetti, "Gender, age and circumstances analysis of flood and landslide fatalities in Italy," *Science of The Total Environment,* vol. 610–611, pp. 867-879, 2018.

[10] "POLARIS- Popolazione a Rischio da Frana e da Inondazione in Italia," The Research Institute for Hydrogeological Protection (IRPI) of the National Research Council (CNR), [Online]. Available: https://polaris.irpi.cnr.it/. [Accessed 16 10 2021].

[11] "World Disasters Report," International Federation of Red Cross and Red Crescent Societies, Geneva, 2020.

[12] Beniamino Russo, Manuel Gomez Valentin, F. Macchione, "Pedestrian hazard criteria for flooded urban areas," *Natural Hazards,* vol. 69, p. 251–265, 2013.

[13] S. N. Jonkman, J. K. Vrijling, A. C. W. M. Vrouwenvelder, "Methods for the estimation of loss of life due to floods: a literature review and a proposal for a new method," *Natural Hazards,* vol. 46, p. 353–389, 2008.

[14] Abt SR, Wittler RJ, Taylor A, Love DJ, "Human stability in a high flood hazard zone," *Water Resources Bulletin,* vol. 25(4), p. 881–890, 1989.

[15] Maruša Špitalar, J. Gourley, C. Lutoff, P. Kirstetter, M. Brilly, Nicholas Carr , "Analysis of flash flood parameters and human impacts in the US from 2006 to 2012," *Journal of Hydrology,* vol. 519, pp. 863-870, 2014.

[16] Hatim O. Sharif, Terrance L. Jackson, Md. Moazzem Hossain, and David Zane, "Analysis of Flood Fatalities in Texas," *Natural Hazards Review,* vol. 16, no. 1, 2015.

[17] Galateia Terti, Isabelle Ruin, Sandrine Anquetin, and Jonathan J. Gourley, "A Situation-Based Analysis of Flash Flood Fatalities in the United States," *Bulletin of the American Meteorological Society,* vol. 98, no. 2, pp. 333-345, 2017.

[18] Katharine Haynes, Lucinda Coates, Felipe Dimer de Oliveira, Andrew Gissing, "An analysis of human fatalities from floods in Australia 1900-2015," Bushfire and Natural Hazards Cooperative Research Centre, 2016.

[19] Sebastiaan N. Jonkman, Ilan Kelman , "An analysis of the causes and circumstances of flood disaster deaths," *Disasters,* vol. 29, no. 1, pp. 75-97, 2005.

[20] Gerry FitzGerald, Weiwei Du, Aziz Jamal, Michele Clark, Xiang-Yu Hou, "Flood fatalities in contemporary Australia (1997-2008)," *Emergency medicine Australasia,* vol. 22, no. 2, pp. 180-6, 2010 Apr;.

[21] D.M.M. Kellar,T.W. Schmidlin, "Vehicle-related flood deaths in the United States, 1995–2005," *Journal of Flood Risk Management,* vol. 5, pp. 153 - 163, 2012.

[22] Ezra Boyd, Marc Levitan, Ivor van Heerden, "Further specification of the dose-response relationship for flood fatality estimation," in *Paper presented at the US-Bangladesh workshop on innovation in windstorm/storm surge mitigation construction. National Science Foundation and Ministry of Disaster & Relief, Government of Bangladesh. Dhaka*, 2005.

[23] Galateia Terti, Isabelle Ruin, Jonathan J. Gourley, Pierre Kirstetter, Zachary Flamig, Juliette Blanchet, Ami Arthur, and Sandrine Anquetin, "Toward Probabilistic Prediction of Flash Flood Human Impacts," *Risk Analysis,* vol. 39, no. 1, 2019.

[24] Paola Salvati, Umberto Pernice, Cinzia Bianchi, Ivan Marchesini, Federica Fiorucci, and Fausto Guzzetti, "Communication strategies to address geohydrological risks: the POLARIS web initiative in Italy," *Nat. Hazards Earth Syst. Sci.,* vol. 16, p. 1487–1497, 2016.

[25] "Po River District Authorty," [Online]. Available: https://www.adbpo.gov.it/. [Accessed 2021].

[26] "Corine Land Cover 2018," Copernicus, [Online]. Available: https://land.copernicus.eu/pan-european/corine-land-cover/clc-2012. [Accessed 2021].

[27] "ISTAT (Istituto Nazionale di Statistica)," [Online]. Available: https://www.istat.it/it/archivio/104317. [Accessed 2021].

[28] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[29] Zhaoli Wang, Chengguang Lai, Xiaohong Chen, Bing Yang, Shiwei Zhao, Xiaoyan Bai, "Flood hazard risk assessment model based on random forest," *Journal of Hydrology,* vol. 527, pp. 1130-1141, 2015.

[30] Dennis Wagenaar, Jurjen de Jong, and Laurens M. Bouwer, "Multi-variable flood damage modelling with limited data using supervised learning approaches," *Nat. Hazards Earth Syst. Sci.,* vol. 17, p. 1683–1696, 2017.

[31] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, 2013.

[32] "RStudio," [Online]. Available: https://www.rstudio.com/. [Accessed 2021].

[33] Leo Breiman and Adele Cutler, Andy Liaw and Matthew Wiener., "CRAN 'randomForest' package," [Online]. Available: https://cran.r-project.org/web/packages/randomForest/randomForest.pdf. [Accessed 2021].

[34] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, Nik Nik Abdullah, "An application of oversampling, undersampling, Bagging and Boosting in Handling Imbalanced Datasets," in *International Conference on Data Engineering*, 2013.

[35] Houtao Deng, Xin Guan, Vadim Khotilovich, "inTrees: Interpret Tree Ensembles," [Online]. Available: https://cran.r-project.org/web/packages/inTrees/inTrees.pdf. [Accessed 2021].

[36] H. Deng, "Interpreting tree ensembles with inTrees," *International Journal of Data Science and Analytics,* vol. 7, pp. 277-287, 2019.

[37] "Risks from floods, storm surges and flash floods," 30 September 2021. [Online]. Available: https://www.munichre.com/en/risks.

[38] Flood risk management: A strategic approach, UNESCO, 2013.

[39] N.E. Klepeis, W.C. Nelson, W.R. Ott, J.P. Robinson, A.M. Tsang, P. Switzer, J.V. Behar, S.C. Hern, W.H. Engelmann, "The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants," *J Expo Anal Environ Epidemiol,* pp. 231-52, May-June 2011.