Hes·so VALAIS WALLIS
Haute Ecole de Gestion
Hochschule für Wirtschaft

Filière Informatique de gestion
Studiengang Wirtschaftsinformatik

Bachelor thesis 2021

Understanding Mobility in Disaster Scenarios for Emergency Communications Design

Student : Dylan Thompson

Professor : Yann Bocchi

Submitted : 13.08.2021

ABSTRACT

The goal of this work was to identify the demand for communication as well as key features of user distribution and movement in the event of a disaster using social media data.

The work was realized in the context of the activities of the research institute of information systems of the HES-SO Valais.

This thesis has multiple planned stages. We conducted a preliminary research phase during which we analyzed the different aspects of disasters and natural disaster management as well as the data sources used in the context of disaster management. Then, we defined a disaster scenario in which we conducted different experiments and data analysis operations with the aim of correlating meteorological measurements with cellular activity data.

Through the different experiments we performed, we were able to demonstrate the influence of Hurricane Sally on cellular activity in Florida. We then interpreted our results and exposed different scenarios in which our experiments could be of use. To deepen our research, we also experimented with the use of a neural network and previously aggregated data, to try and predict possible variations in cellular activity in the context of a hurricane.

Keywords: Emergency communication, Facebook, Disaster management, Data mining, Neural network, Social media data

FOREWORD

This work was done in the context of a bachelor's thesis in Business Information Technology at the HES-SO Valais. It was proposed by the professor Yann Bocchi and Gianluca Rizzo. We were given the task to make use of data from a social network with the aim of characterizing the demand for communication as well as user distribution and mobility in the event of a disaster.

There are currently only a few studies that identify the demand for communication, as well as the key features of user distribution and movement in the event of disaster scenarios. However, the institute of information systems of the HES-SO Valais recently signed an agreement with Facebook under which the institute is authorized to gain access to anonymized data on user mobility and Internet access during a variety of natural and artificial disasters around the world. This gives us the opportunity to use the data accumulated by social media to improve the design of existing emergency communication networks.

The different aspects of this thesis enable us to gain insight on potential improvements in regards to disaster management. However, it is important to note that the work accomplished is centered on research and that the different technological aspects and result interpretation still need further analysis and exploration before being used in a real-world scenario.

The approach taken to accomplish this thesis started by global research on the subject then specific research on one case of a recent natural disaster. Following this, we established the research scenario and finished by implementing the defined scenario and interpreted the obtained results.

ACKNOWLEDGEMENTS

We would like to sincerely thank the people that helped us during the process of the elaboration of this bachelor's thesis for their input and support. Special thanks to:

Mr. Yann Bocchi, the teacher in charge, for having supervised and followed us throughout this work. We are extremely grateful for his ideas, advice and precious feedback.

I would also like to extend my deepest gratitude to Mr. Gianluca Rizzo, for providing invaluable insight and guidance throughout the elaboration of this thesis.

Finally, thanks to all the people who reviewed and corrected this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ABBREVIATIONS

API	-	Application Programming Interface
ASOS	-	Automated Surface Observing Systems
CRISP-DM	-	Cross Industry Standard Process for Data Mining
CSV	-	Comma-Separated Values
HTTP	-	Hypertext Transfer Protocol
IDE	-	Integrated Development Environment
NGO	-	Non-Governmental Organization
NOAA	-	National Oceanic and Atmospheric Administration
SSD	-	Solid State Drive
UI	-	User Interface
WSN	-	Wireless sensor network

INTRODUCTION

In this day and age social media has completely flooded our society, generating millions of data records every second of every hour of every day. This phenomenon could seem overwhelming, but to the data scientist this can mean opportunity. One could use this data for their own benefit but in this paper, we aim to benefit the same society that generated the said records. This brings us to the main research subject:

As of 2021, how can social media data support and improve the design of existing emergency communication networks through the use of modern data mining techniques?

To answer this question, we will first describe different aspects of natural disasters and what current processes exist to manage them. This will give us good insight as to what problems can occur during and after disaster situations. Then, we will examine different data sources that can be of use in the context of disaster management. This will give us insight on what data could potentially be harnessed in combination with social media data for the chosen scenario.

After this research phase, we will overview the different possible use cases and scenarios that we considered and describe the chosen one for our analysis and implementation phase.

During the implementation phase, we will start by describing the different existing works related to our scenario to gain insight on the different challenges and difficulties that we will have to face. Following this, we will oversee the different technological requirements and describe what technologies we chose for this project. Having considered these elements, we will characterize and describe the chosen data sources we will be using in our implementation. We will then perform different data preprocessing actions to be able to perform a thorough analysis of the data.

Finally, we will evaluate the different results we obtained through the data mining process and the possible interpretations of the data that we can link with the design of emergency communication networks. Also, we will see the potential improvements and ideas to pursue the research.

1. NATURAL DISASTERS

Natural disasters have been on the rise all over the world as a result of global warming and environmental destruction. The need for solid planning and disaster management operations are needed more than ever.

1.1. Types of natural disasters and their impact

Natural disasters are naturally occurring events that materialize after some type of natural mechanism or development (IFRC, s.d.). According to the International Disaster Database (EM-DAT, n.d.), there are six groups of natural disasters: biological (epidemic), geophysical (earthquake), climatological (drought), meteorological (storm), hydrological (flood), extraterrestrial (asteroid impact). Each of these different types of natural disasters vary in their impact on society, but economy and health are most often targeted. Each of these different types of natural disasters vary in their impact on society, but the two main categories of impacts are economical and health related. The following citation describes the damages caused by Hurricane Katrina.

> In 2005, Hurricane Katrina devastated New Orleans and the Mississippi gulf coast. In New Orleans alone, more than 200,000 homes were destroyed and over 70 percent of the resident population had to be at least temporarily relocated outside of the greater New Orleans area. In addition, huge sums of federal assistance were necessary to help jump start recovery efforts in the city and surrounding region. Estimates of over $105 to $150 billion in reduced tax revenue, loss of infrastructure, expense of reclamation efforts, and loss of normal revenue were lost to the city (Sharri eff, 2018).

Natural disasters create important health issues. The death toll of natural disasters has been estimated to be on average 60,000 deaths per year or 0.1% of global deaths (Roser & Ritchie, 2019). Other important casualties caused by natural disasters include: “physical trauma, acute disease, and emotional trauma, [...] morbidity and mortality associated with chronic disease and infectious disease” (Giorgadze, Maisuradze, Japaridze, Utiashvili, & Abesadze, 2011).

1.2. Frequency of natural disasters

Furthermore, with the increase of urbanized areas and natural catastrophes related to climate change, the negative impact of natural disasters will likely be on the rise in the future (Loko, 2012). According to an article in the Washington Post from Sarah Kaplan, the global average sea level has risen between eight and nine inches due to human-related activity since the industrial era (Kaplan, 2020). Consequently, rising waters increase the risk of flooding during a hurricane. Another alarming element mentioned in the article is that: “with global sea surface temperature increasing 0.13 degrees Fahrenheit per decade, studies show the chance of a given tropical storm becoming a

hurricane that is Category 3 or greater has grown 8 percent every 10 years" (Kaplan, 2020). These two examples are only the tip of the melting iceberg, with the current tendency, the need for disaster management related resources and actions is of primary importance more than ever.

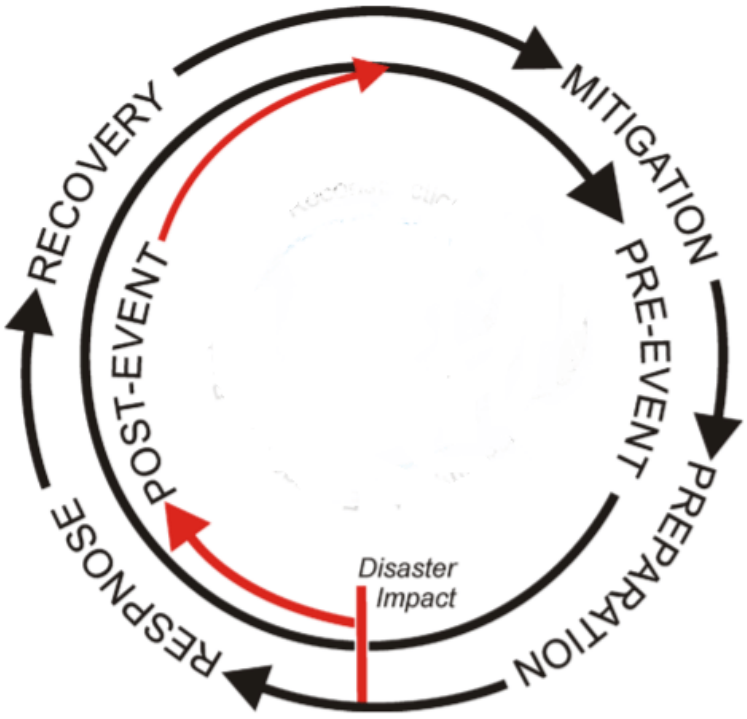
1.3. Disaster management

Now that we have a better idea on how natural disasters impact our society, it is time to talk about how disasters can be managed.

According to the United Nations, disaster management is: "the organization, planning and application of measures preparing for, responding to and recovering from disasters" (United Nations, 2021).

Disaster management has also been formalized into models, like the traditional model that describes two phases: pre-Disaster risk-reduction model phase and post-disaster recovery phase (Blaikie, Cannon, Davis, & Wisner, 2004). This model can be considered as oversimplified because it does not consider the time of the disaster itself (Albtoush, Dobrescu, & Ionescou, 2011). The model proposed by Arifah et al. (2019) integrates the disaster as a parameter and represents a simple but efficient cycle of how a disaster should be managed (Figure 1) (Arifah, Mohd Tariq, & Juni, 2019).

Figure 1: Disaster management phases



Source : https://www.researchgate.net/figure/Four-phases-of-Disaster-Management-Cycle-adapted-from-Herold-Sawada-2012_fig1_334394277

Based on this article from Phengsuwan et al. (2021), the above figure can be described as follows (Phengsuwan, et al., 2021):

- Before the disaster occurs
  - Mitigation: This is the phase where actions are taken to minimize the impact of potential hazards and lowers the risk of them becoming disasters.
  - Preparedness: This is the planification and educational phase, raising awareness in communities and making different plans to overcome the potential impact of a disaster.
- During and after the disaster occurs
  - Response: This phase is the execution of the plans made in the preparedness phase. The goal is to protect people's lives and property against the occurring disaster.
  - Recovery: This phase englobes the actions that are taken to restore properties and public infrastructures (electrical installations, networks) that have been damaged. Treating people's traumas and illnesses also constitutes an important part of the recovery phase.

2. DATA SOURCES USED FOR DISASTER MANAGEMENT

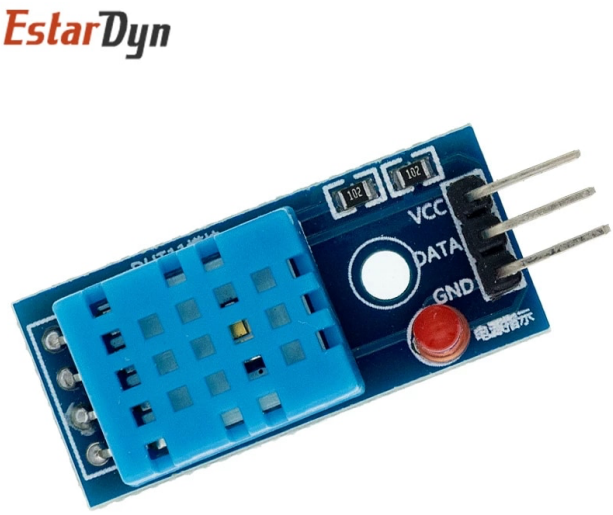
Now that we have talked about natural disasters and how they can be managed, we will explore the different data sources that can potentially be used in the context of a natural disaster, especially in natural disaster management. In the context of disaster management, it is crucial to have data sources that are highly accurate and reliable. Considering the mass of data created by the continuous stream of social media data being created as we speak, it is of the highest importance to select and curate the most relevant data. Furthermore, not only social media data may be used, but also different data sources may be considered and combined for the purpose of disaster management. According to Phengsuwan et al. (2021), the following data sources are useful for effective disaster management and analysis: Sensors, Social Media User, Social Media Platform, and Third Party (Phengsuwan, et al., 2021, pp. 6-9).

2.1. Sensors

2.1.1. Physical sensors

Physical sensors generate earth observation and ancillary data. They generally are small electronic devices that are placed in a region of interest. These sensors can measure for example, temperature, chemicals, vibrations, humidity, pressure, magnetic fields. Figure 2 is an example of a humidity sensor. It cannot operate alone but has to be connected to a microcontroller (like an Arduino ® or a Raspberry Pi ®) to be able to process and emit data.

Figure 2: Digital Temperature and Humidity Sensor



Source : https://www.aliexpress.com/

According to Chen et al. (2013), a good way of managing physical sensors is by using them in WSN's "hundreds or even thousands of sensor nodes can be densely/sparsely deployed in a region of interest without the limitation of cabling and can be easily redeployed/scaled as needed" (Chen, et al., 2013).

In the implementation phase of this thesis, we will be using data gathered by the National Oceanic and Atmospheric Administration (NOAA) 's observation systems that include these different types of physical sensors: Upper air sounding, ASOS, aircraft, buoy, satellite/radar overview.

2.1.2. Human sensors

We call human sensors, people that generate data purposefully using various platforms like a crowdsourcing platform; i.e. geoBingAn app (GeoThings, s.d.). The type of data emitted by users are various, but can be very useful in disaster management, especially in the response phase. For example, a user posts a picture with a geolocation of a landslide that blocks a road on a crowdsourcing platform. This triggers the competent authorities to act accordingly and plan to clear the zone. The advantage is that it does not only inform competent authorities but also other users. If we follow our example of the landslide, a user that was planning to flee an endangered area and has the information of an obstacle on the road will be able to update his route accordingly to increase safety.

2.2. Social Media User

Social media users play an important role in the context of disaster management as they communicate to a wide audience in case of a hazard or risk. Even if the data is not always consistent,

using existing infrastructures and networks enables people to be informed rapidly. As mentioned in Phengsuwan et al. (2021), we can distinguish at least four different types of social media users in the context of communication in a disaster scenario (Phengsuwan, et al., 2021, pp. 6-9):

- Government Authority
  - These are official announcements deployed by governments and associated government entities.
- Research and academic institutions
- Non-governmental organizations
  - They post information through the private sector, like news networks or humanitarian organizations. The information provided can be considered as more accurate than from the public sector and more plentiful than government authorities.
- Public
  - This data source has for origin personal user accounts. It is the source of most of the data that is used in the field of disaster management regarding social media content analysis. It is important to note that the information generated is often of poor quality and needs to be preprocessed and prepared with great care with modern data mining techniques.

2.3. Social media platform

Social media platforms grant (with some conditions) access to their user's data for research purposes. For example, Twitter has an API that enables users having the required permissions to access tweets, users, direct messages, lists, trends, media and places (Twitter API, 2021) via HTTP requests. Another social media company that offers access to their data is Facebook, the company has created a dedicated platform (Facebook Data for Good, 2021) whose goal is to make available some of its data for humanitarian or research purposes. We will extensively cover this data source since it was the main one used in our research. It is important to note that there are two important information dimensions in social media data sources: spatial and temporal.

2.4. Third party

Another data source stated by Phengsuwan et al. (2021) is external companies or organizations that give access to the data that they have collected on social media platforms and processed for public or research use (Phengsuwan, et al., 2021, pp. 6-9). This data source has the advantage of having been preprocessed and organized, which considerably reduces the tasks of filtering, classifying and extracting the data. Unfortunately, the organizations that offer such data sources are scarce.

3. USE CASE ANALYSIS

Now that we have a good idea of the impact of natural disasters and how their management can be organized, as well as the potential sources for data, we can define the main objective of this thesis: the implementation of a set of algorithms using modern data mining techniques using social media data from Facebook Data for Good with the aim of contributing to the design of emergency communication networks and disaster management.

It is important to note that this thesis was fundamentally a research and exploration process. We were given the task to implement a solution that would contribute to disaster management in the context of a natural disaster. Without knowing precisely what problem we needed to solve, we had to explore many different possibilities.

We decided early on to choose as a data source the datasets from Facebook Data for Good. The reason that we selected this data source is that the research Institute of Information systems of the HES-SO Valais was given access to the Facebook Data for Good platform which can be considered a good opportunity since the datasets are not open to the public and the amount of people authorized to access the data is limited. Furthermore, not much research has been done in the field using data from Facebook. Considering the amount of data and the type of information that was made available by Facebook, the choice of the data source was obvious. However, given the substantial quantity of data and different natural disaster scenarios available on the platform, it was critical to carefully select the dataset.

At first, we thoroughly examined the possibility of using Facebook's data to contribute to a thesis written in 2018 made by a student of the HES-SO, Flavien Bonvin: Analysis, design, and implementation of an infrastructure-less situation awareness application (Bonvin, 2018). This thesis had for main objective, the development of a smartphone application that harnessed ad hoc networks (temporary type of Local Area Network (LAN)) as a backup communication system in disaster scenarios. Ad hoc network being an off-grid type of communication infrastructure, a crucial element to take into consideration is the minimum range needed for the nodes of the system to properly communicate with each other. In the case of this application the maximal range for the nodes to communicate properly was 400 meters:

> The floating content radius has five possible distances, 50, 100, 200, 300 and 400 meters. We limited the distance to 400 meters as more length was not necessary in our opinion. Moreover, we did not want to let users have an unlimited distance too, as it makes floating content useless. (Bonvin, 2018)

This constraint brought us to ask ourselves how efficient an application of this type would be in a disaster scenario. If a technology like this one was deployed in a catastrophic situation, would there

be enough devices (nodes) and would they be close enough to be able to have a viable communication in the affected area? To answer this, we examined the possibility of using Facebook's data sets to visualize and measure the proximity between different users in different disaster scenarios. In theory, given the data at our disposal, it was possible. Unfortunately, after carefully reading the documentation from Facebook, we noticed a big drawback, the granularity of the data was not sufficient:

> Our entire architecture is developed around the Bing Tile (Bing Maps Tile System) gridding of the earth's surface. We surface all of our geospatial metrics according to the smallest tile size that can be completed for that metric's pipeline within it's update frequency time period, starting with the minimum tile size allowed for privacy protections - which is Bing Tile Level 16. This is equivalent to roughly 600m on a side near the equator, or the size of two city blocks. (Facebook, s.d.).

This constraint of 600 meters in tile size was very bad news, we could not precisely determine the distance between different user connections during a natural disaster, thus rendering the data from Facebook close to useless for this use case. After having lost enough time on this possible use case, we decided to slightly change direction and started exploring other possibilities to harness Facebook's data in a disaster situation.

We explored the possibility of using the data at our disposal to examine different scenarios, such as:

- Displacement of a population during and after a wildfire or volcanic eruption.
- Population movement and clusters during the COVID-19 pandemic.
- Locating and predicting power outages in different catastrophic situations.

After having explored many different possibilities for research and many discussions with the research team of the institute. We considered the idea of changing data sources, and even completely redefining the direction of the research. Fortunately, after looking through different data sources, o data from National Oceanic and Atmospheric Administration (NOAA) stood out as a good fit. We discovered datasets involving hurricane tracks that could be used in conjunction with the data from Facebook's Data for Good platform. We decided to pursue the idea of using these datasets to measure the impact of the passage of a hurricane on cellular activity in impacted regions. Following this, we defined the following use case:

Would it be possible to use modern data mining techniques to analyze and predict the impact on cellular activity relative to the distance and strength of a hurricane with the goal of strengthening mitigation and preparation actions taken before a hurricane occurs?

4. IMPLEMENTATION

To implement the defined use case, we intend to aggregate two data sources, the first one being a dataset from Facebook Data for Good and the other a dataset from NOAA. Following this aggregation, we will use various tools to model the data and visualize different metrics we can obtain when we correlate hurricane data to social media data. Finally, we will attempt to use a neural network to make predictions on the aggregated dataset.

Before going through the technical details of the implementation, we will begin by analyzing existing approaches that people have elaborated to harness social media data in the context of a natural disaster, to see what is possible and inspire ourselves. Then, we will determine the different technological requirements for this project and select the appropriate software to carry out this project successfully.

As for project management, we decided to inspire ourselves from the CRISP-DM process. You can find more information about this in the part 6.1 of this thesis. As for the project structure you can view it in appendix I. And if you want to run the project locally you can refer to the environment setup in appendix II.

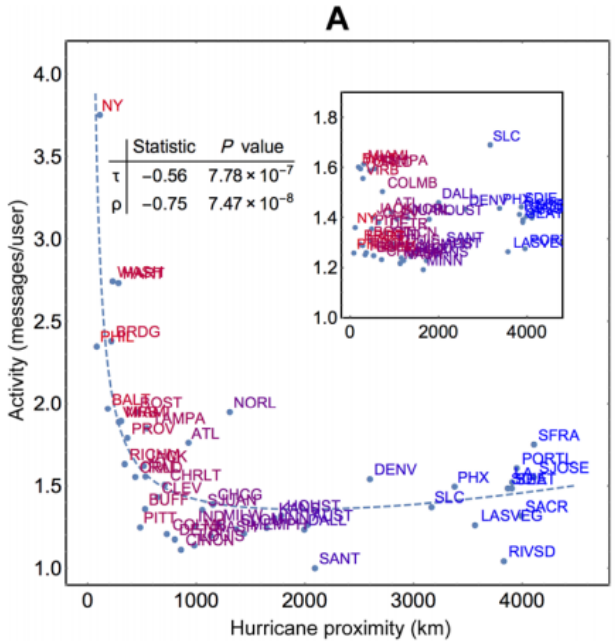
4.1. Analysis of the existing approaches

In the journal article entitled “Rapid assessment of disaster damage using social media activity” by Kryvasheyeu et al. (2016), the authors analyzed tweets that occurred during Hurricane Sandy in 2012 (Kryvasheyeu, et al., 2016). They filtered for the messages posted by users with keywords related to the hurricane (“sandy,” “storm,” “hurricane,” “frankenstorm,” etc.) and determined the following:

> We found that Twitter activity during a large-scale natural disaster—in this instance Hurricane Sandy—is related to the proximity of the region to the path of the hurricane. Activity drops as the distance from the hurricane increases; after a distance of approximately 1200 to 1500 km, the influence of proximity disappears (Kryvasheyeu, et al., 2016).

In Figure 3, we can see that there is a correlation between the proximity of the hurricane and the number of tweets emitted by users containing the word “sandy”.

Figure 3: Characteristic features of Twitter activity mentioning the word “sandy” across locations

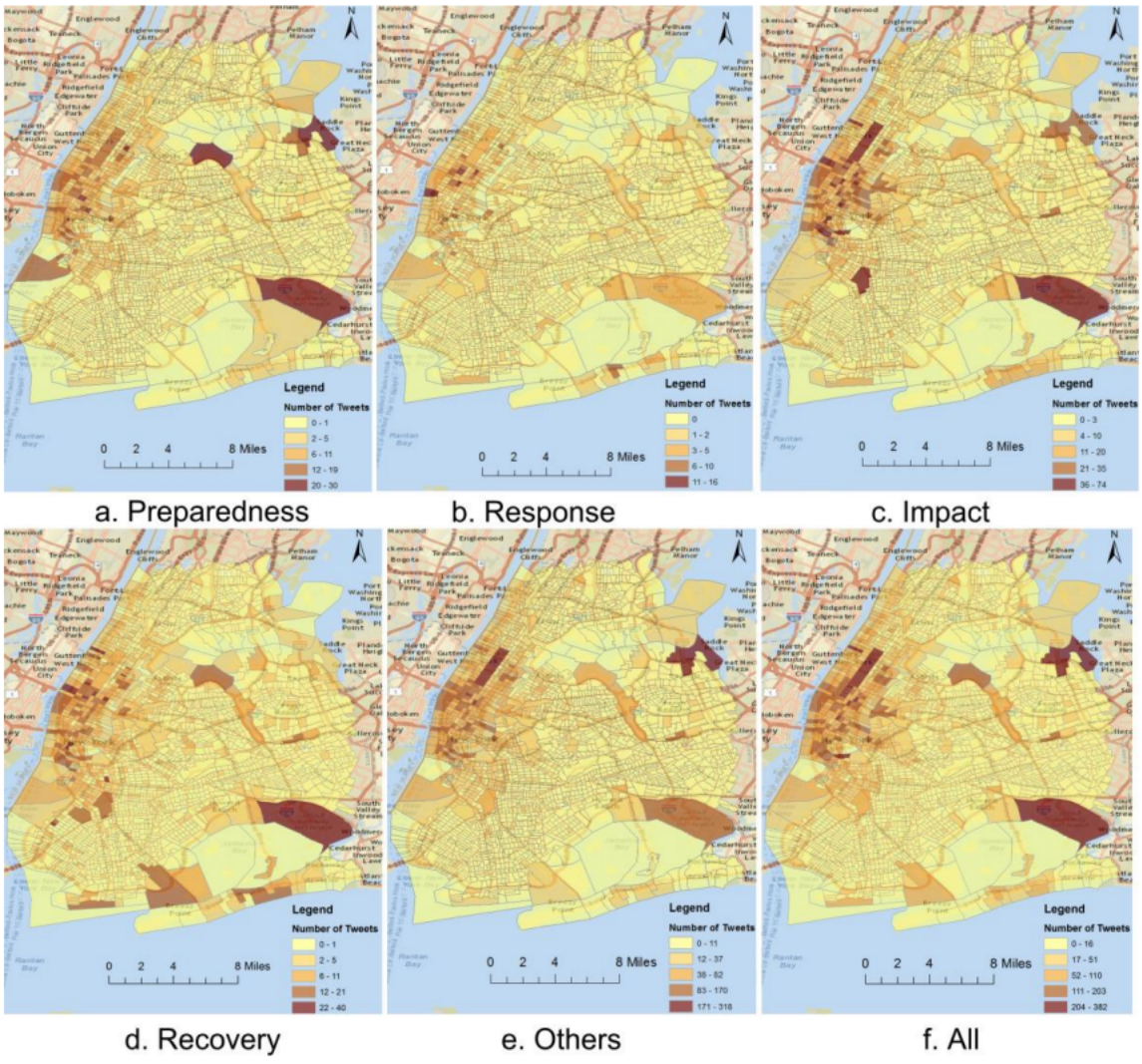


Source : https://www.researchgate.net/

This article is a good example of how social media data can be harnessed to visualize and make statistical assumptions.

Another article came to our attention entitled “Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery” by Huang et al. (2015) (Huang & Yu, 2015). Similarly, to the article by Kryvasheyeu et al. (2016), Twitter is used as a data source. They also analyzed the tweets occurring around the time period of Hurricane Sandy in 2012. But in contrast, they used the contextual information of the tweets, such as hashtags or text, to categorize messages following the disaster management cycle phases described in section 1.4 of this thesis. They were able to investigate Manahan’s citizens’ online behavior in regards to the hurricane by mapping and visualizing tweets of specific themes, as we can see in Figure 4.

Figure 4: The geographical distribution of disaster-relevant tweets within different



Source : https://www.researchgate.net/

With this information, they were able to observe that:

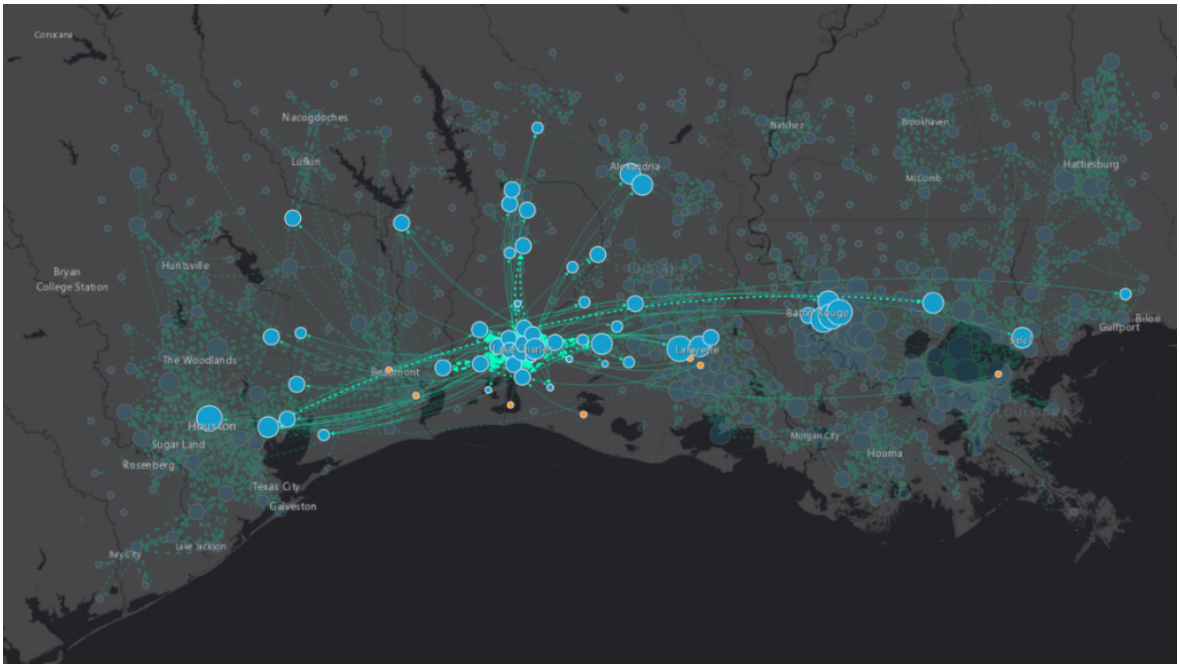
Most tweets were sent from the communities of lower Manhattan, cities within the shore storm surge area such as Hoboken, which lies on the bank of the Hudson River, and Brooklyn on the bank of the East River. These locations were devastated by the storm surge and high winds associated with Hurricane Sandy. Such patterns indicate that Twitter users within impacted neighborhoods are more likely to contribute meaningful data about the disaster (Huang & Yu, 2015).

With this example, we can see how social media content can be associated with different disaster management phases. The article from Huang et al. (2015) also mentions that the information emitted by social media users may not be of the best quality and precision. The reason

for this is that some segments of the population such as people receiving low income, low education or elderly, lack the skills and motivation to use social media in the given context. Furthermore, it is mentioned that in severely damaged areas, participation can be drastically reduced due to damaged infrastructure or people simply fleeing the area. This information is particularly interesting to consider when using social media data for disaster management purposes.

After having considered the previous examples using Twitter messages as data source for disaster impact analysis and management, we remained on a similar topic, but used Facebook's datasets as data source instead. The article by Noah Smith (2020) describes the movement of population before, during and after Hurricane Laura had struck in 2020 (Smith, 2020). The nonprofit humanitarian organization Direct Relief used the data to compare baseline mobility patterns in contrast with patterns that occur before, during and after Hurricane Laura struck Texas. The organization used the data to analyze what portion of the population had received the information of the imminent danger, what areas received the most refugees and where they took shelter. This article is very detailed regarding how the data was processed and used, but the author also generated informative maps like we can see in Figure 5.

Figure 5: Map showing population displacement during the Hurricane Laura in 2020



Source : https://www.directrelief.org/

This map shows: "The average rate of population decrease, relative to baseline, throughout the storm-affected area was 10.6%, with the lowest areas registering a decrease of 32.1%. Blue dots represent outgoing population movement, orange dots represent incoming movement" (Smith, 2020). This type of map enables us to clearly visualize and make assumptions on population displacement during a natural disaster crisis.

While searching for other examples of projects using Facebook's Data for Good data, another interesting article from Akash Yadav (2021) came to our attention (Yadav, 2021). In this article, the primary focus is on network coverage and population density after tropical storm Claudette in Louisiana and Alabama. In the first part, an analysis is performed on the number of days some counties had no network coverage. The interesting part is not about the results obtained, but about the reliability of the data:

> Network Coverage may or may not be directly attributed to the disaster itself and may depend on a number of other factors including whether the cell sites that were used to record the data were used on that particular date or whether there are any Facebook users in that location (Yadav, 2021).

This data consideration is one of the most important to keep in mind while using datasets from Facebook Data for Good. With these different examples, we now have a better understanding on what is possible to do with social media data from Twitter or Facebook regarding disaster management and natural disaster data interpretation.

4.2. Project requirements and technology selection

Being a data mining project, a lot of different software elements are needed to carry out the project successfully. For this project, we evaluated that the following software elements were needed:

- An IDE or a program to orchestrate the flow of the project.
- A development environment or a work directory.
- A program or programming language suited for data mining.
- A program or a library able to output graphs and various visualizations.
- A program or a library that supports machine learning operations.
- A versioning system or a backup system for the project code and data.

These different requirements can of course change or evolve during the project and this is critical to take into consideration. We will now proceed to the evaluation of different options to fulfill the criterion listed above.

After having researched what was available on the market, what was commonly used and what experience we had with technologies for this project, we narrowed down the choice for programs or programming language to three different options: KNIME Analytics Platform, Python and R. We will now analyze these three different technologies with the aim of choosing the best suited one for this project.

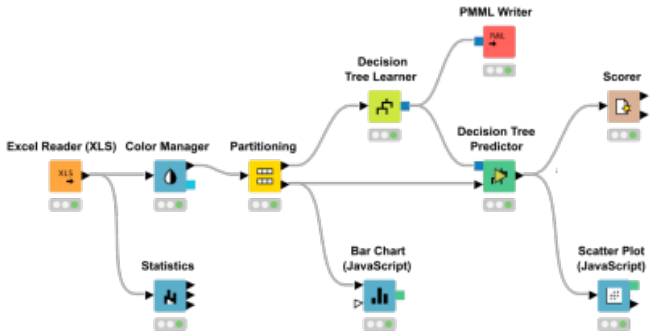
4.2.1. KNIME Analytics Platform

Figure 6: KNIME logo

Source : https://www.knime.com/

KNIME Analytics Platform is an open-source software with which one can perform different data mining and machine learning operations. It is based on the Eclipse and is written in JAVA. The main characteristic of KNIME is that you can build different data science workflows in a visual drag and drop style. Different aspects of a workflow are represented as nodes which are made available by the platform, as we can see in Figure 7.

Figure 7: KNIME analytics Platform workflow example



Source : https://www.knime.com/

KNIME is also built around the principle of collaboration, it is easy to share different workflows with a research team. Furthermore, one can find a lot of information and support on the KNIME community forum. The biggest advantage of KNIME is its ease of use out of the box, with node research functions and different integrated tips and warnings. Unfortunately, ease of use comes with its drawbacks. After having used this platform for more complex operations, such as machine learning and complex graphical representations, we noticed a decline in the ease of use. Often, external libraries or pieces of code are needed.

4.2.2. Python

Figure 8: Python logo

Source : https://www.python.org/

Created in 1980 in, Python has become one of the most popular programming languages in the world. Python can be used for backend development, data science, game development and so much more. Python is truly a multi-purpose programming language.

According to the Python documentation, Python is:

> An interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, very high-level dynamic data types, and classes. It supports multiple programming paradigms beyond object-oriented programming, such as procedural and functional programming. Python combines remarkable power with very clear syntax (Python Software Fundation, s.d.).

An important aspect of Python is not only the programming language itself, but also its impressive community and the high number of libraries and packages available for free. According to (Python Software Fundation, s.d.), a repository of software for the Python programming language, there are almost 320'000 python projects on their platform shared by the python community, all free to use for personal or commercial projects. Furthermore, installing new libraries and packages is easy and straightforward with the pip install command. Overall, Python is an impressive programming language, which can be useful in many situations. It may take a bit of time to learn the syntax and different other subtleties, but it is worth the time spent learning.

4.2.3. R

Figure 9: R logo

Source : https://www.r-project.org/

Created in 1993, R is a programming language built for statistical computing and graphics. R is widely used by data scientists and other scientific fields. The programming syntax and different components of the R programming language are made to be easier to learn and understand, which makes it a bit more accessible than a traditional programming language, unlike Python or JAVA.

According to the r-project website (R fundation, s.d.), the R environment includes the following features:

- An effective data handling and storage facility;
- A suite of operators for calculations on arrays, in particular matrices;
- A large, coherent, integrated collection of intermediate tools for data analysis;
- Graphical facilities for data analysis and display either on-screen or on hardcopy;
- A well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

On the downside, R seems not to be as flexible as some other environments. The fact that R is focused mainly on data analysis, statistics and visualizations, make it an excellent scientific tool but a lot of options like making client applications, integrating to existing projects or deployment seem not to be as straightforward as other environments.

4.2.4. Choice for the project

With information gathered from different sources like (DataCamp Team, s.d.) and personal experience, we elaborated the following decision matrix, with a scale from zero to ten, zero being the lowest grade and ten being the highest, to facilitate the selection of a technology for this project:

Table 1: Technology decision matrix

	Knime	Python / Pandas	R
Personal Experience	6	5	1
Integrability	3	8	4
Ease of use (short term)	8	4	7
Ease of use (long term)	4	6	5
Portability	2	8	4
Complexity	4	5	5
Data visualization	4	5	8
Flexibility	4	8	5
Community / packages	3	6	5
Total	38	55	44

Source: Author

After comparing the advantages and disadvantages of these three technologies and with Python having the highest score in the decision matrix, we decided to use Python, not only because of the experience we have with this technology, but we also may want to be able to easily integrate the project in a larger context (web app or mobile app). And because Python has many packages and support, we gain a lot of different possibilities for future implementations. Nowadays, being flexible and futureproof is critical.

4.2.5. Choice of Python packages

Since this project will be Python based, an important aspect of the implementation will be the libraries and packages we use to carry out this project successfully. We selected the following packages and libraries that we think will be of good use in this project, the descriptions were provided by (Python Software Fundation, s.d.):

Table 2: Python packages used for the project

Python package	Description
Virtualenv	A tool for creating isolated virtual python environments.
Jupyterlab	An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture
Pandas	Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.
Plotly	Ploty is an interactive, open-source, and browser-based graphing library for Python
Numpy	NumPy is the fundamental package for array computing with Python.
Harvesine	Calculate the distance (in various units) between two points on Earth using their latitude and longitude.
Keras	Keras is a high-level neural networks API for Python.
Tensorflow	TensorFlow is an open-source software library for high performance numerical computation. Its flexible architecture allows easy deployment of computation across a variety of platforms (CPUs, GPUs, TPUs), and from desktops to clusters of servers to mobile and edge devices.
Tqdm	Fast, Extensible Progress Meter

Source: https://pypi.org/

4.2.6. Backup technologies of the project and data

Given the size of the datasets and how the project is structured, we decided to follow the 3-2-1 backup strategy. According to Acronis, a global technology company that develops on-premises and cloud software for backup and disaster recovery: to protect efficiently sensitive data from unexpected events or disasters it is recommended to follow the 3-2-1 rule of backup, this rule states that you should create three copies of your data, store two of the copies in at least 2 types of storage media and store the last copy offsite (Acronis, 2021). To follow this rule, and have different versions of our

work, every time we performed significant work on the data or the thesis or reached a milestone, we made three copies of the files and stored one copy on our computer hard drive, one copy on an external SSD and the last copy on Microsoft OneDrive.

4.3. Data gathering and understanding

We started by selecting datasets from Facebook Data for Good with a wide enough time range, so that we would have more flexibility on the choice of a hurricane, we selected the following dataset: "Florida Hurricane Season Disaster Map August 25 2020 id (Facebook traffic)". It is important to mention that the measurements in this dataset were not only for the month of august 2020 but ranged from 26th of august 2020 to the 22nd of November 2020.

After having selected our first dataset we then selected a hurricane that happened from the 11th of September 2020 to the 17th of September 2020, which was Hurricane Sally. Firstly, we chose Hurricane Sally because the dates matched Facebook's Data for Good dataset and secondly because it was very destructive, according to a report made by NOAA, hurricane Sally caused $7.3 billion (USD) in damage in the United States (mainly in Florida and Alabama) and was responsible of four direct and five indirect fatalities (Berg & Reinhart, 2021).

4.3.1. Facebook Data for Good

According to the global digital report, there are now 4.20 billion social media users around the globe (Kemp, 2021). The number of users has grown by 490 million in the past year, which represents a growth of over 13 percent. The number of people that use social media now accounts for more than 53% of the global population.

Other important numbers can be extracted from the worldwide market share of different social media platforms.

Table 3: Social media platforms market share 2021

Platform	Facebook	Twitter	Pinterest	YouTube	Instagram	Tumblr
Share	71.53%	10.01%	7.31%	5.01%	4.28%	0.89%

Source : https://gs.statcounter.com/social-media-stats

With these statistics, we can now see that Facebook is the biggest actor in the field. As a matter of fact, Facebook had roughly 2.85 billion active users worldwide in the first quarter of 2021 (Tankovska, 2021). This is more than 37% of the world's population.

According to an article by Ankush Sinha Roy (2020), Facebook generates around four petabytes of data a day, which corresponds to four million gigabytes of data (Roy, 2020). Fortunately, as part of a humanitarian initiative, Facebook made available part of its data in an aggregated and anonymized way to NGO's and academic institutions, this initiative is called Facebook Data for Good.

Figure 10: Facebook Data for Good logo

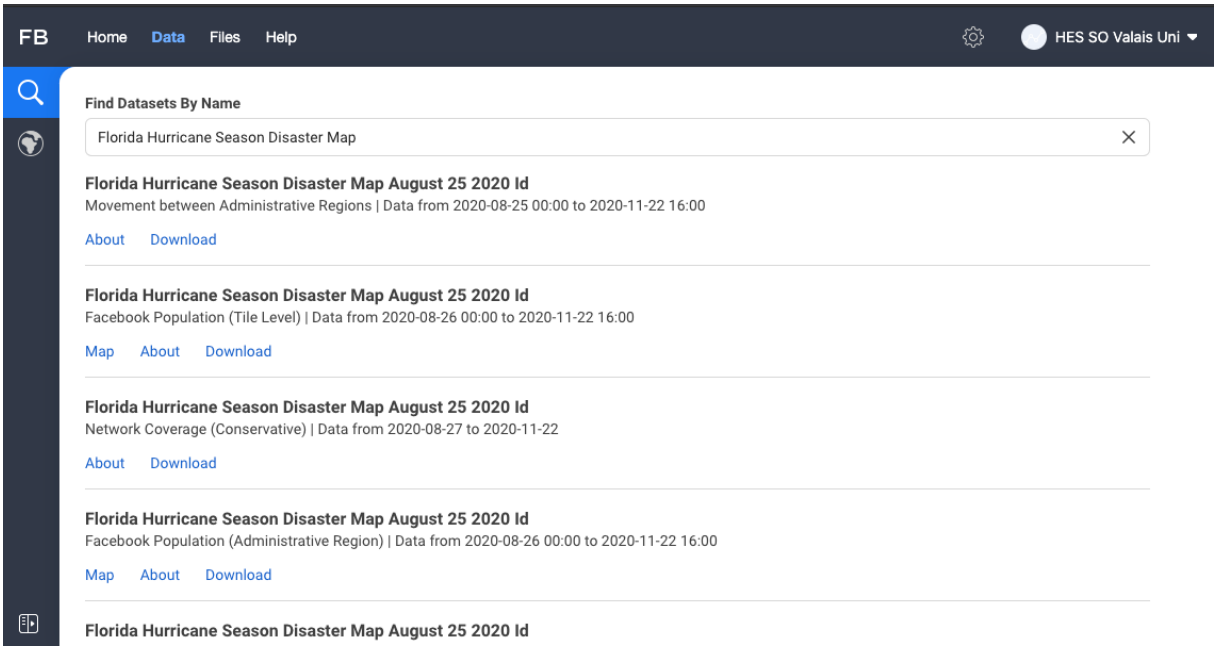
FACEBOOK Data for Good

Source : https://dataforgood.fb.com/

As stated on Facebook Data for Good's website:

We use a variety of technical tools to help our partners access and use data for disaster response, health, connectivity, energy access, and economic growth. Privacy is built into all of our products by default; we use techniques such as aggregation and differential privacy to protect your privacy (Facebook Data for Good, 2021).

Figure 11: UI of the Facebook Data for Good platform



Source : https://partners.facebook.com/

Figure 11 illustrates how the platform is presented. In the top search bar, one can search for different datasets made available by the platform. If one cannot find the desired datasets, there is always the option to ask the Facebook Data for Good team on Slack for complementary datasets. We eventually had to do this for this project because some datasets would not download properly, see appendix III for more details.

To download the dataset in csv format we can simply click on the download link and choose the date range that we are interested in.

For this project, we concentrated ourselves on the disaster map datasets which include the following types:

- Facebook population density maps
- Movement maps
- Power availability maps
- Simple network coverage maps
- Complex network coverage maps
- Displacement maps
- Community help maps
- Commuting zone maps.

We contemplated the idea of using power availability maps, but after having looked and talked to the people of contact at Facebook Data for Good, we realized that power availability maps were not the best choice since they were not always available and contained a lot of noise, you will be able to parts of the conversations with the Facebook team in appendix III.

For this thesis, we concentrated our efforts on the complex network coverage maps, these maps show where Facebook users have cellular connectivity through their mobile device and have location services enabled. With these maps Facebook also determines the speed and latency of the user connecting to their services. As stated in the documentation of Facebook Data for Good, network coverage maps can help to: "Identify cell sites that show increases in user traffic and those that are likely to be down" (Facebook Data for Good, 2021).

We will not go into detail for each type of dataset since we are going to concentrate our attention on the Facebook traffic maps which are a subtype of the complex network coverage maps in this project. As defined on Facebook's Data for Good documentation (Facebook, s.d.), Facebook traffic maps are elaborated in the following manner:

> We poll users that have location services enabled at a consistent rate for the cell sites their phone could communicate with. Each poll returns up to 50 cell sites and the user's current geo-location, and each one such poll response is a ping that we aggregate into grid cells in this map. We calculate the number of pings that originate in each grid cell per day for a baseline-period of 12 weeks before the start of the crisis. We then compare the average ping count per day-of-the-week to the after-crisis ping count (Facebook, s.d.).

In simpler words, Facebook collects the data from users connecting to different cell sites and their geo-location before a crisis and compares them to connections during a crisis.

The traffic maps are composed of the following data fields:

Table 4: Facebook traffic maps data fields

Name	Description	Example
lat	Latitude	26.559050
lon	Longitude	-81.947021
quadkey	Bing tile system id	3202301123112
z_score	Difference in standard deviations from the baseline mean	-1.247856
baseline_mean_num_signals	Mean number of signals observed for this day of the week over a baseline period for up to three months for a tile	14017.51
aftercrisis_num_signals	The number of signals observed for this day for a tile during the crisis.	13738.40
country	Country	US
ds	Date	2020-08-25

Source: Based on the dataset downloaded on Facebook's Data for Good platform

For this project we decided to focus our attention on the latitude, longitude, "z_score" and date. We focused on the "z_score" because Facebook Data for Good defined the "z_score" as:

> Z-Score of the number of pings during the crisis compared to the mean value of signals for the same day of the week in a baseline of up to three months prior to the crisis. The z-score is the difference in standard deviations from the baseline mean (Facebook, s.d.).

The standard deviation will help to show us if data points are close or not to the baseline mean. In case they are far from the mean, we will know that an abnormal peak or drop in the number of cellular connections has occurred, which is an interesting metric when combined with data sources from the NOAA.

4.3.2. NOAA

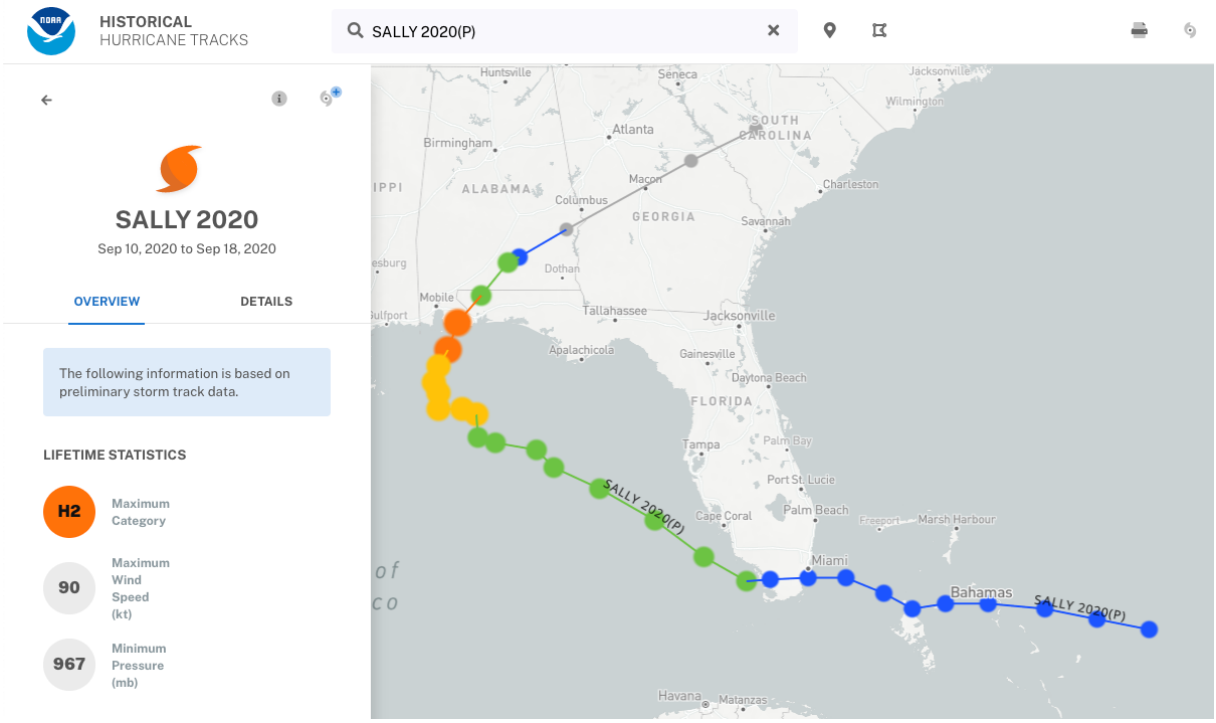
Figure 12: NOAA logo

Source : https://coast.noaa.gov/

NOAA is an American scientific and regulatory agency and part of the United States Department of commerce. They concentrate their efforts on weather forecasting, atmospheric and oceanic condition monitoring as well as the management and protection of marine mammals and endangered species in the U.S (NOAA, 2021).

The agency has a platform that lets you visualize historical information on different hurricanes and tropical storms. In Figure 13, we can observe the tracks of Hurricane Sally that happened in 2020.

Figure 13: Track of Hurricane Sally on NOAA's platform



Source : https://coast.noaa.gov/hurricanes/

The different colors of the data points on the maps represent different wind speeds, gray being the lowest wind speed and orange the highest. The source of the data for this platform comes from another platform of the NOAA called IBTrACS (NOAA, s.d.), where data from the past hurricanes and tropical storms are archived and the records are made available for download in csv, NetCDF and shapefile format. The IBTrACS platform is currently in its fourth version, and the data is updated twice a week. For the purpose of this project, we decided to select the data sources in the csv file format. Furthermore, to stay flexible we decided to download and use the biggest dataset (300 megabytes) of the platform which is composed of all storms and hurricanes available in the IBTrACS record.

As mentioned in part 3 of the thesis our aim is to try correlating the information that we have on past hurricane tracks and the Facebook traffic data that we have from the same time-period. To accomplish this, we selected the "Name", "Description" and "Example" columns of the IBTrACS dataset (Table 5). Each row of the dataset represents a given location and time of the eye of the hurricane.

Table 5: IBTrACS data fields used in this project

Name	Description	Example
NAME	Name of the hurricane or storm	SALLY
ISO_TIME	Date and time of the hurricane or storm	2020-09-10 06:00:00
LAT	Latitude of the data point	24.6000
LON	Longitude of the data point	-73.0000
STORM_SPEED	Speed of the hurricane or storm at that time	10
STORM_DIR	Wind direction of the hurricane or storm at that time	281

Source: Based on the data gathered on the NOAA's platform

4.4. Data preprocessing and preparation

Now that we have a better idea of the origin and structure of the data sources we will be using in this project, we are going to go over the operations of data transformation that were executed to prepare the data to be usable in the analysis phase.

You will be able to find the source code in the file "prepreocessing_transformation_aggregation (facebook traffic + NOAA).ipynb" in the attached files of this thesis.

4.4.1. Data importation

For the importation of the data, we used a function of Pandas library called “read_csv”. For the data from the NOAA this was straightforward:

Figure 14: Importation of the IBTrACS dataset

```
# import data from the IBTrACS dataset and convert them into a pandas dataframe
ibtracs_df = pd.read_csv(r'./datasets/ibtracs.ALL.list.v04r00.csv', low_memory=False, skiprows=[1])
```

Source: Author

We had to specify in the previous figure “low_memory=False” because the size of the dataset, pandas, was unable to process the data with low memory settings. Additionally “skiprows=[1]” was also specified because of table structure, there was a row containing unit measurements beneath the header row that was causing errors.

As for the Facebook traffic data it was a bit more tedious since the datasets were separated by day. We had to loop over the different files and aggregate them into one data frame:

Figure 15: Importation of the Facebook traffic datasets

```
# Specify path of directory where the datasets are located
path = r'./datasets/facebook_traffic'
# Use glob to get an array paths to the datsets
all_files = glob.glob(path + "/*.csv")

fb_traffic_df = pd.DataFrame()
# Iterate over the file paths array and append them to a dataframe
for f in all_files:
    df = pd.read_csv(f)
    fb_traffic_df = fb_traffic_df.append(df,ignore_index=True)
```

Source: Author

The datasets are now respectively stored in “ibtracs_df” and “fb_traffic_df” as data frames.

4.4.2. Data transformation

The data being of different origins, it is sometimes required to format some elements of the datasets to be able to use them together. When this happened, we had to format the dates of both datasets to be able to do operations on them with pandas. We used the “to_datetime” function from the Pandas library to convert the dates in a processable format.

Figure 16: Conversion of the ISO_TIME column

```
ibtracs_df['ISO_TIME'] = pd.to_datetime(ibtracs_df['ISO_TIME'], format='%Y-%m-%d %H:%M:%S')
```

Source: Author

In this case we had to specify the format of the date so that the "to_datetime" function could correctly parse the date. The date from the Facebook traffic data frame was also converted in the same way.

4.4.3. Data filtering

After having transformed the data, we had to filter some of the data to suit our needs. We started by filtering the IBTrACS data frame to obtain the desired hurricane, in our case, hurricane Sally that occurred in 2020. To do this, we harnessed Panda's powerful filtering capabilities, as we can see in the following figure:

Figure 17: Filtering of the IBTrACS data frame

```
# filter the IBTrACS dataframe by year and name to get desired hurricane data
ibtracs_by_year_df = ibtracs_df[ibtracs_df['ISO_TIME'].dt.year == 2020]
hurricane_df = ibtracs_by_year_df[ibtracs_by_year_df['NAME'] == 'SALLY']
# Store the first and last dates of the hurricane for later usage
hurricane_date_min = hurricane_df['ISO_TIME'].min()
hurricane_date_max = hurricane_df['ISO_TIME'].max()
```

Source: Author

We now have a data frame named "df_hurricane" that contains all the information about the tracks of hurricane Sally.

Two variables "hurricane_date_min" and "hurricane_date_max" were used to store the first and last dates of the hurricane, this enabled us to filter the Facebook traffic data frame to only contain data about the time period in which hurricane sally occurred, as you can see in the following figure:

Figure 18: Filtering of the Facebook traffic data frame

```
fb_traffic_df = fb_traffic_df[(fb_traffic_df['ds'] >= hurricane_date_min - datetime.timedelta(1)) &
                              (fb_traffic_df['ds'] <= hurricane_date_max)]
```

Source: Author

4.4.4. Data aggregation

After having imported, transformed and filtered the datasets, the next step was to aggregate the data to be able to measure the influence of the hurricane on cellular connections. To measure the influence of cellular connections, we had to combine each data point of the hurricane track, which corresponds to the place and time of the eye of the hurricane, with all the datapoints of the Facebook traffic dataset at a given date.

Figure 19: Aggregation of the hurricane and Facebook data frames

```
counter = 0
dictionnary = {}

# Iterate over every row or the hurricane dataframe
for i in tqdm(range(len(hurricane_df))):
    track_series = hurricane_df.iloc[i]
    track_series_date = track_series['ISO_TIME'].replace(hour=0, minute=0, second=0)
    # Filter the facebook traffic dataset to only get traffic records corresponding to
    # the date of the hurricane record
    fb_traffic_by_track_date_df = fb_traffic_df[(fb_traffic_df['ds'] == track_series_date)]
    # Iterate over the filtered Facebook traffic data frame and append information
    # on the eye of the hurricane
    for j in range(len(fb_traffic_by_track_date)):
        fb_traffic_by_track_date_series = fb_traffic_by_track_date_df.iloc[j]
        # Calculate the distance from the facebook traffic location to the location
        # of the eye of the hurricane
        distance = haversine((fb_traffic_by_track_date_series['lat'],
                              fb_traffic_by_track_date_series['lon']),
                             (track_series['LAT'],track_series['LON']))
        # Combine the data of the traffic measurement and the eye of the hurricane measurement
        dictionnary[counter] = {'traffic_lat': fb_traffic_by_track_date_series['lat'],
                                'traffic_lon': fb_traffic_by_track_date_series['lon'],
                                'traffic_z_score': fb_traffic_by_track_date_series['z_score'],
                                'traffic_date': fb_traffic_by_track_date_series['ds'],
                                'track_lat': track_series['LAT'],
                                'track_lon':track_series['LON'],
                                'track_storm_speed':track_series['STORM_SPEED'],
                                'track_storm_direction':track_series['STORM_DIR'],
                                'track_distance':distance,
                                'track_date': track_series['ISO_TIME']}
        counter += 1

combined_df = pd.DataFrame.from_dict(dictionnary, "index")
```

```
100%|██████████| 63/63 [04:10<00:00,  3.98s/it]
```

Source: Author

In the previous figure we can see how we iterated over each row of the hurricane data frame to be able to append hurricane information to the corresponding Facebook traffic data point, we then used Haversine to calculate the distance between the eye of the hurricane and the tile of the Facebook traffic measurement. We used a dictionary object to accomplish this, because at first, we tried using the “append” function of the Pandas library to directly append data to a data frame but given the quantity of the data and Pandas’ indexing system, this operation was unable to complete. We even tried to run the operation for more than 72 hours without success. Fortunately, we discovered that, by using a dictionary object to append the data and then converting this object into a data frame, this operation took not more than five minutes. Something that helped us to measure the time and progress of the operation was by using the “tqdm” library, which is the pink bar you can see in Figure 19.

In Figure 20, we can see a preview of the previously generated data frame.

Figure 20: Preview of the aggregated data frame

	traffic_lat	traffic_lon	traffic_z_score	traffic_date	track_lat	track_lon	track_storm_speed	track_storm_direction	track_distance	track_date
0	29.161755	-81.002197	-3.361173	2020-09-10	24.6	-73.0	10	281	941.485238	2020-09-10 06:00:00
1	30.135626	-84.407959	2.430513	2020-09-10	24.6	-73.0	10	281	1282.684653	2020-09-10 06:00:00
2	30.401306	-84.100342	-3.120001	2020-09-10	24.6	-73.0	10	281	1269.694850	2020-09-10 06:00:00
3	30.685163	-85.001221	-1.812275	2020-09-10	24.6	-73.0	10	281	1360.822900	2020-09-10 06:00:00
4	30.306503	-85.155029	0.314988	2020-09-10	24.6	-73.0	10	281	1355.665348	2020-09-10 06:00:00
...	...	...	...	...	...	...	...	...	...	...
1210225	30.628458	-87.330322	-5.692138	2020-09-18	33.9	-81.3	2	82	673.465803	2020-09-18 00:00:00
1210226	25.770213	-80.474854	-1.398438	2020-09-18	33.9	-81.3	2	82	907.477636	2020-09-18 00:00:00
1210227	30.666265	-83.397217	-1.657516	2020-09-18	33.9	-81.3	2	82	410.048048	2020-09-18 00:00:00
1210228	29.602118	-82.891846	-3.196026	2020-09-18	33.9	-81.3	2	82	501.022172	2020-09-18 00:00:00
1210229	27.362010	-81.156006	1.115164	2020-09-18	33.9	-81.3	2	82	727.122554	2020-09-18 00:00:00

1210230 rows × 10 columns

Source: Author

We see now that for each Facebook traffic record we have a corresponding hurricane track record. In the bottom of Figure 20, we also notice that this newly generated data frame contains 1’121’230 rows.

After generating the aggregated data frame, we needed to export the newly generated data into a csv file. Which we used in the following steps. To accomplish this, we simply used the “to_csv” function of the Pandas library. The csv file is saved in the “datasets” directory of the project.

4.5. Data Modeling

After having completed the data preprocessing and preparation steps we could finally start analyzing data and start making assumptions about possible correlations between hurricane tracks and Facebook traffic.

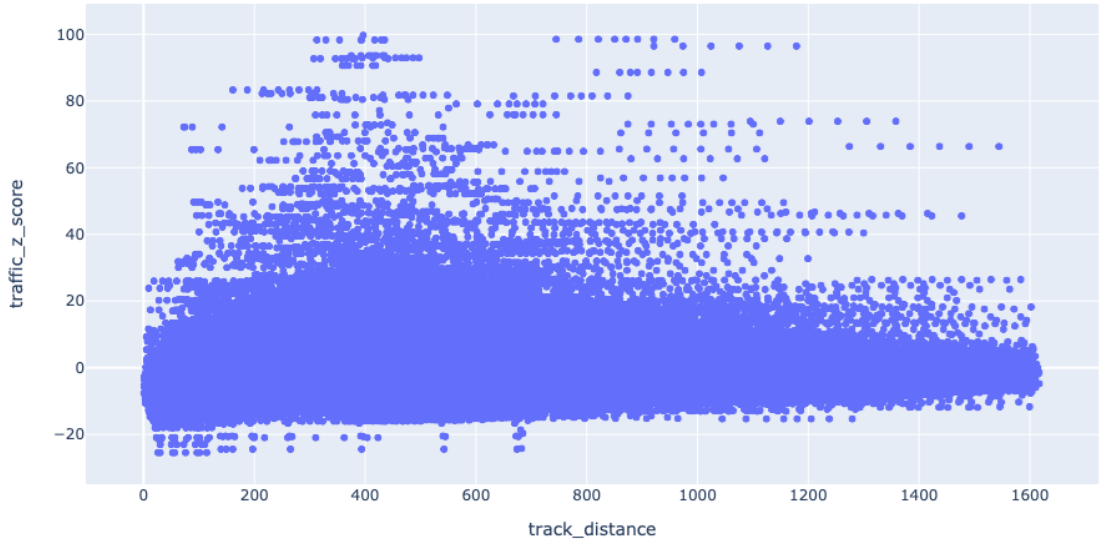
4.5.1. Pre-analysis of the data with Plotly

You will be able to find the source code in the file “dataset interpretation with plotly.ipynb” in the attached files of this thesis.

As mentioned in part 4.3.1 we are using the “z_score” of the Facebook traffic data to determine if there is an abnormal number of connections compared to the mean baseline at a certain location. To accomplish this, we used the Plotly library, a graphing library for python, to try and see possible correlations between outlying cellular connections and the distance from the eye of a hurricane.

After having imported the dataset generated mentioned in part 4.4 of this thesis, we elaborated a scatter plot to visualize graphically if we could see possible outliers in z_scores of the Facebook traffic data compared to the distance of the eye of the hurricane (Figure 21).

Figure 21: Scatter plot comparing the z_score and track distance

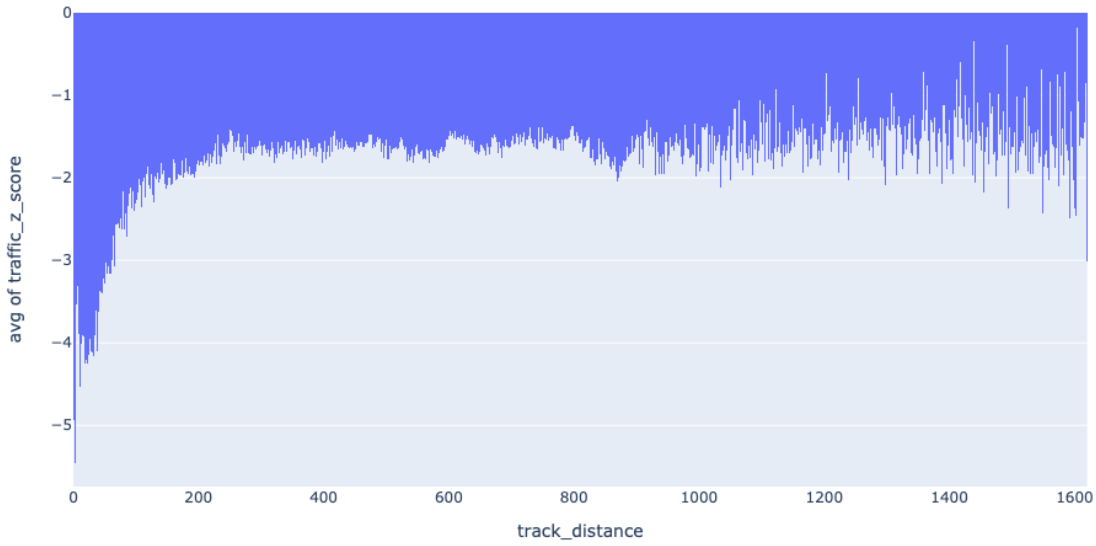


Source: Author

The X axis of the above scatter plot represents the distance in kilometers of the eye of the hurricane compared to the Bing tile in which the "z_score" was measured. The Y axis of this scatter plot represents the "z_score" itself which is the standard deviation of the baseline mean of the number of cellular connections. If we observe this graphic, we can see that between 300 and 600 kilometers there are many data points that have a high standard deviation. With this information we then confirmed our hypothesis of a possible increase of cellular activity compared to the distance of the hurricane.

Although we had confirmed our hypothesis, we were missing information that the scatter plot was unable to give us. Since a scatter plot allows overlapping points, we did not exactly know how many of the data points had a high or low standard deviation. To overcome this, we decided to elaborate a histogram that would correlate the average of the "z_scores" at a given distance (Figure 22).

Figure 22: Histogram comparing average z_scores and the distance of the hurricane



Source: Author

The X axis of the histogram in Figure 22 represents the distance in kilometers and the Y axis represents the average of "z_scores". With this histogram we can see the average of all the "z_scores" at a given distance. To our greatest surprise, the outcome of this histogram was completely different from what we expected. The average of the "z_scores" that were at a short distance from the eye of the hurricane tended to be more negative. This means that the closer the hurricane was from a location, the more there was a drop in the cellular activity compared to the mean baseline in a given tile. This metric tells us that people possibly fled the area or that some network infrastructures were damaged by the hurricane. Another interesting observation we can make is that after 200 kilometers, the average "z_score" is around -1.5, which means that there was a severe drop in the number of cellular connections overall, and not just related to the hurricane. This can be explained by the fact that people left the area before the hurricane season, resulting in a drop in the number of connections to Facebook in the area.

4.5.2. Implementation of a neural network

After having analyzed the data, we asked ourselves how we could possibly use the data at our disposal not only to make observations but also to make predictions on the consequences that a hurricane could have on cellular connections. To accomplish this, we decided to implement a neural network that would have as input the data from our dataset and as output a prediction of whether a "z_score" tends to be close or not to zero.

In the following figure, we imported the previously generated dataset and selected the fields of interest for the inputs and outputs of the neural network:

Figure 23: Importation of the dataset and selection of fields

```
df = pd.read_csv(r'./datasets/fb_traffic_hurricane_track_combined-v3.csv', low_memory=False)
df = df.filter(items=['traffic_lat',
                      'traffic_lon',
                      'track_lat',
                      'track_lon',
                      'track_storm_speed',
                      'track_storm_direction',
                      'track_distance',
                      'traffic_z_score'])
df["outliers"] = df["traffic_z_score"].apply(lambda x: 1 if x < -1 or x > 1 else 0)
```

Source: Author

We now have a data frame containing the desired information to be processed. We also needed to define what our output parameters would be. For this implementation of a neural network the output can only be zero or one. We decided to consider an outlier a "z_score" that was superior to one or inferior to minus one, which is represented as a "1" in the data frame, otherwise if the "z_score" is not considered as an outlier its value is represented as "0". Figure 24 represents a preview of the dataset that will be processed by the neural network:

Figure 24: Preview of the data frame used in the neural network

	traffic_lat	traffic_lon	track_lat	track_lon	track_storm_speed	track_storm_direction	track_distance	traffic_z_score	outliers
0	29.161755	-81.002197	24.6	-73.0	10	281	941.485238	-3.361173	1
1	30.135626	-84.407959	24.6	-73.0	10	281	1282.684653	2.430513	1
2	30.401306	-84.100342	24.6	-73.0	10	281	1269.694850	-3.120001	1
3	30.685163	-85.001221	24.6	-73.0	10	281	1360.822900	-1.812275	1
4	30.306503	-85.155029	24.6	-73.0	10	281	1355.665348	0.314988	0
...	...	...	...	...	...	...	...	...	...
1210595	30.628458	-87.330322	33.9	-81.3	2	82	673.465803	-5.692138	1
1210596	25.770213	-80.474854	33.9	-81.3	2	82	907.477636	-1.398438	1
1210597	30.666265	-83.397217	33.9	-81.3	2	82	410.048048	-1.657516	1
1210598	29.602118	-82.891846	33.9	-81.3	2	82	501.022172	-3.196026	1
1210599	27.362010	-81.156006	33.9	-81.3	2	82	727.122554	1.115164	1

1210600 rows × 9 columns

Source: Author

After having set up the data, we had to convert the Pandas data frame into a Numpy array to be able to process the data. Furthermore, we needed to define the different input and output fields in of the Numpy array:

Figure 25: Conversion to Numpy and input and output selection

```
dataset = df.to_numpy()
X = dataset[:,0:7]
y = dataset[:,8]
```

Source: Author

After having done this, we needed to separate the data into training data and prediction data, it is a common good practice to select only two thirds of the data for training and one third of the data to be used for predictions. To accomplish this, we used the function from a library called scikit-learn:

Figure 26: Separation of training and prediction data

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

Source: Author

The training and prediction data being ready, we then implemented the neural network:

Figure 27: Implementation of the neural network

```
# define the keras model
model = Sequential()
model.add(Dense(12, input_dim=7, activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(2, activation='relu'))
model.add(Dense(1, activation='sigmoid'))
# compile the keras model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])
# fit the keras model on the dataset
model.fit(X_train, y_train, epochs=50, batch_size=10)
# evaluate the keras model
_, accuracy = model.evaluate(X_train, y_train)
print('Accuracy: %.2f' % (accuracy*100))
```

Source: Author

We used the Keras library made for Python. First, we had to define the type of the model, which in our case is sequential. We then implemented the different layers with which the data will be processed. After doing this we needed to compile and fit the model. It is when fitting the model that the neural network is trained based on the previously defined model. Finally, we use the "evaluate" function to see the accuracy of the model on the dataset. It is to be noted that we based this implementation on a tutorial from Jason Brownlee (Brownlee, s.d.).

Figure 28: Neural network training results

```
Epoch 1/5
81111/81111 [==============================] - 54s 658us/step - loss: 0.5207 - accuracy: 0.7870
Epoch 2/5
81111/81111 [==============================] - 57s 704us/step - loss: 0.5175 - accuracy: 0.7872
Epoch 3/5
81111/81111 [==============================] - 62s 765us/step - loss: 0.5175 - accuracy: 0.7872
Epoch 4/5
81111/81111 [==============================] - 60s 741us/step - loss: 0.5175 - accuracy: 0.7872
Epoch 5/5
81111/81111 [==============================] - 60s 735us/step - loss: 0.5175 - accuracy: 0.7872
25347/25347 [==============================] - 14s 528us/step - loss: 0.5175 - accuracy: 0.7872
Accuracy: 78.72
```

Source: Author

When running the training we obtained the results described in Figure 28. We were happy to notice that the accuracy of the model on the dataset was at 78.72%. This meant in theory that when using data of the same nature, we would be able to predict if a “z_score” has a standard deviation above one or minus one in almost 80% of the cases. This indicated that it would be possible to measure the increase or decrease in cellular activity expected in a given location relative to a hurricane track.

Sadly, when trying to use the model to make a prediction with the testing data that we had put aside, the outcome was very different as shown in Figure 29.

Figure 29: Results of predictions based on the neural network model

```
predictions = model.predict_classes(X_test)
for i in range(len(predictions)):
    print('%s => %d (expected %d)' % (X_test[i].tolist(), predictions[i], y_test[i]))
```

```
[30.760718446021, -86.319580078125, 33.8895, -81.3983, 13.0, 63.0, 578.5443649408398] => 1 (expected 0)
[29.602117759256, -81.925048828125, 25.1, -77.3, 7.0, 263.0, 677.5347542536721] => 1 (expected 1)
[29.983486262534, -82.100830078125, 27.0226, -84.0217, 12.0, 300.0, 378.9629839067636] => 1 (expected 1)
[29.983486262534, -82.869873046875, 28.8, -87.5, 4.0, 270.0, 467.4401962052241] => 1 (expected 1)
[28.488004762593, -81.419677734375, 30.4, -87.6, 6.0, 31.0, 634.9827371726709] => 1 (expected 1)
[30.760718446021, -85.792236328125, 25.1, -77.3, 7.0, 263.0, 1044.477704281967] => 1 (expected 0)
[29.964452395086, -83.177490234375, 28.2667, -87.1667, 3.0, 337.0, 431.02457790431851 => 1 (expected 0)
[29.180940841623, -82.342529296875, 28.9349, -88.0301, 3.0, 0.0, 553.4476222700359] => 1 (expected 1)
[29.773913415992, -81.595458984375, 25.5121, -81.2428, 5.0, 266.0, 475.16311135990776] => 1 (expected 1)
[28.642388714794, -82.342529296875, 26.7, -83.4333, 12.0, 304.0, 241.21624562093257] => 1 (expected 1)
[30.533876112077, -87.659912109375, 24.8, -74.1, 10.0, 282.0, 1478.2658089066022] => 1 (expected 0)
[29.773913415992, -81.353759765625, 27.0226, -84.0217, 12.0, 300.0, 402.08053170807887] => 1 (expected 1)
[27.147144408338, -80.277099609375, 28.1667, -86.8, 6.0, 277.0, 652.2794528009407] => 1 (expected 0)
[26.716173334953, -80.079345703125, 28.194, -87.0459, 3.0, 287.0, 706.6476754262372] => 1 (expected 1)
[30.85507882322, -84.869384765625, 25.0, -78.0, 6.0, 279.0, 937.2796289563328] => 1 (expected 1)
[26.735798597235, -81.221923828125, 25.0651, -75.815, 11.0, 275.0, 571.7826021648191] => 1 (expected 1)
[29.831113310205, -81.287841796875, 26.3398, -82.8967, 12.0, 307.0, 419.0565664791207] => 1 (expected 0)
[28.06228556256, -82.210693359375, 29.4425, -88.0725, 3.0, 16.0, 591.6102978804993] => 1 (expected 1)
[30.116621125658, -82.891845703125, 29.6, -88.0, 3.0, 26.0, 495.8994104303076] => 1 (expected 1)
```

Source: Author

When making predictions, the model would never output the correct prediction when expecting an output of 0, the prediction would always be 1. Even after trying different approaches and changing parameters in the model definition, we were never able to make accurate predictions. It was very difficult for us to determine exactly the reason for the inaccurate predictions. Our best guess is that the data was not meaningful enough or that this implementation of the neural network was not the best suited one for this case.

4.6. Evaluation of the results

4.6.1. Technical analysis

If we summarize what has been discovered through the different data models we elaborated, we notice on one hand that when looking at the scatter plot, we can observe an abnormal increase in cellular activity compared to the baseline in areas located 300 to 600 kilometers away from the eye of the hurricane, this could mean that people had a need to communicate about the approaching disaster or find information about Hurricane Sally as it came closer to their current location. On the other hand, with the histogram, we noticed that, on average, cellular activity tended to be lower than usual the closer the users were to the hurricane. This can be explained by the fact that users left the area beforehand or that network infrastructures were damaged resulting in a drop of cellular connections. At first glance, these two observations can seem contradictory, but in fact we can elaborate that the general tendency is that users left the area close to the approaching hurricane or that network infrastructures were damaged, however the users that did not leave the area tended to extensively use the cellular network.

We also attempted to use the datasets we elaborated to predict potential drops or increases in cellular activity using a neural network. The accuracy of the model on the data led us to believe that we would be able to make predictions with an accuracy of about 80%. Unfortunately, when we tried to make predictions with the testing data, we were not able to make predictions accurately. To be able to conduct this type of operation successfully, we would need a better understanding of neural networks and a way of determining the appropriate models for this use case. Furthermore, we would need to determine if the data is simply not suited for this type of operation.

4.6.2. Interpretation

Having performed these different data mining and data analysis operations for this scenario, we are now in position to estimate how the cellular traffic data from Facebook and NOAA can possibly be used in the context of disaster management.

On one hand, we discovered that we could use data processing to analyze the impact on highly increased cellular activity relative to the distance (300 to 600 kilometers) of the hurricane. This could be used to prevent potential overloads or saturation of cellular networks by taking mitigation actions and setting up better network infrastructures in the event of an upcoming hurricane. Knowing where increased cellular activity can occur can also help give incentives to governmental organizations on the need for the population to get access to information about the potential dangers related to an upcoming hurricane. This information could help target specific areas where information is the most needed and set up warnings and advice through different communication channels.

On the other hand, we also discovered that when urbanized areas were very close (0 to a 100 kilometers) to the eye of a hurricane, cellular activity tended to drop compared to the baseline

measurements. As mentioned earlier, this phenomenon can be related to people leaving the area to escape from the wrath of the hurricane. Displacement of population caused by natural disasters is to be taken verry seriously by government authorities. Having incentives on why people have the urge to flee their current location and where they are heading could be of a great help to organize different aspects of population displacement, for example organizing shelter camps strategically or elaborating escape routes. Another aspect of reduced cellular activity can also be related to damaged network infrastructures, having information on where network outages occur can help authorities and organizations know where to intervene and deploy emergency network infrastructures such as the app developed by Flavien Bonvin mentioned in part 3 of this thesis.

Finally, even though we could not predict accurately the potential changes in network activity related to the data we had on hurricanes. It is important to underline the potential of such operations for the mitigation and preparedness stages of disaster management. If one could accurately predict the communication patterns around natural disasters, a lot of the actions stated previously could be done preemptively and in targeted manner thus greatly reducing the impact of natural disasters on society.

5. DISCUSSION

5.1. Results consideration

Although we were able to determine to some extent some tendencies with our implementation, we acknowledge the fact that the results obtained lack precision due to the nature of the data. Without contextual information, Facebook's social media data can be interpreted in many ways. Other examples using Twitter, like in part 4.1, as a data source have the advantage of having contextual information, where Tweets can be classified using words and hashtags to show the intention of the user.

5.2. Future improvements

As mentioned earlier, our neural network failed to make predictions accurately. It would be a big improvement to have a deeper understanding of how neural networks work and be able to elaborate models suited to our use case. After having better results with neural networks, the next logical step would be to simulate hurricanes and predict potential consequences on cellular connectivity.

Another improvement would be to generalize the process and elaborate a way to automate and display different information about Facebook data in correlation with upcoming hurricanes.

5.3. Work retrospective

A lot of aspects of this thesis were centered on research rather than implementation. Being new to the research world, we often deviated too far from the subject and ended up going into rabbit holes, it was difficult to eliminate different research directions and ended up losing a lot of time. Another thing that could have been better from the start, would have been to take a lot more notes, we tended to read a lot but not organize ideas and write down the different research aspects, which hindered the writing process of this thesis.

6. PROJECT MANAGEMENT

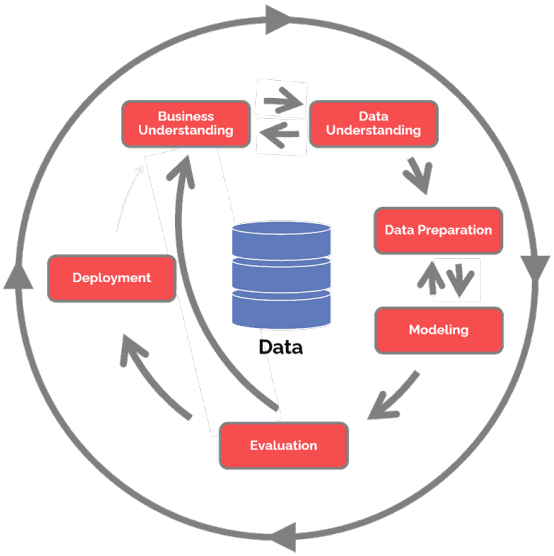
6.1. Methodology

For our project management methodology, we have decided to inspire ourselves from a process model called CRISP-DM. According to Wirth et al. (2000), this process can be summarized in these six steps (Wirth & Hipp, 2000):

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Figure 30 represents these steps as a cycle and project implementation described in Part 4 of this thesis indicates that we followed a similar structure.

Figure 30: CRISP-DM process



Source : https://www.datascience-pm.com/crisp-dm-2/

Although we followed the CRISP-DM process for our thesis, we also integrated notions of agility. We divided our work into periods of time called “sprints” with predefined global objectives, and at the end of each iteration we would make a small retrospective on the work accomplished and the things that needed to be done. It is to be noted that not all the meetings with our teacher in charge and the research team correspond with the end of each iteration, for scheduling and availability purposes.

Conventionally, if we had followed the CRISP-DM process, we would have a deployment phase, but given the nature of this thesis, there was no need for deployment. We presented our results and processes at the end of every iteration and discussed the potential enhancements and future improvements with the research team.

6.2. Work planification & management

As you can observe in the appendix IV, we have a big table that is divided into days and sprints. At the beginning of the thesis, we filled out this table with estimates of hours we could work on the thesis. Doing this was mandatory, because we had to organize our work efficiently around our current employment and scholar work.

As the days went by, we updated the table with the effective hours spent on the thesis and the different subjects on which we worked. It was a simplistic but very efficient way of organizing our work. Furthermore, given the current sanitary situation, this way of dispatching our work and self-discipline was essential.

6.3. Meetings and discussions

Table 6: Meetings

Date	Subject
26.02.2021	Introductory meeting
04.03.2021	Clarify the thesis objectives and initial questions about the problem.
11.03.21	Talk about the aim of the thesis
15.03.21	Meeting to discuss the Facebook datasets.
08.04.21	Finalize and confirm problematic
15.04.21	First demo of using hurricane data (NOAA) and Facebook data with Python/Pandas/GIS.
29.04.21	Demo and review
06.05.21	Demo and review
10.06.21	Demo and review
24.06.21	Structure of the thesis
29.04.21	Progress and review

Source: Author

7. OUTCOME

7.1. Knowledge acquired

Through the process of this research, we acquired a lot of knowledge on natural disasters, data science and machine learning. It was a tremendous opportunity to be able to associate computer science and real-world problems. Even if we would have liked to have gone further in the neural network implementation and be able to make predictions accurately. Furthermore, it was verry interesting to go through the process of a research project, something that we had never done before.

7.2. Difficulties encountered

Gathering and targeting the data was a real challenge, since the objective of the research was broad, we had a lot of different datasets to choose from. Furthermore, sometimes when choosing the desired data there were problems downloading the datasets, we had to communicate with the providers of the datasets to fix the said problem, see appendix III.

The writing of this thesis was also very challenging, it was complicated to organize all the research and the ideas I had and put them down as structured content. Furthermore, with the current sanitary conditions it was sometimes hard to self-organize the work, not to mention consequences on the morale that affected our performance.

CONCLUSION

The purpose of this thesis was to analyze how social media data combined with modern data mining techniques could support and improve the design of existing emergency communication networks. Given the access to Facebook's datasets, we seized the opportunity to perform different data manipulations and analysis on cellular traffic datasets surrounding the time period of a disaster.

Following a preliminary phase of research, we discovered the National Oceanic and Atmospheric Administration (NOAA)'s database and decided to measure the direct impact of Hurricane Sally on cellular activity in Florida by aggregating and analyzing both NOAA's hurricane datasets (IBTrAcs) and Facebook's traffic datasets.

With this analysis, we were able to determine a correlation between cellular activity and the distance of the hurricane. We observed that at close to mid distance, cellular activity would increase considerably, and therefore, describe the increased need of certain areas to access cellular networks. We also determined that on average, at very close range, Hurricane Sally had for effect a sharp decrease of the cellular network usage, which demonstrated potential network outages and population displacement given the proximity of the hurricane.

Although we obtained promising results, we would still need to reiterate the same experiment on the same type of dataset to firmly confirm our hypothesis. But given the nature and availability of the data, there is a time constraint that the elaboration of this thesis cannot possibly support.

Experimenting with neural networks to be able to make predictions on the dataset was one of the ideas we had to truly make a difference in the design of emergency communication networks. Unfortunately, this subject being vast and very complex we were unable to make predictions accurately.

As for potential developments of the research we conducted, we firmly support the idea of further development in the field of machine learning. Furthermore, by including other datasets from Facebook, like movement maps, we could also expand the possible applications of the research we conducted.

All things considered, during the course of this thesis and the multiple experiments it included, we have demonstrated that the usage of social media data plays an essential part in the design of emergency communication networks.

REFERENCES

EM-DAT. (n.d.). *General Classification*. Retrieved 05 03, 2021, from emdat: https://www.emdat.be/classification

Loko, N. L. (2012). Natural Disasters: Mitigating Impact, Managing Risks. *International Monetary Fund*, 5.

United Nations. (2021, 05). *Disaster management*. Retrieved from unddr: https://www.undrr.org/terminology/disaster-management

Kemp, S. (2021, 01 27). *DIGITAL 2021: GLOBAL OVERVIEW REPORT*. Retrieved 05 22, 2021, from DATAREPORTAL: https://datareportal.com/reports/digital-2021-global-overview-report

Tankovska, H. (2021, 05 21). *Number of monthly active Facebook users worldwide as of 1st quarter 2021*. Retrieved from Statista: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/

Chen, D., Liu, Z., Wang, L., Dou, M., Chen, J., & Li, H. (2013). Natural Disaster Monitoring with Wireless SensorNetworks: A Case Study of Data-intensive Applicationsupon Low-Cost Scalable Systems. *Springer Science*, 1-4.

GeoThings. (n.d.). *Helping to strengthen disaster resilience*. Retrieved from geobingan.info: https://geobingan.info/landing

Twitter API. (2021). Retrieved from Developer platform: https://developer.twitter.com/en/docs/twitter-api

(2021). Retrieved from Facebook Data for Good: https://dataforgood.fb.com/

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Hentenryck, P. V., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage usingsocial media activity. *Science Advances*, 11.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. 1-5.

Bonvin, F. (2018). *Analysis, design, and implementation of an infrastructure-less situation awareness application.*

Facebook. (n.d.). *Can I have smaller tiles for more granular versions of the data?* Retrieved from Facebook geoinsights help center: https://www.facebook.com/help/geoinsights/270968402914260l?helpref=typeahead_suggestions&sr=1&query=tiles

Python Software Fundation. (n.d.). *General Python FAQ.* Retrieved from Python documentation: https://docs.python.org/

Python Software Fundation. (n.d.). *Home*. Retrieved from Pypi: https://pypi.org/

R fundation. (n.d.). *What is R ?* Retrieved from r-project: https://www.r-project.org/

DataCamp Team. (n.d.). *Choosing Python or R for Data Analysis? An Infographic*. Retrieved from datacamp: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis?utm_source=adwords_ppc&utm_campaignid=1655852085&utm_adgroupid=61045433422&utm_device=c&utm_keyword=%2Bpython%20%2Br&utm_matchtype=b&utm_network=g&utm_adpostion=&utm_creative=31888

Huang, Q., & Yu, X. (2015). Geographic Situational Awareness: Mining Tweets for Disaster Preparedness, Emergency Response, Impact, and Recovery. *ISPRS International Journal of Geo-Information*, 1-12.

Smith, N. (2020). *Where People Went Before, During, and After Hurricane Laura, Amid Pandemic*. Retrieved from Direct Relief: https://www.directrelief.org/2020/08/responding-to-a-hurricane-in-a-pandemic/

Yadav, A. (2021). *Network Outages and Changes in Population Density After Tropical Storm Claudette in Louisiana and Alabama*. Retrieved from CrisisReady: https://www.crisisready.io/2021/claudette-in-louisiana-and-alabama/

Roy, A. S. (2020). *How does facebook handle the 4+ petabyte of data generated per day? Cambridge Analytica - facebook data scandal*. Retrieved from Medium: https://medium.com/@srank2000/how-facebook-handles-the-4-petabyte-of-data-generated-per-day-ab86877956f4

Facebook. (n.d.). *Disaster Map Dataset Types*. Retrieved from Facebook Data for good: https://www.facebook.com/help/geoinsights/254556701793421/?helpref=hc_fnav&bc[0]=SPACO%20Help%20Center&bc[1]=Disaster%20Maps

NOAA. (n.d.). *International Best Track Archive for Climate Stewardship (IBTrACS)*. Retrieved 2021, from National centers for enviornemental information: https://www.ncdc.noaa.gov/ibtracs/index.php?name=ib-v4-access

Brownlee, J. (n.d.). *Your First Deep Learning Project in Python with Keras Step-By-Step*. Retrieved from Machine Learning magistery: https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

Berg, R., & Reinhart, B. J. (2021, 04 14). *NATIONAL HURRICANE CENTER TROPICAL CYCLONE REPORT*. Retrieved from NHC NOAA: https://www.nhc.noaa.gov/data/tcr/AL192020_Sally.pdf

IFRC. (n.d.). *Types of disasters: Definition of hazard*. Retrieved 2021, from IFRC: https://www.ifrc.org/en/what-we-do/disaster-management/about-disasters/definition-of-hazard/

Blaikie, P., Cannon, T., Davis, I., & Wisner, B. (2004). *At Risk: Natural Hazards, People's Vulnerability and Disasters*. Routledge.

Albtoush, R., Dobrescu, R., & Ionescou, I. (2011). A HIERARCHICAL MODEL FOR EMERGENCY MANAGEMENT SYSTEMS. *UPB*, 55.

Giorgadze, T., Maisuradze, I., Japaridze, A., Utiashvili, Z., & Abesadze, G. (2011, 05). Disasters and their consequences for public health. *Georgian Med News*.

Roser, H., & Ritchie, M. (2019, 11). *Natural Disasters*. Retrieved 05 2021, from Our world in data: https://ourworldindata.org/natural-disasters#natural-disasters-kill-on-average-60-000-people-per-year-and-are-responsible-for-0-1-of-global-deaths

Sharrieff, M. (2018, 07 19). *The Impact of Natural Disasters*. Retrieved 05 05, 2021, from sciencing: https://sciencing.com/impact-natural-disasters-5502440.html

Kaplan, S. (2020, 10 22). *The undeniable link between weather disasters and climate change*. Retrieved from The Washington Post: https://www.washingtonpost.com/climate-solutions/2020/10/22/climate-curious-disasters-climate-change/

Arifah, A. R., Mohd Tariq, M. N., & Juni, M. H. (2019). DECISION MAKING IN DISASTER MANAGEMENT CYCLE OF NATURAL DISASTERS: A REVIEW. *International Journal of Public Health and Clinical Sciences*, 4.

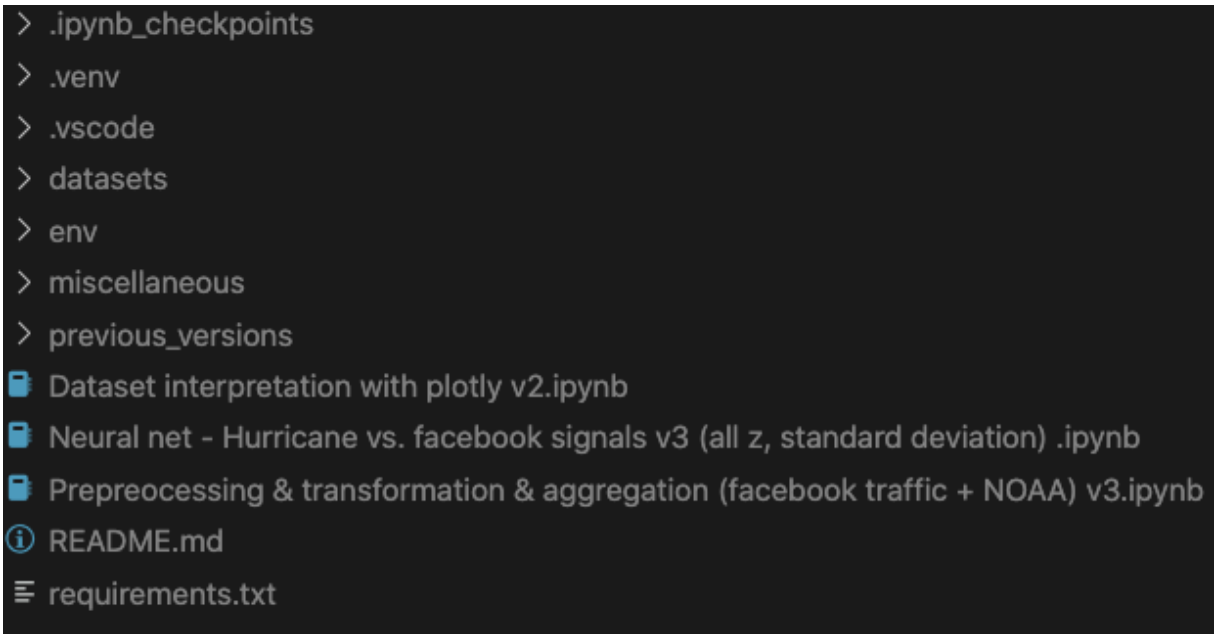
Phengsuwan, J., Tejal, S., Thekkummal, N. B., Wen, Z., Sun, R., Pullarkatt, D., . . . Ranjan, R. (2021). Use of Social Media Data in Disaster Management: A Survey. *Future internet*, 19-20.

NOAA. (2021, 05 25). *About our agency*. Retrieved from National Oceanic and Atmospheric Administration: https://www.noaa.gov/about-our-agency

Acronis. (2021). *The Golden 3-2-1 Backup Rule*. Retrieved 03 2021, from Acronis: https://www.acronis.com/en-eu/articles/backup-rule/

APPENDIX I : Project file structure

The project file structure is the following:



The following directories can be ignored: .ipynb_checkpoints, .venv, .vscode, env. They are directories generated automatically by the tools required for this project.

1. In the datasets directory, you will find all the datasets used in this project.
2. In the miscellaneous directory, you will find test and experiments that were made during the project.
3. In the previous_versions directory, you will find all the previous versions of the project files.
4. At the root of the directory, you will find the Jupyter notebooks that were used for the final form of the project.
5. The Readme and requirements files will be useful for setting up the environment.

APPENDIX II : Environment setup and startup

To run the project locally, you will need to have Python 3 installed and the virutalenv package installed globally in your machine.

1. Create the virtual environment with the following command:

```
python3 -m venv env
```

2. Activate the virtual environment with one of the following commands in the directory of the project:

On Unix or MacOS, using the bash shell:

```
source env/bin/activate
```

On Unix or MacOS, using the csh shell:

```
source env/bin/activate.csh
```

On Unix or MacOS, using the fish shell:

```
source env/bin/activate.fish
```

On Windows using the Command Prompt:

```
env\Scripts\activate.bat
```

On Windows using PowerShell:

```
env\Scripts\Activate.ps1
```

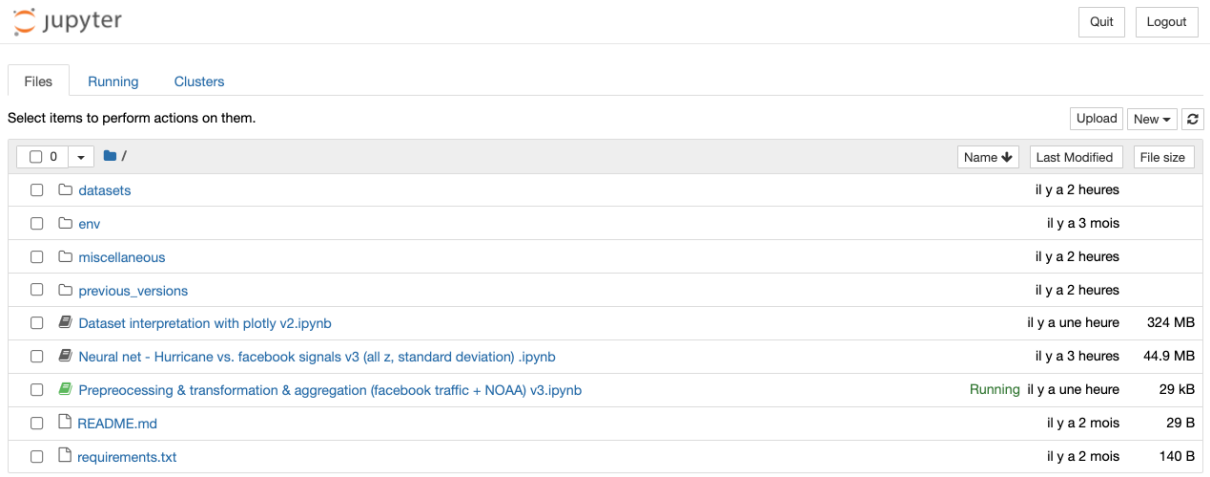
3. Install the Python libraries in your environment with the following command:

```
pip3 install -r requirements.txt
```

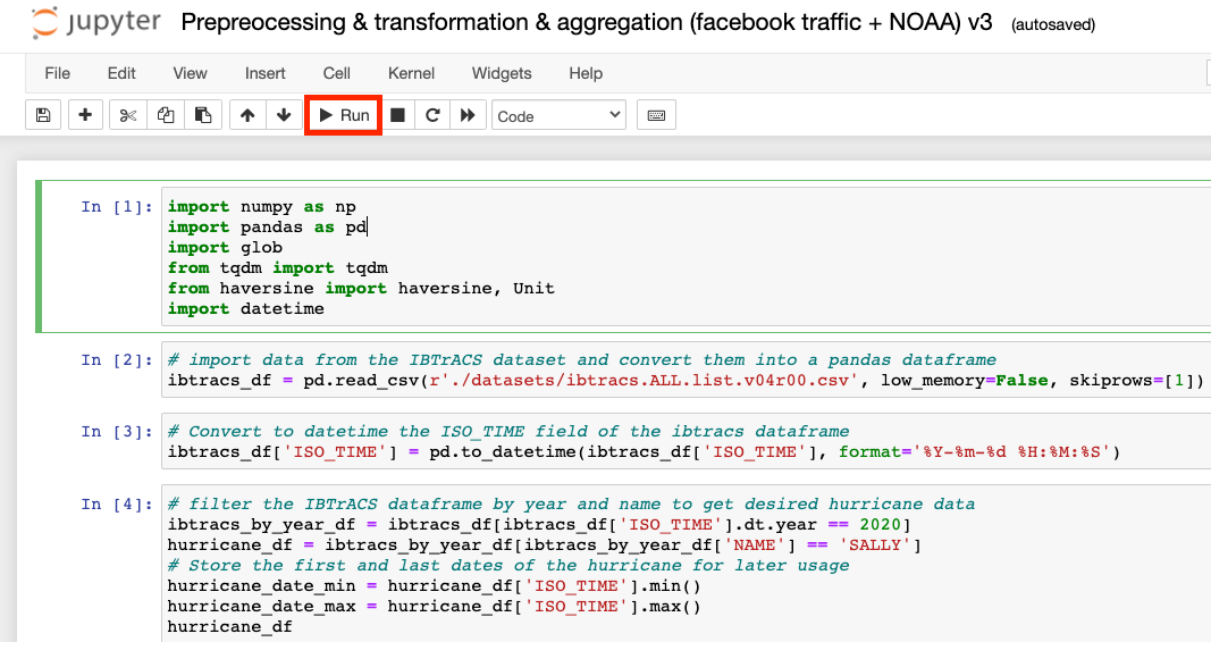
4. You can now launch the Jupyter notebook with the following command:

```
jupyter notebook
```

If everything was properly installed, a browser window should open, and you will be able to see the following:

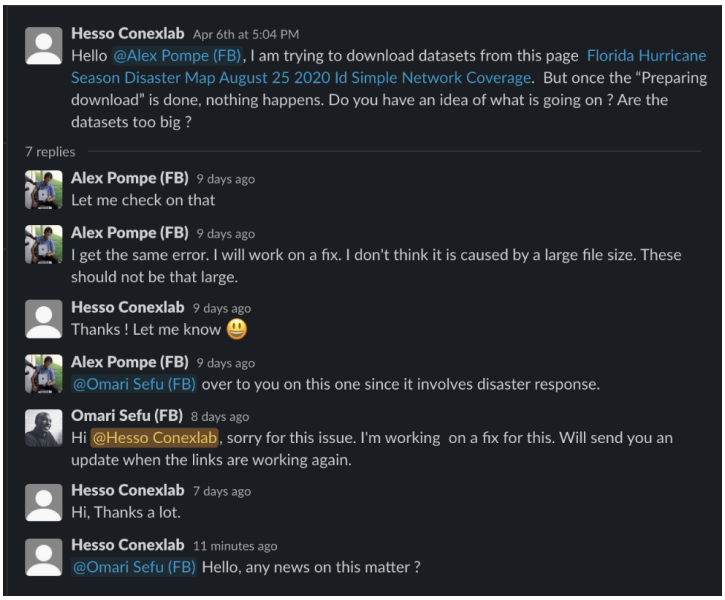


To run a notebook, click on a file with the extension .ipynb, then you will be able to run the cells by clicking on the “Run” button (highlighted in red in the following image):

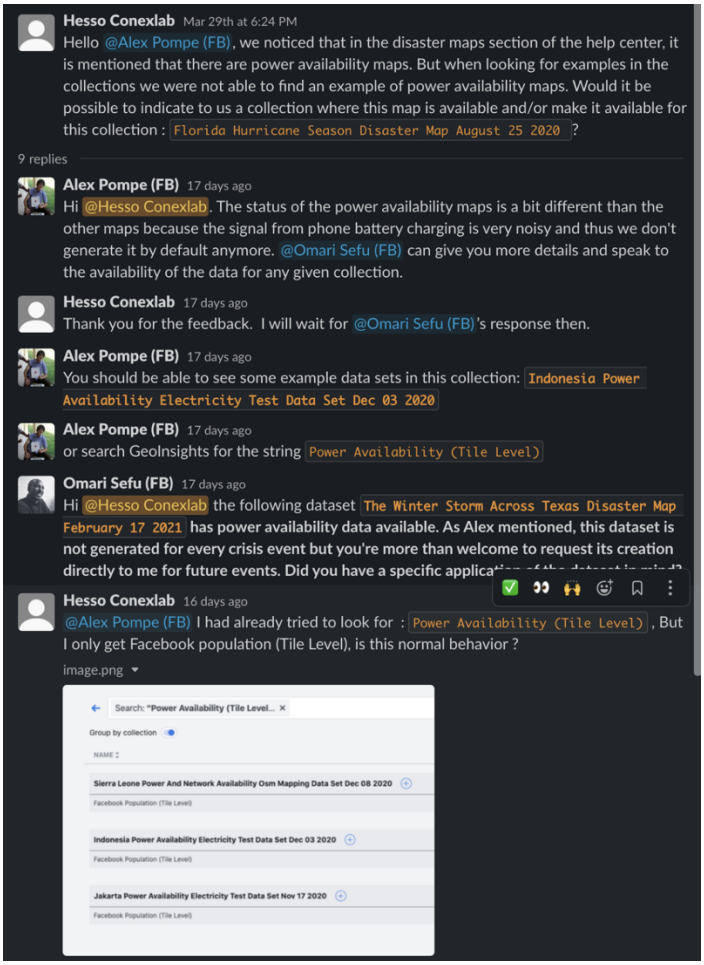


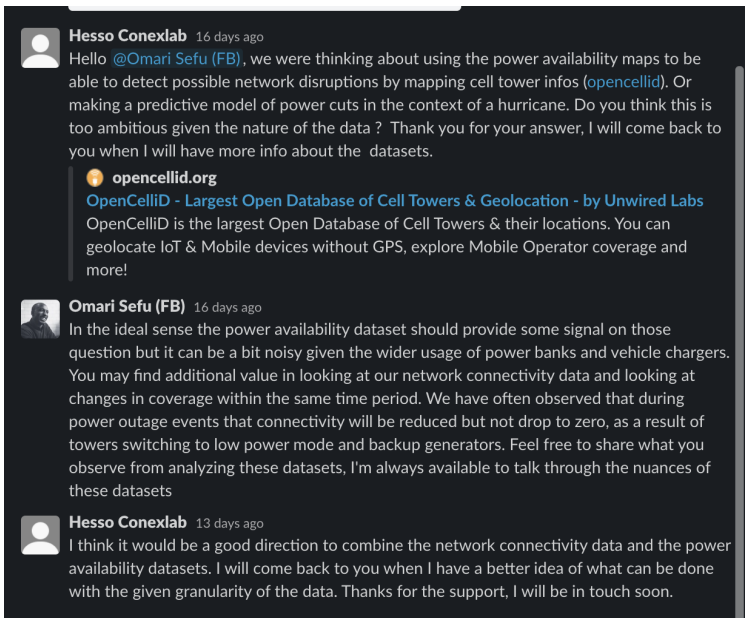
APPENDIX III : Conversation with the Facebook team

Discussion about the download issues:



Discussions about power availability maps:





Hesso Conexlab 16 days ago
Hello @Omari Sefu (FB), we were thinking about using the power availability maps to be able to detect possible network disruptions by mapping cell tower infos (opencellid). Or making a predictive model of power cuts in the context of a hurricane. Do you think this is too ambitious given the nature of the data ? Thank you for your answer, I will come back to you when I will have more info about the datasets.

> **opencellid.org**
> **OpenCelliD - Largest Open Database of Cell Towers & Geolocation - by Unwired Labs**
> OpenCelliD is the largest Open Database of Cell Towers & their locations. You can geolocate IoT & Mobile devices without GPS, explore Mobile Operator coverage and more!

Omari Sefu (FB) 16 days ago
In the ideal sense the power availability dataset should provide some signal on those question but it can be a bit noisy given the wider usage of power banks and vehicle chargers. You may find additional value in looking at our network connectivity data and looking at changes in coverage within the same time period. We have often observed that during power outage events that connectivity will be reduced but not drop to zero, as a result of towers switching to low power mode and backup generators. Feel free to share what you observe from analyzing these datasets, I'm always available to talk through the nuances of these datasets

Hesso Conexlab 13 days ago
I think it would be a good direction to combine the network connectivity data and the power availability datasets. I will come back to you when I have a better idea of what can be done with the given granularity of the data. Thanks for the support, I will be in touch soon.

APPENDIX IV : Project management table

Week 8		
Date	**Task / Event**	**Time (h)**
22.02.2021		0
23.02.2021		0
24.02.2021		0
25.02.2021	Planning	3
26.02.2021	RDV 1	1
27.02.2021	Planning	1
28.02.2021	Planning	1
TOTAL WEEK	6	

Week 9		
Date	**Task / Event**	**Time (h)**
28.02.2021		0
01.03.2021		0
02.03.2021		0
03.03.2021		0
04.03.2021	RDV (2 ème)	4
05.03.2021		0
06.03.2021	Subject understanding	5
TOTAL WEEK	9	

Week 10		
Date	**Task / Event**	**Time (h)**
07.03.2021		0
08.03.2021		0
09.03.2021		0
10.03.2021		0
11.03.2021	RDV 3 + Subject understanding	4
12.03.2021		0
13.03.2021	Subject understanding	5
TOTAL WEEK	9	

Week 11		
Date	**Task / Event**	**Time (h)**
14.03.2021		0
15.03.2021	RDV 4	1
16.03.2021		0
17.03.2021		0
18.03.2021	Natural disaster research	3
19.03.2021	Natural disaster research	8
20.03.2021	Natural disaster research	5
TOTAL WEEK	17	

Sprint	Objective	
Sprint 0	Planning, Understand the subject and start research	
	TOTAL SPRINT 0	41
	SUB TOTAL TB	41

Week 12		
Date	**Task / Event**	**Time (h)**
21.03.2021		0
22.03.2021		0
23.03.2021		0
24.03.2021		0
25.03.2021	Natural disaster research	4
26.03.2021		0
27.03.2021	Natural disaster research	5
TOTAL WEEK	9	

Week 13		
Date	**Task / Event**	**Time (h)**
28.03.2021		0
29.03.2021		0
30.03.2021		0
31.03.2021		0
01.04.2021	Thesis preparation (template)	4
02.04.2021	Thesis preparation (template)	8
03.04.2021	Data source research	5
TOTAL WEEK	17	

Week 14		
Date	**Task / Event**	**Time (h)**
04.04.2021		0
05.04.2021		0
06.04.2021		0
07.04.2021		0
08.04.2021	RDV 5 + Data source research	4
09.04.2021	Data source research	8
10.04.2021	Data source research	5
TOTAL WEEK	17	

Sprint 1

Objective	
Research and template preparation	
TOTAL SPRINT 1	43
SUB TOTAL TB	84

Week 15		
Date	**Task / Event**	**Time (h)**
11.04.2021		0
12.04.2021		0
13.04.2021		0
14.04.2021		0
15.04.2021	RDV 6 + project implementation	4
16.04.2021		0
17.04.2021	Project implementation (testing)	5
TOTAL WEEK	9	

Week 16		
Date	**Task / Event**	**Time (h)**
18.04.2021		0
19.04.2021		0
20.04.2021		0
21.04.2021		0
22.04.2021	Project implementation (testing)	4
23.04.2021		0
24.04.2021	Data source research	5
TOTAL WEEK	9	

Week 17		
Date	**Task / Event**	**Time (h)**
25.04.2021		0
26.04.2021		0
27.04.2021		0
28.04.2021		0
29.04.2021	RDV 7 + project implementaion	4
30.04.2021		0
01.05.2021	Project implementation (testing)	5
TOTAL WEEK	9	

Sprint 2

Objective	
Start project implementation	
TOTAL SPRINT 2	27
SUB TOTAL TB	111

Week 18		
Date	**Task / Event**	**Time (h)**
02.05.2021		0
03.05.2021		0
04.05.2021		0
05.05.2021		0
06.05.2021	RDV 8 + Data source research	4
07.05.2021		0
08.05.2021	Natural disaster research	5
TOTAL WEEK	9	

Week 19		
Date	**Task / Event**	**Time (h)**
09.05.2021		0
10.05.2021		0
11.05.2021		0
12.05.2021		0
13.05.2021	Existing projects research	8
14.05.2021	Project implementation (testing)	8
15.05.2021	Existing projects research	5
TOTAL WEEK	21	

Week 20		
Date	**Task / Event**	**Time (h)**
16.05.2021		0
17.05.2021		0
18.05.2021		0
19.05.2021		0
20.05.2021	RDV 9 + Project management	4
21.05.2021		0
22.05.2021	Project implementation (testing)	5
TOTAL WEEK	9	

Sprint 3

Objective

State of the art research and implementation

TOTAL SPRINT 3	39
SUB TOTAL TB	150

Week 21		
Date	**Task / Event**	**Time (h)**
23.05.2021		0
24.05.2021	implementation (data preprocessing)	8
25.05.2021		0
26.05.2021		0
27.05.2021	implementation (data preprocessing)	4
28.05.2021		0
29.05.2021	implementation (data preprocessing)	5
TOTAL WEEK	17	

Week 22		
Date	**Task / Event**	**Time (h)**
30.05.2021		0
31.05.2021		0
01.06.2021		0
02.06.2021		0
03.06.2021	Thesis redaction	4
04.06.2021		0
05.06.2021	implementation (data preprocessing)	5
TOTAL WEEK	9	

Week 23		
Date	**Task / Event**	**Time (h)**
06.06.2021		0
07.06.2021		0
08.06.2021		0
09.06.2021		0
10.06.2021	RDV 10 + Data source research	4
11.06.2021		0
12.06.2021	implementation (data data modeling)	5
TOTAL WEEK	9	

Sprint 4

Objective

Data preprocessing and start redacting the thesis

TOTAL SPRINT 4	35
SUB TOTAL TB	185

Week 24		
Date	**Task / Event**	**Time (h)**
13.06.2021		0
14.06.2021		0
15.06.2021		0
16.06.2021		0
17.06.2021	implementation (data data modeling)	4
18.06.2021		0
19.06.2021	implementation (data data modeling)	5
TOTAL WEEK	9	

Week 25		
Date	**Task / Event**	**Time (h)**
20.06.2021		0
21.06.2021		0
22.06.2021		0
23.06.2021		0
24.06.2021	RDV 11 + implementation	4
25.06.2021		0
26.06.2021	implementation (data data modeling)	5
TOTAL WEEK	9	

Week 26		
Date	**Task / Event**	**Time (h)**
27.06.2021		0
28.06.2021		0
29.06.2021		0
30.06.2021		0
01.07.2021	implementation (data interpretation)	4
02.07.2021		0
03.07.2021	Sprint 5	5
TOTAL WEEK	9	

Sprint 5

Objective	
Data modeling	
TOTAL SPRINT 5	27
SUB TOTAL TB	212

Week 27		
Date	**Task / Event**	**Time (h)**
04.07.2021		0
05.07.2021		0
06.07.2021		0
07.07.2021		0
08.07.2021	implementation (data interpretation)	4
09.07.2021	implementation (data interpretation)	8
10.07.2021	implementation (data interpretation)	5
TOTAL WEEK	17	

Week 28		
Date	**Task / Event**	**Time (h)**
11.07.2021		0
12.07.2021		0
13.07.2021		0
14.07.2021		0
15.07.2021	Thesis redaction	4
16.07.2021	Thesis redaction	8
17.07.2021	Thesis redaction	5
TOTAL WEEK	17	

Sprint 6

Objective	
Data interpretation	
TOTAL SPRINT 6	34
SUB TOTAL TB	246

Sprint 7

Week 29		
Date	**Task / Event**	**Time (h)**
18.07.2021		0
19.07.2021		0
20.07.2021		0
21.07.2021		0
22.07.2021	Thesis redaction	4
23.07.2021	Thesis redaction	8
24.07.2021	Thesis redaction	5
TOTAL WEEK	17	

Week 30		
Date	**Task / Event**	**Time (h)**
25.07.2021		0
26.07.2021		0
27.07.2021		0
28.07.2021		0
29.07.2021	RDV 12 + Thesis redaction	4
30.07.2021	Thesis redaction	8
31.07.2021	Thesis redaction	5
TOTAL WEEK	17	

Objective	
Redaction	
TOTAL SPRINT 7	34
SUB TOTAL TB	280

Sprint 8

Week 31		
Date	**Task / Event**	**Time (h)**
01.08.2021		0
02.08.2021	Thesis finalisation	8
03.08.2021	Thesis finalisation	8
04.08.2021	Thesis finalisation	8
05.08.2021	Thesis finalisation	8
06.08.2021	Thesis finalisation	8
07.08.2021	Thesis finalisation	5
TOTAL WEEK	45	

Week 32		
Date	**Task / Event**	**Time (h)**
08.08.2021		0
09.08.2021	Thesis finalisation	8
10.08.2021	Thesis finalisation	8
11.08.2021	Thesis finalisation	8
12.08.2021	Thesis finalisation	8
13.08.2021		0
TOTAL WEEK	32	

Objective	
Redaction	
TOTAL TB	357

AUTHOR'S STATEMENT

I hereby declare that I have carried out the attached bachelor's thesis alone, without any other assistance than that duly indicated in the references, and that I have used only the sources expressly mentioned. I will not give any copy of this report to a third party without the joint authorization of the RF and the professor in charge of the follow-up of the bachelor's work, including the applied research partner with whom I collaborated, apart from the persons who provided me with the main information necessary for the writing of this work and whom I quote below: Yann Bocchi, Gianluca Rizzo.

Sierre, on the 13th of August 2021

_____

Dylan Thompson