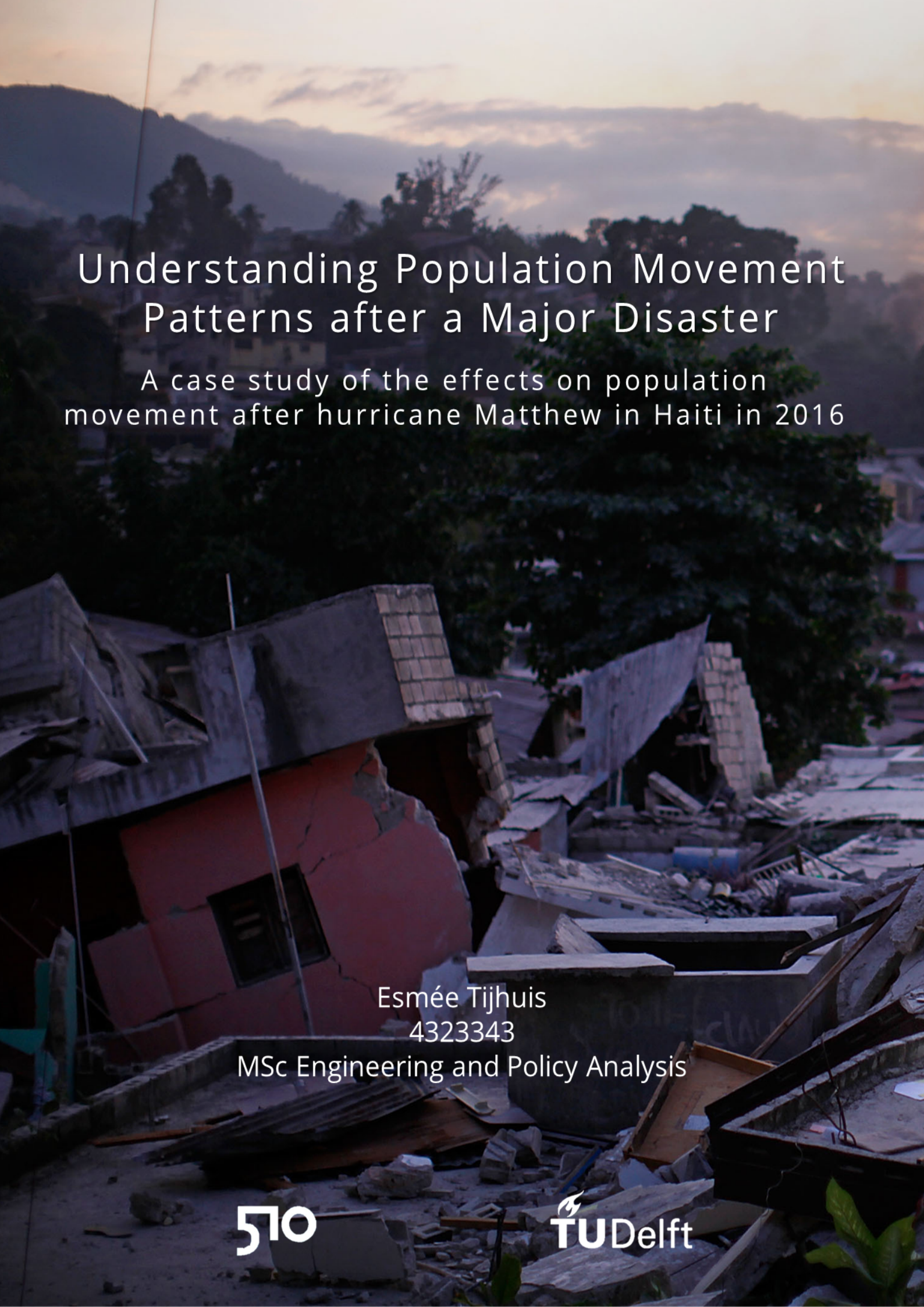


Understanding Population Movement Patterns after a Major Disaster

A case study of the effects on population movement after hurricane Matthew in Haiti in 2016

Esmée Tijhuis
4323343
MSc Engineering and Policy Analysis

510

TU Delft

This page was intentionally left blank

UNDERSTANDING POPULATION MOVEMENT PATTERNS AFTER A MAJOR DISASTER

A case study of the effects of Hurricane Matthew in Haiti in 2016

Thesis submitted in fulfilment of the requirements for the degree of Master of Science in Engineering Policy Analysis

by

Esmée Tijhuis

Student number: 4323343

July 30, 2021

To be defended in public on August 13th 2021

Graduation Committee:

Chairperson	: Prof.dr. Martijn Warnier,	Systems Engineering
First Supervisor	: Dr.ir. Trivik Verma,	Policy Analysis
Second supervisor	: Prof.dr. Martijn Warnier,	Systems Engineering
External Supervisor	: Dr. Marc van den Homberg,	Scientific Lead

FLOWMINDER.ORG

TU Delft Delft University of Technology

510 AN INITIATIVE OF THE NETHERLANDS RED CROSS

Preface

Dear reader,

With the dawn of the last period of my study, I felt an enormous drive to work on a research topic which is closely related to society. Inevitably, I was searching for additional incentives to draw motivation from my social environment in the pandemic. Those crucial aspects made 510 the perfect organization to work for during my thesis. I was able to combine some of my scientific interests into one research topic, which made it comfortable to work on it on an everyday basis. The endmost period of my study remained to be a challenging one but enriched me in numerous ways. Without a doubt, it was one of the most rewarding things I have done in life and cleared the way for new ambitions I set for my future life. Fortunately, I was surrounded with amazing people who guided me in differing ways to support me.

First of all, I would like to thank my graduation committee. Trivik, your input has been of incredible value for my thesis. Our weekly meetings were inspiring and fruitful, I have incorporated many of your suggestions, and you were always available for last-minute meetings. I have gladly made use of your incredible expertise in the field. Martijn, we have been meeting several times, but all those meetings made a profound difference on the way I looked at my thesis and the problems I faced. Our discussions were delightful, and your enthusiasm for my project made my enthusiasm grow too. Marc, I admire how involved you have always been in my project while working crazy hours and supervising a dozen other students. During every meeting, you were up to date, gave me great advise and showed me what dedication for your job is. All three of you, thanks for the inspiration and motivation!

Also, thanks to my family and friends for supporting me throughout the project. A special thanks to Roos, Boris and Can. Our endless study-sessions have kept me sharp and happy during this Covid-graduation. I am sure that you have made all the difference in the quality of this report. Also, thank you papa, mama, Benthe, Jade and Joël! I know that I have not been very involved in our lifes together, but you have always been supportive and present when I needed you. Thank you Evi, Esmee, Ysanne, Irene, and all others, for our study-work sessions together and letting me babble.

This report marks the end of my student career. I hope you will enjoy reading this thesis as much (or more) than I had writing it.

Esmée Simone Tijhuis

Amsterdam, July 30, 2021

Executive Summary

Natural disasters pose serious threats to people living in disaster prone areas. With the growing number of disasters, the number of refugees grows steadily along-side it. Fortunately, many humanitarian organisations work tirelessly to provide the aid that is needed in times of disasters, preferably when and where is it needed most. However, the agencies involved need information in order to execute these essential tasks. Current methods of collecting this information on, for example, the needed food and supplies, rescue missions and warning signals, incorporate surveys or headcounts carried out during the emergency relief phase. Sharing this information among humanitarian agencies and governmental organisations is often time consuming and inefficient. The agencies, although experienced, work with the tools that are available, such as data standardisation on online platforms that are sometimes build. A major problem however, is that the agencies are often not aware of where refugees are and where they are going. In short: what the mobility behaviour of this highly vulnerable group is.

The importance of understanding human mobility behaviour has been recognised by many researchers, resulting in numerous studies that have provided insight in the matter. Whilst patterns in human mobility behaviour have been recognised, even during a major disaster, the predictability of these patterns has received insufficient attention. A major break-through in the predictability of population movement patterns came with the application of mobile location data of large numbers of users.

The advantage of using mobile location data seems clear, the data is already collected by Mobile Network Operators and the information it contains could potentially reveal patterns that could not have been uncovered using other methods. A major problem here is that the data contains sensitive information that violates the privacy of mobile phone users. As a solution, the mobile location data is aggregated over the time- and space-dimension. This results in the loss of details that could be valuable and make us question the suitability of using the data for prediction of population movements. This poses a knowledge gap, based on which this study is proposed. The objective of the study states: *"Gaining more insight in the predictability of mobility patterns during the disaster recovery phase in order to advice humanitarian agencies on where to focus their support almost real-time"*.

During the study, the answer to the following main research question is pursued:

How do populations move during a major disaster and can the destinations of the moving population be determined using data that is available to humanitarian aid organisations?

A case is introduced as illustration of the possibility of pattern recognition: the aftermath of Hurricane Matthew in Haiti in 2016. During this devastating disaster, an estimated number of 2.1 million people were affected. Many of these people became refugees right after the hurricane hit land and tried to find a safer place to stay. The mobility patterns of these refugees are tracked by their mobile location in Call Detail Records

(CDR). The data is expected to allow for researching the predictability of the population movements, which will essentially function as valuable information for humanitarian agencies on where to best focus their money and efforts.

It is clear that, during the aftermath of the disaster in Haiti, political tension and a lack of data management led to miscommunication and confusion among the involved agencies. As a result, especially the affected population from rural areas that had no homes to return to did not receive the assistance they needed. And furthermore, could efficient data collection have prevented the loss of the already limited amount of money and efforts by humanitarian aid organisations.

Data provided by the organisation Flowminder allows for a different approach towards the investigation of the population's movements. It includes also those individuals that did not end up in shelters or refugee camps and could therefore provide more insight in the extend to which the consequences reach. For the purpose of protecting the privacy of the users, the aggregated data retrieved from the original CDR's has been used.

As some of the identified drivers behind the mobility behaviour of individuals were not present in the dataset, the geo-locations have been estimated. This made it possible to add missing drivers, but it also showed that the spatial aggregation of the data could at least to some extend be undone. Using the obtained geo-locations of the displacements within the dataset, a predictive model is build. The model incorporates a k-Nearest Neighbour classifier and showed a maximum accuracy score of 67%. An important reason for the underperformance of the model is the clustering of values that has taken place in the pre-processing of the data.

Overall, some important lessons are learned about the predictability of population movements in a situation where data is scarcely available. The aggregation of the displacement data shows that there is a trade-off between the protection of sensitive user information and the potential added value that research around CDR data provides. As a result, some recommendations regarding the data handling were formulated. The most important recommendation is that the involved parties should try to work towards stronger mutual trust, so that less aggregated CDR data could be used in future research. Before this is established, the data preparation practices could be revised, so that the data becomes more useful while still protecting the users' privacy. It is advised that the time-frames that are separated are resampled so that they are of the same length. And that the last filter used to identify Internally Displaced Persons within the larger CDR dataset is not applied, as this filter excludes the displacements of people that had no working phone connection during the first week. This subset of displaced persons is expected to be highly vulnerable and the exclusion means that their movements are not represented in the analysis.

Contents

List of Figures

List of Tables

Acronyms

CDR Call Detail Record

DDM Dynamic Disaster Model

DTM Displacement Tracking Matrix

EPA Engineering and Policy Analysis

EWEA Early Warning Early Action

HDX Humanitarian Data Exchange

IARP Innovative Approaches in Response Preparedness

IDP Internally Displaced Person

IFRC The International Federation of Red Cross and Red Crescent Societies

IOM International Organisation for Migration

KNN k-Nearest Neighbour

MNO Mobile Network Operator

NB Naive Bayes

NGO Non-Governmental Organisation

PMT Protection Motivation Theory

SVM Support Vector Machine

UN United Nations

WHO World Health Organisation

Introduction

1

"All across the world, in every kind of environment and region known to man, increasingly dangerous weather patterns and devastating storms are abruptly putting an end to the long-running debate over whether or not climate change is real. Not only is it real, it's here, and its effects are giving rise to a frighteningly new global phenomenon: the man-made natural disaster."

Barack Obama - 2006

Many people are affected by natural hazards and disasters worldwide, and the numbers are growing every year. The UN Refugee Agency estimates the number of forcibly displaced persons to have grown to almost 80 million in 2020 (Walles, 2020), an amount higher than ever measured. The part of the internally displaced refugees is around 57%, meaning that more than half of the refugees stay in their own country. These so-called Internally Displaced Person (IDP) 's are considered highly vulnerable, as the camps that they often end up staying in do not meet the needed health standards (WHO, 2012). In some cases the refugees flee for the government that is supposed to protect them and in other cases the government does not have the means to help refugees within their country. In addition to the poor conditions in these camps, they are usually located in areas that are difficult to reach, making it challenging to deliver humanitarian assistance (UNHCR, 2020). Also, the exact number of people staying in such camps is difficult to estimate due to the high rate of in- and outflow of refugees and limited methods to count the number of people living in the camps effectively.

Fortunately, many organisations are concerned with the health and well-being of refugees and are committed to the improvement of disaster response and accommodation of this vulnerable group. One of these organisations is The The International Federation of Red Cross and Red Crescent Societies (IFRC). Not only are IFRC volunteers often among the first to provide emergency relief after a disaster, the foundation also focuses on the reduction of suffering by engaging before a disaster strikes (Red Cross, 2008). The IFRC has found that the impact of disasters can effectively be mitigated during the phase right before the blow. This phase is called the Early Warning Early Action (EWEA) period. Examples of these activities include the warning of farmers to harvest their crops or help strengthen houses that are located in risk areas (Red Cross, 2020).

In order to improve the speed and (cost-)effectiveness of humanitarian aid, an IFRC initiative called 510 was initiated (510.global, 2021b). The goal of 510 is to effectively use data to better understand opportunities that emerge around humanitarian aid. As a result of this, the organisation aims to better prepare for or cope with disasters and crises (510.global, 2021a). 510 is involved in four phases around a disaster, including the EWEA phase. Before the EWEA period, 510 engages in disaster preparedness which includes digital risk assessment (510.global, 2021b). Moreover, following the disaster, disaster response and recovery phases are applied by volunteers of the IFRC. Figure 1.1 shows a visual representation of the four phases in a time-frame with their complementary approach.

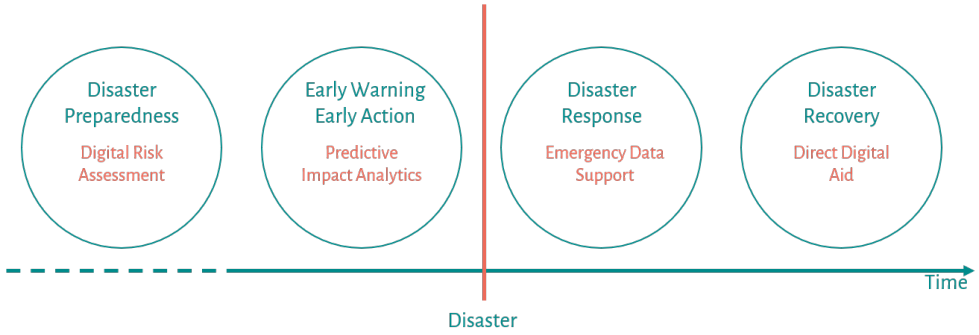


Figure 1.1: The 510.global disaster phases

Currently, emergency relief activities for refugees are emphasised during the disaster response period, partly due to the lack of information on the Spatio-temporal movement of the affected population. Methods to gain some insight into the number of people staying in known displacement sites include headcounts and surveys. Collecting this data is time-consuming, while a quick response is essential during a (natural) disaster. Also, these methods leave out the refugees staying in other places, such as with family, in abandoned buildings or even in caves (Mooney & Yemen, 2012). Therefore, these people are often not provided with the aid they need after fleeing their homes. Improving the information communication on refugees' location could potentially lead to far more fitting humanitarian aid than is the case now. Essential requirements would be to gain insight into how many people leave their homes before, during and after a disaster. What locations they move to and how they decide on their destination. In other words, the mobility behaviour of the population. This way, humanitarian agencies will have the opportunity to prepare camps with supplies on locations that IDP 's are likely to arrive.

These methods of surveys and headcounts do however not satisfy the extensive data collection that is needed in order to recognise patterns in the IDP 's mobility behaviour. Not only is the data collection often chaotic and of lower priority during a disaster (imagine having to fill in a survey while being in major distress), this data also is mainly essential to analyse emergency relief measures. Real-time data collection would therefore be of great value if it could be analysed in real-time too. Preferably even without the input of time and effort from emergency workers, so that their primary focus can be on the people who need them most.

1.1 Problem definition

Humanitarian organisations cooperate with (local) governments to distribute their collected data as efficiently as possible, given the disaster circumstances. Understanding the current situation helps describe the issues that data collection during a disaster brings and the solution required to tackle these issues. It is essential that human behaviour during disaster situations is analysed, clarifying the sort of information that the data should

at the least contain to represent the IDP's movements. It is then possible to compare the current data collection with potentially more accurate data available during a disaster.

1.1.1 Current methods of data collection during a disaster

Collecting accurate data during a disaster is essential to ensure effective disaster response (Morton Hamer, 2011; Kubo et al., 2019). However, the chaotic nature of the situation during a disaster does not allow for extensive data collection. In addition, there might be political and economic challenges that prevent the collection of accurate data (Morton Hamer, 2011). Then there is the endless need for information, the most critical factors of which have been identified by The Sphere Project (UNHCR, 2011) to be mortality rate, infrastructure assessment, medical supply necessities, food and shelter needs. All of these factors are related to or retrievable from the movements of the population. Also, the sharing of information among humanitarian agencies, government organisations and financial support agencies requires timely communication, which is often tricky during emergency activities. The World Health Organisation (WHO) too stresses the importance of improvement of the current data-sharing tools on health data. The data collection tools used by humanitarian organisations mainly focus on the medical needs of refugees in camps and the mapping of facilities and potential locations for new shelters where food and medical assistance can be provided.

In some cases, the efficiency of data sharing is upgraded significantly by the efforts of Universities or Research Groups. During the earthquake in Haiti in 2010, Universities Harvard and Boston joined forces into building a data portal that incorporated all disaster-related data to facilitate data sharing (Harvard University, 2010). Simple forms ensured the standardisation of the data storage and the possibility of fast data sharing among different agencies. Nevertheless, the humanitarian volunteers had no information on how many new refugees were expected to show up or leave the camps.

Although human mobility plays a critical role in disaster response and emergency relief, not much real-time data is collected that contains patterns of a population's movements during a disaster. Understanding human movement patterns is vital for three main reasons as described by Wang and Taylor (2016). Firstly, it determines the effectiveness of evacuation activities. Overcrowding, crushing and accidents are likely to happen during an evacuation (Pan, Han, Dauber, & Law, 2007), while a deeper understanding of the fleeing behaviour of people helps to prevent these kinds of unnecessary injuries and even deaths. Secondly, population movements effectively impact the efficiency of information communication. Often during disasters, communication networks are damaged, making the information infrastructure dependent on peer-to-peer connections. The part of the population that displaces themselves increases the width and speed of information diffusion in this situation (Kleinberg, 2007). Critical information on dangers, injuries and evacuation routes often find their way in these communication routes (Alessandro, 2010). And thirdly, population movement prediction could potentially save lives since localising individuals that are still staying in dangerous zones might be guided to safety (Wang & Taylor, 2016). In addition, prediction of the movements of the population allows humanitarian agencies with information on where to set up camps and how many people to expect over the following days, weeks or months.

Even though the importance of understanding population movement behaviour is recognised in these studies, the application of the obtained insights on movement patterns during disasters is still minor (Wang & Taylor, 2016). There are no traces of any tool or platform on which humanitarian agencies can see forecasts of population movements during a disaster, nor are there predicting models available on these movements. Studies on the phenomenon of movements during a disaster and studies on mobility patterns in normal conditions help us understand the motivations behind people's decision to change their location.

1.1.2 Human mobility behaviour during disasters

Previous research has shown that patterns can be recognised regarding population movement after sudden-onset disasters, such as the earthquake in Haiti in 2010 (Lu, Bengtsson, & Holme, 2012), the earthquake in Nepal in 2015 and Hurricane Matthew in Haiti in 2016 (Li, Dejby, Albert, Bengtsson, & Lefebvre, 2019a), and a cyclone in Bangladesh in 2012 (Lu et al., 2016). The primary finding in these studies is that mobility patterns after sudden onset disasters are far better predictable than was expected (Lu et al., 2012; Li et al., 2019a). First of all, Li et al. (2019a) found that half of the number of IDP 's have returned to their homes after four to five weeks after a disaster. Secondly, these studies show that refugees are likely to go to locations that they have visited before the disaster (Li, Dejby, Albert, Bengtsson, & Lefebvre, 2019b; Lu et al., 2012, 2016). These are expected to be places where family or friends live where they can stay in safety. This observation also complies with the study of Mooney and Yemen (2012) where surveys were used to determine the destinations of relocating residents. Lastly, a thorough analysis of the earthquake in Nepal in 2015 has shown that mobile location (or Call Detail Record (CDR)) data is suitable for near real-time assessment of refugee destinations (Wilson et al., 2016). Assessment of the obtained data has shown the in- and outflow of different areas within Nepal after the earthquake in unprecedented detail, which proved to be of great value for humanitarian organisations already during the disaster (Wilson et al., 2016).

Such detailed analysis on the Nepal earthquake by Wilson et al. (2016) revealed patterns in population movement that were not only valuable at the time but also provide insight in IDP behaviour after sudden disasters as a whole. However, earthquakes are a type of natural hazard that, in many cases, have a warning time of only a few seconds (Rafiei & Adeli, 2017) in contrast to floods, where the warning time could be days or even weeks before the flood happens. In some countries, certain rivers are known to flood following a seasonal pattern. The more extended warning period of floods gives the population living in risk areas some time to leave their homes for a safer place. In addition, it provides humanitarian agencies with the opportunity to apply warning measures, such as guiding the population of an affected area to higher ground or help reinforcing houses and other important buildings. Therefore, it is of value for these agencies to gain some more insight into refugees' moving behaviour during the critical period before the disaster happens and after. On the other hand is the damage caused by a flood proceeding for a longer period, making it hard to determine what the losses will be in the long term.

Collecting real-time data is essential in this context. The role of social media in collecting real-time data has increased significantly as, during the last decade, the data sets within these platforms were becoming more open and accessible. Especially location stamps added by users on Twitter and Facebook provided unprecedented levels of population movements during disasters (Wang & Taylor, 2016; Liu, Yang, Ye, An, & Chen, 2021; Barchiesi, Preis, Bishop, & Moat, 2015). Using these data sources do, however, raise concerns of ethics and moral validation behind the usage of these data sets (R. F. Hunter et al., 2018; Maher et al., 2019). Primarily humanitarian aid organisations are concerned with their public approval using data that is not anonymous and of which the ownership is not clear (Maher et al., 2019). This means that there is a need for ethically obtained, real-time data on population movements during a disaster.

1.1.3 Data availability

Mobile location data could pose as a solution for this ethically obtained, real-time data requirement. This data type is collected by the local phone providers for billing purposes and is called CDR data. Every record represents a call or SMS routed through a cell tower (assumed the closest one), thereby providing the user's approximate location. The record is made anonymous after data collection but includes the demographic information of the person using the phone. This kind of CDR data is called event-driven as an event (read: call or text message sent) causes the trigger to store the location of a user. Opposed to network-based CDR data, which requires an internet connection to provide the user's location every several minutes (Oliver, Matic, &

Frias-Martinez, 2015). During the analysis of CDR data, it is assumed that every phone is used by a single person, meaning that the location of an individual can be tracked. Detailed CDR data is however hard to obtain, as this data is valuable as an asset for telephone providers and it is sensitive for personal privacy validation. Anonymisation of the data is therefore essential.

CDR data has the potential to help us gain insight into where refugees are located almost real-time, as is done after the Nepal earthquake in 2015 (Li et al., 2019a). Here, the CDR data provided a surprising level of information on the movement of the population and showed patterns that are nearly impossible to be exposed using only surveys and counting (Wilson et al., 2016). A better understanding of these patterns are not only precious for humanitarian organisations during emergency relief but might also serve as a predictive measure during the disaster recovery phase. The Spatio-temporal information of a high number of mobile phone users can reveal so-called predictors that could help humanitarian organisations determine where aid is needed shortly after or even before a disastrous event takes place. The scale of the CDR data collection allows for pattern recognition and describing changes in these patterns during a disaster.

1.1.4 Knowledge Gap

Among others, an organisation called Flowminder has already researched the prediction of population movements during particularly devastating disasters and found that CDR data is highly relevant to build a predictive model upon (Li et al., 2019a). However, as the data collection and sharing is somewhat problematic, seeing the sensitive nature of the individual information that it contains, the usability of CDR data needs to be put into a broader perspective. The data in itself might be incredibly valuable, but the need for anonymisation could lead to considerable loss of information. On top of that is CDR data hard to obtain, and do only a handful of researchers have access to it. This study, therefore, focuses on the suitability of using CDR data to predict population movements during major disasters compared to other data. The main emphasis lies on the usage by humanitarian organisations and their access to different data sources at the time of the disaster. As far as is known, this kind of study had not yet been performed.

1.2 Research scope

This research is conducted in cooperation with 510, an initiative of the Netherlands Red Cross. 510 works together with the Red Cross National Societies and their local partners to improve the speed, quality and cost-effectiveness of humanitarian aid through data processing and modelling. Digital transformation and understanding humanitarian data are at the core of their mission and vision. One of the programs that 510 is engaged in is called Innovative Approaches in Response Preparedness (IARP). This program focuses on reducing the impact of climate change through forecast-based financing, data preparedness and cash transfer programming. Within the program, 510 engages with another organisation called Flowminder in a feasibility study of population movement. The purpose of this project is to assess the feasibility of predicting population movements both pre- and post-disaster.

Flowminder is a Swedish non-profit foundation with solid expertise in the usage of mobile operator data, geospatial data and survey data. Their primary focus is on assisting the most vulnerable populations in low- and middle-income countries, where their efforts have led to close relations with local Mobile Network Operator (MNO) 's. Their network of MNO's not only opened up the possibility of retrieving mobile location data of large numbers of phone users for research conducted by Flowminder themselves, but it also brought the option of sharing data with other parties to study. The mobile location data is sensitive, and therefore Flowminder has designed several models that allow the pre-processing of the data to be done before it leaves the servers of the MNO.

The cooperation between 510 and Flowminder allows for a unique combination of expertise and networks. The joint forces around data analysis and modelling and within the humanitarian field provide that the population movements during and after a disaster can be studied in a different way than has been done before. Not only can already obtained data be used to find patterns in the population movements, it also offers the possibility of near real-time assessment of the data, when the relations with MNO 's within disaster-prone countries are initiated, strengthened or maintained.

The project of 510 and Flowminder assessed the feasibility of predicting population movements both pre- and post-disaster, but this study is applied to the post-disaster part of the project only. The potential application of the findings pre-disaster, based on the findings, will however be discussed too. The findings are expected to potentially make a difference in humanitarian aid operations from the EWEA-phase until the Disaster Recovery-phase. Since there are disasters that do not have an EWEA-phase, such as earthquakes that do often only have a warning period of a few seconds, these are out of the scope of this study.

1.3 Research setup

Movement patterns of the affected population during a disaster have been widely researched. However, the analyses are performed (sometimes years) after the disaster has taken place. Getting a better understanding of the relocating behaviour of a population during a disaster is already particularly valuable for humanitarian aid organisations and governments. The possibility of providing real-time data in combination with a more complete overview of displacement behaviour is expected to support humanitarian agencies to focus their energy and effort effectively and efficiently. This research will assist the 510 data analysis team and reflect on how the data sharing could be improved.

The main objective of the research is formulated as follows: *Gaining more insight in the predictability of mobility patterns during and after a disaster, in order to advise humanitarian agencies on where to focus their support almost real-time*. This study will focus on the EWEA phase, the emergency relief phase the recovery phase within the disaster phases as shown in fig. 1.1. The emergency relief phase differs from the other two, as it is typically very chaotic. All involved actors focus on the immediate danger zones and minimise the number of casualties. These teams contain military forces, local and international medical personnel and rescue mission teams. The application of population movement models is thus different during the emergency relief phase. Although often the teams that play a significant role in the recovery phase are already present on-site during the emergency relief phase, their focus is somewhat different. Their efforts are concentrated on providing aid and supplies in areas where IDP 's are still in need of it and help reorganise and rebuild communities. This phase is where the understanding of population movement is essential and where considerable differences can be made with EWEA practices too.

Given the research objective, the main question of the research will be:

Main research question

How do populations move during a major disaster, and can the destinations of the moving population be determined using data that is available to humanitarian aid organisations?

Sub-questions leading to the answering of the main question are the following:

1. What are currently identified main drivers for change in mobility behaviour during and after an expected disaster?
   (a) What does the literature say about the changes in mobility patterns after disasters?
   (b) Which are the most important lessons that can be learned from this literature?
2. How do emergency response activities trigger the movements of a population?
   (a) Which characteristics of the population should be considered?
   (b) Which warning signals were in place before and during the flood?
   (c) Can a change in mobility be recognised that indicate a follow up on warning signals by the local population?
3. Which changes in population movements can be recognised after a flood in comparison to normal conditions?
   (a) How is the change in mobility behaviour due to the disaster detected?
   (b) How can the explanation of the change in mobility behaviour be strengthened?
   (c) What are the key drivers behind a change in mobility due to a disaster that could be used to predict mobility behaviour?
4. What are the most important requirements for the specification of a predictive model?
   (a) Which methods represent the real-world working of the system best?
   (b) How can the model be optimised?
5. To what extent can the population movement caused by a disaster be predicted?
   (a) How well does the model perform?
   (b) Which features are most important for the prediction?
   (c) Does the model change over time?

1.4 Research approach and structure

The research approach that will support the answering of these questions is a case study. As a case, the consequences of Hurricane Matthew in Haiti in 2016 will be analysed, as this disaster resembles such a destructive event that it is almost inevitable that the mobility patterns will change significantly (Wang & Taylor, 2016). The approach fits the purpose of gaining a better understanding of real-world population movements in future disastrous events. Performing a case study does however come with challenges. The most important limitations of case studies are their validity, and generalisation (Yin, 2013). Caused by the fact that there are often just a few cases identifiable, validation of the found results has to be done differently than when many cases can be studied. Predominantly because the significance of the results cannot be determined through sample size. This property of the method poses a major challenge, but some suggestions to tackle the problem are proposed. One of those suggestions is to collect data from observations in the actual local setting (Erickson, 2012). In this proposed study, that means that real-time and local data is collected in Haiti. Overcoming this generalisation problem means that the research should not search for an abstract theory. Instead, it should aim for concrete conclusions that are applicable to the selected case (Yin, 2013).

The five introduced sub-questions are subsequently addressed as shown in fig. 1.2. The first sub-question is answered by means of a literature review in chapter 2. Based on the literature review, the research design in outlined in chapter 3, which serves as an overarching chapter that puts together the steps that are taken to fulfill the research objective. Until this chapter, the scope of the research is converging. The methodology is the central part of this study, from which the scope will be diverging by applying the methodology to a case. The first step, introducing the case, consists of a theoretical exploration and is addressed in chapter 4. The second step is the data preparation, which is described in chapter 5. The third and fourth step are the feature selection (in chapter 6) and model selection (in chapter 7). Based on the methodology, the results are presented in chapter 8. A thorough interpretation of the results and the methodology as a whole is discussed in chapter 9, followed by the main conclusions in chapter 10.

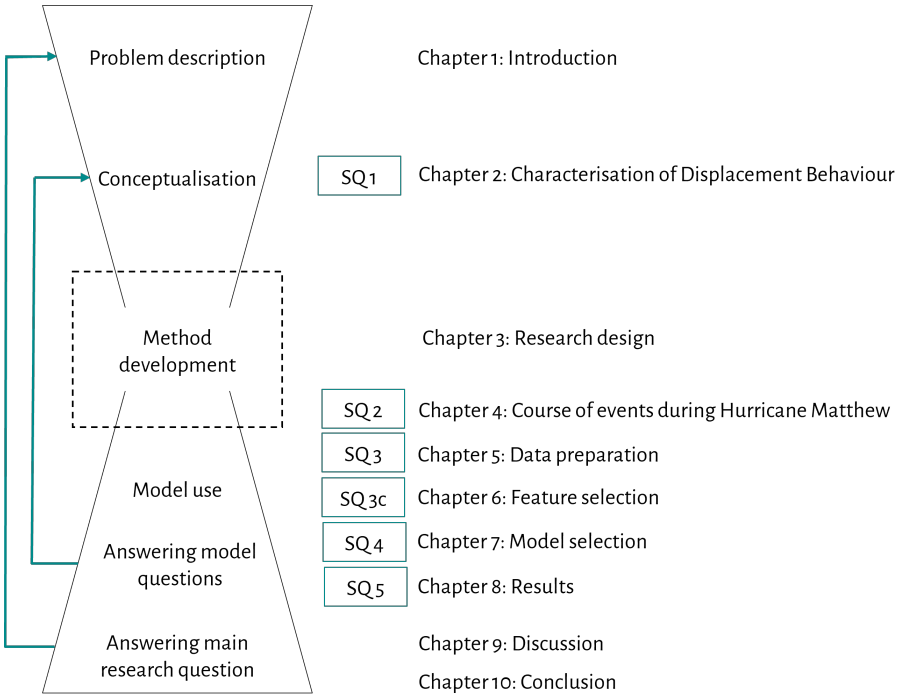


Figure 1.2: Structure of the report including the chapters and the research questions that are addressed

Characterisation of Displacement Behaviour

2

"No one leaves home
unless home is the mouth of a shark."

Warsan Shire
- 2015

To identify the main drivers for changes in population movements during a disaster, a better understanding of the process of decision making by the people within that population must be established. The addressing of individuals' movements is specifically focused on their mobility behaviour. Once we can identify what makes a person decide to leave their homes or take a different route to work, the changes in movement patterns in the system as a whole can be grasped and analysed. A definition of core concepts (in section 2.1) is used to form a base on which the main drivers behind human behaviour during a disaster are discussed. To identify these drivers, a literature review is conducted in section 2.2. This review is focused on human behaviour during disasters, and more specifically, expected disasters such as floods. The most important lessons from the literature review are summed up in section 2.3, which will form the basis for further data exploration.

To characterise population movements and the patterns that it shows, the first research question will be addressed in this chapter:

Research question 1

What are currently identified main drivers for change in mobility behaviour during and after an expected disaster?
(a) What does the literature say about the changes in mobility patterns after disasters?
(b) What are the most important lessons that can be learned from this literature?

2.1 Definition of core concepts

To understand population movements, it is essential to understand the mobility behaviour of the individual's that the population consists of. Some core concepts might sound conflicting or ambiguous and therefore require further explanation. The difference between population movements and mobility behaviour is explained and the difference between Internally Displaced Persons and refugees. Using various definitions, the use of the terms is outlined so that the applied definitions within this study become clear.

The difference between population movements and mobility behaviour is mainly caused by a difference in perspective on the system. *Population movement* can be defined as "the movement of people (in case of the human population) from one place to another with intentions to settle in the new location either temporarily or permanently" [1]. Mobility behaviour describes "how individual humans move within a network or system" (Keyfitz, 1973). When translating the population movements to individuals, a likely translation would be to speak of individual movements. Nevertheless, this term does not quite grasp the essence of what is studied here. Individuals move, but it is the change in their behavioural patterns that is interesting in this case. In normal conditions, individual movement is quite predictable as it follows daily, weekly, and seasonal patterns (Noulas, Scellato, Lambiotte, Pontil, & Mascolo, 2012), but a disaster disrupts these patterns. In other words, the change in the mobility behaviour of many individuals is what causes the population movements (see fig. 2.1).

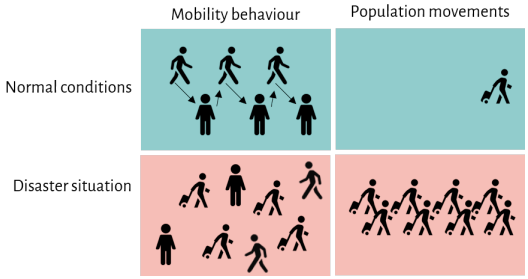


Figure 2.1: In normal conditions, mobility behaviour is rather predictable and population movements are minimal, during disaster situations the mobility behaviours of individuals change and a larger proportion of the population moves.

The difference between an IDP and a refugee is more apparent. As the UN Refugee Agency describes it: "Refugees are people who have fled war, violence, conflict or persecution and have crossed an international border to find safety in another country"[2], while "Internally displaced people (IDPs) have not crossed a border to find safety. Unlike refugees, they are on the run at home"[3]. This study, refugees are not considered. While their relevance is to no extent diminished, refugees often seek permanent displacement instead of temporal, meaning that they often do not return to their homes. In addition, refugees are often lost from sight when crossing the country's borders, as they then also leave the borders in which humanitarian aid organisations operate. When assisting internally displaced persons, one of the main priorities is their return home and their resilience during potential future disasters. This also enhances the continuation of local development.

2.2 Literature Review

This literature review covers the theoretical understanding of a part of human behaviour during a disaster. In a situation where the consequences of an expected disaster form major security risks, individuals' decision-making process involves many perceptions and presumptions. For example, if the perceived risk of staying at home is high and the options of leaving for a safer place are available, it may lead to the decision to resettle for a while. On which variables the decision to leave or stay depends, and what the destination of the displacement will be, are widely researched. Mostly because the relevance of understanding the decision making process

[1] https://www.aresearchguide.com/population-movement.html
[2] https://www.unhcr.org/what-is-a-refugee.html
[3] https://www.unhcr.org/internally-displaced-people.html

of people in distress is reflected upon by the success or failure of humanitarian aid actions. As the results of research around the fleeing behaviour of a population during a disaster implicate the actions taken by humanitarian agencies, finding drivers is highly relevant. Once agencies understand when and where people might relocate, they can focus their efforts on these hotspots more accurately.

Therefore, this literature review aims to identify the variables that influence a person's decision to stay at home or relocate during a disaster. The approach to this review is shown in fig. 2.2.

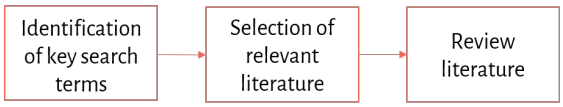


Figure 2.2: Approach for the literature review

2.2.1 Selection of keywords

Identifying key search terms is a process partly done by trial-and-error and partly the result of reviewing experience. Using peer-reviewed research papers and other sources such as news articles, agency reports, web pages, and conference papers, a broader sense of the research field has been obtained. An unstructured exploration of the available information has resulted in the selection of keywords that are presented in table 2.1. Research methods that have been used are snowballing and finding related studies. The primary sources for the literature review are Google (Scholar/Books) and Scopus.

Table 2.1: Selected keywords

Human Mobility Behaviour	**Disaster Response**	**Flood**
Human displacement	Disaster response	Flood
Displacement behaviour	Humanitarian aid	Flooding
Mobility	Humanitarian assistance	Hurricane
Mobility behaviour	Early warning signals	Flood Risk
Population movement	Evacuation	Storm surge
Origin Destination	Risk perception	

2.2.2 Relevant papers

A combination of the selected keywords generated a fair amount of publications, of which some have been selected for the literature review. Based on their relevance for this study, the publications in table 2.2 are chosen. The table outlines the concepts discussed too, to provide an overview of the coverage that the selected keywords offer.

2.2.3 Review

Like in other research fields, the generally increasing availability of data on the behaviour of a population has offered the possibility for more detailed analyses. In the last decade, the use of big data, machine learning and choice behaviour modelling has enhanced the understanding of the driver's behinds individuals' decision to displace or stay at their residence. Previous to the availability of large datasets and enough computing power to run heavy models, primarily qualitative research was done using data collected by surveys. This method is still widely used during the disaster recovery phase and to study human behaviour afterwards. It does provide

Table 2.2: Selected literature with keywords

Citation	Human mobility behaviour	Disaster response	Flood
(Meekan et al., 2017)	✓		
(Haataja, Hyvärinen, & Laajalahti, 2014)		✓	
(Poussin, Botzen, & Aerts, 2015)			✓
(Wilson et al., 2016)	✓	✓	
(Smith & McCarty, 2009)	✓	✓	
(Do, 2019)	✓	✓	
(Lazrus, Morss, Demuth, Lazo, & Bostrom, 2016)	✓	✓	
(Few & Tran, 2010)	✓	✓	
(Sultana, 2010)		✓	✓
(Kaewkitipong, Chen, & Ractham, 2016)		✓	✓
(Allaire, 2016)		✓	✓
(Tim, Pan, Ractham, & Kaewkitipong, 2017)		✓	✓
(Botzen, Aerts, & Van Den Bergh, 2009)		✓	✓
(Haynes, Tofa, Avci, van Leeuwen, & Coates, 2018)	✓	✓	✓
(Aerts et al., 2018)	✓	✓	✓
(L. M. Hunter, 2005)	✓	✓	✓
(Hamilton, Demant, Peden, & Hagger, 2020)	✓	✓	✓
(Bempah & Øyhus, 2017)	✓	✓	✓

a thorough understanding of information that individuals took into consideration when deciding to leave. Big data, on the other hand, has a broader focus. It not only involves people who stayed in refugee camps or temporary shelters (often surveys are held on these locations) but also takes those into account that stays with family or friends. Alternatively, those that decided not to return to their home location, although this proportion of the population is often tiny (L. M. Hunter, 2005). In other words, big data represents a larger part of the population, but with more shallow information as it is often quantitative. In contrast, qualitative data from surveys and stakeholder interviews represent a smaller population with more detailed information.

Studying the determinants that accurately describe the decision-making process that people go through during a disaster allows for many different approaches. Poussin et al. (2015) uses the Protection Motivation Theory (PMT) to describe the decision making process. This theory uses the threat appraisal, meaning in this case that residents take adaptive measures in order to protect themselves from expected floods, but only when they perceive the threat to be high. If they perceive the protective measures to be effective, affordable and easy, they will consider taking protective measures (Poussin et al., 2015). However, the perception of the flood probability seems to be affecting the perceived affordability strongly, and therefore the analysis is highly dependent on this variable. The study also took place in France, so one could argue that a stable political situation and economy are key determinants for the outcome of the application of the PMT model.

Another approach is the Dynamic Disaster Model (DDM). It describes the progress of evacuation activities and human behaviour during a disaster from a psychological perspective. The cognitive and emotional states of an individual that is affected by a major disaster change over time (Vorst, 2010). Therefore, the model incorporates three disaster phases: the pre-impact phase, the impact phase, and the post-impact phase. During the pre-impact phase, the threat stage and warning stage occurs. Pre-impact is also the phase in which evacuation happens. Risk estimation is very low, and thus, not everyone sees displacement as a necessary action. The impact phase involves heavy stress and sometimes denial of the intensity of the disaster, both likely to hin-

der effective evacuation(Vorst, 2010). Only during the post-impact phase are survivors ready to evacuate or be evacuated (Vorst, 2010). This psychological approach to the mental stages of affected populations is essential for the understanding of human behaviour, and Vorst claims that evacuation models that take these factors into account represent reality far better than models that do not.

Even though the PMT and the DDM seem to grasp some of the key drivers behind a person's decision, analyses of disaster data do provide us with much insight too. Hamilton et al. (2020) performed a thorough literature review on human behaviour around flood water. People decide to leave their homes because they seek safety or receive evacuation warnings and instruction (Adeola, 2009). Other determinants that can be found are demographic variables, people's prior experience with disasters, the population's information, and the possibilities of relocating. For example, the number of social contacts in other regions and the financial situation do significantly widen their options of relocating. The listed determinants and their implications for population movements are investigated.

Demographic variables

Demographic variables play a role in the determination of people relocating or staying. Disaster studies in Fukushima and Florida indicated that males and older people were less likely to evacuate (Do, 2019; Smith & McCarty, 2009) while families with young kids often did leave their homes during a disaster (Do, 2019). Family composition too played an essential role during flash flood evacuations in Bangladesh, where the population perceived adolescent girls as highly vulnerable in public shelters. Especially households where no male family member was present stayed home and tried to sit out the disaster (Alam & Rahman, 2014). Although it seems that these measures suggest the protection of girls, another study in Bangladesh indicates that the consequences of these actions are far stretching. Sultana (2010) found that suffering during a disaster is unevenly distributed between males and females, that gender related-issues often intensify during a disaster, and that perceived benefits from humanitarian aid do not include the genders evenly. It thus appears that gender, age, and family composition play a role in the relocating behaviour of affected people and that the distribution of humanitarian aid requires that the demographic differences in IDP's are considered.

Education levels within a population provide contradictory results in mobility behaviour after a flood. Sometimes they make people move, and other times they make them stay (Do, 2019). As speaks to intuition, education might contribute to higher income, opening up possibilities of relocating compared to households with lower income. The combination of high education and the decision not to evacuate is by some considered irrational, but there are apparently more influential forces in place than education levels.

A low socio-economic status in many cases leads to slow response and less evacuation activity (Bempah & Øyhus, 2017). Wisner, Blaikie, Blaikie, Cannon, and Davis (2004) too found that social vulnerability is a crucial determinant of disaster risk and its impacts on an individual or household. Education and income do contribute to the socio-economic status of a person. However, also cultural values of a population seem to explain some of the seemingly irrational attitudes that people sometimes show in disaster situations (Fielding, 2018; Bempah & Øyhus, 2017). For example, a study in Ghana showed that communities often perceive disasters as acts of God, leading to lower intentions to take action during a disaster (Bempah & Øyhus, 2017), as is also the case in other developing countries. Failure of entrusted authorities strengthens the inactions by the population. In these cases, it is observed that previous experiences with flood rationalise these beliefs and encourage people to take the needed measures (Adeola, 2009).

Prior experience with disasters and risk perception

People learn from previous experiences, meaning that prior floods also help strengthen their risk perception (Aerts et al., 2018). This does not only affect the decision of relocating or staying but also the mitigation activities and flood-preparedness behaviour (Tversky & Kahneman, 1992; Botzen et al., 2009), lowering the impact of a disaster significantly. Communities also hold prior experience of floods as a sort of collective memory, which in turn intensifies the feeling of responsibility to take mitigating measures against floods as a group. These measures include early warning systems and training, constructing levees to protect critical infrastructure and enforcing buildings. Over time it has become clear that people learn from flood experiences and update their risk perception with it. Floods that took place further away in the past tend to lose their impact on risk perception within a community, as can be seen in The Netherlands (Botzen et al., 2009). Here, even though the population of the Netherlands is aware of the increasing flood risk and the history of devastating storm surges, the overall perception of risk is relatively low (Botzen et al., 2009).

Obtained information

Although the prior experience of a person was found to be significant in predicting the odds of evacuation, it appeared less important than friends' and family members' influence in determining evacuation behaviour (Adeola, 2009). Due to their social environment's incentives, people in (to be) affected flood areas are inclined to take evacuation measures more seriously, in contrast to when these incentives come from governments. Something to consider here is that the effectiveness of warning signals and evacuation instructions from governments depends on the population's trust in their policy. For example, some residents from an area affected by Hurricane Katrina reported that they did not relocate because of a false evacuation alarm that took place only a few months before the Hurricane hit (Adeola, 2009). After that, they did not take the warning signals seriously. On the other hand, do flood prevention measures sometimes leave residents with a false sense of security, as no one anticipates breaches of levees (Adeola, 2009; Botzen et al., 2009).

The information that people receive around the risk of a (coming) disaster is thus not always taken seriously or an incentive to take protective measures. The source of information might even be a better determinant for one's risk perception than prior experience. During a catastrophic flood disaster in Thailand, internet-based information sources, mainly social media, has shown to be an accurate information distributor (Kaewkitipong et al., 2016; Tim et al., 2017). Social media offered information that was not found in other online sources. They appeared to be more localised and near real-time (Allaire, 2016), enabling residents to significantly reduce their losses because they had time to move valuable possessions to higher ground.

While social media was widely used to spread information during a major flood in Thailand, receiving this information might be tricky. During floods, internet connection and electricity are not always as reliable as needed, and devices used need to be charged en connected to properly working communication infrastructures. In a Finnish study, official warnings were given through radio broadcasts, and it was perceived as a suitable medium for communication from authorities to the public (Haataja et al., 2014). Considering the high level of development and connectivity in Finland, radio broadcasts might even make a more substantial impact in developing countries.

Another important reason for people to relocate is that their source of income has ceased. On the other hand, people often decide to stay because they believe it safer to stay at home or they disregard official warnings due to 'warning fatigue' (Haynes et al., 2018). Also, the intention to protect property was observed in case studies in Vietnam; people with a lower income were prone to protect their property in contrast to protecting their health (Few & Tran, 2010). Ajibade et al. (2015) as well found that income and coping strategies were the best predictors for determining flood impact in Nigeria. At the same time, the consideration for personal

health was not a good predictor.

Destination selection

Social networks are considered the most influential factor in deciding on a destination during relocation (Do, 2019). Especially places where people have family and friends are essential destinations for displaced persons. Social networks are important in migration patterns (Warner et al., 2009). However, the prominence of the factor really came to light in the study of migration patterns after the nuclear disaster in Fukushima (Do, 2019). The most important destination were the residents of friends or family. The second most important reason for choosing a destination then and there were recommended shelters by the government (Do, 2019).

2.3 Key findings of Chapter 2

The aim of doing the literature review is to answer research question 1, which was introduced at the start of this chapter: ***What are currently identified main drivers for change in mobility behaviour during and after an expected disaster?*** Considering previous research on human behaviour around floods has led to many insights that form the basis for further analysis.

The already identified main drivers for a change in mobility behaviour after a flood are: intensity of the disaster at the home location, family composition, socio-economic status, risk perception (strengthened by prior experience, the source of information, the taken mitigating measures and influenced by the social network of a person), options for relocating (what does a family leave behind versus what will they expect to find at their destination).

From these drivers, some observations are derived that could be tested with the case study in chapter 6. Most importantly, it seems clear that the heavier the impact of a disaster, the more people will relocate. A major consequence of heavy disasters is that buildings and crops are damaged or destroyed, leading to the immediate need for shelter and income. Secondly, different approaches can describe the underlying rationale of individuals that decide to relocate. Incorporating the findings that other methods show, such as the PMT and the DDM, can add valuable insights to data analysis. These methods are better focused on the thought processes of affected people and can explain the subsequent behaviour quite well. Some other important observations from the literature review are:

1. The heavier the impact of a disaster, the more people will relocate.
2. We can expect that people without many contacts living close-by will travel larger distances than people who have contacts living close-by.
3. People whose houses have been destroyed are expected to leave their homes to seek safety and opportunities.
4. Families that lose their source of income due to a disaster are expected to seek other sources of income by relocating.
5. Families with lower incomes are expected to relocate less and later than families with higher incomes.
6. People that live in areas where major disasters happen more often are expected to have a higher risk perception, which might lead to a rapid reaction on evacuation orders or the adoption of preventive measures.
7. Social networks are expected to influence the choice of destination when relocating strongly.

8. Community driven mitigating measures are expected to be more successful in areas where disasters are still fresh in memory. In addition, they are expected to be more successful when trust in national authorities is low.

9. The source of which the information comes that people receive determines their risk perception and thus influences their decision to leave or stay. Social media might play a significant role in information spreading, but the circumstances around online communication must be ideal. Seeing that, this way to work, people depend on the availability of internet and electricity. Therefore, radio broadcasts could be considered too.

All and all, many factors need to be considered when studying the decision-making process of an individual to relocate or stay home during a disaster. These factors are used to determine if the population movements after a disaster can be predicted. How this is done is elaborated upon in chapter 3. Translating the individual displacement behaviour to displacement patterns within a population appears very dependent on the socio-demographic composition of the population. Therefore, also a thorough analysis of the Haitian population and the warning signals in place during Hurricane Matthew in 2016 is conducted, of which the findings are described in chapter 4.

Research design

3

From the literature reviews, several drivers have been deducted that influence the relocating behaviour of people that are affected by a disaster. The found drivers can be used to check if their behaviour is predictable. In order to do so, a methodology has been developed that includes the drivers and translates them to features using available data. A case will be applied to the methodology to illustrate the working of the initiated steps.

The main goal of the methodology is to gain more insight into the predictability of mobility patterns during and after a disaster to advise humanitarian agencies on where to focus their support almost real-time. Thus, this method uses data available to humanitarian aid organisations when a disaster occurs. It is relevant to the organisations to determine the locations that people move to and when they move. Ultimately, a time-related origin-destination matrix would help to find the critical patterns of population movements. However, this data is not available yet, and therefore the de-aggregation of the data is a vital step towards the predictability model. How to manage this is described in four steps: case selection, data preparation, feature selection and model selection. Finally, the model is applied to explore the patterns that can be recognised that tell us something about the population movements. In this chapter, the four-step methodology is introduced in section 3.1, followed by an overview of the method in section 3.2.

3.1 A four-step methodology

The development of the methodology has been done through a process of tackling problems. The ultimate outcome of an origin-destination matrix from aggregated data made it possible to align the steps towards that goal. Since using CDR data for researching human behaviour is still under development, this study also focuses on the challenges and opportunities that this data type holds. Firstly, a fitting case is selected and introduced. The information about the case must be sufficiently rich so that the found drivers in the literature reviews can be tested. Since the behaviour of a population during a disaster is highly dependent on the environment of that population, the properties of the country and the disaster that is chosen are discussed. After the case selection, data preparation takes place. The characteristics that are found within the case are used to set up the origin-destination matrix. Some of the characteristics will also be used to measure the predictability of the population movements, which are selected in the third step. Here, also the data exploration takes place to see if some movement patterns can already be uncovered. In the last step, a model is selected that fits the purpose of finding the drivers behind population movements. The performances of different models are tested and optimised before selecting the most suitable one.

In the ideal situation, there are no obstacles, and the steps are thus sequential. In practice, these steps often take place more iteratively. An overview and short introduction of the four steps is shown in fig. 3.1.

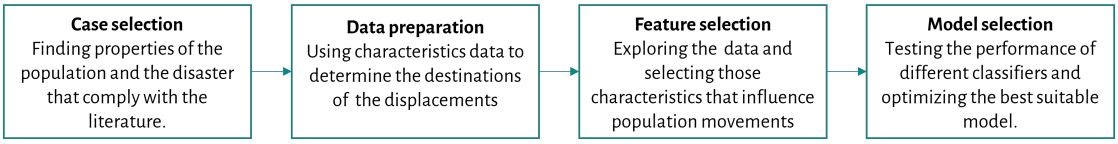


Figure 3.1: An overview of the four-step method that is applied within this study.

3.1.1 Case selection

This research of population movement after a major disaster is done by means of a case study. Now that the drivers behind individual displacement behaviour have been presented, these will be applied to the event of Hurricane Matthew. The disaster that evolved around the Hurricane caused major hazards, primarily in the Caribbean. The country that was hit the hardest by this natural disaster is Haiti. Due to the high wind speeds and extreme rainfall, nearly half the country was left in devastation. This case fits the purpose of studying population movements because the after-effects of the Hurricane unfolded over a more extended period. As a result, the population had the chance to decide to relocate over multiple days, while humanitarian aid agencies anticipated the possible strategies to help the most vulnerable among the survivors. In addition, studying flood response in Haiti is highly relevant, as floods are recurring events. Learning about the strategies that worked during the aftermath of Hurricane Matthew, might contribute to good practices in the future.

3.1.2 Data preparation

The data available in this case study includes information about the changes in individual mobility behaviour due to Hurricane Matthew. The dataset that is at the core of this study is an aggregated version of CDR data and the characteristics that are added to this data. The CDR data from the population affected by Hurricane Matthew in Haiti is generated by telecom provider Digicel. Digicel has the largest market share of 73.59% in the country and is one of only two providers in the country (Conseil National des Telecommunications, 2016). Because of the larger market share, Digicel is assumed to be representative of the population in Haiti (Dejby, Li, Albert, & Lefebvre, 2019). The period of which CDR data has been obtained is from April 5th 2016, until April 3th 2017. The data, therefore, represents 26 weeks before the disaster and 26 weeks after.

During the emergence relief period after Hurricane Matthew, the organisation called Flowminder contributed to the humanitarian response with real-time datasets on population movements (Dejby et al., 2019). This dataset contained data on all calls made or received by Digicel subscribers and aggregated spatially to show the inflow and outflow of subscribers in the affected regions on several days. In addition, the comparison to the inflow and outflow of these regions in normal conditions were incorporated. As a typical individual makes between 2 and 11 calls daily and given the regularity of human behaviour, this provides an excellent basis to identify displaced persons among the subscribers (Dejby et al., 2019).

The preparation of the CDR data includes a distinct analysis of the preprocessing of the raw records, which includes the identification of displaced persons among all the callers within the dataset and the addition of characteristics. Of these characteristics, some are expected to be essential drivers in determining an individual relocating or staying at their home. However, the literature suggests that other important characteristics are not present in the dataset. Such as variables that say something about a person's socioeconomic status and risk perception. If the geographic location of the homes and destinations of the movements within the aggregated CDR data were known, these missing characteristics could be added.

The addition of the locations can be done by de-aggregation of the dataset. Some of the added characteristics can be combined and traced back to an area within Haiti that they correspond to. That way, the characteristics of the individual's home location can be assumed to be the same as the average value for the area's

characteristic that the home location is in. Even though this is considered a major simplification, it does at least tell something about the drivers identified within the literature.

Besides the option of adding some characteristics based on the geo-location of the displaced persons, de-aggregation of the data provides inside in the security of the aggregated CDR dataset. The limited availability of this data type stems from its sensitive nature, as it contains private information of mobile phone users. When de-aggregation of the data could be done, the feasibility might be affected. De-aggregation of the data could thus also be used to recommend ways to share such sensitive data securely.

In order to de-aggregate the data successfully, overlapping characteristics need to be found that are available for the aggregated data, as well as for the areas within the country (Haiti in this case). The more data is available, the stronger the de-aggregation will be, as multiple comparisons can be made. For every data point within the aggregated dataset, an area that is best represented by the characteristics is selected. As a method, a supervised classifier can be applied. For this classifier, the overlapping characteristics are used to identify the locations within the aggregated data.

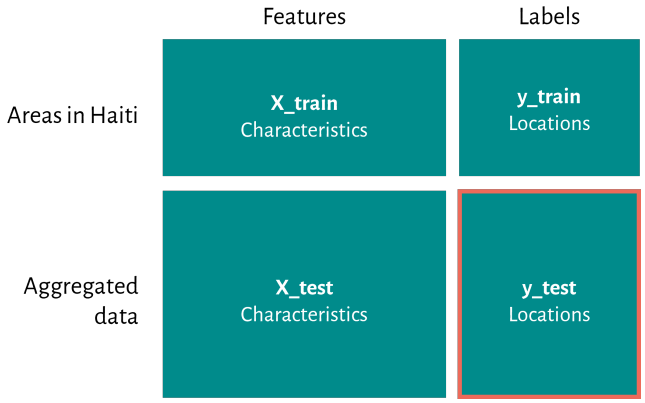


Figure 3.2: Using a supervised classifier, the locations within the aggregated data can be determined, using the characteristics within the dataset.

Validation of the found locations is challenging. Typically, after a classifier has been applied, the classifier's performance can be measured based on the difference between the training set labels and the testing set labels. Here, the testing set is the pursued information, meaning that the mislabelled data cannot be identified. To validate the findings in another way, case-specific information can be used as well as the Euclidean distances.

3.1.3 Feature selection

The characteristics within the original aggregated CDR dataset and the characteristics that can be added after the determination of the locations can be used as features within a classification model. Using data exploration, the characteristics are tested to influence the mobility behaviour of the individuals. The ones that do will be included in the model as features. To channel the data exploration, the findings within the literature review in chapter 2 are used.

The geo-locations are also used as input for the classes within the supervised classification model. Although information about the travelled distance is available within the aggregated CDR dataset, this information is not used for the prediction of population movements. The most important reason for this decision is

that the distances provide an unjustifiable level of precision. And although these distances could be rounded off or grouped, geo-locations automatically contain more information than the distances travelled only. In addition, using the actual locations might help with directing humanitarian aid where their effort is needed most. This study plays a more exploratory role to determine the opportunities that the aggregated CDR data holds for the prediction of population movements. However, incorporating the geo-locations could serve as a stepping stone towards the prediction of where larger groups of people can be expected.

3.1.4 Model selection

The model selection is based on the form in which the data is used. When the features are used to determine a continuous predictee, a regression can be considered. In this study, however, classes will be used as predictees, meaning that for the model selection, several classifiers are tested.

To be able to compare the performance of the different classifiers, the F1-score is used. This accuracy score is calculated after a supervised classifier has been applied, testing the number of correctly predicted classes. The terms True Negative, True Positive, False Negative and False Positive are used to check if the predictions were right or wrong. Their meaning is displayed in fig. 3.3. The higher the proportion of True scores, the higher the model performance.

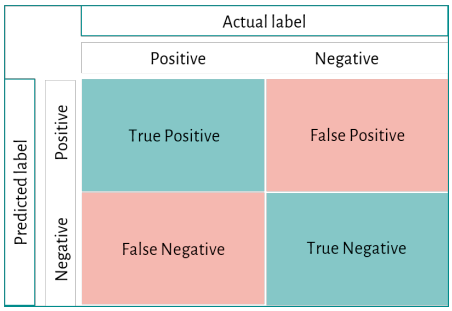


Figure 3.3: Visual representation of the true and false labelling of the data, with in green the rightly labelled data and in red the mislabelled data

The precision score indicates the accuracy of positive predictions by calculating the percentage of True Positives from the total of positive predictions (which is the True Positive plus the False Positives), see eq. (3.1). The recall score is the fraction of True Positives of the True Positives and False Negatives together; see eq. (3.2). The F1-score is then calculated, using eq. (3.3). Support shows the number of displacements that were considered for the score; in this case, the total number of displacements per dataset.

$$\text{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{3.1}$$

$$\text{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{3.2}$$

$$\text{F1 Score} = 2^{*}(\text{ Recall} * \text{ Precision}) / (\text{Recall} + \text{ Precision}) \tag{3.3}$$

A basic application of the models is used for the first selection of supervised classifiers, and their F1-scores are compared. The applied classifiers are: Logistic Regression, Decision Tree, K-nearest neighbour, Naive Bayes

and Support Vector Machine. These methods fit the purpose of classifying data based on multiple features. In case other kinds of datasets are available, other methods could be more suitable.

The classifiers are supported by a package in Python called Sci-kit Learn. Within this package, for all the classifiers, optimisation practices are available too. So, after comparing the accuracy scores for the basic implementation of the classifiers, different techniques are implemented. This strengthens the confidence that the most fitting classifier is found eventually.

When the F1-scores start to deviate, the choice can be made to optimise a selection of the classifiers. In the ideal situation, all the models are optimised, and all their outcomes (meaning not only their F1-score) are analysed. Due to limited time and resources, only one or two classifiers are selected for optimisation in this study. The optimisation techniques are used to strengthen the models' performance, but they do not provide higher accuracy scores. Techniques such as Cross Validation, Grid Search and standardisation of the data lead to stronger confidence that the found accuracy will be found again when repeating the study.

3.1.5 Model application

The eventual choice of a classifier then depends on the goal of the study. In this case, a deeper understanding of the change in population movements after a disaster is pursued. The feasibility of using individual movement data that are retrieved from Call Detail Records play a central role. Classification of the destination of individuals based on the characteristics of those connected to their home location is thus essential.

By applying the most fitting classifier, the results need to be interpreted and analysed. The different steps should be interpreted to be iterative rather than subsequent. Learning more about the data by analysing the results might provide new insights that can help to redesign the used classifiers. This process thus contains potential feedback loops.

3.2 Overview of the methodology

How the four steps are connected to each other, which deliverables are expected in each step and which methods are applied in each step is shown in fig. 3.4. Throughout this report, the flowchart is filled in so that the application of the methodology is straightforward.

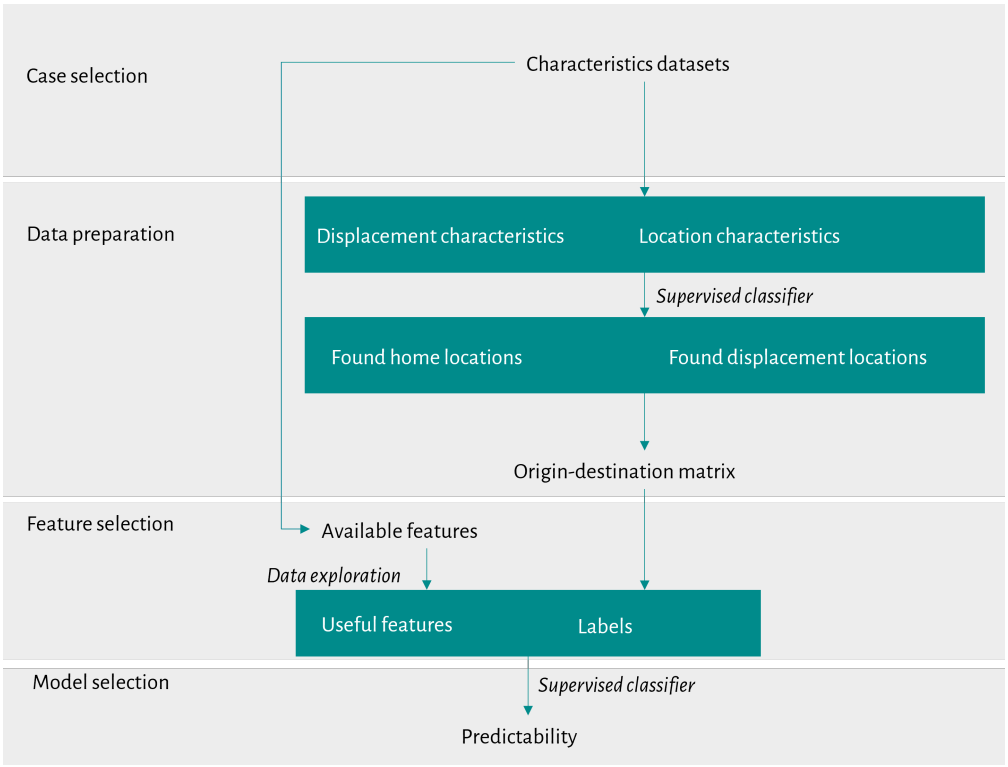
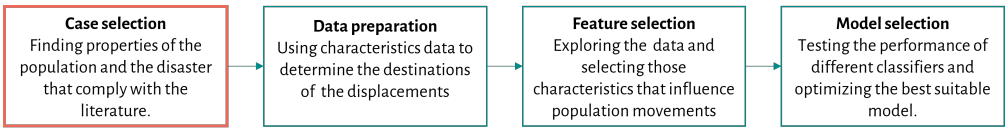


Figure 3.4: A visual representation of the methodology that is used in this study, with all the steps and methods included

Course of events during Hurricane Matthew

4



As a first step, the case for the study is selected. The course of events that took place around the Hurricane is explored, including the activities that humanitarian aid agencies initiated. That way, the drivers that emerged from the literature review can be applied, and possible flaws can be taken into account when advising organisations on where their efforts are needed most.

In this chapter, the development of Hurricane Matthew is introduced in section 4.1. Then in section 4.2, the demographic composition of the Haitian population is analysed in order to gain a better understanding of the conditions that most people live in and answer sub-question (a). Thereafter, the information that is already gathered around the population movement and disaster response activities after Hurricane Matthew is discussed in section 4.3 (addressing sub-question (b)). Then, a summation of the key findings in section 4.4 (sub-question (c)). The aim is to answer the following research question to better grasp what happened in Haiti in 2016 and 2017.

Research question 2

How do emergency response activities trigger the movements of a population?
(a) Which characteristics of the population should be considered?
(b) Which warning signals were in place before and during the flood?
(c) Can a change in mobility be recognised that indicate a follow up on warning signals by the local population?

4.1 Case study introduction

Hurricane Matthew was a devastating category 5 Atlantic storm that caused much damage on its course. The Hurricane took off from the Westcoast of Africa, and on September 25th 2016, it was already clear that the storm could potentially gain much power during its travel across the ocean. The path and intensity of the Hurricane when it neared land are shown in fig. 4.1. Two days later, on the 28th, the storm was officially upgraded to a Hurricane and named Matthew. On September 30th, the Hurricane's rapid intensification took place, where the peak wind speed went up to 240 km/h within a day. Only four days later, on October 4th, Hurricane Matthew made landfall on Haiti. Until then, the storm had avoided land in South America. On the same day, predictions

were made that the Hurricane would also go over the Southeast coast of the US. The following days, the Hurricane went over the Bahamas, inflicting punishing blows on the islands.

Meanwhile, residents of the Southeast coast of the US are urged to move inland to avoid the Hurricane. On October 7th, the Hurricane arrived at the coast of the US, but fortunately the eye of the storm stays about 120 kilometers away from the beach. Therefore, the damage in Florida, North- and South Carolina is less than expected; power outages, flooding and damaged buildings cause an estimated 10 billion dollars in damages in the US. By this time, Matthew has become a Category 1 Hurricane. In the following days, from October 12th on, the Hurricane turned eastwards and ceased to exist.

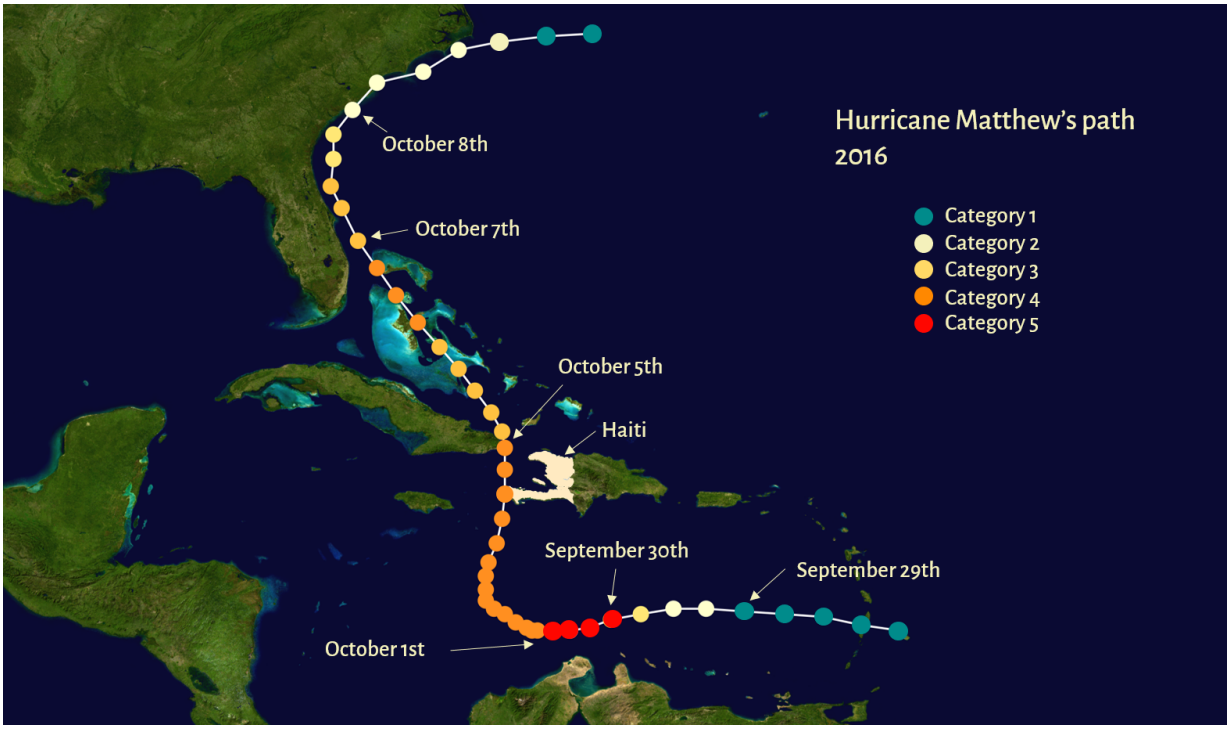


Figure 4.1: Route and intensity of Hurricane Matthew

4.2 Demographics of Haiti

To understand the behaviour that people in Haiti show during a disaster of this magnitude, it is essential to understand the socio-demographic composition of the population. Haiti is a developing country ranked 170 out of 189 on the Human Development Index by the World Health Organisation. Political instability, vulnerability to natural disasters, and a low GDP make the country remain among the world's poorest countries. The poverty rate is around 58%, and the distribution of wealth is highly polarised, with the wealthiest 20% of the population holding more than 64% of the countries income, compared to less than 2% held by the poorest 20% of the population (World Bank, 2021). Infant and maternal mortality rates have remained high in the last decade. The Human Capital Index estimates that a child born in Haiti today will grow up to be only 45% as productive as they could be when they would receive full education and health care (World Bank, 2021).

Haiti has about 11.2 million inhabitants. 96% of the population has been exposed to severe natural disasters, such as earthquakes and Hurricanes (World Bank, 2021). Climate change is expected to intensify the disasters in Haiti, and as most people live in the coastal areas, they are particularly vulnerable to floods. Mountains characterise the inland areas, making it challenging to construct a proper infrastructure. About 60% of the population live in urban areas (CIA.gov, 2021), most of which live in Port-au-Prince, the capital city. Although the sanitation and availability of drinking water have been improved in the last decade within the urban areas, the health care within the country stays at an alarmingly low level, making the population highly vulnerable for infectious diseases (CIA.gov, 2021).

About 62% of the population can read and write, further complicating information distribution (CIA.gov, 2021). There are over 6 million sim-card subscriptions, leading to a so-called teledensity of 58 per 100 inhabitants. Only roughly 32% of the population had access to the internet in 2018. The telecommunication infrastructure is among the least developed in Latin America, and it received an extra blow during Hurricane Matthew in 2016, as it causes $35 million of damage to the communication infrastructure.

The GDP per capita in 2016 in Haiti was 1266$US. Most money is made in services, and around 22% of the GDP comes from agriculture. There is widespread unemployment and underemployment in the country; over 40% of the population is unemployed. Moreover, two-thirds of the population do not have a formal job (CIA.gov, 2021). On top of that, the inflation rate in 2016 was around 14% (Statistica, 2021), meaning that the purchasing power had gone down significantly. Since that year, inflation rates went up even higher, topping at 22% in 2020 (Statistica, 2021). Due to these circumstances, an estimated 58% of the population lives below the poverty line (CIA.gov, 2021).

The country is prone to extreme weather because it is located on the path of seasonal Hurricanes. Due to deforestation, however, Hurricanes do have a more devastating impact, as landslides and floods are given way. Before Hurricane Matthew, the country also suffered from multi-annual drought, outbreaks of cholera and food insecurity (Grünewald & Schenkenberg, 2017).

It is important to understand the desperate situation that a large proportion of the population lives in to comprehend the magnitude of the disasters' consequences in Haiti. Natural disasters, such as Hurricane Matthew, leave the people in the country in peril; there are no financial reserves or social safety measures available for households that are left without a roof, without a source of income or the necessary healthcare.

4.3 Consequences of Hurricane Matthew in Haiti

As a result of the poor conditions that the average Haitian lives in, the options that they perceive when considering relocation are expected to be very limited, the individual drivers that are found in section 2.3 might, to some extend, explain the patterns found during and after the disaster. The International Organisation for Migration (IOM) took the lead in mapping the population movements the very day that the Hurricane hit Haiti (International Organisation of Migration, 2016). On October 5th, their situation report states that 8000 households have been affected and that five deaths, six injuries and one missing person had been reported. They counted 15.623 displaced individuals. In the Grand Anse department, all communication networks were down, and a key bridge connecting the southern peninsula to the rest of the country had collapsed the previous day. It follows that access to the most affected areas is blocked. In the days that followed, more deaths, injuries and displacements were reported. On October 20th, it was clear that at least 546 people had died (although international media estimated this number to be over 900), 80% of the electricity network was destroyed, and 90% of the homes in the South and Grande Anse department were damaged or destroyed by the Hurricane. By this time, 175.000 displaced persons were staying in 307 temporary shelters, while remote rural areas were still being assessed due to the difficult access. Also, the number of cholera cases went up as many people stayed

close to each other within temporary shelters and camps. A more detailed description of the events that took place right before, during and after Hurricane Matthew made landfall follows in the following subsections.

The hazards that the Hurricane caused, especially in the southern part of the country, were predominantly a consequence of the extreme wind speed and heavy rainfall. Figure 4.2 provides an overview of the areas that are hit hardest.

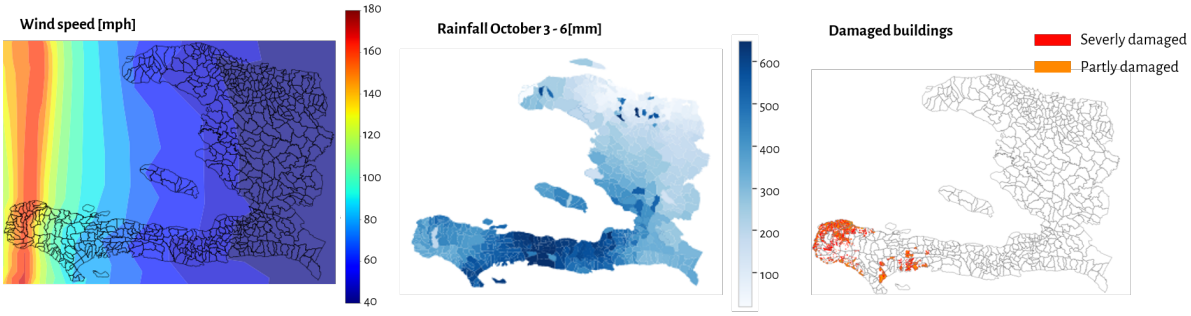


Figure 4.2: Wind speed, rainfall and damaged buildings after Hurricane Matthew

4.3.1 Early response activities

The first signals of a developing tropical storm came from the US National Hurricane Center. The Center engages in the monitoring and forecasting of tropical Hurricanes and cyclones. In their report on April 7th 2017, they summarised the history of forecasts and warning signals during the development of Hurricane Matthew, between September 28th and October 9th 2016 (Stewart, 2017). Their forecasting performances have been relatively high in the last decade, and so they were during Hurricane Matthew. The prediction error decreased linearly over time to a mean error of only 32 km at a forecast period of 12 hours. Up to 24 hours in advance, it was not definite that the Hurricane would cross Haiti in its path, as it could still head to the west crossing Jamaica and Cuba. Because the Center has built up a reputation of providing accurate predictions, the warning signals are often taken seriously.

The watch and warning actions around Hurricane Matthew in Haiti were the following (Stewart, 2017):

- **September 30th - 9PM.** Tropical Storm Watch issued at the southern border of Haiti.
- **October 1st - 3PM.** Hurricane Watch issued at the southern border of Haiti.
- **October 1st - 9PM.** Hurricane Watch changes to Hurricane Warning of the southern border of Haiti and Hurricane Watch issued at the northern border of Haiti.
- **October 2nd - 9AM.** Hurricane Warning issued in the whole country of Haiti.
- **October 4th - 7AM.** Hurricane Matthew makes landfall at the southern border of Haiti.
- **October 5th - 3PM.** Hurricane Watch changed to Tropical Storm Warning in the whole country of Haiti.
- **October 5th - 9PM.** Tropical Storm Warning discontinued in Haiti.

Although these warning actions did reach large parts of the population, many Haitians did not take the evacuation announcements very seriously due to the so-called "Cry Wolf" syndrome. However, the evacuation

was considered a better success than during similar disasters in the past (Grünewald & Schenkenberg, 2017). The uncertainty of the exact path that Hurricane Matthew would follow caused confusion among residents in the southern peninsula. However, the devastating earthquake in Haiti in 2010 had still provided its lessons for disaster preparation in the future. This led to a stronger disaster preparedness on the community level and helped streamline the early response activities. Before there were signs of a Hurricane arriving, emergency supplies were distributed to facilities in all departments, and local emergency responders received annual emergency training. Grünewald and Schenkenberg (2017) called it the "return on investment in disaster preparedness".

The rapid mobilisation of international organisations contributed to a quick setup of emergency relief activities. Figure 4.3 gives a more detailed insight into the course of events around the Hurricane and its many involved parties. Because many telephone cell towers were down, communication between different areas was complicated. Despite the efforts made by telecom provider Digicel, many areas remained unreachable for up to 10 days. Satellite phones provided by scouts, Non-Governmental Organisation (NGO)'s and the United Nations (UN) became crucial in the communication of the overall response to the Hurricane (Grünewald & Schenkenberg, 2017).

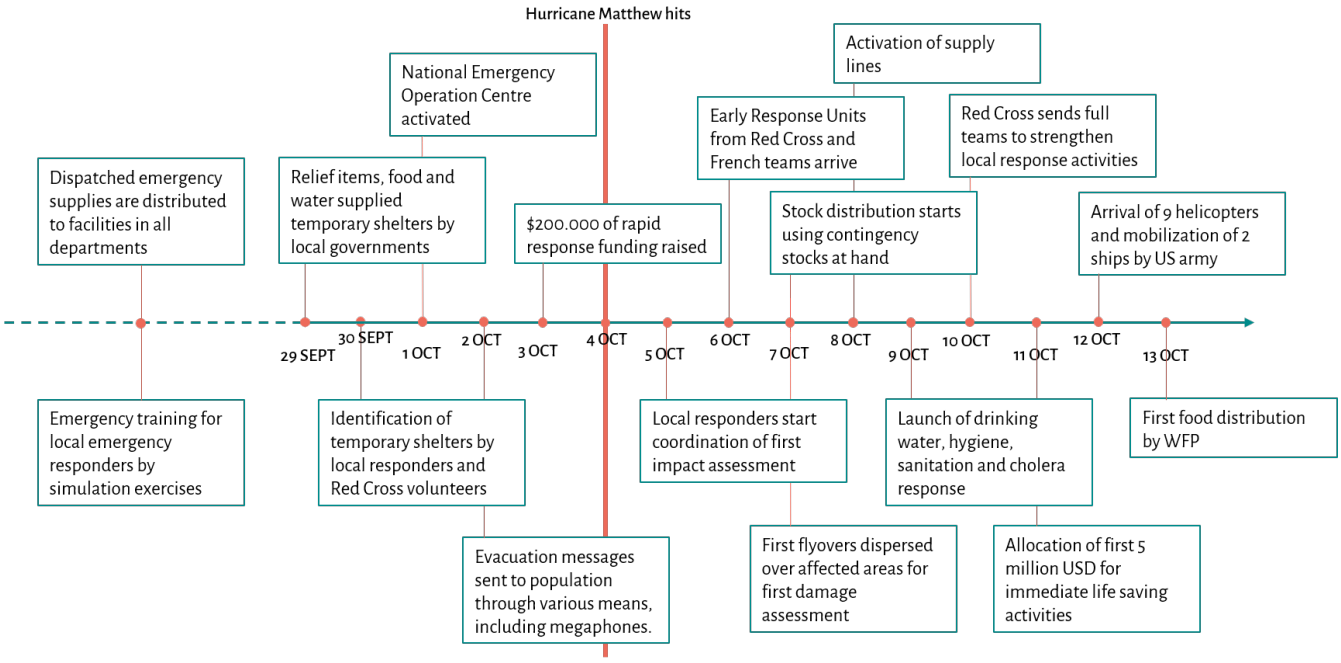


Figure 4.3: Timeline of early action operations around Hurricane Matthew

Data collection and management appeared to be erratic during the Hurricane response. Lack of coordination and quality of the data collection and assessment caused a vast decline in confidence in presented results. A direct consequence was that until five weeks later, there were areas where people had received little to no assistance (Grünewald & Schenkenberg, 2017). Essentially, three main factors made that data usage during the disaster response phase was catastrophic. Firstly, the earliest collected data contained an unjustified level of precision. Referring to unrealistically specific terms wrongly give the impression that data is sufficiently available and have a counter-productive effect on the quality of response. Secondly, many outdated figures were used. Not only were demographic figures from years before the Hurricane applied, but the collected data was

also not updated until six weeks later. This led to the presented shelter needs falling well below the actual needs. Moreover, there were competing data collection initiatives. A lack of coordination made that organisations only considered their own aim. When surveys were distributed without authorisation by the Haitian government, they felt that the aid agencies bypassed them. None of the data gathered and reported by these initiatives got formally published because the government did not authorise them. Especially the number of deaths became highly sensitive. In this environment, confusion, misunderstanding and tensions dominated the first weeks of the response (Grünewald & Schenkenberg, 2017).

4.3.2 Continued response activities

The data collection improved after a few weeks, partly because the IOM took the lead in tracking population movements. By systematically tracking the people staying in shelters, detailed information about the displaced population was gathered. In the Displacement Tracking Matrix (DTM), the departments of Grande Anse, Sud and Nippes were assessed by the IOM and the UN Migration Agency. Their report from February 2017 summarises the findings of the different rounds of investigation that the IOM performed in November and December 2016 (IOM - The UN Migration Agency, 2017).

A methodology for displacement tracking was carried out through interviews at bus stations, where travelling individuals provided information on their displacement (IOM - The UN Migration Agency, 2017). The most important finding was that 58.1% of the interviewees indicated that their destination was in the Ouest department, while 40.4% indicated Grande Anse, Sud and Nippes to be their destination. The displacement tracking map has lead to the visualisation shown in fig. 4.4.

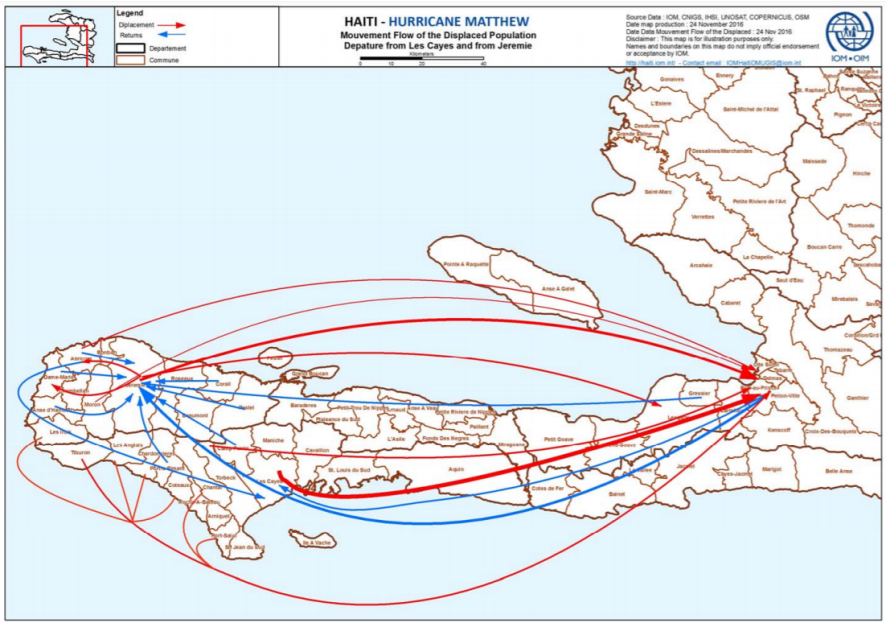


Figure 4.4: Population Movement DTM RD2

Another applied method was shelter profiling, which has been done by consolidating information gathered by humanitarian partners, governmental organisations and national institutions (IOM - The UN Migration Agency, 2017). A total of 711 evacuation shelters were inventoried, of which 423 shelters were profiled in

more detail. The approaches used are field visits, observations, physical counts and interviews with key informants. These 432 shelters housed 10.531 households, equivalent to 43.584 individuals. On February 14th 2017, 47 shelters had remained open, hosting 7.015 individuals (representing 1.564 households). Of these 47 sites, 23 sites have been registered by DTM teams. Here resided 4.596 individuals, representing 1.071 households, of which 1.281 were identified as vulnerable. Some of the most important factors for considering individuals as vulnerable are a single-female head of house (30.9% of the reported population), orphaned children (26.7%), people suffering from chronic illness (12.8%), and unaccompanied elderly persons (9.2%).

Of the 1.071 registered households in February 2017, 66.1% indicated that their residence was completely destroyed, while 26.1% stated that their homes were severely damaged (IOM - The UN Migration Agency, 2017). Compared to the 59.8% destroyed homes in November 2016, this observation suggests that people of which their homes have been completely destroyed have stayed longer in the camps. 86.2% of the households indicated that their crops were damaged. And in February, 77.4% of the households came from rural areas, and the remaining 22.6% came from urban areas. While in November 2016, 53.2% of the households came from rural areas and 46.8% from towns. An overview of this data is shown in fig. 4.5. This leads to the observation that people from urban areas returned to their homes sooner than people from rural areas. Of the households, a majority of 62.1% indicated that their homes were made of wood, 16.3% stated it was made of brick, and the rest had houses made of straw, metal sheets or concrete.

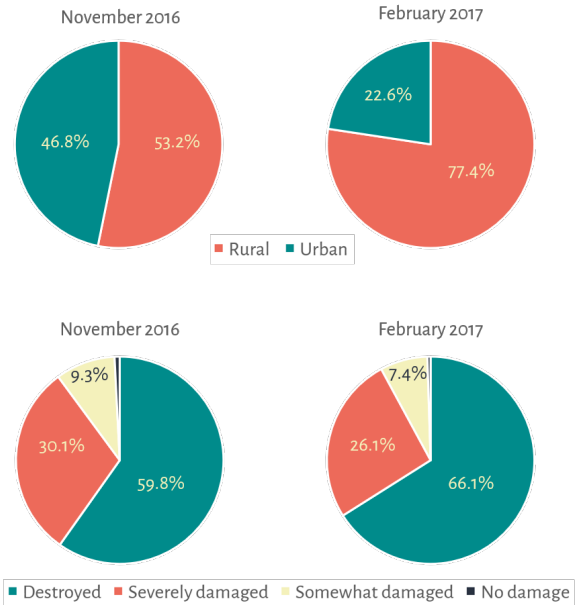


Figure 4.5: Changes in housing information of the displaced population in 3 months

It is clear from this analysis that the most severely affected part of the population stayed in the remaining shelters four months after Hurricane Matthew hit landfall. Households from urban areas returned home sooner than people from rural areas. The proportion of individuals indicating that their houses were completely destroyed increased too. This could be explained by assuming that houses in rural areas were more severely damaged than houses in urban areas and that people without houses to return to stay longer in the camps. It is also clear that the most vulnerable people stayed in the camps longer.

4.4 Key findings of Chapter 4

A trend analysis performed by the IOM stated some important findings in the population movements (International Organisation of Migration, 2016). Firstly, many people moved from highly urban areas (Ouest Department) towards the south to verify the whereabouts of loved ones, to assist relatives, and to show solidarity for the affected population. By rounding up food, clothes and hygiene articles, they contributed to the relief activities happening in the severely damaged southern departments of Grande Anse and Sud. Within those areas, the trend to move from rural areas to urban areas was observed. Especially to the cities Jeremie in Grande Anse and Les Cayes in the Sud department, where also most of the assistance was located (International Organisation of Migration, 2016). As large parts of the rural areas now only contained lost crops, sources of income were lost, which could explain the trend of moving to urban areas.

In the following months, the movement patterns changed. The majority of people now relocated away from the southern peninsula towards the Ouest department. These changes have been attributed to the following factors. Grande Anse and Sud are mainly agricultural departments, and the loss of crops and income resulted in the loss of perspective of recovery. Therefore, many people left for the capital to seek alternative income courses. Also, the activity by emergency relief agencies decreased significantly over time. For many affected households, their return home was stimulated, causing the closure of many evacuation centres. The trend of movements to urban areas was still observed.

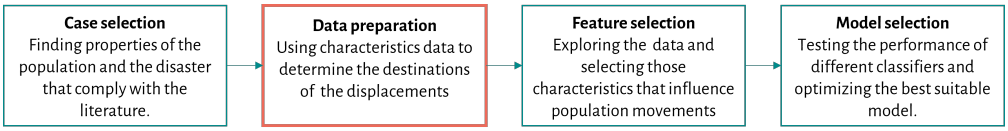
The Real-Time Evaluation report of (Grünewald & Schenkenberg, 2017) states that one of the most important improvements that involved parties should make is the coordination and management of data collection and processing. Since the country was in the middle of an election campaign, the figures became a politically charged issue (Grünewald & Schenkenberg, 2017). Organisations used different data collection forms, and competing needs assessment initiatives were ongoing. With a constrained funding environment, data sharing and collaborative planning among operating agencies were limited. The available real-time data lacked reliability, as can be concluded from revision projects later on.

A significant limitation recognised by most of the relief reports is that emergency assistance was focused on urban areas. Most analyses did not incorporate the population that was affected by the Hurricane but did not stay at emergency shelters. Although it seems logical that humanitarian and governmental organisations put emphasis on the strategy with which the most remarkable improvements can be made, many suffering households were left in devastating circumstances that are hard to recover from. The political tension that the scheduled elections brought did only worsen the faith of the most vulnerable households, as the presented numbers of successful relief operations reflected upon the leading government. Saving small amounts of households in a more extended period does not stimulate this success rate enough.

Nevertheless, did ongoing humanitarian aid in Haiti in the years before Hurricane Matthew pay off during the relief activities. Long term strategies of training local emergency response organisations made it possible to identify suitable shelter locations and extend the chances of survival in areas that were hard to access. Strengthening houses or public buildings might also have prevented many of them to collapse. This best practice should be considered to continue in the future.

Data preparation

5



An evaluation of the emergency relief activities that were in place in Haiti after Hurricane Matthew, has uncovered some flaws. Most importantly, the cooperation between different organisations was strained, partly due to the politically tense situation. This led to erratic data collection and management, causing confusion and sometimes a false sense of success. Furthermore was the focus of humanitarian aid was on urban centres. While people from rural areas, of whom many lost their income due to damaged crops, insufficiently assisted or not assisted at all.

Channelling data collection and processing during the crucial first period after a major disaster is typically a challenge that 510 would face. In cooperation with the organisation Flowminder, 510 aims to improve their understanding of population movements that might ultimately help predict or forecast the found patterns. As part of the IARP program, the two organisations work together in close relation to exchange their expertise.

The preparation of the available datasets forms the setup for the predictive model. Firstly, the available datasets are introduced in section 5.1. Then, the data that is used to spatially track individuals that are moving during and after Hurricane Matthew is introduced in section 5.2. Following up, two additions take place: the addition of spatial information in section 5.3 and of extra variables i section 5.4. Understanding the process is essential to effectively interpret the results eventually. In this chapter, also sub-question 3 (a) and (b) will be answered. Sub-question 3 (c) is addressed in chapter 6.

Research question 3

Which changes in population movements can be recognised after a flood in comparison to normal conditions?
(a) How is the change in mobility behaviour due to the disaster detected?
(b) How can the explanation of the change in mobility behaviour be strengthened?
(c) What are the key drivers behind a change in mobility due to a disaster that could be used for the prediction of mobility behaviour?

5.1 Available datasets

There is a lot of information available on the demographic information of Haiti and around Hurricane Matthew. Tools such as the Humanitarian Data Exchange (HDX) form a good start for the search of valuable data. Because the data collection by concerned organisations that were present in Haiti at the time was somewhat uncoordinated (Grünewald & Schenkenberg, 2017), the credibility of the datasets needs to be examined closely.

Most importantly to this study, an origin-destination matrix of population movements is pursued, but this type of information is not publicly available. However, an aggregated version of individual movement patterns is created by Flowminder. Flowminder is an organisation that is concerned with the analysis of scaled data on the movements of a population through their mobile location. By collaborating with local telecom providers, they obtained CDR data. This data involves the location data of mobile users, in which the time, date and location of the presumably closest cell tower are documented. By using these elements, detailed information on the movements of the users can be collected. The telecom providers use this data for billing purposes and are reluctant to share such privacy sensitive information with other parties. Anonymisation of the data is therefore inevitable, leading to a highly aggregated dataset that does not contain geospatial information and of which the time-stamps have been grouped.

In order to provide some information on the movement patterns of the individuals, characteristics of the home and displacement locations of the IDP's have been added to the aggregated dataset. Based on these characteristics, the data can partly be de-aggregated. Therefore, the sources of these characteristics are used to find corresponding areas in Haiti that match the combinations best. Table 5.1 Shows an overview of the data incorporated in the aggregated displacement dataset of Flowminder, and the datasets that are used to connect it to the areas in Haiti. These areas are based on an administrative 2 level, as the information is distinguishable between the areas while still providing the location data on a detailed level. Administrative 2 level areas can be interpreted as communes, while level 1 areas are departments and level 3 are sections.

Table 5.1: All datasets included in the analysis with their source, description and numerical information

	Description	Dimension	Rows	Mean	Minimum	Maximum
Flowminder						
Median calls	Median number of weekly calls made with the simcard pre-disaster	Time	74992	10.8	2.0	132.0
Radius of Gyration	Standard deviation of the distances between each visited location and the centroid	Time	74992	4.6	0.0	92.2
Entropy	The frequency at which a simcard's daily location changes and the number of distinct locations visited	Time	74992	1.3	0.0	5.6
Contact locations	Persons that were contacted at least once per month in the pre-disaster period and whose number of call days with the IDP (simcard) rank among the top 10	Time	73368	0.6	0.0	1.0
Known locations	All routed cell tower clusters by the simcard. A proportion of the locations visited pre-disaster within 10km of home or displacement	Time	74992	0.5	0.0	1.0
Gust	Wind speed, footprint in miles per hour	Time	74992	58.5	45.0	155.0
Property damage	Proportion of damaged property within a radius of 3km	Time	13982	0.3	0.1	0.75
Rain	Accumulated precipitation between October 3th and 6th on administration 3 level	Time	74992	496.6	125.0	675.0
Distance to main road	Shortest distance as a crow flies to the closest primary or secondary road	Time	74992	7.3	0.1	61.4
Distance to urban center	Shortest distance as a crow flies to the closest urban center, which is based on the population density	Time	74992	7.6	0.5	72.9
Population density	The population density within administration 3 level	Time	74992	121.7	1.4	283.3
Distance to home location	Shortest distance as a crow flies from displacement location to reference location	Space	74992	36.84	0.6	310.7
NASA Global Precipitation Measurement						
Rain	Accumulated precipitation between October 3th and 6th on administration 3 level	Space	140	369.0	126.0	657.8
University College London						
Gust	Wind speed, footprint in miles per hour	Space	140	59.8	40.0	155.0
International Organisation of Migration (IOM)						
Property damage	Proportion of damaged property within a radius of 3km	Space	140	0.1	0	0.74
Worldpop 2015						
Population density	The population density within administration 3 level	Space	140	7	0.7	234.9
Humanitarian Data Exchange (HDX)						
Haiti administration 2	Shapefile of the administative level 2 areas within the country with respective name	Space	140	NA	NA	NA
Risk Evaluation Report (Red Cross)						
GDP per capita	Gross Domestic Product per capita in 2008 as an average of every admin 2 area in Haiti (source: World Bank)	Space	570	590.4211	3	35206
Flood Exposure	Global estimated risk index for flood hazard per admin 2 area in Haiti (source: World Bank)	Space	523	0.0067	0	1
Travel time to nearest city	Estimated travel time to the nearest city measured for 50.000 or more people worldwide in the year 2000. The average travel-time per admin 2 area in Haiti	Space	570	160	3.8	593.7

5.2 Displacement tracking: the aggregation of CDR data

In order to guarantee the proper handling of this sensitive information by all involved parties, Flowminder thus cleaned and prepared the dataset containing CDR's and aggregated it. As a result, IDP's are identified without providing their spatial location. Instead, variables that describe the home location and displacement location of the sim cards that most likely belong to IDP's are provided. These variables could be used to de-aggregate the data so that geo-locations can be estimated.

The original CDR data comes from the Mobile Network Operator (MNO) called Digicel. Their market share in Haiti, at the time of the Hurricane, was 73.59% (Conseil National des Telecommunications, 2016). In the half-year before the disaster, the penetration rate of mobile users within the Haitian population is 61.09%. Of these mobile users, the vast majority uses prepaid call minutes (60.37%) and only a tiny proportion pays later (0.72%). Before the disaster, the average subscriber of Digicel talks on the phone for an average of 44 minutes per month. The users of the other telecom provider, called Natcom, call substantially longer; monthly approximately 106 minutes (Conseil National des Telecommunications, 2016).

A description of the environments in which these data have been collected is given in section 5.2.1, followed by a description of the applied method for identifying IDP's in section 5.2.1. Then the implementation of characteristics of the IDP's within the dataset are discussed in section 5.2.2. Lastly, the limitations of using CDR data are summed up in section 5.2.3. Sub-question (a) *How is the change in mobility behaviour due to the disaster detected?* is answered in this section.

5.2.1 Identification of internally displaced persons

Several methods have been applied to extract a subset of persons that have been internally displaced due to the disaster. These are primarily based on the movement trajectories of a subscriber prior to the disaster and the changes in their movements after. The goal is to select the individuals that are believed, with high confidence, to be displaced following the disaster and determine their Spatio-temporal mobility disruption (Dejby et al., 2019). This means that not only the time of relocating is determined, but also the point at which the person returned to 'normal' (Dejby et al., 2019).

The principle of Flowminder's method is that a level shift takes place within the distance curve of a subscriber (Dejby et al., 2019). A distance curve has been described by Flomwinder to be "the sequence of points, over time, made up of the distance of someone's 'normal' location" (Dejby et al., 2019). Detection of such level shifts requires 1) clean input data, so that 2) the distance curves can be constructed and 3) a level shift could be detected:

1. The spatial and temporal resolution of the data must be up to an appropriate level, meaning that only the subscribers that call regularly are selected. Regular is defined to be at least four call-days prior to the disaster, at least two call-days in the week of the disaster and at least four call-days after that (Dejby et al., 2019). A call-day is a day on which a call is made.
   Also, some cell towers are clustered together as their locations are so close to each other that they do not accurately represent the callers location. In general, a call is routed through the cell tower closest to the caller. However, there are situations in which the maximum capacity of the tower is reached. Localising the caller based on the routing cell tower is usually done by constructing a Voronoi tessellation and assigning the caller to a Voronoi cell. Nevertheless, this causes boundary issues when the caller is situated right between multiple towers. While it is essential to the study to determine if a subscriber has moved or not. Therefore, towers close together are clustered, implementing a method that minimises the total within-cluster variance (Dejby et al., 2019). As a result, towers close together are grouped, while isolated towers remain ungrouped.

2. The construction of the distance curves involves the following steps. First, a reference location is calculated, presumably the home location of the subscriber. For the population movement study of Flowminder, it is not essential to determine the home location, but the location that the individual is commonly seen (Dejby et al., 2019). The reference location is determined to be where the subscriber is observed the most days during the pre-disaster period. To reduce the possibility of falsely identifying a traveling person to be an IDP, only a subset of the reference locations is used: locations within the departments that are affected (Artibonite, Grand Anse, Nippes, Nord Ouest, Ouest, Sud, Sud-Est). Secondly, the scalar distances between observed locations to the reference location of an individual are calculated. The minimum distance from the reference location is chosen as a data point for that particular call-day because it is crucial to determine if a subscriber is at their reference location or not. As a result, even if the subscriber has travelled a lot during a day, their distance is zero if they are seen at their reference location. This leads to a piecewise-constant signal with regular spaced sampling points (Dejby et al., 2019).

   Lastly, an iterative median filter is applied to reduce noise in the piecewise-constant signal. The filter is applied until the filtered dataset is identical to the unfiltered dataset.

3. A step detection algorithm detects the points at which a level shift occurred. Here, Flowminder has chosen to identify sudden changes in mean levels within the time series. To strengthen the confidence of identifying subscribers that actually were displaced as IDP's, only those that comply with the final filters are considered. Filter 1: the individuals have a step change lasting at least three days, in which a minimum of two calls has been made away from the reference location. Filter 2: the individuals have at least three call days at the reference location in the week before the disaster. And filter 3: the subscribers spent half of the pre-disaster period at their reference location, and this is verified by determining that they have at least one call day in at least 90% of the weeks. This way, their reference location is considered to be stable.

The generated subset of subscribers are labeled to be IDP's and consists of 37.839 individuals. The method ensured the prioritisation of a pure sample with a low false positive rate (Dejby et al., 2019).

5.2.2 Characteristics of IDP's within the dataset

The movements that the identified IDP's make after the disaster are aggregated over time, which result in three datasets: week 1, week 2 to 5, and week 6 to 26. Every row within the datasets contains a movement, and the different movements are not connected to the IDP's, meaning that from the datasets, it cannot be determined which exact paths individuals took. If someone is displaced twice, two rows are added to the datasets. As there are 74.992 rows in the three datasets combined and only 37.839 IDP's have made the movements, this means that on average, every IDP made two displacement steps. The datasets use the reference location and the destination, which could mean that for every displacement, the represented IDP might have made a step already or make one after. When addressing the information in the dataset, the rows will be called *displacements*, which are thus not the same as IDP's.

For every displacement, a set of characteristics is provided. These variables could determine the main drivers behind the displacements and indicate the change in mobility due to the disaster. The values of these characteristics all stem from the Hurricane. The data might give the illusion that IDP's that displaced from week 2 on still had to endure the heavy rain and wind speeds that are present in the datasets, but these represent the original rain and wind-speed during the Hurricane which had passed after week 1. The variables refer to the reference location (also called *home location*) and destination of the displacement, and of these two locations, the variables during (and right after) hurricane Matthew are shown. An example of the dataset is shown in fig. 5.1 This figure contains dummy data and merely provides an impression of what the data looks like.

	Home location variables		
Displacement	Windspeed	Rainfall	Damaged_buildings
42	90 mph	550 mm	80%
43	120 mph	120 mm	30%
44	65 mph	80 mm	0%
45	90 mph	320 mm	50%
46	75 mph	500 mm	90%

Figure 5.1: An example of what the data of the home location looks like, using dummy data

5.2.3 Limitations of CDR data

The usage of CDR data comes with many limitations that need to be acknowledged in order to interpret insights that the set provides rightly. As already discussed, the number of displacements is not the same as the number of IDP's, but the displacement steps are not connected, meaning that the dataset does not provide information on the trace of an IDP. Furthermore, the dataset contains variables with information on the origin and destination during hurricane Matthew, meaning that the IDP's did not have to endure the rainfall and wind speeds that the datasets from week 2 to 26 contain. These merely represent the conditions around the reference location (or home location) of the IDP at one point in time, although the dataset suggests differently. It is important to keep in mind while analysing the results. In addition to these limitations, there are four others.

First of all, mobile location data allows the approximation of the user's location and is not information that should be considered exactly right. The routing tower varies depending on signal strength, traffic intensity and, most importantly, network outrages (Dejby et al., 2019). In the days following Hurricane Matthew, the communication network of Digicel was severely damaged, meaning that there was almost no mobile communication on the southern peninsula in the first week after the disaster. In addition, a lower tower density means that the location specificity is still even lower; this is true especially in more rural areas. This leads to a skewed dataset that primarily represents individuals that were less severely affected by the Hurricane.

Secondly, CDR data could be skewed towards wealthier and less vulnerable people (Dejby et al., 2019). Since a large part of the Haitian population lives under the poverty rate, and about 40% of the population has no phone, it seems logical that the most inferior part of the residents is not represented within the dataset. Also, the identification of IDP's among other subscribers led to the drop of data representing less used sim cards. Seeing the high rate of prepaid call users in Haiti, the subscribers with less money to spend might thus be excluded from the dataset.

Thirdly, other datasets used to characterise IDP's home, and displacement locations might be outdated. And data that describe the home and displaced location appear to be very detailed, while the reliability is not expected to be as strong as it appears. The datasets used are relatively coarse. For example, the wind speed of different areas is expressed in speeds of tenths of miles per hour. This means that when zooming in, two neighbours would have experienced very different wind speeds while living next to each other. On the other hand, the population density data suggests very detailed information, while the data is clustered to ten levels. The level of specificity of the clustered data can be misleading and must therefore be considered when interpreting the results.

And lastly, but foremost importantly: CDR data refer to sim cards and not to people (Dejby et al., 2019). Multiple people can use the same sim card; one person can use multiple sim cards, the sim cards might change owner or be disposed of. All of these scenarios are very likely in disaster situations and are therefore fundamental to be considered and understood when interpreting the results from the data exploration.

The usability of the dataset is, however, still very much present. As the DTM by the IOM incorporates individuals and families in refugee camps, this method only provides information on the most vulnerable part of the IDP's. CDR data is probably skewed towards less vulnerable people, but overlooking their needs during a disaster would be wrong. And also, the people that do not end up in camps are still represented within the CDR dataset. When the usage of this type of data appears to be fruitful during disaster situations, the possibilities of CDR data sharing will potentially grow with it.

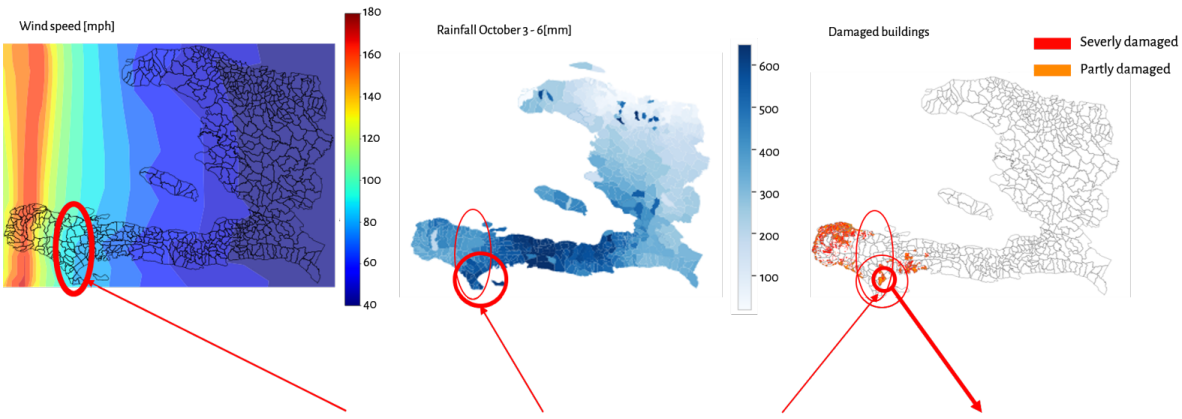
The variables that Flowminder has added to the dataset do contain some of the drivers that have been identified in chapter 2: the contacts living close-by, the intensity of the disaster and the population density. However, some important drivers cannot be derived from the dataset, such as the family's income, the experience with floods or evacuations and how far they are located from the nearest city. In order to still test the latter three, spatial information of the displacements origin and destination is estimated, based on the variables describing the Hurricane. The following sections will elaborate on the method of adding the missing variables.

5.3 Addition of spatial information

Although the detection of a change in mobility behaviour of a population is described in the displacement dataset of Flowminder, some variables have been identified as important drivers for a person or family to decide to displace. However, the findings in chapter 2 show that some important variables are missing. Therefore, the second sub-question (b) *How can the explanation of the change in mobility behaviour be strengthened?* will be explored in this section. Because of the level of aggregation that the dataset holds, spatial information of the displacements is not available. However, de-aggregating the dataset can be done using the identification process that led to the aggregation in the first place. It is then possible to add the missing variables and improve the analysis of why, how and when population movement occurs.

It is important to de-aggregate the data, even though Flowminder already has the untouched data internally available. First of all, because Mobile Network Operators might not want to share less aggregated data in the future. When the movements of a population can be described by variables that are available instead of by CDR data, the humanitarian organisations could benefit from that information. Also, if the aggregated data becomes more widely available, researchers with different backgrounds could direct their expertise on more detailed data. Possibly, the usage of tools or expertise not at hand within Flowminder or 510 might shine a different light on the dataset. De-aggregation of the data could furthermore uncover distortions that the aggregation process has caused. Due to the clustering of some of the variables, essential details might have gone lost. Parties such as Flowminder might benefit from the insights that the data's de-aggregation offers that helps improve similar dataset in the future.

Four datasets will be used to extract the approximate location of the home and destination of every displacement: population density, total rain between 3-5 October 2016, average wind speed and the percentage of damaged buildings. The general idea of de-aggregating the data is shown in fig. 5.2. The used datasets will too have to be aggregated, as some of the data in the dataset provided by Flowminder contains information within a 3km radius. The aim is to imitate the data in the displacement dataset as closely as possible. An overview of the steps taken to add spatial information is provided in fig. 5.3.



Displacement	Windspeed_h	Rainfall_h	Damaged_ buildings_h	Home location
42	90 mph	550 mm	80%	"name"

Figure 5.2: Visualisation of how the de-aggregation of the spatial data is conducted with three variables instead of four

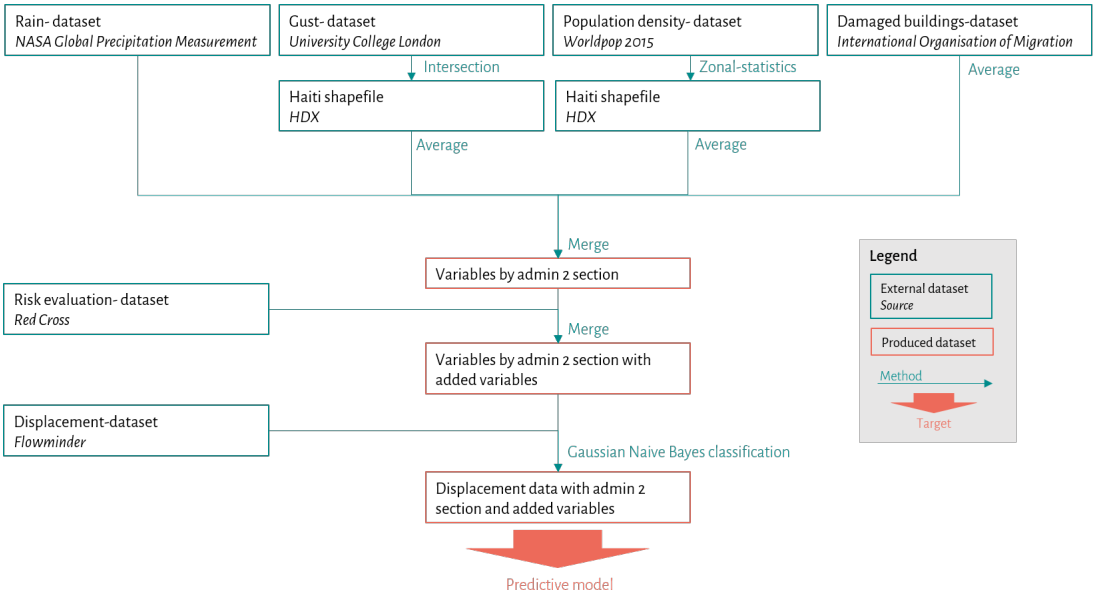


Figure 5.3: Overview of the datasets processing

5.3.1 Dataset matching

The datasets that Flowminder has used to add as variables are shown in table 5.1. The same datasets are used to de-aggregate the data. For some of the variables, the datasets must be tweaked to make them compatible with Flowminder's values. For the data matching, therefore, the preparation of the datasets is shortly elaborated upon. The most important focus is on the corresponding range of the data between the two datasets.

The **population density** within a 3km radius around the locations is derived from the Worldpop dataset of 2015, in which a raster is used to describe the population density. The raster contains squares of 300m by 300m, indicating the population density in that area. To get the data sized to administration level 2 (admin 2), the data is aggregated using a shapefile of Haiti. For every admin 2 area, the average population density is calculated, considering that the raster squares are cut at the edges of the admin 2 areas. This leads to the assumption that the population is evenly distributed within the raster squares and aggregated to an even distribution within the admin 2 areas. As shown in fig. 5.4, the population density within the displacements dataset does not represent the areas in Haiti, as can be expected when comparing density data. A large proportion of the displacements came from the densely populated area of the capital Port-au-Prince. Furthermore, the population density data within the displacements dataset has been clustered to ten values, causing some discrepancies.

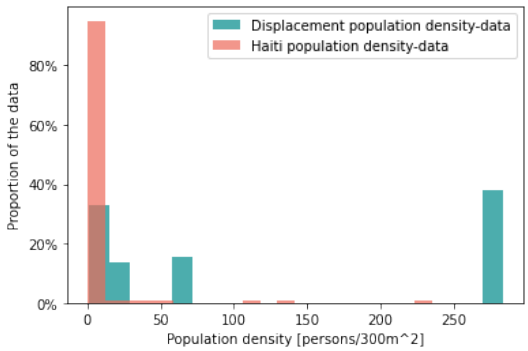


Figure 5.4: Population density data comparison of the displacement data with the area data

The **total rain** that has fallen between 3 and 5 October 2016 is a dataset is originally produced by the Nasa Global Precipitation Measurement tool that uses satellite imagery to estimate the rainfall in different areas (NASA, 2021). The IFRC has uploaded the precipitation data of Haiti around hurricane Matthew on the Humanitarian Data Exchange (HDX), already differentiated to admin 2 areas. Figure 5.5 shows that many displacements took place in the areas that endured more rain, which complies with the idea that more heavily affected areas generate more population movement.

The **average wind-speed** dataset consisted of polygons representing different wind speeds. Initially, the dataset was created by University College London, but the British Red Cross Maps Team uploaded the dataset to the HDX. Translating the polygons to admin 2 areas meant that the intersection of every two polygons was calculated. If two different wind speeds were measured within one admin 2 area, the weighted average of the two was assigned to that area. As the Flowminder dataset contains the average wind speed of every displacements location within a 3km radius, the wind speed is again assumed to be evenly distributed within the admin 2 areas. It appears that a large proportion of the displacements came from areas that endured wind speeds between 40 and 60 mph (see fig. 5.6), contradicting the idea that the most severely hit places gener-

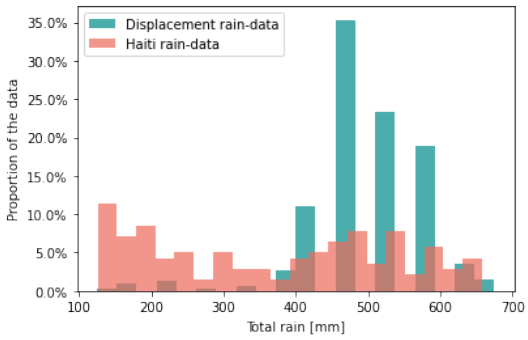


Figure 5.5: Total rainfall between 3 and 5 October 2016, data comparison of the displacement data with the area data

ate more movement. A wind speed of 45mph was measured in Port-au-Prince, partly explaining why so many displacements took place in areas that withstood relatively low wind speeds. The other part can be explained by the fact that the wind speed caused many hazards during the Hurricane, but the displacement data represents movements of six months. The wind did not play a significant role in the decision of moving when the proportion of damaged buildings (likely also caused by the wind-speed) is taken into consideration as well.

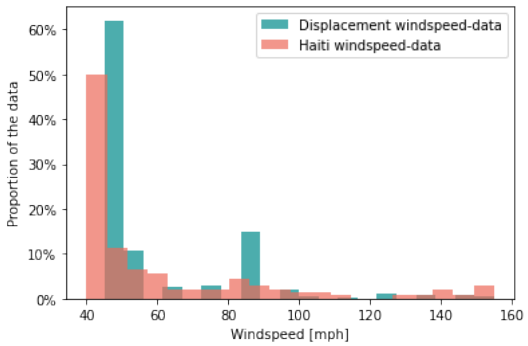


Figure 5.6: Wind-speed data comparison of the displacement data with the area data

The **percentage of damaged buildings** dataset already contains information on admin 2 level. The level of damage is divided into critical damage, severe damage and partial damage. Only the critical damage is considered, as this appears to match the data in the displacement datasets of Flowminder (see fig. 5.7). The displacement dataset represents the data of the damaged buildings within the 3km radius around the location of the displacement. This aggregation level cannot be accomplished when only administrative areas are represented. Therefore, the average proportion of damaged buildings is used for the data-matching. Also, only a small part of the country has been inspected for damage, using satellite and drone imagery. This leads to the impression that all buildings remained intact in the rest of the country, which is a false impression. For data matching purposes, however, the other areas and displacements have a damaged building value of zero.

As a result, the plots in fig. 5.8 show how the variables are distributed within the admin 2 areas in Haiti.

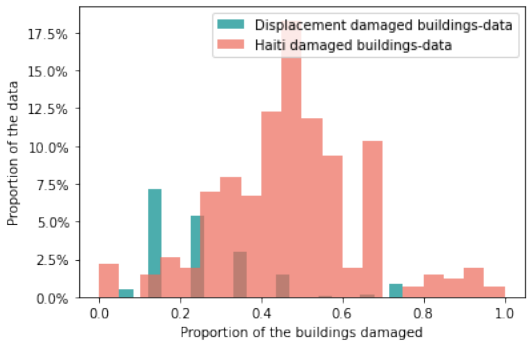


Figure 5.7: Proportion of damaged buildings data comparison of the displacement data with the area data

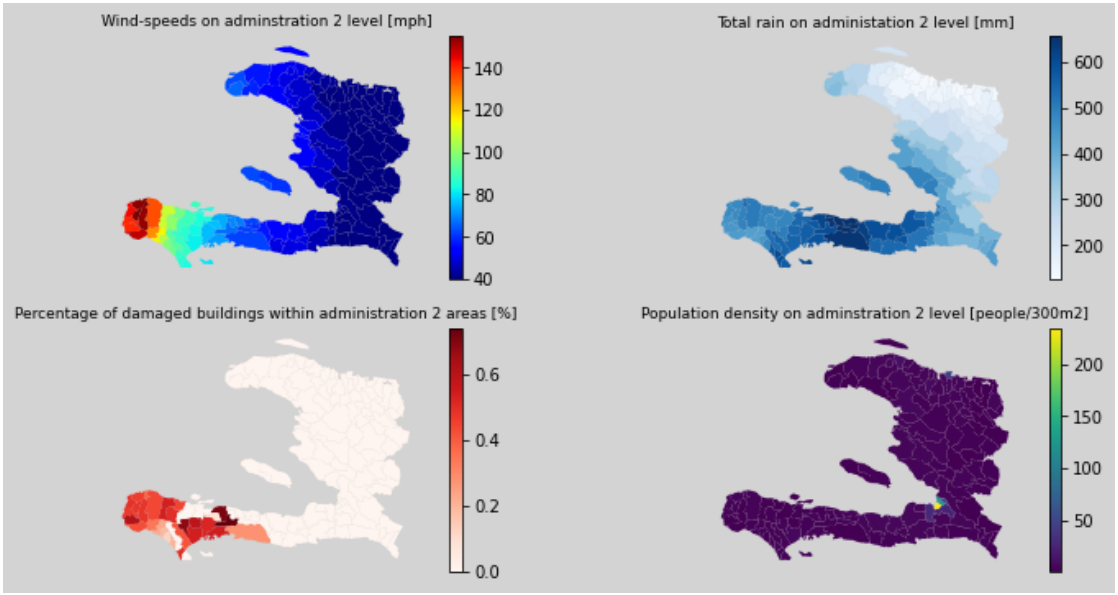


Figure 5.8: Visualisation of the windspeed, rain, percentage of damaged buildings and population density after processing

5.3.2 Naive Bayes classifier

In order to fit the characteristics of every displacement to the characteristics of the admin 2 areas in Haiti, a Naive Bayes (NB) classifier is applied. All NB methods are based on the Bayesian Theorem, which uses the prior probability, likelihood and evidence to calculate the posterior probability (Wu et al., 2008). The function used to calculate the posterior is shown in eq. (5.1) and eq. (5.2). Naive Bayes classification fits this data, as it is a parametric model as opposed to methods such as the K-Nearest Neighbour classifier. In addition, it represents non-linear data as opposed to methods such as Logistic Regression. Moreover is the classifier fast and easily understood. The latter is essential in this case, as the data is not labelled and therefore is not easily validated.

$$p\left(\mathbf{x_i} \mid \mathbf{y}\right) = \frac{p\left(\mathbf{x_i}\right) p\left(\mathbf{y} \mid \mathbf{x_i}\right)}{p(\mathbf{y})} \tag{5.1}$$

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}} \tag{5.2}$$

In order to fit the characteristics of every displacement to the characteristics of the admin 2 areas in Haiti, a Multinomial NB and Gaussian NB classifier are applied. The important difference between the two is that Gaussian NB assumes the input variables to be normally distributed. Three different normality tests have been applied, namely Shapiro-Wilk, D'Agostino's K^2, and Anderson-Darling; all three confirming that the data is not normally distributed. Multinomial NB takes not-normally distributed input data too. The input is parametrised by vectors $\hat{\theta}_{yi} = (\hat{\theta}_{y1} + \hat{\theta}_{y2} ... \hat{\theta}_{yn})$ for each label y (the areas in this case, see fig. 5.9) with n number of features (four in this case: wind-speed, rain, population density and proportion damaged buildings). The probability $p\left(xi \mid y\right)$ is represented by the vector $\hat{\theta}_{yi}$ and for every feature i in the sample belonging to label y.

Multinomial NB uses a smoothed version of the maximum likelihood (see eq. (5.3)). Here, $N_{yi} = \sum_{x \in T} x_i$ is the number of times feature i appears in class y in the training set T. $N_y = \sum_{i=1}^{n} N_{yi}$ is the total count of all features in class y and α represents the smoothing prior.

$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n} \tag{5.3}$$

Multinomial NB classification is typically used for training data that contains texts of which the number of appearances is estimated. The count-function for N within the equation supports this application and is also used to find the closest class for the features in the displacement dataset. To make sure that this method fits the purpose best, also the Gaussian NB classifier is applied to the dataset. The Gaussian NB classifier assumes all input variables to be normally distributed, but these variables are not. With the assumption that adding more data leads to a normal distribution, the mean and standard deviation could be used to estimate the class for every displacement (see eq. (5.4)). Here, μ_y and σ_y are estimated using the maximum likelihood for every feature y and thereby offers the option that the features within the input variables are not the same as the features of the labels. As a result, for every displacement, the admin 2 area that contains feature-values closest to its own is selected to most likely be the area that the displacement came from or went to.

$$p\left(x_i \mid y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{\left(x_i - \mu_y\right)^2}{2\sigma_y^2}\right) \tag{5.4}$$

The strategy used to label the displacement with their most likely admin 2 area is visualised in figure fig. 5.9. Typically, labels are found by training a model and then the accuracy can be tested by measuring how many times the model assigned the correct label. In this case, there are no labels yet to compare the model performance with, leading to a more challenging validation process.

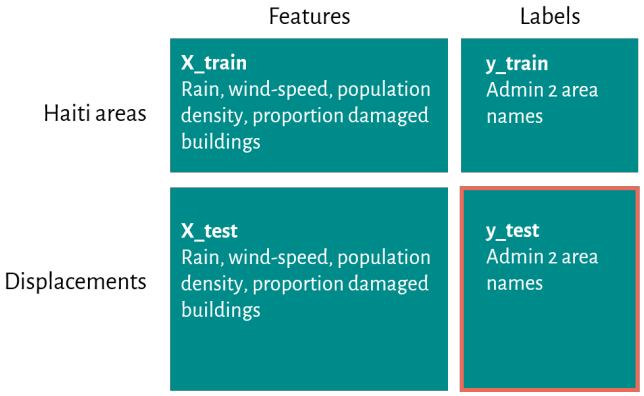


Figure 5.9: Display of the features and labels within the training and testing datasets with the characteristics included

5.3.3 Validation of the model

As a result of the Naive Bayes classifiers, the locations of the home and destination are determined for every displacement. Validation options of these results are limited, as the actual locations for the displacements are not public. As validation practices, some simple tests have therefore been applied: checking if the capital location is rightly assigned, backtracking the feature data, checking if there is no conflict with the information provided by Flowminder and the calculation of the Euclidean distance.

Capital city check. The capital of Haiti, Port-au-Prince, is the only area with a population density of over 200 people per 300 m^2. Therefore, the displacements that indicate the population density around the home location and/or destination of this scale are expected to be in the capital city. Table 5.2 shows that the Gaussian NB classifier mislabelled a considerable number of displacements wrongly, while the Multinomial NB classifier has not mislabelled any of the displacements. The label of the capital city is the only one that holds some value for validation, as it is the only label of which we know for sure that the population density is higher than 200 persons/300m^2. Although the Gaussian NB classifier mislabelled a tiny proportion of the displacement data, we know for sure that those displacements are wrongly labelled.

Table 5.2: Validation of the right application of the capital-area

	Week 1		Week 2 to 5		Week 6 to 26	
	GNB	MNB	GNB	MNB	GNB	MNB
Home locations	41	0	30	0	86	0
Displacement locations	122	0	102	0	276	0

On average, 1.28% of the labels that should have been classified as Port-au-Prince were not labelled as such. On the whole, this means that we know for sure that 0.43% of the labels within the whole displacement dataset is not applied correctly.

Backtracking data. By backtracking the data, the four variables within the admin 2 areas and their representation by the displacements are analysed. For this process, the features of the found labels within the

displacement data are grouped, and the mean for every feature is then compared to the mean of the same feature within the administrative areas in Haiti. It, therefore, represents how much the found average for every label is off of the real average because that is assumed to indicate mislabelling.

The average differences are compared in percentage, and result in the following variation:

- Gaussian NB classification: **3.8%**
- Multinomial NB classification: **21.6%**

It is clear that the found averages are much more off when the Multinomial NB classifier is applied. Sampling some of the rather larger differences do not indicate a simple explanation for it. It seems that the Multinomial classifier mislabels many of the displacements, as the differences are sometimes larger than 100%.

No conflict with Flowminder's information. Within the Flowminder technical report, some information is provided that hints at the locations of the IDP's (Dejby et al., 2019). It is verified that the locations assigned by the classification model are not contradictory to this information. Even though the information in the technical report cannot confirm if displacements have been labelled rightly, it can also not deny that they are.

Euclidean distance calculation. The distance between the area features and the displacement features is determined using a Euclidean distance equation in eq. (5.5) for the standardised data. This validation method is mostly used to verify the right choice of NB classifier.

$$d\left((x_1, x_2, x_3, x_4), (y_1, y_2, y_3, y_4)\right) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 + (x_4 - y_4)^2} \tag{5.5}$$

Here, the d represents the distance, the x_i represent the values of the four features i for the areas in Haiti, and y_i represents the values of the four features i for the displacements. The distances calculated for the Gaussian and Multinomial NB classifiers show differences (see fig. 5.10), and the Multinomial NB classifier has a larger overall distance than the Gaussian.

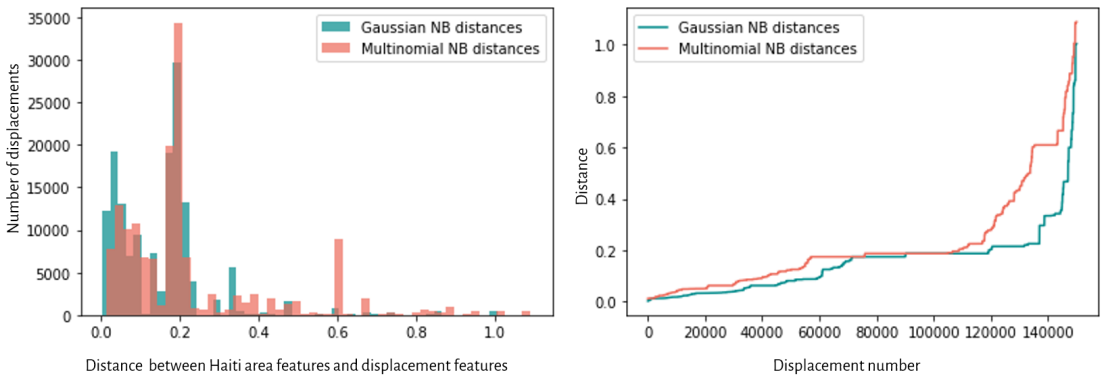


Figure 5.10: Calculated distances for the Gaussian and Multinomial Naive Bayes classifiers

Although the validation practices cannot rule out the mislabelling of several displacements, it strengthens the belief that the found labels are correct to some extend. The three applied methods indicated that the labels are at the least plausible to have been applied rightly. The Multinomial NB classifier appears to meet the validation practices worse than the Gaussian NB classifier has, despite the certain mislabelling of some of the displacements from or to the capital city. The effect of this mislabelling is, however, not as great as the

larger distances found in the features that the Multinomial NB classifier has labelled. Because this raises the suspicion that the Multinomial classifier mislabelled more displacements than the Gaussian classifier. The Gaussian NB classifier is therefore selected to label the displacement data. The mislabelled displacements for the capital city and the displacements with a Euclidean distance greater than 0.3 have been removed as an extra measure. This serves the goal of cleaning the dataset from mislabelled data and results in the exclusion of 10.1% of the original displacement data.

5.4 Extra variables

The labelling of the displacement data has led to the opportunity to add extra variables linked to admin 2 areas. The literature review indicated the importance of more drivers than could be found in the original displacement dataset, which can now be added too. The found drivers behind a person's decision to leave their home are the intensity of the disaster at the home location, family composition, socio-economic status, risk perception and the options for relocation (see chapter 2). These drivers can be translated to measurable variables that are now available in the dataset and will be used to test the observations formulated in chapter 2. With the found locations, the displacements within the dataset can be displayed. The majority of the displacements appear to have taken place within the province that also contains the capital Port-au-Prince.

The extra variables were derived from the Red Cross dataset that contains information on the different areas within Haiti. Using variables available to the Red Cross means that at least one humanitarian aid organisation has this information. That simplifies the variable selection too. The variables are:

- GDP per capita
- Travel-time to the nearest city
- Flood exposure

These three variables are identified to represent some of the drivers found in chapter 2 of which a representing variable was still missing. The socio-economic status of a household can be simplified by using the *GDP per capita* in the area that they come from. Risk perception is affected by experience with previous disasters, and therefore the variable *flood exposure* within every area determines the risk perception. And the options for relocation is simplified by the variable *travel-time to the nearest city* as a city offers the opportunity for other sources of income and humanitarian aid organisations appear to be more active in urban areas. All three variables are greatly simplified and build upon numerous assumptions, which must be considered when deriving conclusions from the analyses.

5.5 Key findings of Chapter 5

The data preparation was mainly focused on the rightful identification of IDP's within a large population of moving individuals. This led to the exclusion of a large part of the population. There are, however, still some displacements in the dataset that raise the suspicion that still some IDP's have falsely been labelled so. Foremostely, because a considerable part of the displacements took place within the capital city, Port-au-Prince, even weeks after the disaster, while the disaster effects were substantially more dramatic in the Southern Peninsula of the country. In addition, the filters that were applied to decrease the number of falsely identified IDP's within the data make that the areas that lost connection are excluded, while these expectantly contain many IDP's. It could thus be considered to provide the unfiltered dataset so that the filters can be applied that serve the purpose of the study that it is used for.

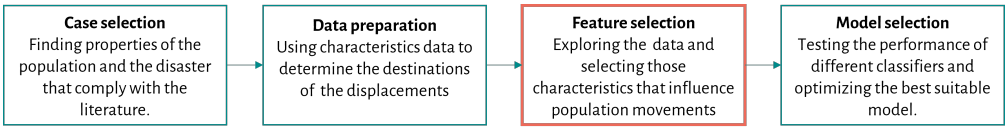
Also, this case shows that the aggregated data can be de-aggregated to at least some extend. For aggregated datasets that include derivable variables, this technique could help with the unravelling of certain information that was aimed to be covered. In this case, the validation of the found locations is not strong enough to state that they certainly comply with the actual locations of the displaced persons. Furthermore, the original aggregation also meant the loss of details in time-steps. As a result, three displacement datasets were formed: those in week 1, week 2 to 5 and week 6 to 26. The loss of valuable details in the data as opposed to the protection of the people represented by this sensitive data is a tension field of which the trade-off has been made. Owners of sensitive but informative data should thus consider this trade-off and take the necessary measures to prevent de-aggregation and prevent the loss of critical data.

One of the critical obstacles that the aggregation of the data has brought is that the displacements are not linked, and therefore, it is unclear how far a destination actually was from home. This information might lead to entirely different results, as the displacements over short distances might belong to the same person. The quality of the aggregated data would expectantly increase considerably, as the actual mobility behaviour of the individual is far better represented that way than is the case in the current dataset.

The aggregation of timely and spatial information in the same dataset also affects the analyses. The timeframes are not the same size, and the third data set represents 20 weeks of displacements. It is questionable what information could be distracted from this information. It took between 6 and 26 weeks for a large part of the IDP population to return to their reference location, but the vast difference between 6 and 26 weeks does weaken the conclusions that can be drawn from it. Aggregating the time component in the data while also not providing geo-locations makes the datasets' utility limited.

Feature selection

6



Now that potential features have been identified and available data has been prepared, the features can be selected. The selection will take place through semi-structured data exploration. From the data exploration, the features and labels that are used in the predictive model are selected.

A major interest during the feature selection procedure is the capability of describing population movements. The answer to sub-question 3 (c) will be pursued in this chapter. First of all, the data exploration is shown in chapter 6, where the literature review forms the basis of the process. The effect of the characteristics that are included in the data on the population movement patterns is explored. Selecting the fitting features follows from the exploration and the availability of certain datasets when a disaster occurs. An extensive explanation of this procedure is described in section 6.2.

Research question 3

Which changes in population movements can be recognised after a flood in comparison to normal conditions?
(a) How is the change in mobility behaviour due to the disaster detected?
(b) How can the explanation of the change in mobility behaviour be strengthened?
(c) What are the key drivers behind a change in mobility due to a disaster that could be used for the prediction of mobility behaviour?

6.1 Data exploration

From the data exploration, including the extra variables, some significant findings can be derived, and the observations formulated in chapter 2 can be analysed. In this section, identifying key factors that are needed for the prediction of mobility behaviour will be explored. Because data exploration is a somewhat unstructured process, the method is simple: the observations specified by the literature review in chapter 2 form the basis for the exploration, and interesting or remarkable findings appear during the process are mentioned too.

6.1.1 Observation checking

For the observations, a substantiation can be found in the literature. As the CDR data represents individual movements, the drivers behind an individual's decision to seek a safer location or stay at their home are researched. For the formed observations, a translation to the available data has been done in order to test if they hold in the case of this CDR dataset around hurricane Matthew too. For every check, the differentiation of the three time-steps is taken into consideration, too, seeing if the mobility behaviour changes over time as well.

Observation 1: The heavier the impact of a disaster, the more people will relocate. Translation: because there is no information about the population that is not relocating, this observation cannot be tested. The effect of the wind speed and rain on the distances travelled by the identified IDP population is therefore shown.

In fig. 6.1 it becomes clear that the distances travelled are affected by the wind speed and rainfall. In both figures, a trend is visible, showing that the travel distance is negatively correlated with wind speed and rainfall in week 1. The correlation becomes positive later in time for wind speed, while the correlation becomes less negative later in time for rainfall. The average distance travelled in the three time-steps increases: week 1 - 29 km, week 2 to 5 - 40 km, week 6 to 26 - 42 km. Probably this is because the time-steps become longer too. The effect of wind speed and rainfall declines over time, but the consequences (landslides, property damage) stay. These consequences could explain the trends seen in fig. 6.1, the situation becomes more unbearable over time.

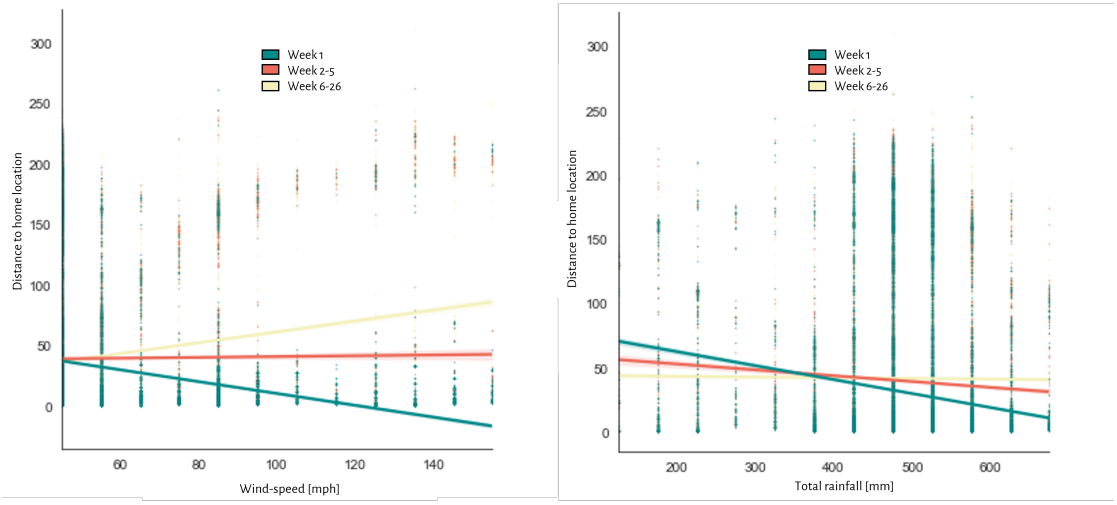


Figure 6.1: The effect of wind-speed (left) and total rainfall (right) on the distances travelled in three time-steps, showing their datapoints and the found linear regression lines.

Observation 2: We can expect that people without many contacts living close-by will travel larger distances than people that do have contacts living close-by. Translation: we expect a negative correlation between the proportion of contacts living within 10km of the home location and the distance travelled.

The social network has impressive effects on the decision of a person to leave for a safer place. For week 1 and week 2 to 5, the proportion of contacts living nearby appears to be negatively correlated to the distance that is travelled during the displacement. For week 6 to 26, there is a minimal negative correlation between the two. Ideally, data of the people that did not move would be available to test whether the people that did not

leave their home also have more contacts living close by. In this dataset, however, the importance of contacts around the home location seems to decrease over time.

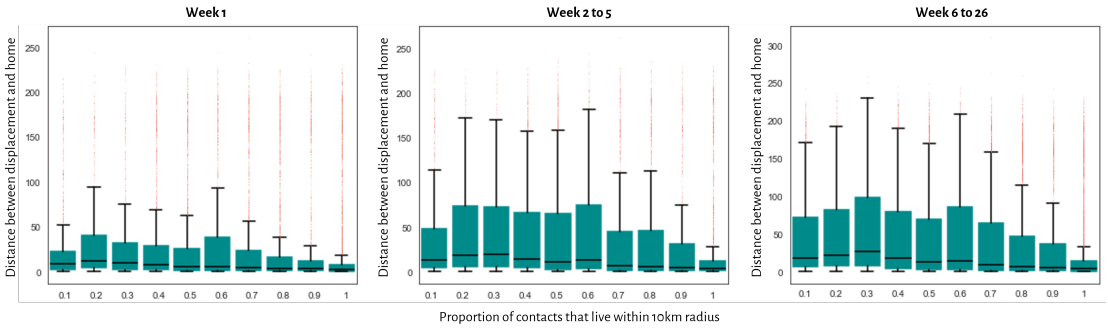


Figure 6.2: Boxplots of the proportion of contacts living close-by and the distance travelled in three time-steps

Observation 3: People of which their houses have been destroyed are expected to leave their homes to seek safety and opportunities. Translation: when many buildings have been destroyed, more IDP's will be identified.

The building assessment has only been done in several admin 2 areas, but that does not mean that there is no damage in the other areas. With the information that we have, we can determine that most displacements within the dataset took place from areas that were not heavily damaged. When comparing the distances that were covered by their time-step, we see that the people from areas that endured a lot of damage did not travel much in the first 5 weeks, but substantially more after week 6 (see fig. 6.3). An explanation for this finding could be that the damage also affects the road network or that people stay longer to help rebuild or relieve in heavily damaged areas. There is, however, no foundation for this explanation, as it is strongly dependent on the identification process of the IDP's within the Haitian population. It is, however, clear that the proportion of damaged buildings does affect the population movement.

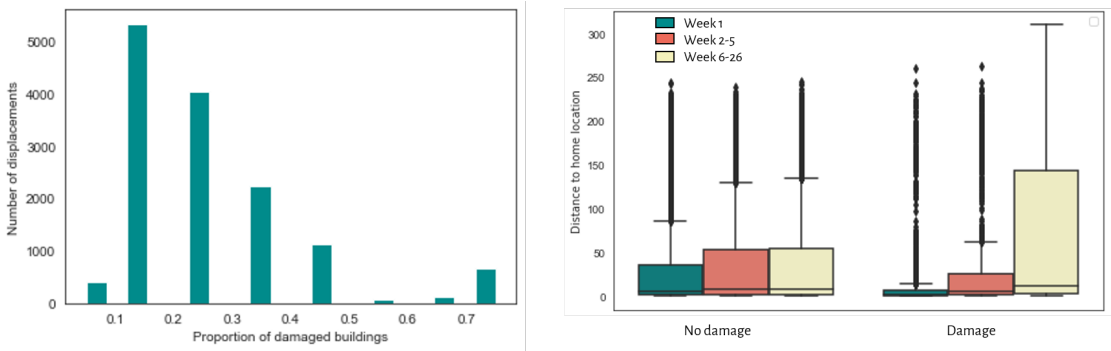


Figure 6.3: The effect of damaged property on the displacement distance of the population in three time-steps

Observation 4: Families that lose their source of income due to a disaster are expected to seek other sources of income by relocating. Translation: since it is unclear which displacements have taken place due to loss of income, this observation check is focused on farmers. Many families in the outskirts lost their source of income when their crops were destroyed. This observation is tested by checking if there is a positive correlation between population density and distance to home location.

It is actually assumed that everyone who loses their income source due to a disaster will seek other sources of income. However, for farmers, this almost automatically means relocating, which can be measured. Therefore, the observation is tested for farmers, assumed to be living further away from city centres in less densely populated areas. The distinction between a farmer and not a farmer is made based on the travel time to the nearest city, of which the spread offers a distinct separation as shown in fig. 6.4. The distances travelled by people from rural areas increase over time, while the distances travelled by people from urban areas decrease. This could be explained by the opportunities that urban areas offer when it comes to finding a source of income. But again, no substantiation for this statement can be derived from the data. It is thus merely concluded that the travel-time to the nearest city from the home location of an individual does affect their movements after a disaster.

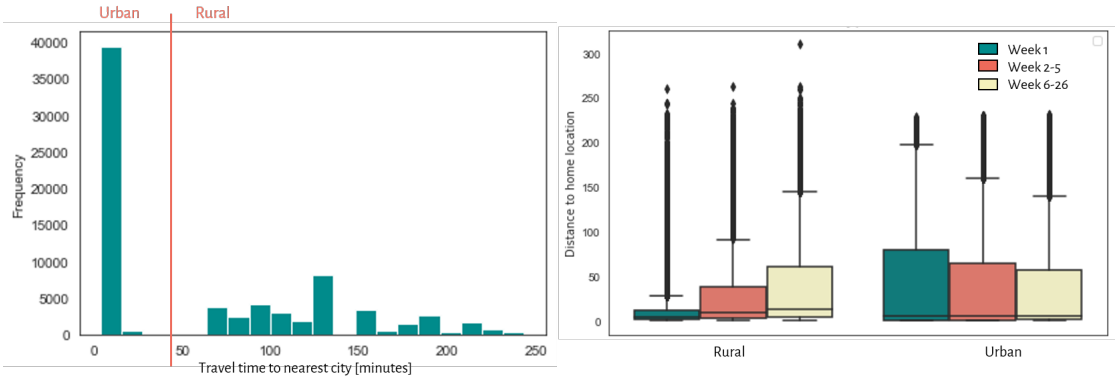


Figure 6.4: The effect of coming from rural or urban areas on the displacement distance in three time-steps

Observation 5: Families with lower incomes are expected to relocate less and also later than families with higher incomes. Translation: a negative correlation between GDP per capita and distance travelled.

The effect of a higher versus a lower GDP per capita appears to be very dependent on where the boundary between high and low is set. Figure 6.5 Shows the distribution of GDP per capita within the displacement datasets and where the boundaries have been set. The effect on the displacement distance is shown in fig. 6.6. It appears that the people with a higher GDP cover less distance from their home location in week 1 to 5. Later in week 6 to 26 the distance travelled compared to the GDP is highly dependent on what is considered a higher GDP. There are multiple interpretations of these results possible, including the lack of necessity to move because of more substantial reserves or the more displacement possibilities that people with more money are presented with. What we can derive from this is that the GDP per capita does affect the distance travelled.

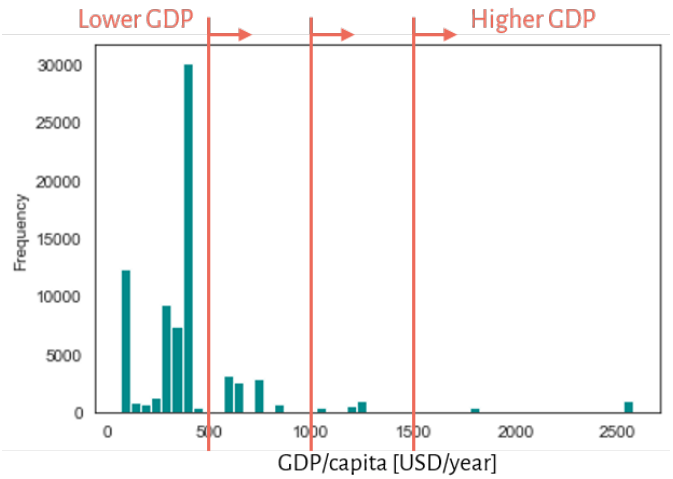


Figure 6.5: The three boundaries between a low and high GDP per capita that are tested.

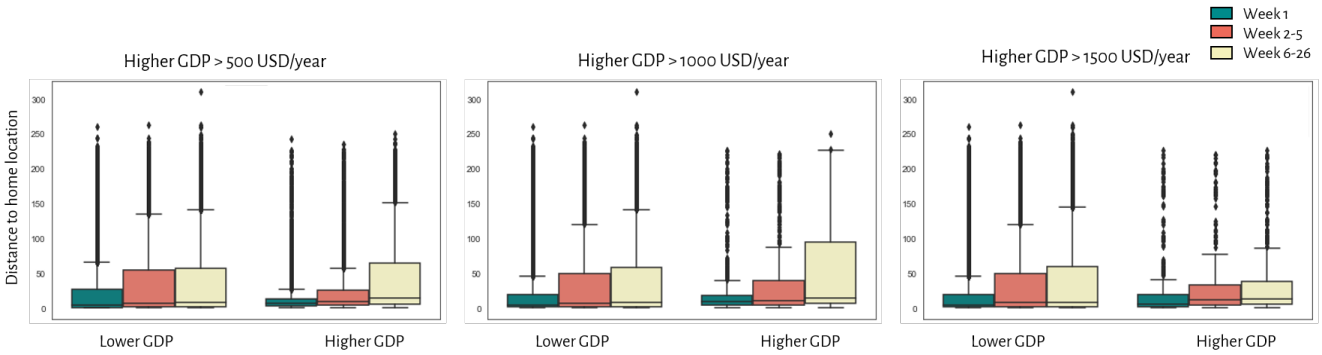


Figure 6.6: The effect of a low and high GDP per capita on the displacement distance on three different boundaries for high GDP in three different time-steps

Observation 6: People that live in areas where major disasters happen more often are expected to have a higher risk perception, which might lead to a rapid reaction on evacuation orders or to the adoption of preventive measures. Translation: a positive correlation between flood experience and distance travelled, especially in the first week.

The Red Cross Foundation has used the data of flood exposure to determine which areas in Haiti are more vulnerable to floods compared to others. In their assessment, the areas that have been exposed to floods appear to result in larger distances travelled in the first week after the disaster. This does confirm the observation. Still, the explanation for this result might be coincidental or dependent on the population. Therefore, we can only conclude that a difference in distance travelled can be noticed between areas that endure flood exposure and areas that do not.

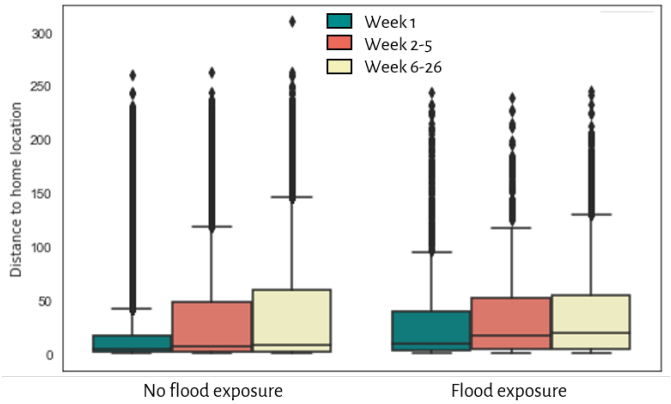


Figure 6.7: The effect of flood exposure on the displacement distance in three time-steps.

Observation 7: Social networks are expected to influence the choice of destination when relocating strongly. Translation: people with a low proportion of contacts at their home location go to a destination with a larger proportion of contacts.

The proportion of contacts located within a 10km radius of the home location and the displacement location are both tested for their effect on the displacement distance. As shown in fig. 6.8, the displacements for which there are more contacts at the displacement location than at the home location appear to move further away in the first week. However, only 10.3% of the displacements have a higher proportion of contacts at the displacement location than at the home location. This disproportionate distribution should be taken into account when analysing the effect of contacts. It can, however, be concluded that the proportion of contacts close to the locations have some effect on the distance travelled.

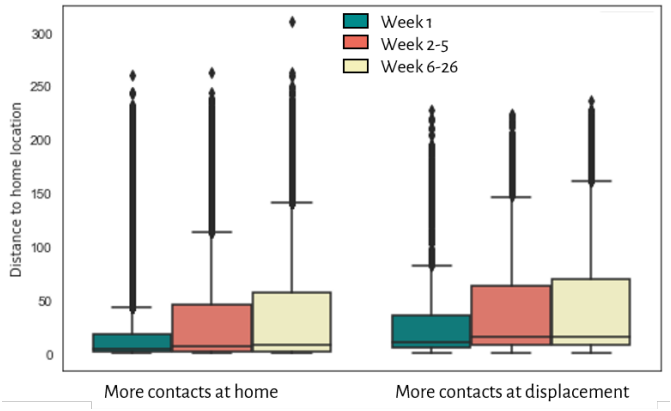


Figure 6.8: The effect of having contacts residing close-by on the displacement distance in three time-steps.

Observation 8: Community-driven mitigating measures are expected to be more successful in areas where disasters are still fresh in memory. In addition, they are expected to be more successful when trust in national authorities is low.

Even though this observation cannot be tested with the available datasets, the evaluation report of Grünewald and Schenkenberg (2017) indicate that the efforts of community-based measures were successful during the aftermath of Hurricane Matthew. Therefore, based on the literature, it is advised to continue the focus on local involvement when preventive and reactive measures are introduced.

Observation 9: The source from which the information comes that people receive determines their risk perception and thus influences their decision to leave or stay. Social media might play a significant role in information spreading, but the circumstances around online communication must be ideal. Seeing that, this way to work, people depend on the availability of internet and electricity. Therefore, radio broadcasts could be considered too.

The means of information spreading is not represented within available datasets. From the literature, it is clear that the communication network was severely damaged during the Hurricane (International Organisation of Migration, 2016). It followed that communication between organisations took place using radio and satellite phones (Grünewald & Schenkenberg, 2017). It does not become clear from the literature how the moving population received their information. The activity on social media, such as Twitter and Facebook, was way higher than in normal conditions (Martín, Li, & Cutter, 2017). A study about the use of Facebook in emergence campaigns showed that the Haitian population reacts actively towards sharply articulated Facebook messages (Arroyo-Almaraz, Calle Mendoza, & Van Wyk, 2018). In the Haitian population, Facebook is used to a significantly larger extend than Twitter, 89% and 4.1% respectively [1]. Since the political situation in Haiti was, and still is, relatively unstable (Grünewald & Schenkenberg, 2017), it is understandable that the influence of social media on the risk perception of the population is strong.

6.1.2 Results of the exploration

From the observation checks, it can be concluded that the following variables affect the distances within the displacements: proportion of contacts, the percentage of damaged buildings, the travel time to the nearest

[1] Source: https://gs.statcounter.com/social-media-stats/all/haiti.

city, the GDP per capita, flood exposure and variables that indicate the severity of the disaster (wind speed and rainfall in this case).

Apart from the observations checks, some other findings were interesting for analysing the mobility behaviour after Hurricane Matthew. fig. 6.9 shows that most displacements that took place were only between 0 and 5 km from the home location. These distances seem to follow an exponential decline trend. Also, most displacements took place within the same administrative area or further away than neighbouring areas (see fig. 8.6). Only a small proportion of the displacements took place between neighbouring areas.

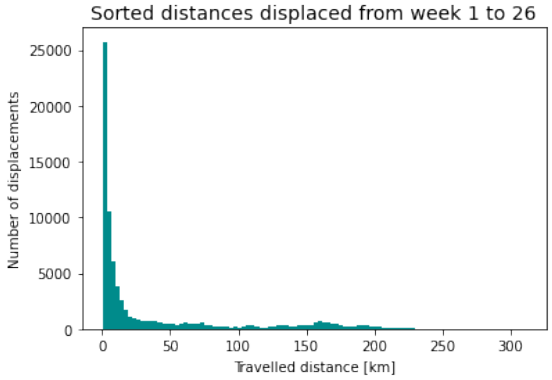


Figure 6.9: Total distances travelled between week 1 and 26 after hurricane Matthew

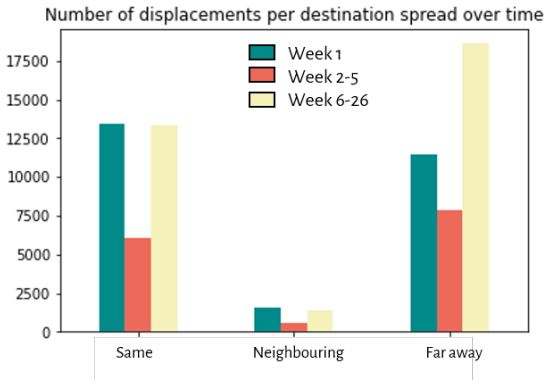


Figure 6.10: Displacement destinations and their different counts in the three time-frames

The case of Hurricane Matthew in Haiti confirms that the theoretical foundation in literature can, to at least some extent, explain a change in population movement after a disaster. Although some of the trends can be found in the data, the underlying drivers remain guessed. The variables that appear to influence the population movement the greatest are the ones that were expected too, meaning that this case follows previously observed displacement tendencies.

6.2 Features and classes

The features and classes are selected from the produced dataset in chapter 5. The choices that underlie the selection of fitting features and classes are elaborated upon in this section, as every choice comes with the necessary implications for the results. First, the feature selection is presented, then the class selection, followed by an exploration of their combined workings.

6.2.1 Feature selection

Features within the available datasets are considered, and the selection of the to be included features is based on two considerations: the data for the feature must be available at the time of the disaster. They must at least have shown any indication that they influence the movements of a population in a disaster situation. These considerations make it that the displacement location variables are not included, as the displacement location is unknown at the time of the disaster.

The features that have been shown to affect the distance covered within the displacements are displayed in table 6.1, as well as the variables available at the time of the disaster or closely after. In the table, the variables

for the home location are shown with their connectivity to the travel distance and their availability around the start of the disaster.

Table 6.1: The availability of features compared to the connection to travel-distance found

Home location features	**Connected to travel-distance**	**Available at disaster start-time**
Median calls		X
Known locations close-by		X
Radius of gyration		X
Entropy		X
Contacts close-by	X	X
Windspeed during disaster	X	X
Proporty damage during disaster	X	X
Rainfall during disaster	X	X
Distance to main road		X
Distance to urban center	X	X
Population density	X	X
GDP per capita	X	X
Flood exposure	X	X
Travel time to nearest city	X	X
Timestep	Pursued	Pursued
Distance to home location	Pursued	Pursued

Several features contain the same sort of information within the table: the distance to an urban centre, population density, and travel time to the nearest city all say something about the location being in a rural, suburban or urban area. The population density could arguably be most detailed, however, in this dataset, the values are clustered, and the details are thus lost. For the choice between the distance to the nearest urban centre or the travel time to the nearest city, the availability of the data is highly valued and therefore the data that the Red Cross foundation already holds is selected [2]. This leads to the following selection of features with their feature name:

- Contacts close-by the home location - **Contacts**
- Windspeed at the home location - **Windspeed**
- Rainfall at the home location - **Rain**
- GDP per capita in the admin 2 area of the home location - **GDP**
- Flood exposure of the home location - **Exposure**
- Travel time to the nearest city from the home location - **City-travel**

6.2.2 Class preparation

Both the time and distance to the home location are pursued features, meaning that they are the classes that are fitted to the feature data. A distinction is made based on the time-steps, meaning that the datasets of week 1, week 2 to 5, and week 6 to 26 are separately analysed. And the distance to the home location is converted to

[2] The influence of replacing the travel-time with the population density has been tested too, but the difference on the model performance was considered insignificant

class data instead of continuous data. The most important justification for this choice is that the continuous data gives the impression of specificity that is not expected to represent the measured level of detail. The high fallout rate of cell towers, combined with the finding that most displacements were around 5km away from the home location, makes the reliability not match the detail level. The assigned classes are the following: the displacement takes place from and to the same admin 2 area (*same*), to a neighbouring area (*neighbour*) or to a location that is further away than the neighbouring areas (*far*). The conversion of continuous data to class data is done by using the assigned locations to the displacement data; an example is presented in fig. 6.11. The areas that share a boundary with the area of the home location are considered neighbouring areas.

Home location	Displacement location	Classname	Class
Abricots	Abricots	Same	0
Abricots	Bonbon	Neighbour	1
Abricots	Port-au-Prince	Far	2

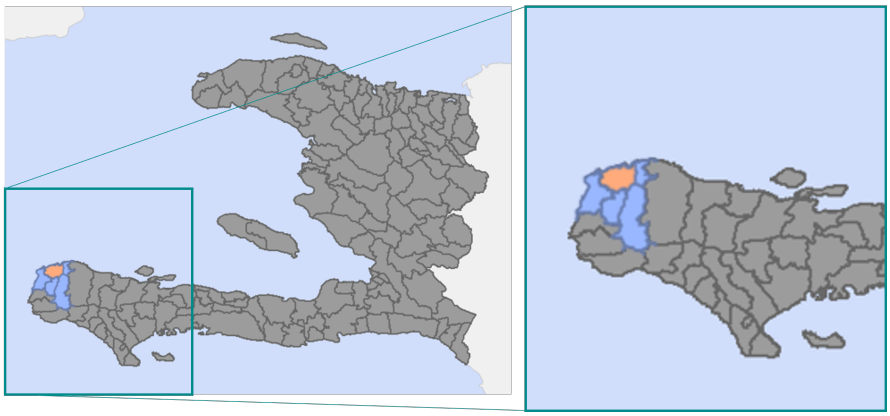


Figure 6.11: Example of class assignment for displacements from Abricot

6.2.3 Feature and class exploration

As a first test, the correlation between the selected features is tested. We see that they are all poorly correlated to each other (see fig. 6.12), and it can thus be expected that a linear classifier will not perform well on this data.

A first check on the features within every destination class reveals some interesting patterns already. From table 6.2, it seems that the displacements that take place within the same administrative area have a higher proportion of close contacts nearby and a higher GDP per capita. The latter could be because the people who come from areas with a higher average GDP per capita might be more willing to stay in their own area than leave for an area with possibly lower GDP per capita. And evidently, the distance travelled is correlated to the destination, as expected.

The same check is applied to grouping the time frames. Table 6.3 shows that the features show less differentiation between the groups than the destinations. An interesting difference, however, seems that the distance travelled increases over time, however, this display is misleading, considering the difference in the size of the time frames. The first time frame represents one week, the second represents four weeks, and the last represents 21 weeks. The distance that people can travel within a larger time frame is logically greater.

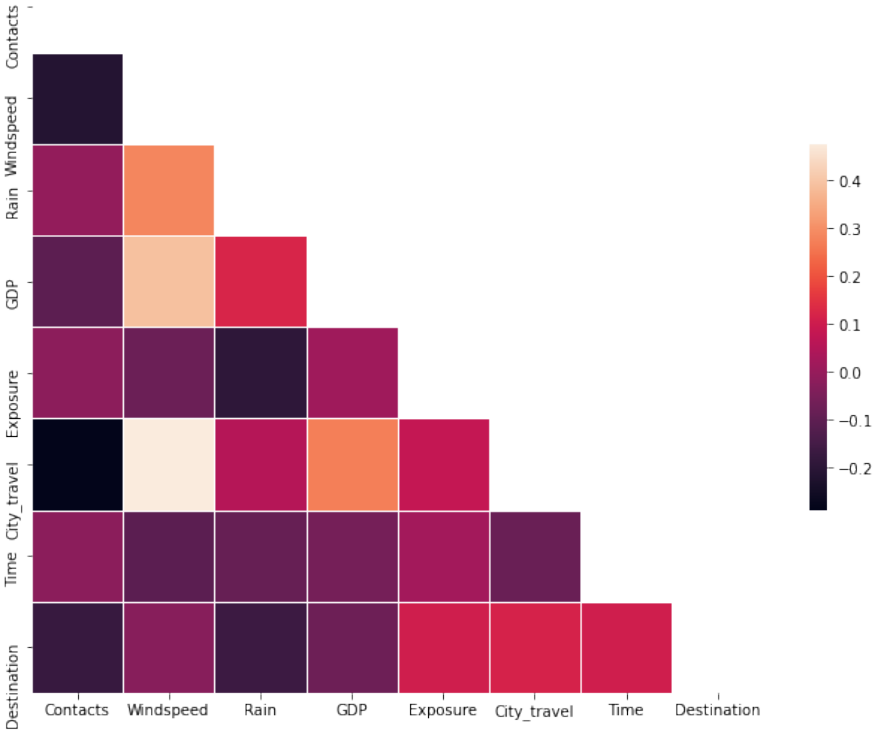


Figure 6.12: Correlations between all included features

Table 6.2: Average values for the features distinguished by destination class

Destination	Contacts	Windspeed	Rain	GDP	Exposure	City_travel	Distance
0	0.68	59.08	511.19	415.50	0.00030	64.72	3.79
1	0.53	61.78	519.42	374.00	0.00089	131.06	15.08
2	0.57	57.68	481.98	382.58	0.00148	82.05	67.00

Table 6.3: Average values for the features distinguished by time step

Time	Contacts	Windspeed	Rain	GDP	Exposure	City_travel	Distance
1	0.62	61.74	506.44	415.13	0.00080	84.57	29.18
2	0.61	57.95	492.62	389.59	0.00094	75.05	39.55
3	0.61	56.17	490.49	385.04	0.00105	71.41	41.69

6.3 Key Findings of Chapter 6

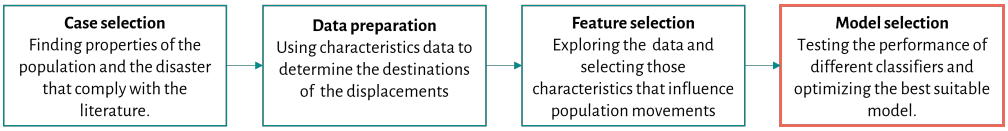
Even though some of the observations from the literature review in chapter 2 do not show the expected trends, the characteristics that were expected to affect population movements appeared to do so. From the data exploration that followed the data preparation, the affecting characteristics have been selected to be features. These features will be used in the predictive model: contacts, wind speed, rain, GDP, Exposure and City-travel.

The ones that were identified are used to classify population movements based on their destination. The information that the travel distance contained has been used extensively in the data exploration, but the continuous variable is not suitable for the predictive model. Firstly, because the distances in the data give the impression of specificity, that is not likely to be achieved seeing the data collection process. And secondly, because the distance data does not contain geolocation data, while the classes do. Lastly, the destinations of the displacements are of vital importance for humanitarian aid organisations that want to know more about where and when to expect IDP's that need shelter.

A distinction of three classes have been initiated are: the destination is the same admin 2 area, the neighbouring admin 2 area or further away. Because many of the features are spatially correlated, the first two classes could also be put together. In this study, they will, however, be kept separately. The chosen features and classes have shown low correlation scores, meaning that it is expected that linear classifiers will not perform well on this data. In chapter 7 the best performing classifier will be pursued.

Model selection

7



"All models are wrong,
some are useful"

George Box - 1976

An analysis of the effects of different drivers on the population movements can confirm expected patterns or even reveal new ones. A conceptualisation of the CDR data that is available for Hurricane Matthew has done so already, but the usage of CDR data is complex and challenging. The sensitivity of the data complicates data sharing, and the found solution for this problem - aggregation of the data - leads to the inevitable loss of information. This means that it is not only important to derive as much information as we can from the available CDR data, it is also potentially rewarding to see if other datasets could stand-in for the CDR data.

In this chapter, the conceptualised model will be specified. This basically means that the conceptual model is translated to a computational model that represents real-world patterns to the best of its ability. Therefore, the predictability of CDR data is studied, using several features that have been identified in the previous chapter. The purpose of the model is to recognise patterns in the mobility behaviour of a population during crises and find which features are needed to find these patterns. Lessons learned from data exploration and the data preparation in chapter 5 are intensively applied during the process. Building a predictive model requires many choices, all of which could have far-reaching consequences for the outcome. Research question 4 guides the specification of the predictive model. In the first section, section 6.2, the first sub-question (a) is addressed, followed by the second sub-question (b) in section 7.1. The third sub-question (c) is discussed in section 7.2. In section 7.3, research question 4 is answered.

Research question 4

What are the most important requirements for the specification of a predictive model?
(a) Which methods represent the real-world working of the system best?
(b) How can the model be optimised?

7.1 Method selection

For the classification of labelled data, a supervised classifier is pursued. Some suitable methods are applied to the data to test their performance in order to make an informed decision of which model to apply. The different classifiers are introduced, described and applied. The first application sets the basis for the performance comparison between the classifiers, as many practices can be used to improve the performance of a classifier.

7.1.1 Method introduction

Logistic regression. Although the name includes 'regression', Logistic Regression is a classification model. It uses a logistic function to label classes, with the probability as output. The method acts somewhat similar to a linear regression function, with the difference that a stashing function is used over the regression output. In default, the logistic regression uses a Sigmoid function to determine one of two possible outcomes. An example is shown in fig. 7.1, which visualises the distribution of a Sigmoid function, where the z value represents the linear regression function eq. (7.1). In the case of Hurricane Matthew, there are three possible outcomes, and we, therefore, need a multinomial logistic regression function to classify the features. The Sci-kit Learn package in Python offers the possibility of using a multi-class logistic regression model. Another strategy for the inclusion of multiple classes is the One-vs-Rest extension.

$$z_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i \tag{7.1}$$

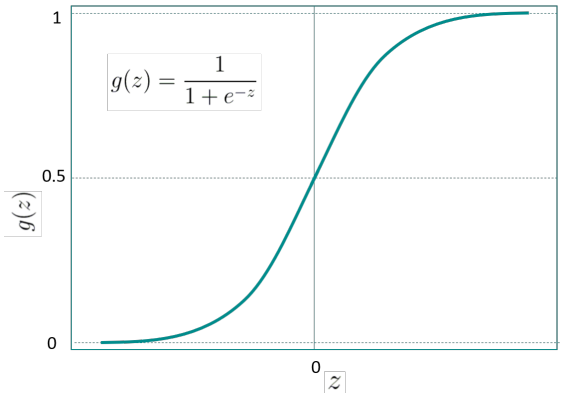


Figure 7.1: Example of the Sigmoid function

The advantages of logistic regression are that they are easy to interpret, run fast and take the direction of the features towards the classes into consideration. However, it cannot be applied to non-linear classification problems, meaning that the classifier is not expected to result in a high accuracy score for this case.

The needed assumptions for the method are the following:

- The dependent and independent variables are linearly related.
- The variance of the errors in the training data needs to be constant, at least to some extend.
- The independent variables should not be co-linear.

Decision Tree. The decision tree classifier is used for the classification of independent variables through nodes. Every node contains a condition that determines the next node until a leaf node is reached (see fig. 7.2). In the leaf nodes, the output is predicted. These diagrams are well known, as they are used in the most diverse decision situations. The crux here is to find the right condition to shorten the paths towards the leaf nodes, and the algorithm can use the entropy to maximise the information gain within every decision node. Entropy is calculated with eq. (7.2), where p_i is the frequency probability of class i. The information gain within a node is calculated by eq. (7.3), which simply expresses that the greater the reduction in entropy, the more information is gained about Y from X. A visual representation of the workings of entropy for the decision nodes is shown in fig. 7.3.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{7.2}$$

$$IG(Y, X) = E(Y) - E(Y \mid X) \tag{7.3}$$

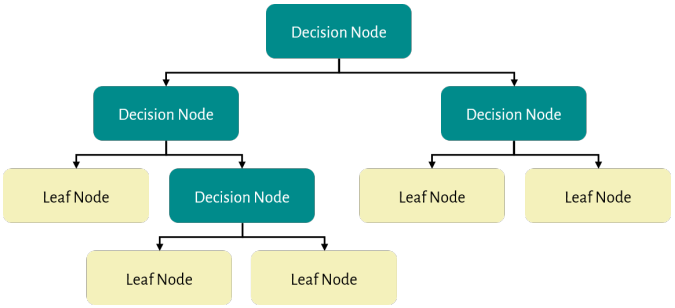


Figure 7.2: Visualisation of the paths in a decision tree

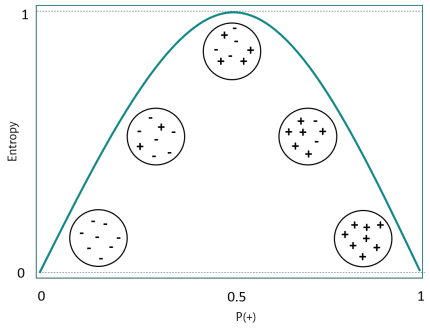


Figure 7.3: Visualisation of the entropy for the probability of + data

The advantage of using a decision tree classifier is that no assumptions on the data distribution are needed. It can also handle co-linearity and the predictions are relatively understandable. On the other hand is there the disadvantage of overfitting, as the sweet spot between purifying the output in the leaf nodes and the usability of the model is hard to find. A decision tree quickly grows very large, but the maximum depth can prevent this from happening. An optimisation practice is to discover the optimal value for the maximum depth or apply a random forest in which many decision trees determine the output of a set of features through majority voting.

K-nearest neighbour. The k-Nearest Neighbour (KNN) classifier uses the distance between features to predict the output. To do so, it uses a local approximation to find the class of a new data point and assumes that the test points closest do describe the datapoint best. It looks at the k number of closest points and uses majority voting to determine the class. A visual representation of the effect of choosing k is shown in fig. 7.4; for $k = 3$ the black dot will be assigned to class Red-Cross, while for $k = 7$ the black dot is classified as Green-Triangle.

The advantage of using a KNN classifier is that it is an easy and simple model, and the only parameter that can be tuned is the number of k's that are included. This automatically brings us to the disadvantage: the choice for k potentially changes the classification completely. It also takes a lot of computation time when the sample size is large. And the inclusion of different features requires proper scaling, as the distances are otherwise potentially unfairly treated.

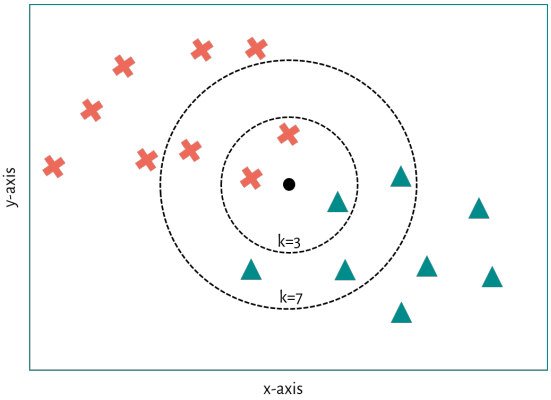


Figure 7.4: Visualisation of the effect of choosing k in the KNN-classifier; for k = 3, the black dot is classified as "red cross", while for k = 7 the dot is classified as "green triangle".

Naive Bayes. The NB classifier was introduced in chapter 5, as it was used to assign the geo-locations to a set of features. The model is based on Baye's rule of probability, as shown in eq. (7.4), where $p\left(x_i \mid y\right)$ is the posterior probability, $p(x)$ and $p(y)$ are the probability for x_i and y and $p\left(y \mid x_i\right)$ is the likelihood of y when we know x_i is true. The most applied methods within NB are the Gaussian classifier (assuming that the features follow a Gaussian distribution) and the Multinomial classifier (assuming a multinomial distribution). The difference in outcome that these two classifiers bring about is thoroughly analysed in chapter 5. For this test, the Multinomial classifier is applied.

$$p\left(\mathbf{x_i} \mid \mathbf{y}\right) = \frac{p\left(\mathbf{x_i}\right) p\left(\mathbf{y} \mid \mathbf{x_i}\right)}{p(\mathbf{y})} \tag{7.4}$$

The assumption for the NB classifier is that all features are independent, meaning that the other features can be excluded without consequence. This is a theoretical assumption since features cannot be completely independent. An advantage of using NB is that it can be pretty accurate with limited training data. Furthermore, it converges faster than other discriminative models when the independence condition is satisfied, and it supports binary and multi-class classification problems. A major disadvantage is that the independence of the features means that the classifier cannot represent the real world correctly.

Support Vector Machine. The Support Vector Machine (SVM) technique can be used for regression and classification and can represent both linear and non-linear problems. In this case, a non-linear SVM classifier is applied, as the correlation between the features appeared weak in section 6.2. This method uses a kernel function to derive a new hyperplane so that the training data will be linearly separable, such as in fig. 7.5. Then, the linear curve classifies the labels in the hyperplane.

The advantages of the SVM classifier are that it is possible to solve complex problems, and it handles outliers quite easily. A disadvantage is that the training time takes longer for larger datasets. For a multinomial problem, such as is the case here, the degree can determine the fit of the hyperplane. For degree = 1, the kernel is linear, while higher values for the degree make the kernel more flexible. The choice for the degree can also be used to optimise the accuracy of the classifier.

In the case of the displacement data for Hurricane Matthew in Haiti, the four different kernels have been

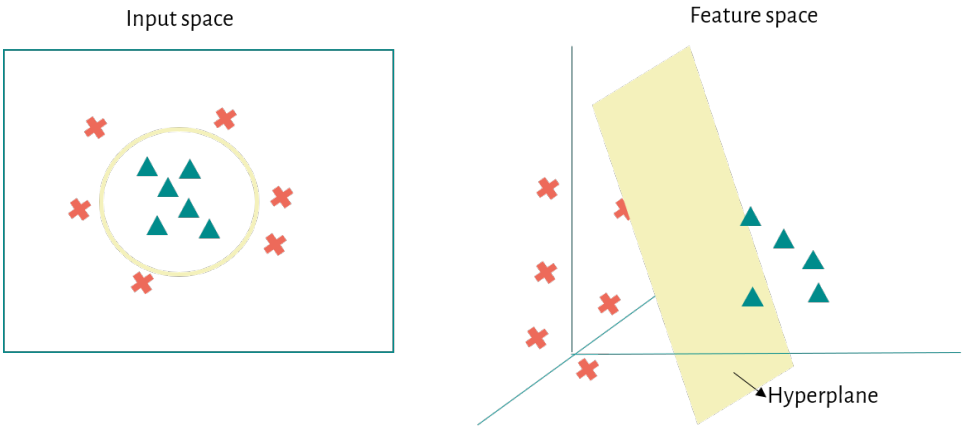


Figure 7.5: Visualisation of a non-linear hyperplane for the Support Vector Machine classifier

tested: linear, polynomial, Gaussian and Sigmoid. As the run-time for the linear, Gaussian, and Sigmoid kernels are quite long, the comparison between the four kernels has been conducted on a smaller training set. The accuracy scores for the classifiers including all features are the following: linear kernel, 55.4%; Gaussian kernel, 58.1%; Sigmoid kernel, 43.5%; polynomial kernel, 59.2%. The differences are not spectacular, but the polynomial kernel is the highest and therefore selected for the additional reason that the kernel seems to fit the data better than the other kernels.

For the polynomial kernel, the degree needs to be determined as well. When including all the features, the optimal value for the degree is 3 or 4 (table 7.1). The accuracy score may, however, vary when more or fewer features are included. Therefore a combination of the two is ran, in order to see if patterns can be found. The result is shown in section 7.1.1, and it shows that the accuracy score stays relatively low, but the degree and choice of included features can improve the accuracy by 12.5%. Overall, the trend is visible where the higher the degree and number of features, the higher the accuracy score. The optimum is found when including all features and using 5 of 6 degrees and including five features while using 9 and 10 degrees. As the increase of degrees does simultaneously increase the chance of overfitting, the model comparison will include six features and 5 degrees in the polynomial kernel of the SVM classifier.

7.1.2 Comparison of the methods

To test the accuracy between the classifiers, the F1-score for accuracy score. This score checks the proportion of mislabelled displacements within the test set, as shown in eq. (7.7). All combinations of the features are tested to see if they are all needed to predict the destination. The tables, including all accuracy scores, are shown in appendix A. From the comparison, it can be concluded that the Decision Tree classifier and the KNN classifier perform best, and the Decision Tree performs even a bit better than the KNN classifier. Although a Decision Tree is faster than a KNN classifier, the outcomes of the KNN classifier are better generalisable, and the results are more easily interpreted.

$$\text{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{7.5}$$

$$\text{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{7.6}$$

$$\text{F1 Score } = 2^{*}(\text{ Recall } * \text{ Precision}) \,/\, (\text{Recall } + \text{ Precision}) \tag{7.7}$$

Table 7.1: The effect of degree on the accuracy of the Support Vector Machine including all features

Degree	Accuracy
1	43.6
2	45.1
3	59.2
4	59.2
5	56.8
6	56.8
7	58.7
8	58.7
9	58.7
10	58.7

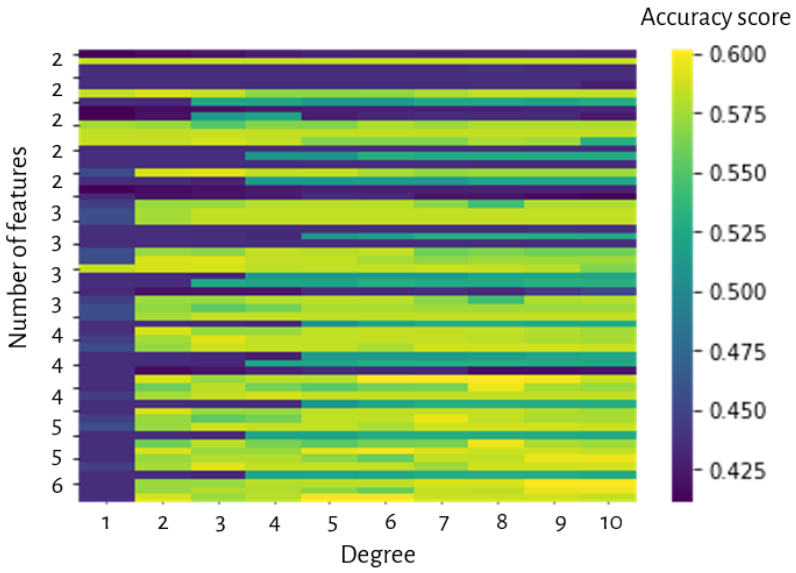


Figure 7.6: The effect of degree on the accuracy of the Support Vector Machine including all features

7.2 Model optimisation

To strengthen the justification for a choice of model, the Decision Tree classifier and the KNN classifier are optimised to a further extend. Ideally, all the classifiers are optimised, which could be considered in future research.

7.2.1 Decision Tree optimisation

For the Decision Tree, this means that first, the criterion on which the conditions in the decision nodes is based is changed, from "Entropy" to "Gini". The Gini index calculates how well the data points are mixed together; when a set is unequally mixed, the Gini index will be maximum. The Gini criterion, however, does not lead to an improvement of accuracy using the data of week 1, but a slight improvement is seen when applied to the data of week 2 to 5 and week 6 to 26 (see fig. 7.7).

Also, the maximum depth value was tested, leading to the conclusion that the accuracy is optimised when there are between 9 and 11 decision levels. This leads to an almost impossible interpretation of the choices that were made within the nodes to maximise the output accuracy. To illustrate the effect of adding more levels, a visual representation of the decision and leaf nodes is provided in fig. 7.8 and fig. 7.9, showing that the readability is problematic.

Lastly, a Random Forest classifier is applied. This is an ensemble model that uses a combination of Decision Trees to strengthen the output. It, therefore, becomes more robust and accurate and is better armed against overfitting. The principle of a Random Forest is that multiple Decision Trees are obtained, of which the outcome for the classification is determined by majority voting. The interpretation of the predictions from a Random Forest is more complex even than from a Decision Tree. It appears that the accuracy score of a Random Forest consisting of 100 Decision Trees remains around the same values found by single trees, meaning that the precision gain is minimal. However, the confidence in the accuracy is strengthened.

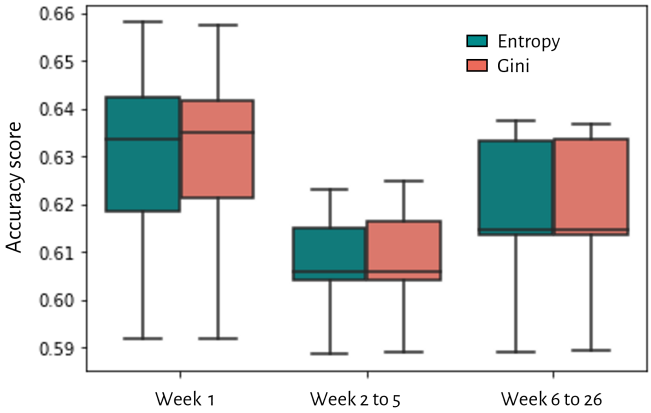


Figure 7.7: Boxplot of the accuracy scores of the Decision Tree classifier with the distinction of the used criterion: Entropy or Gini.

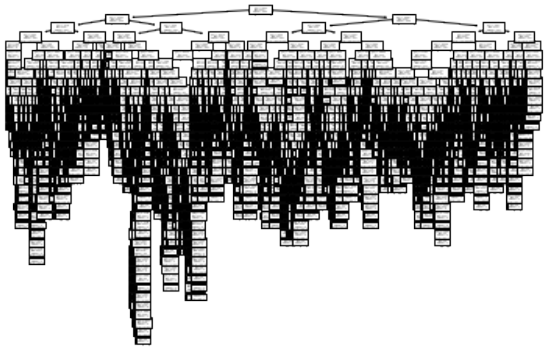


Figure 7.8: Visualisation of a Decision Tree without a maximum depth setting

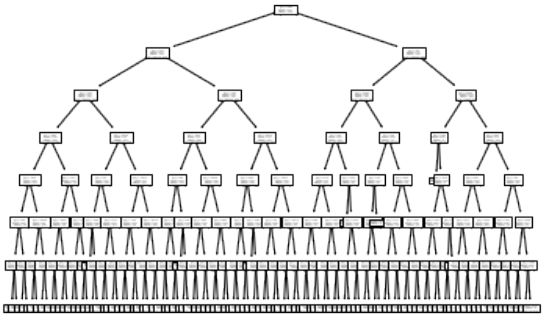


Figure 7.9: Visualisation of a Decision Tree with a maximum depth of 7 levels

7.2.2 K-Nearest Neighbour optimisation

For the KNN classifier, one major optimisation practice is the determination of k. Testing the effect of k on the accuracy score has been done by including all the available features, as the score for 6 features was among the highest for the three-time distinguished datasets. A range from $k = 0$ to $k = 100$ results in the accuracy growth as shown in fig. 7.10. The optimum for the data of week 1 was found for $k = 35$ and led to an accuracy score of 65.8%.

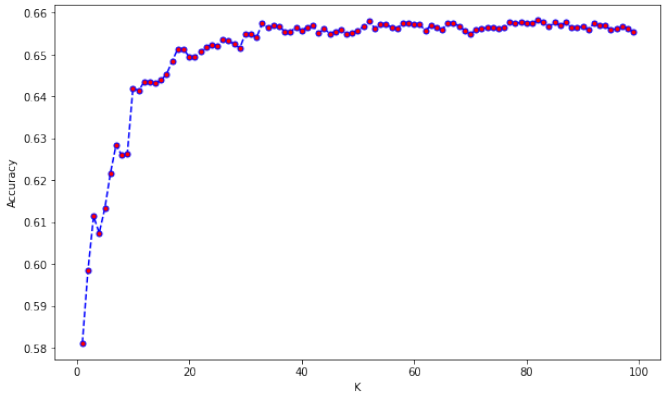


Figure 7.10: K versus accuracy for all features of the week 1 dataset included in the KNN classifier

Using the previously determined number of neighbours, a K-Fold Cross-Validation technique is applied to strengthen the accuracy of the KNN model. The model is usually trained on a part of the dataset and tested on the remaining part. In order to make sure that the accuracy is not found by some lucky split between the training and testing set, both are randomly selected, and the model is run multiple times. A visual representation of how the technique works is displayed in fig. 7.11. In this case, ten iterations are used with $k = 35$. The mean accuracy for the cross-validation resulted in 66.0%, which is similar to found accuracy in one iteration. This could be expected, as the dataset is rather large and the alterations within the training and testing set when splitting the data differently are thus minimal.

As a formalisation, also a technique called Grid Search was applied. This technique combines the previous two optimisations by searching for the optimal value for k while using Cross-Validation too. The algorithm shows slightly different results. The optimal number of neighbours is determined to be 45, and the accuracy score then results in 67.2%.

An accuracy score of 67.2% is not considered very high. Mainly because there is a large dataset available for training, and the labels are only distinguished in three options. Seeing that the KNN classifier considers the distance between data points, the large values (such as for rain and wind) might distort the classification. Hence, three different scalers are tested: standardisation, Min-Max scaling and Robust scaling.

Standardisation. The Standard scaler of the Sci-kit Learn package within Python removes the mean and scales the data to unit variance. The scaler does not work optimally with outliers, as they influence the mean and standard deviation. By shrinking the range of the data, these details might thus be lost. Using a standard scaler results in a maximum accuracy score of 66.4%.

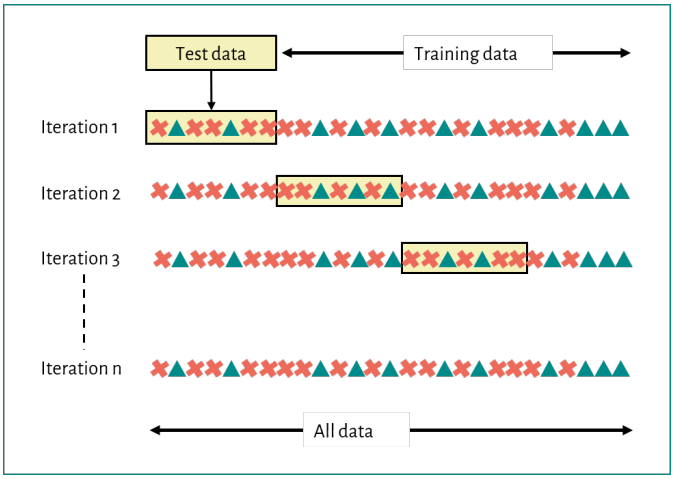


Figure 7.11: Visualisation of the workings of K-Fold Cross Validation where the subset of test data is differed in each iteration.

Min-Max scaling. The Min-Max scaler too is very sensitive for outliers, as this scaler rescales the data between 0 and 1. When outliers are present, the effect is that inliers are all compressed into a narrow range. Usage of the Min-Max scaler results in a maximum accuracy score of 66.6%.

Robust scaling. A robust scaler is based on percentiles and is thus less influenced by outliers. This results in a larger range than the Standard scaler and the Min-Max scaler but still give similar results. This is also reflected in the maximum accuracy score, which has come to 66.3%.

7.3 Key findings of Chapter 7

The highest model performance was produced by the KNN classifier, which, therefore, will be selected. But this is not the only reason for the selection: the KNN classifier is also easier to interpret and has a lower chance to be overfitting the data. Several optimisation techniques have led to an accuracy score of 67.2% for a model including all the features. This is a moderate score for a predictive model for which large training datasets are available.

The limitations of the model performance are possibly posed by the level of aggregation within the datasets, leading to identical feature values representing different labels. When this occurs, the training of the model can only go to the point where it calculates the chance of a label appearing. A solution to this problem could be to add more features to classify the data, as this would expectantly increase the differences in feature combinations. It could increase the accuracy score significantly.

Extra features that could be considered are the IDP's Entropy and Radius of Gyration. These features were not considered as this data is strongly connected to the CDR data and cannot serve as a replacement for it. In the case that CDR data is available, however, the effect of these extra features could be interesting. And although no clear connection was found between the known locations close-by the home location of the IDP and their mobility changes, this information could still be included in the predictive model.

The most critical requirements for the specification of a predictive model appear to be a highly descriptive set of features and a lot of computing power. The model is not fit to predict population movements to a high

extend, but the process of method selection has also provided more profound insights into the data requirements. The level of detail required in the data to perform a classification is considerably high. Clustered data is, therefore, less fitting for this application and should thus be avoided.

Results

8

After the specification of the best performing model for this data, the application of the classifier produces results. In this chapter, the results are presented and discussed so that they help better to understand the predictability of population movements during disasters. In the first section, sub-question (a) is addressed. The second section is dedicated to the importance of feature inclusion; sub-question (b) and (c) are addressed. Here, the model results using a subset of the features is analysed. It might help to verify or reject the expectation that the population movement patterns do change over time. Research question 4 is addressed in this chapter.

Research question 5

To what extent can the population movement caused by a disaster be predicted?
(a) How well does the model perform?
(b) Which features are most important for the prediction?
(c) Does the model change over time?

8.1 Model performance

The overall performance of the KNN classifier is shown in table 8.1, where the accuracy f1-score shows to be 67% for the first time-frame (week 1 after hurricane Matthew). The accuracy score is 62% for week 2 to 5 after the disaster and 64% for week 6 to 26 after the disaster. An interesting observation is that the precision of label 1, the destination is a neighbouring area, is lower than that of label 0 and 2; respectively the destination is the same area and further away than the neighbouring area. The support is lower because the data is imbalanced: in the first time-frame, only 297 displacements were made from an area to a neighbouring area, while over 2000 observations represent the other two labels.

We see that the model does not well classify overall the displacements to neighbouring areas. The class is relatively underrepresented, and therefore the model is not well trained in distinguishing the class from the other two. This is also confirmed in the confusion matrices in fig. 8.1. It also is notable that for the second time-frame, week 2 to 5, not one displacement was classified to a neighbouring area (class 1). The reason could be that in the training set, no representatives were present for this class. As a solution, the selection of the training set could be changed, for example, by using a K-Fold Cross Validation algorithm.

Table 8.1: The precision, recall and F1 score of the KNN classifier for the three separate time-frames (T1 = week 1, T2 = week 2 to 5, T3 = week 6 to 26

		Precision			Recall			F1-score			Support		
		T1	T2	T3	T1	T2	T3	T1	T2	T3	T1	T2	T3
Destination	0	0.67	0.58	0.59	0.77	0.55	0.59	0.71	0.56	0.59	2667	1200	2676
	1	0.55	0	0.41	0.22	0	0.03	0.31	0	0.06	297	111	278
	2	0.67	0.64	0.67	0.61	0.72	0.72	0.64	0.68	0.69	2323	1592	3728
	Accuracy							0.67	0.62	0.64	5287	2903	6682

Another solution for this problem could be to oversample or undersample the input data. That way, the difference in representations by the three classes is accounted for. For both techniques, several methods can be applied. Oversampling means the creation of dummy data that does not change the measurements within the dataset. For example, the mean, standard deviation, and presence of outliers are simulated. Undersampling often means that a subset of the overrepresented classes is used to train the model. A disadvantage of undersampling is that important information can be lost when only using a small proportion of the available data. In appendix B, oversampling and undersampling techniques are applied to the KNN classifier; it shows that the model performance is not improved by it. Therefore, these techniques are not recommended in this case but could be helpful to consider in other cases.

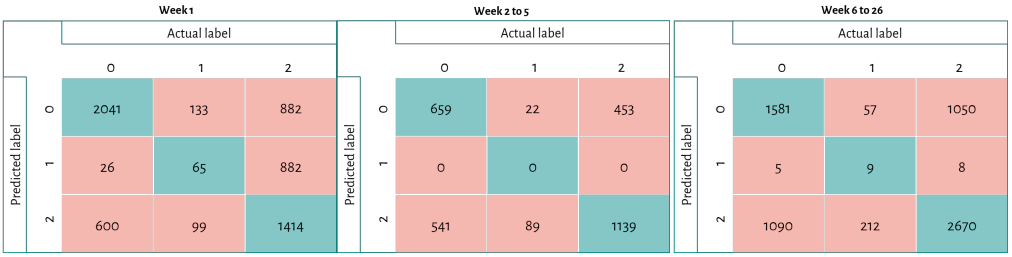


Figure 8.1: Confusion matrices of the KNN classifier of all three datasets

8.2 Importance of feature-inclusion

The model takes all six selected features into consideration when determining the fitting label, but they might not all be of the same importance. Some features might even distort the classification. That is why the model is run with all possible different combinations and numbers of features.

For the displacement data of all three time-steps, it is clear that the inclusion of all features does not provide the optimal outcome (see table 8.2, table 8.3 and table 8.4). In addition, the order of importance differs between the time-frames:

- **Week 1:** Contact = City-travel > Wind speed > Exposure > Rain > GDP
- **Week 2 to 5:** Contacts = GDP > Wind speed = City-travel > Rain > Exposure
- **Week 6 to 26:** Contacts = City-travel > Wind speed = GDP > Rain > Exposure

Apparently, all models show that some features do not add to the accuracy score or even lower it. In week 1 after the disaster, the wind speed, exposure to floods and rainfall do not add information on the population movements. And the GDP per capita even decreases the accuracy score. In week 2 to 5, the flood exposure adds no value to the accuracy, and the features wind speed, GDP per capita and rainfall only add a minimal amount of accuracy to the model. In week 6 to 26, flood exposure is not adding any information on the destination of the population, and the rainfall and wind speed only add a minimal amount to the accuracy score. Most important are the proportion of contacts living close-by and the travel time to the nearest city.

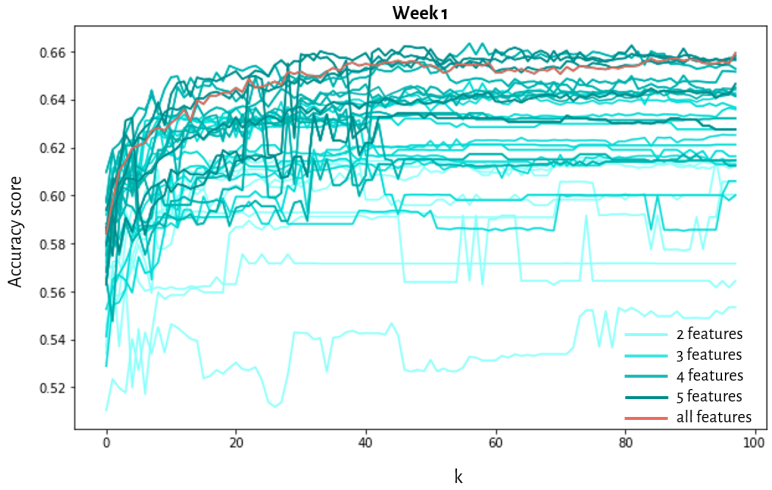


Figure 8.2: The effect of feature inclusion on the accuracy score for different values of k in week 1.

Table 8.2: The highest accuracy scores for every number of features included in week 1

Number of features	Features included	Highest accuracy score
2	Contacts, City-travel	64.7%
3	Contacts, wind speed, City-travel	66.3%
4	Contacts, wind speed, Exposure, City-travel	66.3%
5	Contacts, wind speed, Rain, Exposure, City-travel	66.3%
6	Contacts, wind speed, Rain, GDP, Exposure, City-travel	65.9%

When maximising the accuracy score while minimising the number of features, the models change somewhat. For week 1, the included features are: Contacts, Wind Speed and City-travel. For week 2 to 5 they are: Wind Speed, Rain, GDP and City-travel. For week 6 to 26, the features included are: Contacts, Wind Speed, Rain and GDP.

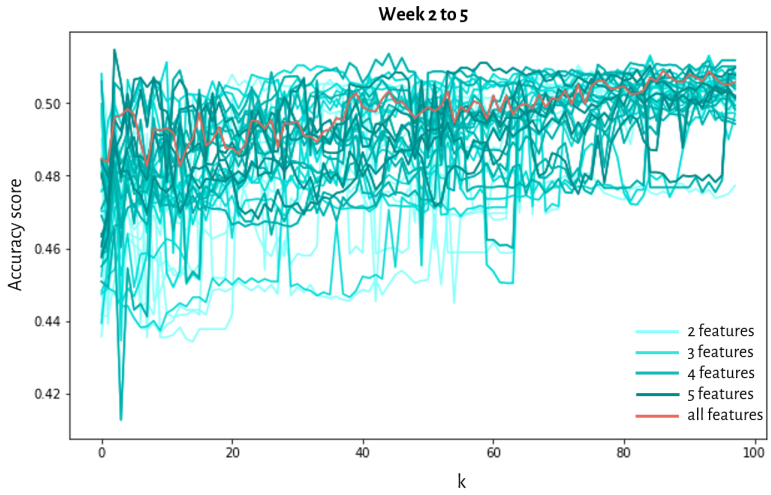


Figure 8.3: The effect of feature inclusion on the accuracy score for different values of k in week 2 to 5.

Table 8.3: The highest accuracy scores for every number of features included in week 2 to 5

Number of features	Features included	Highest accuracy score
2	Contacts, GDP	51.3%
3	wind speed, GDP, City-travel	51.4%
4	wind speed, Rain, GDP, City-travel	51.5%
5	wind speed, Rain, GDP, Exposure, City-travel	51.5%
6	Contacts, wind speed, Rain, GDP, Exposure, City-travel	50.9%

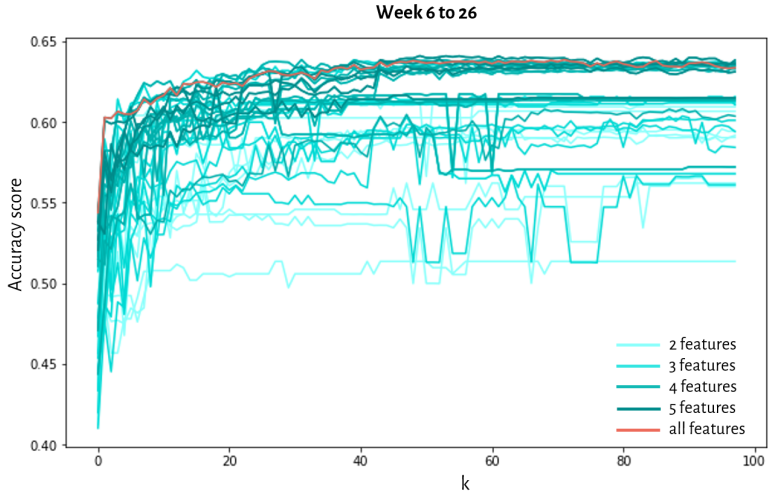


Figure 8.4: The effect of feature inclusion on the accuracy score for different values of k in week 6 to 26.

Table 8.4: The highest accuracy scores for every number of features included in week 6 to 26

Number of features	**Features included**	**Highest accuracy score**
2	Contacts, City-travel	63.8%
3	Contacts, wind speed, GDP	63.9%
4	Contacts, wind speed, Rain, GDP	64.1%
5	Contacts, wind speed, Rain, GDP, Exposure	64.1%
6	Contacts, wind speed, Rain, GDP, Exposure, City-travel	63.8%

8.3 Effect of the extra variables

The addition of the extra variables has a mixed effect on the accuracy of the movement prediction. For the original variables, only those are selected that were already present in the aggregated dataset: contacts, wind speed, rain and population density. The extra variables are considered those that have been added based on the geographic locations. section 8.3 shows that the addition of these variables has a positive effect on the accuracy when the number of features is low. However, the difference is less pronounced when more features are added. Except for the displacement prediction in week 2 to 5, the original features add a decent percentage to the accuracy.

Table 8.5: A comparison of the highest scores when including 2, 3 or 4 features between the original variables(Contacts, wind speed, Rain and Population density) and the extra variables (original variables plus GDP per capita and Flood exposure)

	Week 1		Week 2 to 5		Week 6 to 26	
Number of features	Extra	Original	Extra	Original	Extra	Original
2	64.7%	60.4%	51.3%	57.8%	63.8%	55.1%
3	66.3%	64.2%	51.4%	60.0%	63.9%	59.0%
4	66.3%	66.0%	51.5%	60.5%	64.1%	59.5%

8.4 Addition of spatial information

When plotting the spatial population movements of the first week, the potential of mapping the geographic locations of IDP's becomes apparent. As shown in fig. 8.5, most displacements after the disaster took place within the Ouest department. In this department, also the capital Port-au-Prince is situated. The consequences of the disaster were felt hardest in the areas Grande Anse, Nippes and Sud. Figure 8.5 shows that the inflow of IDP's is very low in these areas, but that there is still a large proportion moving within these areas instead of towards other areas where the consequences might be less inflicting.

To give an impression of the potential application of movement models, the destination maps in the three time-steps have been printed (see fig. 8.6. It is clearly visible that the number of displacements increases over time, but this is because the time-steps do not represent the same number of weeks. In the first week, the movements within the areas on the west side of the southern peninsula are more distinct than in the following weeks. And as expected, the movements are not only present in the areas that are hit hardest by the Hurricane; the consequences are visible throughout the whole country.

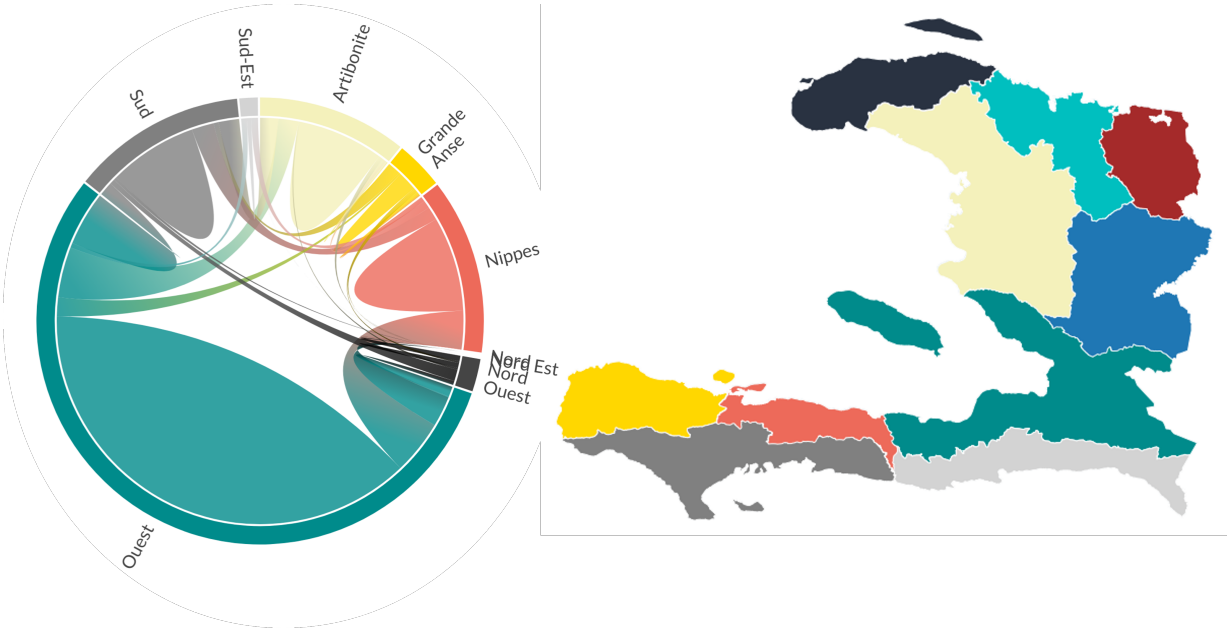


Figure 8.5: The population movements in the six months after Hurricane Matthew visualised in space using administration 1 areas.

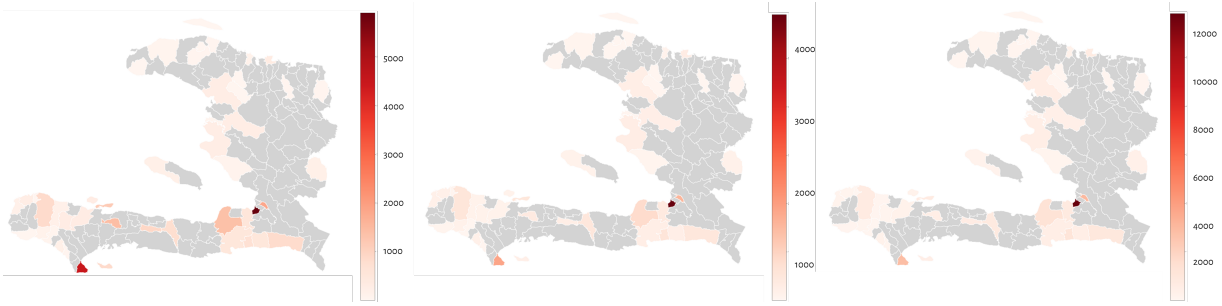


Figure 8.6: The number of IDP's that are expected to arrive at the admin 2 areas differentiated by the three time-steps: week 1 (left), week 2 to 5 (middle) and week 6 to 26 (right)

Discussion

9

This chapter is dedicated to the critical reflection and interpretation of the results presented in chapter 8. Processes prior to the results pose inevitable limitations to the study that are essential to consider when interpreting the results correctly. In section 9.1, the research setup is recaptured, to emphasise again what the main objective and research question are. The limitations of the study are addresses in section 9.3, followed by the academic and the societal contribution in respectively section 9.4.1 and section 9.4.2. Based on the found implications of the study, recommendations are formulated for the involved parties in section 9.5.

9.1 Recapture of the research setup

This study is set up to bridge the knowledge gap in research on the feasibility of the prediction of population movement patterns after a disaster. Many humanitarian aid organisations and researchers have invested their efforts in gaining a better understanding of how mobility patterns change due to a disaster and this is important work as it can safe lives. Why people decide to relocate, when they do so, and where they go are essential factors within the research field. There are two kinds of research recognised within the research field: (1) the qualitative studies using smaller datasets that contain detailed information and (2) quantitative studies using large datasets containing less detailed information. The share of the latter has increased considerably over the last decade as the availability of data has increased considerably too. A potential source for large descriptive data are Call Detail Records (CDR). The records contain time and location data of mobile phones and are collected for billing services, however their potential contribution to an improved understanding of human behaviour is widely recognised.

The application of CDR data in disaster situations is expanding, but the suitability of this type of data is questionable. As the data contains highly sensitive information of individuals, the data is aggregated and processed before usage. This study thus focuses on the suitability of CDR's to describe and predict population movements during a disaster, taking into consideration that only a hand-full of researchers have access to such data. The main research question is formulated as: ***How do populations move during a major disaster and can the destinations of the moving population be determined using data that is available to humanitarian aid organisations?***.

Because the applicability of the data is highly valued, this study pursues the following research objective: ***Gaining more insight in the predictability of mobility patterns during and after a disaster, in order to advise humanitarian agencies on where to focus their support almost real-time***. Relief practices that take place after a disaster are often focused on immediate danger and somewhat chaotic. Many involved parties have their own agenda and some collected data might be (politically) sensitive. Data collection and sharing practices that require min-

imum effort and money are therefore desirable.

To answer the main question, a methodology was introduced to test the predictability of population movements, containing four steps. The first step involves the case selection, where the consequences of Hurricane Matthew in Haiti have been addressed. Because the behaviour of a population appears to be dependent on the socio-demographic characteristics, a thorough analysis of the country and the disaster are essential. The second step is the data preparation. An origin-destination matrix was pursued, but the location data of the displacements of the population were unavailable. Using Naive Bayes classifier, the area in Haiti that was most likely to be the origin and the one that was most likely the destination have been determined. Bringing back the location-information means a de-aggregation of the displacement data. Based on the found locations, extra information could be added too. During the third step, feature selection, the variables that appeared to be influential on population movements, and are available to humanitarian organisations have been selected. Here, also the classes are initiated. The fourth step involves the model selection, in which different classifiers are applied to the features and classes. Based on their accuracy score, the best performing model is optimised and used. Application of the methodology has provided insights in the feasibility of predicting population movements. These insights are discussed below.

9.2 Interpretation of the results

A K-Nearest Neighbour (KNN) classifier has been used to classify features of a displacement. The predicted classes are based on the destinations of a displacement. Although the outcomes of the applied KNN classifier are strongly validated by the implementation of Grid Search and K-Fold techniques, the predictive power of the model is limited. The accuracy score does not transcend 67% and a large proportion of the classes are mislabelled by the model.

The predictions of population movements after a disaster thus still need improvement before they can be applied onto a real case. This study does however provide a strong basis on which the further development of a predictive model could be build. Insights have been uncovered so that a continuation of improvements can take place.

First of all, the performance of the model differs between the different time-frames. The strongest accuracy is found when using the data in week 1 after the Hurricane: 67%. In week 2 to 5, the accuracy score is lowest, namely 62%. And in week 6 to 26, the accuracy score is 64%. This is probably due to the fact that week 2 to 5 less data is available, and because the class representation of the people moving to neighbouring areas is low. It thus appears that the population movements are less predictable with the included features in week 2 to 5 than in week 1, and week 6 to 26. The addition of more variables or the distinction of unique values within the features instead of using clustered values.

Secondly, not all the included features appear to be of the same value for the prediction of population movements. The choice for the features is based on literature, but the usage appears not to be that strong. In the 6 months after the disaster, the proportion of contacts living close-by the home location appear to have the most pronounced predictive power on the mobility behaviour. In week 2 to 5, the GDP per capita and rainfall influence the accuracy of the classification, while in week 1 and week 6 to 26 their influence is lower. Also, in the week after the disaster, the wind-speed is rather important for the predictability, while from week 2 to 26 this feature increases the accuracy score only minimally.

Thirdly, many displacements took place within the same area, confirming that many IDP's do not travel far from their homes to find safety. The largest proportion of displacements in week 1 took place from and to the capital city. Since the IOM concluded that many people travelled towards the affected area to help, show support or check upon family and friends, it is likely that the people moving from the capital city are not seeking

safety. Incorporating that the humanitarian organisations focus on the most vulnerable part of the population, future research could exclude those IDP's that are likely not affected by the disaster.

In theory, if these IDP's are assumed to represent the whole Haitian population, an extrapolation of these numbers could be used to predict the number of expected IDP's in every area. It is however not clear why these people are moving and how desperate their situation is. The focus of the humanitarian organisations is often directed at those that are most vulnerable and in the most dire conditions. Application of the model, testing the results and more research towards the validity of these predictions are essential.

9.3 Limitations

The strength of the study is most visible in the methodology application. The limitations of the developed methodology are therefore discussed, so that the implications of the study in a broader sense can be reflected upon. Even though the progression of the method is based on the scarce availability of detailed data, the included steps do lead to a strong exploration of the feasibility of population movement predictions. The limitations of the methodology are discussed, based on two pillars: the used data sources, the applied methodology.

9.3.1 The used data sources

The most prominently used dataset came from Flowminder, who aggregated data that originally came from Call Detail Records (CDR). The use of CDR's already brings some limitations, seeing that the data is not representative for the Haitian population as a whole. Also, the aggregation practices that were applied to the original data in order to anonymise the data pose limitations. These two implications are discussed, followed by the addressing of the extra feature data that was added.

The Mobile Network Operator that collects the CDR's has the user data of 45% of the Haitian population, while only 61% of the population has a mobile phone (Conseil National des Telecommunications, 2016). It is therefore expected that the population represented in the data is skewed towards the wealthier part of the population. This does comply with the suspicion that some of the identified IDP's are actually helping or checking upon family and friends, as the wealthier part of a population is more likely to have the resources to do so.

It is also important to keep in mind that CDR's refer to sim-cards and not to individuals. That means that the recorded displacements should be considered as moving phones, as the user of the sim-cards could change easily. Multiple people could use the same sim-card, one person can use multiple sim-cards, the sim-card might change owner or be disposed. Especially during disaster situations, all these scenarios are very likely to take place. It is thus clear the that representation of the data towards Internally Displaced Persons must not be taken to literally. We can still use the data for finding important drivers for individuals to change their location during a disaster however. Flowminder has connected some important characteristics to the displacements to indicate how the behaviour of phone users is affected by, in this case, hurricane Matthew. Not only does this involve the assumption that the data is accurate, also the level of detail must be considered. The variables for wind-speed, property damage, rain, the distance to main roads and urban centers, the population density and the number of IDP's close-by have been clustered. Therefore, the level of detail is limited. In addition, all the variables are assumed to be evenly distributed to the area in which the sim-card, and so the person using it, is situated.

Something to keep in mind here is that one of the stronger properties of CDR data is that the behaviour of an individual can be analysed. The patterns emerging when the behaviour of many individuals are combined is what makes studying them interesting. In other words, the translation of individual choices to the outcome for a whole system is what we try to accomplish when studying CDR data. However, the addition of characteristics based on the individuals' locations diminish the power of the data somewhat. Given the privacy of the

sim-card users, using the location-based characteristics are probably the best way to still say something about the person, but in a perfect world, personal data would be added to the CDR data to describe the characteristics. For example, variables such as the GDP per capita based on location are considerably less powerful than using the actual income of the person. Or, take the percentage of damaged property. Naturally, the higher the percentage of damaged property around the person, the higher the chance that the property of that person is damaged too. But the actually interesting information is whether the property of the person in question has been destroyed, damaged or is intact, and look at how that affects their behaviour.

The addition of extra variables is also based on the location in which the people are situated. The used datasets for these variables as well as some the Flowminder used for the aggregation are outdated. The GDP per capita, traveltime to the nearest city and the flood exposure are all in need of updating, as the data is outdated.

9.3.2 The applied methodology

The limitations within this study are found throughout different steps in the methodology, and therefore the limitations per step are discussed. Not only will the limitations that are directly connected to this study be repeated, also the consequences that stem from these limitations are important to take into account. Limitations are found in the four methodology steps are addressed in their order of occurance.

Step 1: Case selection

The selected case is one that gives challenges and opportunities. A major challenge is that the part of the population in possession of a phone is rather low. Also, the mobile network does not cover all parts of the country and is vulnerable to the blows of a Hurricane. This means that this type of data collection during and after a disaster is a risk; the communication infrastructure could be so damaged that a large proportion of the phone users has no connection. In addition is the usability of a phone dependent on available charging points. A power outage could cause a distortion in the data collection and outages are common during (natural) disasters.

Great opportunities that are connected to the case are that Hurricanes and other natural disasters are recurring in Haiti. Gaining a better understanding in the behaviour of the affected population thus means that the insights can be applied in following relief practices. Besides that is the political situation in Haiti still unstable, meaning that the data collection and usage can easily become a sensitive subject. Data collection filters that are installed onto the servers of Mobile Network Operators (MNO's) bypass the political issues. And data can be collected about people that are still residing in areas that are hard to reach.

9.3.3 Step 2: Data preparation

The most apparent limitation within the data preparation is that the characteristics in the aggregated dataset are clustered. The sources of the characteristics in the aggregated dataset have been provided, but the exact processing practices are not available. This means that the de-aggregation is based on data matching methods, which are focused on the spreading of the values. Especially for the characteristic *proportion of damaged property in a radius of 3 km* it is unclear if the data preparation has led to a fitting match between the aggregated dataset and the area-data. This could easily lead to a mis-labelling during the de-aggregation. Moreover, for the characteristic *population density in a 3 km radius*, the values had been clustered in the aggregated dataset, making the data-matching practices challenging. The characteristics *rainfall* and *windspeed* were conveniently transformed to area-data, as their values were not clustered and the pre-processing practices were clear. Also, the wind-speed and rainfall were based on data that was not available on a higher resolution than the administrative level 2 areas, while that of the property damage and population density were, and thus needed to be modified.

Another major limitation is that the validation of the found geo-locations is based on theoretical substantiations and measurements that are based on the data matching pre-processing. This means that a lot of emphasis is thrown on the data matching, which, as discussed, is not bulletproof. The application of a Naive Bayes classifier does not add to the clarity of the labelling, as the method that would expectantly perform better did not do so. Within NB classification, a Gaussian and a Multinomial method can be applied. Since the data is not normally distributed, the Multinomial classifier was expected to better represent the characteristics. However, the Gaussian classifier appeared to label the data rightly with a higher confidence.

Also the expectantly wrongly identified IDP's have been included in the study. A large proportion of the displacements came from the capital city Port-au-Price, while the most devastating consequences of the Hurricane were felt on the southern peninsula of the country. It could thus be considered not to include these displacements in future research.

Step 3: Feature selection

For the feature selection two requirements were used: there would have to be an indication that the population movement patterns are influenced by the particular feature and the data would have to be available at the time a disaster takes place. Because the study is led by substantiations in literature, only those features that are justifiably influential have been considered. There are however some characteristics that Flowminder added to the displacement data that are not used, leading to a major limitation of this study.

As a conceptualisation, the unused characteristics in the displacement datasets have been ran through the K-nearest neighbour classifier with the destination classes, leading to an increased accuracy score. The added characteristics are:

- Proportion of known locations within a 10km radius of the home location.
- The median number of calls per week per user.
- Radius of gyration of a user in the 6 months before the disaster.
- Entropy: the frequency of which the locations of a user change in the 6 months before the disaster.

Including these characteristics in the KNN classifier, an accuracy improvement of 8% is found. This is a rather large improvement and the application of these extra variables should be included in future research. In that case, the methodology can be implemented again, with the main difference of not sticking to literature substantiation. It is possible that another classifier will even improve this score or that different optimisation techniques lead to better results, so these should be considered too. Due to the limited time available in this research, the inclusion of these features has not been incorporated. It could however be a good starting point from which future research can take over.

Checking the observations that were made in the literature review formed a vast basis of the feature selection, but these checks were mere simplifications. The translation from the theoretical observations to quantified variables meant that some observations were represented by variables that did not cover their implication. For example, observation 1 states that the heavier the impact of the disaster the more people relocate. As a measurement for the impact, the wind-speed and rainfall were used. However, their impacts differ completely by location and there is no information available about the people that are not moving. The simplifications are rather influential in the rest of the study, as the feature selection is based on these observations.

Lastly, the transposition of a continuous predictee (the displacement distance to the home location) to classes, caused limitations too. With this decision, information has gone lost especially about the displacements further away. This loss is not that problematic, as the direction in which the displacements took place

is unknown. But this also means that a lot of confidence is placed on the correct labelling of the locations during the data preparation. Also, the only information that can still be derived from the population movement prediction is if the people are staying close to their own area or not. Once they leave the area in which they live, their tracks are lost. For future research, it could be considered to cluster the areas to see what kinds or features make that IDP's are attracted to certain areas. For example, literature suggests that many people relocated towards urban areas. This would mean that a cluster of areas in Haiti with higher population density would attract more people.

Step 4: Model selection

During the model selection, five classifiers have been tested. They are applied in their most basic form and only the two best performing models, based on their accuracy score, have been optimised. The initial selection is composed of classifiers that are often used in basic classification problems. This does not mean that there is not a better model available for this problem. In addition could optimisation of all the models have led to the selection of a better performing model. Due to time limitation however, only two models have been optimised.

Optimisation practices too are not as thorough as they could be. Only within the used coding package, Scikit Learn, the optimisation possibilities are endless. Strongly relying on the accuracy score narrows the scope of these possibilities and thus a classifier that shows a better representation of reality could be found.

The model that has been applied in the end is the k-Nearest Neighbour classifier. Limitations of this classifier is that it is easily overfitted. The methodology partly accounts for this limitation, as the inclusion of a model comparison means that the outcomes are always case specific. It does however complicate generalisation of the found results.

9.4 Implications of the study

Taking the model results and the limitations, the implications of the study become clearer too. The developed methodology has been effective for the achievement of the research goal, even though the data availability was scarce and contained limited information. Seeing that the results of the to be interpreted with the consideration of the limitations, a lot of research still needs to be done. The development of the

effectiveness, usefulness, potential ethical concerns literature implications

9.4.1 Academic contribution

The academic contribution of this study is mainly based in the exploration of the suitability of CDR data to describe or predict population movements during disastrous events. Although the data was found to be of great potential in fulfilling this purpose, this study found some major objections towards using the data in its current aggregated state. The trade-off between the protection of the privacy of mobile phone users who the data represents and the loss of information that this protection caused forms an imbalance that is addressed in this study.

The most important information within CDR data are the time and location of the records. For privacy protection, both these dimensions have been aggregated to the level that the data has lost its strength. The strength within CDR data is that it can potentially show the mobility behaviour of many individuals in a time-space continuum, providing a level of detail that is unprecedented in other data sources. That the data must be aggregated is inevitable, the misuse of the data is almost guaranteed if individuals could be tracked using this information. However, aggregation over both the time and space factor breaks down the strength of this data significantly. In addition, the variables provided to still derive information from, without using the location of

the displacement (i.e. contacts living close-by, the population density, et cetera) contain clustered values that suggest a higher level of specificity than they present.

In this study, the CDR's were de-aggregated to the extend that the locations of the displacements were extracted to an administrative level 2 area. Seeing that the real locations are not public, the validation of the findings has been troublesome, but they show that there is reason to believe that the found locations are at least to some extend right. The labels that were used in the predictive model could also make up for mislabelling neighbouring areas, by merging label 0 (displacements within the same area) and label 1 (displacements between neighbouring areas) together. The areas that are close together are expected to contain similar features for the spatially coherent for the variables wind-speed and rain, meaning that the found areas might fit as good as the neighbouring areas.

Another important finding is that the clustered values of the features used in the predictive model prevent a high accuracy score of the model. The classifier cannot train on data that contains different labels for the same sets of features. The details within the data make that the drivers behind displacements can be identified. For a working predictive model, more detailed data is essential.

Seeing the problems that the data sources and pre-processing have brought, the academic contribution is the methodology that has been developed to overcome these serious obstacles. The thorough considerations of applicable classifiers and the design of a validation method to still ensure some correct applications challenge existing methodologies around data scarce problems. The steps within the methodology have been formulated in a way that they are applicable in other feasibility studies too.

Also an academic contribution to the specific case has been achieved. Application of the predictive models could lead to a change in the way we look at humanitarian aid all together. The identification of more influential variables and missing information to kick-start future research around this topic provide a good basis for building a suitable predictive model. This study might form just the first step, but the exploration of the potential that aggregated CDR data has could prove to be influential in identifying locations that IDP's are likely to displace to.

9.4.2 Societal contribution

The societal contribution lies in the potentially saved and improved lives that might follow from a better understanding in mobility behaviour during disasters. If humanitarian aid organisations can get real-time location data of large parts of the population, their efforts can be pointed towards the most vulnerable population that is in need of assistance. It might in addition improve the efficiency with which supplies are distributed throughout affected areas and could improve Early Warning Early Action practices that lead to the mitigation of disaster consequences.

Information about the displacing population could after one week already potentially determine some of the destinations in that can be expected in the following weeks. This study showed that there are ways to build a useful predictive model once more detailed data is available, as the addition of the geo-locations provided new insights in the population's movements. Further development of of the data sharing agreements between involved parties is key to let CDR data serve them to its fullest potential.

When more insight has be obtained about the difference between the people relocating and the people staying at their homes, such predictive models could be made in real time. This does require many validation rounds, but already in this stage it is valuable to know which part of the population is likely to stay within their own area. Combining demographic information and disaster forecasts could potentially lead to the prediction of population movements in advance. This could significantly reduce the suffering by people that have to leave their homes, and/or lose their source of income.

9.5 Recommendations for involved parties

As briefly introduced, CDR data has not reached its full potential in helping us understand the changing patterns of population movement during a disaster. Even when considering the privacy of sim-card holders, there is much to be gained by implementing some recommendations. These are twofold: there are recommendations for the data preparation and for the improvement of data sharing. Applying these practices is expected to bring considerable improvement in the quality of research around CDR data.

9.5.1 Improvement for data preparation practices

For the data preparation, the usability of the data would considerably improve if the time-frames were separated in even lengths. The current three datasets contain information on respectively week 1, week 2 to 5 and week 6 to 26. Especially the latter time-frame is so broad that the findings within this dataset are not strong enough to help us get a better understanding in how a population moves. Datasets containing weekly data, but still aggregated, would expectantly contain valuable information that could be used in modelling the population movements. Humanitarian aid organisations would benefit from this less aggregated information too, as they are more specific. Even datasets containing 3-weekly data would simplify the interpretation of the results in a predictive model.

The usability would also improve if some information about the displacement locations is revealed. Researching the effect of using administrative level 2 or rather level 3 on the privacy is essential, but would be incredibly fruitful when proven adequate. With this, also the details of the variables would be improved, given that the values do not have to be clustered anymore.

In relation to the identification of IDP's within the CDR data, the quality of the information the data contains could improve when the last filter is not applied. The last filter ensures that the people that had no working connection in the first days after the disaster are excluded from the IDP subset. Their movements might help us more in understanding mobility behaviour after a disaster than the movements of those that came from less affected areas. Providing the data without the last filter enables the researcher to filter the data they see fit for the purpose of their research. And in addition, the connection of displacements that were made by the same sim-card would provide valuable information too. That would allow for exploring the populations movement patterns in a way that represents the real-world better.

9.5.2 Improvement for sensitive data sharing

Already studying highly aggregated data is expected to have a positive effect on the future of sharing sensitive data. Once the potential of this data is recognised, other parties might contribute to overcoming the challenges that handling this sensitive data pose. The most influential parties are Mobile Network Operators, Flowminder and researchers. It is recommended that the procedures of data sharing are explored and that it is researched how the mutual trust between the involved parties can be enhanced. The fact that the owner of this kind of data has commercial interests complicates the matter, so in the case that enhancing mutual trust proves to be difficult, humanitarian aid organisations could consider collecting the data themselves. For example by using applications for which the user give consent that their data is used in research.

Since the mobile network of Haiti has shown to be vulnerable, given the seasonal hazards that the island endures, further research of the replacement of CDR data by other variables might pay off as well. CDR data in itself contains information of a level that cannot be achieved using qualitative methods such as surveys. But the application of CDR data in a country where people are used to regular flooding, humanitarian activities and the occasional hurricane could mean that the lessons to be learned are finite.

Conclusion

10

The considerable limitations of the results have been addressed in chapter 9, and these contain important information on the interpretation of the results presented in this chapter. Firstly, the main conclusions of the research are presented in section 10.1, where also the sub-questions and the main research question are answered. Then the link to the MSc program that this thesis has been conducted for is given in section 10.2.

Main research question

How do populations move during a major disaster, and can the destinations of the moving population be determined using data that is available to humanitarian aid organisations?

Sub-question 1: What are currently identified main drivers for change individual mobility behaviour during and after a flood?

Literature described different drivers behind the change in mobility behaviour due to a disaster. The most important ones are shortly described. Firstly, the intensity of the disaster at the home location, as the consequences of the disaster are often determined by the strength of the disaster. Here, the Protection Motivation Theory (PMT) could be applied, because the change in mobility behaviour is also dependent on how influential the consequences are for the every-day life of the individual or family. Secondly, the family composition has appeared to influence the decision whether to relocate or stay. For example, families with young children are often more inclined to seek a safe location than individuals. Thirdly, the socio-economic status of a person, since a higher status opens up possibilities of relocating, and less dependence on immediate income provides some room for different options. The risk perception of a family or individual is also a vital driver behind the mobility behaviour of the affected population. It is strengthened by prior experience with similar disaster situations, faith in the information source, previously taken mitigating measures, and influenced by the social network of the family.

These drivers do not influence a person individually, but are highly co-dependent. Also, the importance of all of these drivers differs from person to person, and other forces might be at play that are not considered in literature or in this review.

Sub-question 2: How do emergency response activities trigger the movements of a population?

The Haitian population is among the poorest on earth. The desperate situation that many residents endure makes that the continuous presence of humanitarian aid organisations is inevitable. While the population is

recovering from the first disaster, another already comes along. Luckily, an ascending trend in the effectiveness of disaster preparedness measures is seen. By implementing relief activities on a community level, suitable shelters could be identified and stocked days before the disaster made landfall. Still, reports that were made at the time showed that many people ignored the evacuation alerts. Partly due to a phenomenon called *"warning fatigue"*, as many of these alerts are sent out yearly. But also because it is hard to reach everyone, given the low teledensity and poor conditions of many households.

After the Hurricane hit, the population movements became more pronounced. Many people came from the capital city to the southern peninsula to check upon their family and friends, or to assist in the emergency relief activities. A vast majority of the safety seeking population went to urban areas, probably to replace lost sources of income. From the people staying in camps or temporary shelters, those who came from rural areas and those whose houses had been completely destroyed stayed longest. Even after five months, around 7000 individuals were residing in such camps and shelters.

Humanitarian organisations try to stimulate a rapid return to the home locations, and from the people receiving assistance the return rate is fairly high. Others have however been off these organisations' radars.

Sub-question 3: Which changes in population movements can be recognised after a flood in comparison to normal conditions?

Based on Call Detail Record (CDR) data, the change in population movements after Hurricane Matthew have been analysed. The drivers that were found in the literature review have been used as a foundation in the exploratory research. The CDR data incorporates displacements of Internally Displaced Persons (IDP's) in the six months after the disaster, and some patterns can be recognised. First of all, it appears that the proportion of contacts that live close by the home of the the IDP's influences their displacement behaviour: the lower the proportion, the further away the IDP's relocate. Secondly, it seems that IDP's that come from areas with a higher average GDP per capita move earlier than those from areas with a lower average GDP per capita. Also, the intensity of the Hurricane has the expected impact, as the stronger the wind speed and the higher the rainfall, the more people relocate. Furthermore, the people living in more rural areas displaced more over time, while the people living in urban areas displaced less over time. And lastly, people from areas that are often exposed to floods displace earlier that people from areas that are not exposed to regular floods. This complies with the literature, stating that the risk perception is higher for people with experience in a similar situation.

Sub-question 4: What are the most important requirements for the specification of a predictive model?

Since the included features in the model have been selected based on their influence on population movement patterns and literature, only a small selection of features have been included. It is expected that the inclusion of more features will significantly increase the accuracy of the classification of population movements. The classes that are pursued are based on the destination of the displacements.

This method thus requires the inclusion of several features that expectantly influence the movement patterns of a population, with the corresponding origin and destination. That way, a model can classify the destination, based on the features and the origin. The predictive power will improve when the features are highly descriptive, so the data collection and pre-processing practices influence the performance of the classification model substantially.

Sub-question 5: To what extent can the population movement caused by a disaster be predicted?

The population movements can to some extend be predicted. In the model, the destinations of IDP's have been classified, based on the features connected to the origin of the displacement. The classification makes that the predictive power is limited, since a distinction between three classes has been made: the destination

is (1) within the same area, (2) a neighbouring area or (3) further away. The exact destination is thus not sought an this diminishes the predictive power.

In addition, the classifier performed with an maximum accuracy of 67%. The predictability is dependent on the time after the Hurricane and the included features. Although the predictive power not strong enough is to apply this model on a real-life disaster situation yet, it forms a good basis on which future research can build.

10.1 Main conclusion

Call Detail Record (CDR) data enables the analysis of human behaviour on a large scale and the information that it contains can be promising. Not only does it allow us to track the movements of many individuals throughout time, it uncovers patterns in a persons decision making process that potentially tell us a lot about the effects of different interventions. The opportunity of finding new information on human behaviour has been noticed in several research fields, but every researcher eventually finds the same blockade: privacy. The data represents a detailed track of individuals and therefore these individuals must give approval (almost certainly lowering the amount of data that can be collected), or the data must be aggregated to the point that user anonymity is guaranteed.

As a consequence of aggregated data, potentially important information could be lost. Especially in the case that both the dimension of location and time are aggregated, as these two could be considered as the essence of the CDR data. There are however techniques that increase the aggregation level, by de-aggregating the data. Naive Bayes classification has shown to be a functioning method within Machine Learning to de-aggregate a dataset that has incorporated information on at least one of the two essential dimensions; location in this case. By using the same variables to describe the administrative areas within the country that were used to describe the rows within the data, Naive Bayes classification can find the area that is most likely to fit the row. Matching the variables of the areas to the variables within the displacement dataset represents the backbone of the process, as the de-aggregation is driven by the closeness of datapoints between the two datasets.

The de-aggregated dataset offers the possibility to use more detailed analyses as more data could be added. In this study, the extra data that was added formed the basis of analysing population movements after hurricane Matthew. Not only did the analysis confirm that many displacements took place within the same administrative area, it has also shown that the population movement could to some extend be predicted when CDR data is absent. The variables of GDP per capita, the proportion of contacts staying close-by and the distance from the home location to the nearest city present an accuracy score of 65% in week 1, 51% in week 2 to 5, and 64% in week 6 to 26 after the disaster. The addition of wind-speed information increases the score with mostly 2%. This means that it is to some extend possible to imitate the population movements with these available variables when reliable CDR data is not.

The results have also taught some lessons about the movements of a population in a disaster situation. First of all, the movements are not initiated immediately after the disaster. Until long after the disaster has taken place, parts of the displaced population have not returned to their homes yet. Secondly, most displacements were only 5km from home, meaning that humanitarian organisations should stay close to the areas where the disaster hits hardest. And thirdly, the displacements that took place longer after the disaster were covering larger distances.

10.2 Link to EPA program

Within the Engineering and Policy Analysis (EPA) program, students are encouraged to look beyond the technical implications of their model results. The program is designed to teach students to structure complex problems within two fundamental themes: policy and politics, and analytics, modelling and simulation. The consideration of the environment in which a study takes place makes that EPA-students are able to place the results in context. This study has fulfilled the expectations of using a technical solution for a social problem, while incorporating external factors that might effect the usage of the model. The exploration of population movements has been done using data analytics and modelling, but the interpretation of the results does not merely involve the model performance. The implications, challenges and opportunities of the method and the result are considered too. This interdisciplinary approach adds considerable value to the research produced within the EPA program.

References

510.global. (2021a). *510 Mission and Vision*. Retrieved from https://www.510.global/510-mission-vision/

510.global. (2021b). *What we do*. Retrieved from https://www.510.global/what-we-do-3/

Adeola, F. O. (2009). Katrina cataclysm: Does duration of residency and prior experience affect impacts, evacuation, and adaptation behavior among survivors? *Environment and Behavior*, *41*(4), 459–489.

Aerts, J. C., Botzen, W. J., Clarke, K. C., Cutter, S. L., Hall, J. W., Merz, B., … Kunreuther, H. (2018). Integrating human behaviour dynamics into flood disaster risk assessment. *Nature Climate Change*, *8*(3), 193–199.

Ajibade, I., Armah, F. A., Kuuire, V. Z., Luginaah, I., McBean, G., & Tenkorang, E. Y. (2015). Assessing the biopsychosocial correlates of flood impacts in coastal areas of lagos, nigeria. *Journal of Environmental Planning and Management*, *58*(3), 445–463.

Alam, K., & Rahman, M. H. (2014). Women in natural disasters: a case study from southern coastal region of bangladesh. *International journal of disaster risk reduction*, *8*, 68–82.

Alessandro, V. (2010). Complex networks: the fragility of interdependency. *Nature*, *464*(7291), 984–985.

Allaire, M. C. (2016). Disaster loss and social media: Can online information increase flood resilience? *Water Resources Research*, *52*(9), 7408–7423.

Arroyo-Almaraz, I., Calle Mendoza, S., & Van Wyk, C. (2018). Efficacy in communication of dngos. the use of facebook in emergency campaigns. *Revista Latina de Comunicación Social*, *73*, 765–789.

Barchiesi, D., Preis, T., Bishop, S., & Moat, H. S. (2015, 8). Modelling human mobility patterns using photographic data shared online. *Royal Society Open Science*, *2*(8). Retrieved from http://dx.doi.org/10.1098/rsos.150046orviahttp://rsos.royalsocietypublishing.org. doi: 10.1098/rsos.150046

Bempah, S. A., & Øyhus, A. O. (2017). The role of social perception in disaster risk reduction: Beliefs, perception, and attitudes regarding flood disasters in communities along the volta river, ghana. *International journal of disaster risk reduction*, *23*, 104–108.

Botzen, W., Aerts, J., & Van Den Bergh, J. (2009). Dependence of flood risk perceptions on socioeconomic and objective risk factors. *Water resources research*, *45*(10).

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

CIA.gov. (2021). *Haiti - The world factbook*. Retrieved from https://www.cia.gov/the-world-factbook/countries/haiti/

Conseil National des Telecommunications. (2016). *Tableau De Bord du Secteur de la Téléphonie Mobile*. Retrieved from http://conatel.gouv.ht/sites/default/files/TABLEAU_DE_BORD%5B4%5D%20%281%29.pdf

Dejby, J., Li, T., Albert, M., & Lefebvre, V. (2019). Contributing to a better understanding of human mobility in crisis - technical report. *Flowminder Foundation*.

Do, X. B. (2019). Fukushima nuclear disaster displacement: How far people moved and determinants of evacuation destinations. *International Journal of Disaster Risk Reduction*, *33*, 235–252.

Erickson, F. (2012). Comments on causality in qualitative inquiry. *Qualitative Inquiry*, *18*(8), 686–688.

Few, R., & Tran, P. G. (2010). Climatic hazards, health risk and response in vietnam: Case studies on social dimensions of vulnerability. *Global Environmental Change*, *20*(3), 529–538.

Fielding, J. L. (2018). Flood risk and inequalities between ethnic groups in the floodplains of england and wales. *Disasters*, *42*(1), 101–123.

Grünewald, F., & Schenkenberg, E. (2017). Independent real time evaluation: Response to hurricane matthew in haiti. *Groupe URD*.

Haataja, M., Hyvärinen, J., & Laajalahti, A. (2014). Citizens' communication habits and use of icts during crises and emergencies. *Human technology*, *10*(2).

Hamilton, K., Demant, D., Peden, A. E., & Hagger, M. S. (2020). A systematic review of human behaviour in and around floodwater. *International journal of disaster risk reduction*, *47*, 101561.

Harvard University. (2010). *Haiti Earthquake Data Portal | Center for Geographic Analysis*. Retrieved from `https://gis.harvard.edu/haiti-earthquake-data-portal`

Haynes, K., Tofa, M., Avci, A., van Leeuwen, J., & Coates, L. (2018). Motivations and experiences of sheltering in place during floods: implications for policy and practice. *International journal of disaster risk reduction*, *31*, 781–788.

Hunter, L. M. (2005). Migration and environmental hazards. *Population and environment*, *26*(4), 273–302.

Hunter, R. F., Gough, A., O'Kane, N., McKeown, G., Fitzpatrick, A., Walker, T., … Kee, F. (2018, 3). Ethical issues in social media research for public health. *American Journal of Public Health*, *108*(3), 343–348. Retrieved from `http://ajph.aphapublications.org/` doi: 10.2105/AJPH.2017.304249

International Organisation of Migration. (2016). *IOM - DTM documents Haiti*. Retrieved from `https://haiti.iom.int/dtm-documents`

IOM - The UN Migration Agency. (2017). *Hurricane Matthew Response; Displacement Tracking Matrix (DTM) - Haiti*. Retrieved from `https://haiti.iom.int/sites/haiti/files/documents_files/DTM_Matthew_Report_Round%204.pdf`

Kaewkitipong, L., Chen, C. C., & Ractham, P. (2016). A community-based approach to sharing knowledge before, during, and after crisis events: A case study from thailand. *Computers in Human Behavior*, *54*, 653–666.

Keyfitz, N. (1973). Individual mobility in a stationary population. *Population studies*, *27*(2), 335–352.

Kleinberg, J. (2007). The wireless epidemic. *Nature*, *449*(7160), 287–288.

Kubo, T., Yanasan, A., Herbosa, T., Buddh, N., Fernando, F., & Kayano, R. (2019, 3). Health data collection before, during and after emergencies and disasters—The result of the kobe expert meeting. *International Journal of Environmental Research and Public Health*, *16*(5). doi: 10.3390/ijerph16050893

Lazrus, H., Morss, R. E., Demuth, J. L., Lazo, J. K., & Bostrom, A. (2016). "know what to do if you encounter a flash flood": Mental models analysis for improving flash flood risk communication and public decision making. *Risk analysis*, *36*(2), 411–427.

Li, T., Dejby, J., Albert, M., Bengtsson, L., & Lefebvre, V. (2019a, 8). Detecting individual internal displacements following a sudden-onset disaster using time series analysis of call detail records. *arXiv*. Retrieved from `http://arxiv.org/abs/1908.02377http://dx.doi.org/10.5281/zenodo.3349848` doi: 10.5281/zenodo.3349848

Li, T., Dejby, J., Albert, M., Bengtsson, L., & Lefebvre, V. (2019b, 8). Estimating the resilience to natural disasters by using call detail records to analyse the mobility of internally displaced persons. *arXiv*. Retrieved from `http://arxiv.org/abs/1908.02381http://dx.doi.org/10.5281/zenodo.3349848` doi: 10.5281/zenodo.3349848

Liu, X., Yang, S., Ye, T., An, R., & Chen, C. (2021, 2). A new approach to estimating flood-affected populations by combining mobility patterns with multi-source data: A case study of Wuhan, China. *International Journal of Disaster Risk Reduction*, *55*, 102106. doi: 10.1016/j.ijdrr.2021.102106

Lu, X., Bengtsson, L., & Holme, P. (2012, 7). Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(29), 11576–11581. doi: 10.1073/pnas.1203882109

Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., … Bengtsson, L. (2016, 5). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environmental Change*, *38*, 1–7. Retrieved from http://dx.doi.org/10.1016/j.gloenvcha.2016.02.002 doi: 10.1016/j.gloenvcha.2016.02.002

Maher, N. A., Senders, J. T., Hulsbergen, A. F., Lamba, N., Parker, M., Onnela, J. P., … Broekman, M. L. (2019, 9). *Passive data collection and use in healthcare: A systematic review of ethical issues* (Vol. 129). Elsevier Ireland Ltd. doi: 10.1016/j.ijmedinf.2019.06.015

Martín, Y., Li, Z., & Cutter, S. L. (2017, 7). Leveraging Twitter to gauge evacuation compliance: Spatiotemporal analysis of Hurricane Matthew. *PLOS ONE*, *12*(7), e0181701. Retrieved from https://dx.plos.org/10.1371/journal.pone.0181701 doi: 10.1371/journal.pone.0181701

Meekan, M. G., Duarte, C. M., Fernández-Gracia, J., Thums, M., Sequeira, A. M., Harcourt, R., & Eguíluz, V. M. (2017). The ecology of human mobility. *Trends in ecology & evolution*, *32*(3), 198–210.

Mooney, E., & Yemen, U. (2012). *IDPs Outside of Camps* (Tech. Rep.).

Morton Hamer, M. J. (2011). Challenges in Disaster Data Collection during Recent Disasters Mild Traumatic Brain Injury in the Emergency Department View project EMS and Emerging Infectious Diseases View project. Retrieved from https://www.researchgate.net/publication/51821412 doi: 10.1017/S1049023X11006339

NASA. (2021). *Global Precipitation Measurements - Research Topics*. Retrieved from https://gpm.nasa.gov/science/research-topics

Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A Tale of Many Cities: Universal Patterns in Human Urban Mobility. Retrieved from www.plosone.org doi: 10.1371/journal.pone.0037027

Oliver, N., Matic, A., & Frias-Martinez, E. (2015, 8). Mobile Network Data for Public Health: Opportunities and Challenges. *Frontiers in Public Health*, *3*, 1. Retrieved from http://journal.frontiersin.org/Article/10.3389/fpubh.2015.00189/abstract doi: 10.3389/fpubh.2015.00189

Pan, X., Han, C. S., Dauber, K., & Law, K. H. (2007). A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. *Ai & Society*, *22*(2), 113–132.

Poussin, J. K., Botzen, W. W., & Aerts, J. C. (2015). Effectiveness of flood damage mitigation measures: Empirical evidence from french flood disasters. *Global Environmental Change*, *31*, 74–84.

Rafiei, M. H., & Adeli, H. (2017). Neews: A novel earthquake early warning model using neural dynamic classification and neural dynamic optimization. *Soil Dynamics and Earthquake Engineering*, *100*, 417–427.

Red Cross. (2008). *Early warning Early action* (Tech. Rep.). Retrieved from www.ifrc.org

Red Cross. (2020). *Disaster preparedness for older people—Activities—Red Cross EU Office*. Retrieved from https://redcross.eu/projects/disaster-preparedness-for-the-elderly

Smith, S. K., & McCarty, C. (2009). Fleeing the storm (s): An examination of evacuation behavior during florida's 2004 hurricane season. *Demography*, *46*(1), 127–145.

Statistica. (2021). *Haiti - Statistics Facts*. Retrieved from https://www.statista.com/topics/4617/haiti/#dossierSummary

Stewart, S. R. (2017). Tropical cyclone report - hurricane matthew. *National Hurricane Center*.

Sultana, F. (2010). Living in hazardous waterscapes: Gendered vulnerabilities and experiences of floods and disasters. *Environmental Hazards*, *9*(1), 43–53.

Tim, Y., Pan, S. L., Ractham, P., & Kaewkitipong, L. (2017). Digitally enabled disaster response: the emergence of social media as boundary objects in a flooding disaster. *Information Systems Journal*, *27*(2), 197–232.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, *5*(4), 297–323.

UNHCR. (2011). *Humanitarian Charter and Minimum Standards in Humanitarian Response The Sphere Project*. Retrieved from www.sphereproject.org

UNHCR. (2020). *Internally Displaced People.* Retrieved from https://www.unhcr.org/internally-displaced-people.html

Vorst, H. C. (2010). Evacuation models and disaster psychology. *Procedia Engineering*, *3*, 15–21.

Walles, E. (2020). *UNHCR: Numbers of displaced people in world passes 80 million - InfoMigrants.* Retrieved from https://www.infomigrants.net/en/post/29030/unhcr-numbers-of-displaced-people-in-world-passes-80-million

Wang, Q., & Taylor, J. E. (2016, 1). Patterns and Limitations of Urban Human Mobility Resilience under the Influence of Multiple Types of Natural Disaster. *PLOS ONE*, *11*(1), e0147299. Retrieved from https://dx.plos.org/10.1371/journal.pone.0147299 doi: 10.1371/journal.pone.0147299

Warner, K., Ehrhart, C., Sherbinin, A. d., Adamo, S., Chai-Onn, T., et al. (2009). In search of shelter: Mapping the effects of climate change on human migration and displacement. *In search of shelter: mapping the effects of climate change on human migration and displacement.*.

WHO. (2012). Displaced people. *WHO*.

Wilson, R., Erbach-Schoenberg, E. Z., Albert, M., Power, D., Tudge, S., Gonzalez, M., … Bengtsson, L. (2016, 2). Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake. *PLoS Currents*, *8*(Disasters). Retrieved from /pmc/articles/PMC4779046/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4779046/ doi: 10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c

Wisner, B., Blaikie, P., Blaikie, P. M., Cannon, T., & Davis, I. (2004). *At risk: natural hazards, people's vulnerability and disasters*. Psychology Press.

World Bank. (2021). *The World Bank in Haiti.* Retrieved from https://www.worldbank.org/en/country/haiti/overview

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … others (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1–37.

Yin, R. K. (2013, 7). Validity and generalization in future case study evaluations. *Evaluation*, *19*(3), 321–332. Retrieved from http://journals.sagepub.com/doi/10.1177/1356389013497081 doi: 10.1177/1356389013497081

Comparison of classifiers

On the following pages, a comparison of the accuracy scores of five different classifiers is shown. In the tables, a distinction is made between the time-steps of which the datasets have been used. Table A.1 shows the accuracy comparison of the classifications of displacements in week 1. In table A.2, those of week 2 to 5 are shown, and in table A.3, those of week 6 to 26 are shown. From these tables, it is concluded that the Decision Tree (DT) classifier and the K-Nearest Neighbour (KNN) classifier perform best.

Table A.1: Accuracy scores for all combinations of features and methods in week 1

	Features	LR	DT	KNN	NB	SVM
2	Contacts, Windspeed	0.491	0.614	0.573	0.326	0.518
2	Contacts, Rain	0.467	0.612	0.565	0.349	0.575
2	Contacts, GDP	0.405	0.637	0.594	0.323	0.515
2	Contacts, Exposure	0.383	0.544	0.495	0.525	0.491
2	Contacts, City-travel	0.421	0.639	0.611	0.349	0.544
2	Windspeed, Rain	0.335	0.616	0.589	0.317	0.563
2	Windspeed, GDP	0.482	0.632	0.609	0.399	0.517
2	Windspeed, Exposure	0.366	0.589	0.588	0.108	0.515
2	Windspeed, City-travel	0.460	0.632	0.606	0.429	0.540
2	Rain, GDP	0.427	0.620	0.597	0.317	0.557
2	Rain, Exposure	0.379	0.584	0.530	0.108	0.575
2	Rain, City-travel	0.477	0.615	0.575	0.406	0.575
2	GDP, Exposure	0.310	0.618	0.564	0.525	0.515
2	GDP, City-travel	0.454	0.618	0.547	0.395	0.535
2	Exposure, City-travel	0.447	0.618	0.572	0.108	0.544
3	Contacts, Windspeed, Rain	0.455	0.636	0.599	0.329	0.575
3	Contacts, Windspeed, GDP	0.491	0.654	0.630	0.396	0.515
3	Contacts, Windspeed, Exposure	0.478	0.616	0.587	0.357	0.517
3	Contacts, Windspeed, City-travel	0.473	0.654	0.620	0.432	0.533
3	Contacts, Rain, GDP	0.477	0.640	0.599	0.317	0.542
3	Contacts, Rain, Exposure	0.460	0.615	0.575	0.368	0.575
3	Contacts, Rain, City-travel	0.483	0.637	0.565	0.406	0.575
3	Contacts, GDP, Exposure	0.416	0.637	0.594	0.364	0.515
3	Contacts, GDP, City-travel	0.454	0.638	0.584	0.395	0.536
3	Contacts, Exposure, City-travel	0.451	0.640	0.611	0.391	0.544
3	Windspeed, Rain, GDP	0.469	0.632	0.602	0.372	0.557
3	Windspeed, Rain, Exposure	0.335	0.616	0.597	0.325	0.575
3	Windspeed, Rain, City-travel	0.485	0.633	0.592	0.403	0.559
3	Windspeed, GDP, Exposure	0.485	0.632	0.609	0.414	0.515
3	Windspeed, GDP, City-travel	0.472	0.632	0.605	0.395	0.525
3	Windspeed, Exposure, City-travel	0.472	0.632	0.606	0.429	0.533
3	Rain, GDP, Exposure	0.427	0.620	0.597	0.317	0.542
3	Rain, GDP, City-travel	0.483	0.620	0.584	0.412	0.555
3	Rain, Exposure, City-travel	0.477	0.615	0.575	0.406	0.575
3	GDP, Exposure, City-travel	0.456	0.617	0.547	0.397	0.536
4	Contacts, Windspeed, Rain, GDP	0.468	0.653	0.614	0.359	0.544
4	Contacts, Windspeed, Rain, Exposure	0.468	0.636	0.601	0.341	0.575
4	Contacts, Windspeed, Rain, City-travel	0.495	0.657	0.583	0.405	0.559
4	Contacts, Windspeed, GDP, Exposure	0.477	0.654	0.630	0.415	0.529
4	Contacts, Windspeed, GDP, City-travel	0.472	0.654	0.608	0.395	0.530
4	Contacts, Windspeed, Exposure, City-travel	0.484	0.655	0.620	0.429	0.521
4	Contacts, Rain, GDP, Exposure	0.466	0.640	0.599	0.317	0.554
4	Contacts, Rain, GDP, City-travel	0.493	0.639	0.602	0.412	0.547
4	Contacts, Rain, Exposure, City-travel	0.487	0.637	0.565	0.406	0.559
4	Contacts, GDP, Exposure, City-travel	0.474	0.638	0.584	0.397	0.536
4	Windspeed, Rain, GDP, Exposure	0.478	0.632	0.602	0.317	0.544
4	Windspeed, Rain, GDP, City-travel	0.492	0.633	0.596	0.412	0.579
4	Windspeed, Rain, Exposure, City-travel	0.493	0.633	0.592	0.403	0.559
4	Windspeed, GDP, Exposure, City-travel	0.486	0.632	0.605	0.397	0.530
4	Rain, GDP, Exposure, City-travel	0.481	0.620	0.584	0.412	0.547
5	Contacts, Windspeed, Rain, GDP, Exposure	0.458	0.653	0.614	0.325	0.545
5	Contacts, Windspeed, Rain, GDP, City-travel	0.461	0.655	0.613	0.412	0.568
5	Contacts, Windspeed, Rain, Exposure, City-tra...	0.492	0.657	0.583	0.403	0.559
5	Contacts, Windspeed, GDP, Exposure, City-travel	0.483	0.654	0.608	0.397	0.536
5	Contacts, Rain, GDP, Exposure, City-travel	0.494	0.639	0.602	0.412	0.549
5	Windspeed, Rain, GDP, Exposure, City-travel	0.488	0.633	0.596	0.412	0.568
6	Contacts, Windspeed, Rain, GDP, Exposure, Cit...	0.478	0.655	0.613	0.412	0.557

Table A.2: Accuracy scores for all combinations of features and methods in week 2 to 5

	Features	LR	DT	KNN	NB	SVM
2	Contacts, Windspeed	0.392	0.597	0.545	0.340	0.539
2	Contacts, Rain	0.427	0.605	0.574	0.327	0.543
2	Contacts, GDP	0.364	0.629	0.590	0.310	0.539
2	Contacts, Exposure	0.348	0.589	0.489	0.401	0.571
2	Contacts, City-travel	0.386	0.633	0.594	0.352	0.502
2	Windspeed, Rain	0.424	0.574	0.552	0.332	0.563
2	Windspeed, GDP	0.422	0.615	0.587	0.356	0.539
2	Windspeed, Exposure	0.363	0.560	0.552	0.469	0.539
2	Windspeed, City-travel	0.375	0.610	0.504	0.383	0.502
2	Rain, GDP	0.518	0.620	0.542	0.262	0.597
2	Rain, Exposure	0.512	0.572	0.493	0.469	0.543
2	Rain, City-travel	0.416	0.620	0.566	0.384	0.555
2	GDP, Exposure	0.516	0.616	0.596	0.469	0.539
2	GDP, City-travel	0.414	0.616	0.524	0.364	0.459
2	Exposure, City-travel	0.404	0.616	0.485	0.469	0.502
3	Contacts, Windspeed, Rain	0.438	0.620	0.583	0.328	0.563
3	Contacts, Windspeed, GDP	0.385	0.630	0.601	0.361	0.539
3	Contacts, Windspeed, Exposure	0.397	0.602	0.566	0.398	0.539
3	Contacts, Windspeed, City-travel	0.403	0.631	0.612	0.383	0.482
3	Contacts, Rain, GDP	0.447	0.632	0.598	0.262	0.562
3	Contacts, Rain, Exposure	0.426	0.605	0.578	0.417	0.543
3	Contacts, Rain, City-travel	0.418	0.630	0.603	0.384	0.555
3	Contacts, GDP, Exposure	0.364	0.628	0.590	0.407	0.539
3	Contacts, GDP, City-travel	0.403	0.630	0.590	0.364	0.455
3	Contacts, Exposure, City-travel	0.395	0.634	0.594	0.378	0.502
3	Windspeed, Rain, GDP	0.493	0.618	0.546	0.335	0.605
3	Windspeed, Rain, Exposure	0.399	0.579	0.554	0.344	0.563
3	Windspeed, Rain, City-travel	0.419	0.618	0.572	0.344	0.542
3	Windspeed, GDP, Exposure	0.418	0.615	0.587	0.368	0.539
3	Windspeed, GDP, City-travel	0.415	0.610	0.524	0.396	0.565
3	Windspeed, Exposure, City-travel	0.395	0.610	0.504	0.383	0.482
3	Rain, GDP, Exposure	0.517	0.620	0.542	0.262	0.562
3	Rain, GDP, City-travel	0.431	0.620	0.569	0.372	0.594
3	Rain, Exposure, City-travel	0.410	0.620	0.566	0.384	0.555
3	GDP, Exposure, City-travel	0.380	0.616	0.524	0.364	0.455
4	Contacts, Windspeed, Rain, GDP	0.448	0.626	0.605	0.335	0.601
4	Contacts, Windspeed, Rain, Exposure	0.445	0.616	0.587	0.344	0.563
4	Contacts, Windspeed, Rain, City-travel	0.441	0.632	0.602	0.344	0.542
4	Contacts, Windspeed, GDP, Exposure	0.390	0.631	0.601	0.365	0.539
4	Contacts, Windspeed, GDP, City-travel	0.383	0.633	0.604	0.396	0.565
4	Contacts, Windspeed, Exposure, City-travel	0.422	0.632	0.612	0.383	0.457
4	Contacts, Rain, GDP, Exposure	0.449	0.632	0.598	0.262	0.585
4	Contacts, Rain, GDP, City-travel	0.441	0.631	0.595	0.372	0.591
4	Contacts, Rain, Exposure, City-travel	0.412	0.630	0.603	0.382	0.573
4	Contacts, GDP, Exposure, City-travel	0.418	0.630	0.590	0.364	0.485
4	Windspeed, Rain, GDP, Exposure	0.499	0.618	0.546	0.338	0.601
4	Windspeed, Rain, GDP, City-travel	0.438	0.618	0.575	0.372	0.605
4	Windspeed, Rain, Exposure, City-travel	0.404	0.618	0.572	0.376	0.542
4	Windspeed, GDP, Exposure, City-travel	0.405	0.610	0.524	0.396	0.565
4	Rain, GDP, Exposure, City-travel	0.436	0.620	0.569	0.372	0.591
5	Contacts, Windspeed, Rain, GDP, Exposure	0.415	0.626	0.605	0.338	0.574
5	Contacts, Windspeed, Rain, GDP, City-travel	0.398	0.634	0.603	0.372	0.605
5	Contacts, Windspeed, Rain, Exposure, City-tra...	0.416	0.632	0.602	0.369	0.542
5	Contacts, Windspeed, GDP, Exposure, City-travel	0.434	0.634	0.604	0.396	0.565
5	Contacts, Rain, GDP, Exposure, City-travel	0.406	0.631	0.595	0.372	0.580
5	Windspeed, Rain, GDP, Exposure, City-travel	0.403	0.618	0.575	0.372	0.605
6	Contacts, Windspeed, Rain, GDP, Exposure, Cit...	0.411	0.634	0.603	0.372	0.588

Table A.3: Accuracy scores for all combinations of features and methods in week 6 to 26

	Features	LR	DT	KNN	NB	SVM
2	Contacts, Windspeed	0.339	0.592	0.519	0.302	0.521
2	Contacts, Rain	0.449	0.597	0.550	0.310	0.556
2	Contacts, GDP	0.377	0.630	0.593	0.503	0.561
2	Contacts, Exposure	0.346	0.593	0.520	0.444	0.590
2	Contacts, City-travel	0.402	0.630	0.585	0.349	0.543
2	Windspeed, Rain	0.351	0.564	0.551	0.226	0.559
2	Windspeed, GDP	0.347	0.610	0.569	0.332	0.522
2	Windspeed, Exposure	0.496	0.563	0.516	0.444	0.521
2	Windspeed, City-travel	0.379	0.610	0.586	0.377	0.529
2	Rain, GDP	0.347	0.609	0.547	0.251	0.522
2	Rain, Exposure	0.382	0.562	0.529	0.444	0.556
2	Rain, City-travel	0.385	0.609	0.550	0.365	0.517
2	GDP, Exposure	0.404	0.609	0.581	0.444	0.561
2	GDP, City-travel	0.363	0.609	0.572	0.361	0.540
2	Exposure, City-travel	0.385	0.609	0.601	0.108	0.547
3	Contacts, Windspeed, Rain	0.451	0.603	0.535	0.363	0.556
3	Contacts, Windspeed, GDP	0.374	0.632	0.589	0.336	0.522
3	Contacts, Windspeed, Exposure	0.396	0.594	0.551	0.392	0.521
3	Contacts, Windspeed, City-travel	0.405	0.631	0.595	0.378	0.547
3	Contacts, Rain, GDP	0.433	0.632	0.574	0.281	0.523
3	Contacts, Rain, Exposure	0.452	0.603	0.544	0.397	0.556
3	Contacts, Rain, City-travel	0.410	0.631	0.577	0.365	0.546
3	Contacts, GDP, Exposure	0.380	0.630	0.593	0.478	0.561
3	Contacts, GDP, City-travel	0.408	0.633	0.543	0.361	0.541
3	Contacts, Exposure, City-travel	0.414	0.631	0.585	0.375	0.548
3	Windspeed, Rain, GDP	0.469	0.610	0.558	0.365	0.535
3	Windspeed, Rain, Exposure	0.421	0.566	0.544	0.405	0.556
3	Windspeed, Rain, City-travel	0.421	0.610	0.550	0.365	0.550
3	Windspeed, GDP, Exposure	0.402	0.610	0.569	0.332	0.522
3	Windspeed, GDP, City-travel	0.385	0.610	0.564	0.359	0.582
3	Windspeed, Exposure, City-travel	0.417	0.610	0.586	0.377	0.545
3	Rain, GDP, Exposure	0.390	0.609	0.547	0.278	0.523
3	Rain, GDP, City-travel	0.425	0.609	0.541	0.392	0.587
3	Rain, Exposure, City-travel	0.383	0.609	0.550	0.365	0.546
3	GDP, Exposure, City-travel	0.390	0.609	0.572	0.361	0.541
4	Contacts, Windspeed, Rain, GDP	0.437	0.634	0.576	0.365	0.524
4	Contacts, Windspeed, Rain, Exposure	0.378	0.606	0.545	0.417	0.556
4	Contacts, Windspeed, Rain, City-travel	0.447	0.632	0.594	0.365	0.555
4	Contacts, Windspeed, GDP, Exposure	0.376	0.632	0.589	0.335	0.522
4	Contacts, Windspeed, GDP, City-travel	0.408	0.632	0.549	0.359	0.529
4	Contacts, Windspeed, Exposure, City-travel	0.438	0.632	0.595	0.378	0.551
4	Contacts, Rain, GDP, Exposure	0.436	0.632	0.574	0.319	0.520
4	Contacts, Rain, GDP, City-travel	0.449	0.632	0.569	0.392	0.608
4	Contacts, Rain, Exposure, City-travel	0.389	0.631	0.577	0.365	0.568
4	Contacts, GDP, Exposure, City-travel	0.434	0.633	0.543	0.361	0.551
4	Windspeed, Rain, GDP, Exposure	0.429	0.610	0.558	0.374	0.524
4	Windspeed, Rain, GDP, City-travel	0.422	0.610	0.547	0.390	0.579
4	Windspeed, Rain, Exposure, City-travel	0.415	0.610	0.550	0.365	0.555
4	Windspeed, GDP, Exposure, City-travel	0.429	0.610	0.564	0.359	0.529
4	Rain, GDP, Exposure, City-travel	0.427	0.609	0.541	0.392	0.608
5	Contacts, Windspeed, Rain, GDP, Exposure	0.416	0.634	0.576	0.373	0.530
5	Contacts, Windspeed, Rain, GDP, City-travel	0.421	0.634	0.570	0.390	0.577
5	Contacts, Windspeed, Rain, Exposure, City-tra...	0.444	0.632	0.594	0.365	0.568
5	Contacts, Windspeed, GDP, Exposure, City-travel	0.447	0.632	0.549	0.359	0.528
5	Contacts, Rain, GDP, Exposure, City-travel	0.443	0.632	0.569	0.392	0.593
5	Windspeed, Rain, GDP, Exposure, City-travel	0.443	0.610	0.547	0.393	0.577
6	Contacts, Windspeed, Rain, GDP, Exposure, Cit...	0.451	0.634	0.570	0.393	0.551

Oversampling and Undersampling

B

The skewed performances of the labels can be corrected using oversampling and undersampling methods. As the classes are not evenly represented (see fig. B.1), an oversampling technique is applied. This technique is called Synthetic Minority Over-sampling Technique (SMOTE) and has the potential to perform as well as under-sampling the majority class (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). In contrast to under-sampling, this technique makes sure that no important information is lost and that also interesting outcomes are included by using a random sampler and imitating outliers. The sampler must be ran several times to make sure that the created data does not influence the model performance. The undersampling technique involves the random selection of data from the over-represented labels (label 0 and 2 in this case).

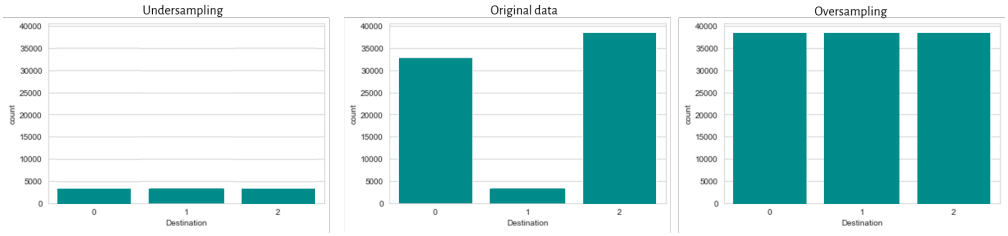


Figure B.1: The sizes of the three destination classes within the original dataset (middle), after undersampling (left) and after oversampling (right)

The effects of the sampling techniques are clearly not beneficial for the model outcomes. The model using oversampled data performs slightly worse for all three time-frames and the model using undersampled data performs significantly worse. It is clear from fig. B.2, fig. B.3, fig. B.4, fig. B.5, fig. B.6 and fig. B.7 that the original data outperforms the sampled data. The undersampled data probably does not contain enough information to train the model accurately, meaning that a rather large amount of data is needed to bring up the model performance. On the other hand does the oversampling also distort the information within the data. Probably mostly because the data is so imbalanced that the oversampling technique cannot accuractely replicate the existing data for label 2.

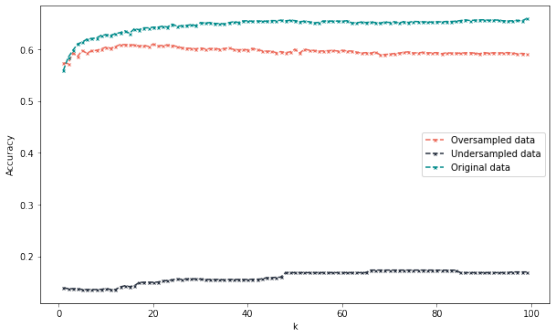


Figure B.2: The accuracy score by choice of k for data that is oversampled and undersampled in week 1

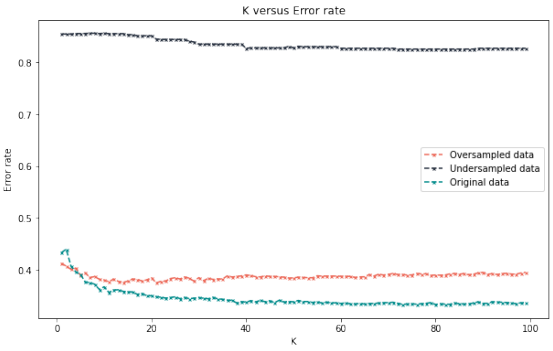


Figure B.3: The error rate by choice of k for data that is oversampled and undersampled in week 1

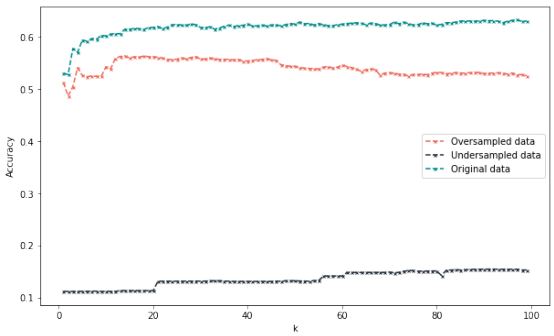


Figure B.4: The accuracy score by choice of k for data that is oversampled and undersampled in week 2 to 5

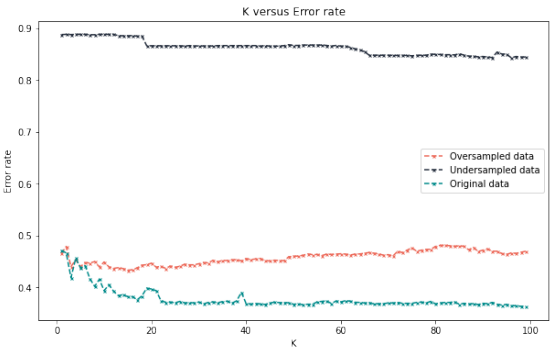


Figure B.5: The error rate by choice of k for data that is oversampled and undersampled in week 2 to 5

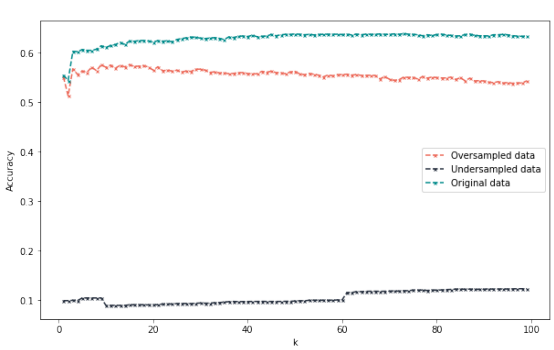


Figure B.6: The accuracy score by choice of k for data that is oversampled and undersampled in week 6 to 26

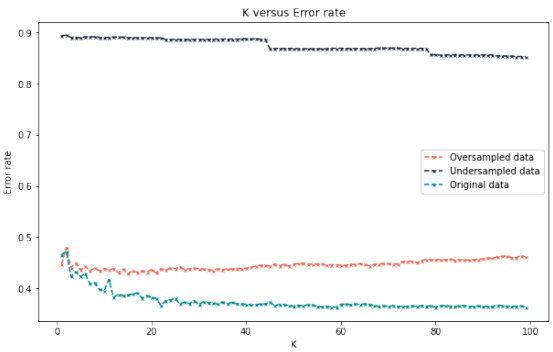


Figure B.7: The error rate by choice of k for data that is oversampled and undersampled in week 6 to 26