SD201 PROJECT REPORT

Mathilde Froger, Apolline Isaia, Laurybe Moyse, Solal Urien

TABLE OF CONTENTS

INTRODUCTION

Our project will aim at understanding the type, evolution, and tendencies of the disasters in the United States of America by analysing a data set constructed by the Federal Emergency Management Agency (FEMA). Furthermore, we will ask ourselves the following question:

- ✓ Are some locations in the USA particularly dangerous, in the sense that they are especially affected by disasters in regard to the rest of the country?
- ✓ What can we do to improve the management and allotment of the relief forces?
- ✓ How can we interpretate the location of the disasters and their evolution in space?
- ✓ Can we observe the effects of climate change on the number, size, and duration of natural disasters in the US?
- ✓ How is Covid-19 considered in this dataset, as it is an unprecedented disaster?

In the entirety of this report, we will have to make the following **hypothesis**: *the number of declarations received by the FEMA since 1953 is uncorrelated to human biases*, i.e., it is a good approximation of the real number of disasters that happened since 1953. This hypothesis – although a strong assumption – will allow us to consider the dataset a reliable source, on which the influence of technological and social innovations can overlooked. Grounding this hypothesis is the sheer nature of what is considered a "disaster": it is a phenomenon that "*overwhelms the resources of local and state authorities*". A state of emergency must be declared by the affected state, which leads us to believe that a very small portion of the data might have been skewed by the differences between administrative policies.

Context

The sheer size and geographic diversity of the United States means that the country experiences a variety of different natural disasters on a frequent basis. The number of total disasters since 1953 is therefore particularly high, and all the more so as the FEMA dataset is not limited to natural disasters. Indeed, it also includes among other things biological disasters, dam breaks or terrorist attacks.

Another aspect of our analysis will be the context of global warming. The climate change induces an increasing of global surface temperatures. This fact leads to the possibility of more droughts and increased intensity of storms. Furthermore, as more water vapor evaporates into the atmosphere, and as the ocean surface temperatures rise, tropical storms become more powerful due to their increased wind speeds. Moreover, the rising sea level exposes locations not usually subjected to the power of the sea and to the erosive forces of waves and currents, leading to a likely increase of floods.

Therefore, the number of natural disasters and their size is undoubtedly increasing (and has almost certainly already risen since 1953). We might most likely observe this fact during our analysis of the FEMA's dataset.

Description of the dataset

The FEMA is the American governmental organization in charge of helping people before, during, and after disasters by coordinating disaster response and providing relief funds. This dataset is a high-level summary of all federally declared disasters since 1953. The author downloaded it from the FEMA website and applied a few simple data cleanings.

The dataset starts in 1953 and is updated regularly (the last data are from October 1st, 2022). Therefore, it includes the latest data on Covid-19, which is extremely important to notice, as we will see later.

The dataset in itself is an .csv document composed of 23 columns and 63.699 rows, i.e. about 1.5 million entries.

You will find underneath a brief description of the most useful columns (the entire description can be found with the bibliography):

- **fema_declaration_string**: Agency standard method for uniquely identifying Stafford Act declarations.
- **disaster_number**: Sequentially assigned number used to designate an event or incident declared as a disaster.
- **state**: US state, district, or territory.
- **declaration_type**: Either DR (major disaster), EM (emergency management), or FM (fire management)
- **declaration_date**: Date the disaster was declared.
- **fy_declared**: Fiscal year in which the disaster was declared.
- **incident_type**: Type of incident (e.g “Fire”,”Flood”…). The incident type will affect the types of assistance available.
- **declaration_title**: Title for the disaster. Can be a useful identifier such as "Hurricane Katrina" or "Covid-19 Pandemic".
- **incident_begin_date**: Date the incident itself began.
- **incident_end_date**: Date the incident itself ended. This feature has about 14% NA entries.
- **disaster_closeout_date**: Date all financial transactions for all programs are completed (98% NA entries)

Description of the data cleaning

We had to **clean the dataset**, as it included some issues for the analysis. As an instance, some of the disasters were reported at more than one place at a time, leading to more than one entry for a sole disaster, a fact that can be problematic for some of the analysis (on evolution and types of disaster, for example) but that is welcome for other analysis (e.g., the efficiency of interventions per period). Therefore, after cleaning the database (between other things deleting the duplicates data), we observe that the total number of accidents is *4709* (meaning on average *68 disasters per year*) and not *63.699*. This induces that a lot of accident were reported more than once, on average *13.5 per disaster*, however, as the USA have more than 50 states, a nationwide disaster such as a hurricane might be reported more than 50 times, while a disaster that took place in a single town will only be reported once, hence a certainly high standard deviation.
We then build the **correlation matrix** to analyse the relationship between each column and how they depend on each other and to diagnose unwanted relation. We detect only two correlations here: the one between the columns *fy_declared* and *year*, which is perfectly coherent, and the one between *disaster_number* and *year*, which is understandable as the disasters are numbered starting from year 1953.

	disaster_number	fy_declared	ih_program_declared	ia_program_declared	pa_program_declared	hm_program_declared	fips	place_code	declaration_request_number	year
disaster_number	1.00	0.81	0.04	-0.66	0.15	0.29	-0.01	0.20	-0.54	0.81
fy_declared	0.81	1.00	0.16	-0.66	0.06	0.34	0.01	0.39	-0.64	1.00
ih_program_declared	0.04	0.16	1.00	0.12	-0.29	0.29	0.00	0.09	-0.21	0.16
ia_program_declared	-0.66	-0.66	0.12	1.00	-0.37	-0.13	-0.01	-0.06	0.47	-0.66
pa_program_declared	0.15	0.06	-0.29	-0.37	1.00	-0.00	0.02	-0.10	-0.10	0.06
hm_program_declared	0.29	0.34	0.29	-0.13	-0.00	1.00	-0.01	-0.04	-0.45	0.33
fips	-0.01	0.01	0.00	-0.01	0.02	-0.01	1.00	-0.03	0.06	0.01
place_code	0.20	0.39	0.09	-0.06	-0.10	-0.04	-0.03	1.00	-0.31	0.39
declaration_request_number	-0.54	-0.64	-0.21	0.47	-0.10	-0.45	0.06	-0.31	1.00	-0.65
year	0.81	1.00	0.16	-0.66	0.06	0.33	0.01	0.39	-0.65	1.00

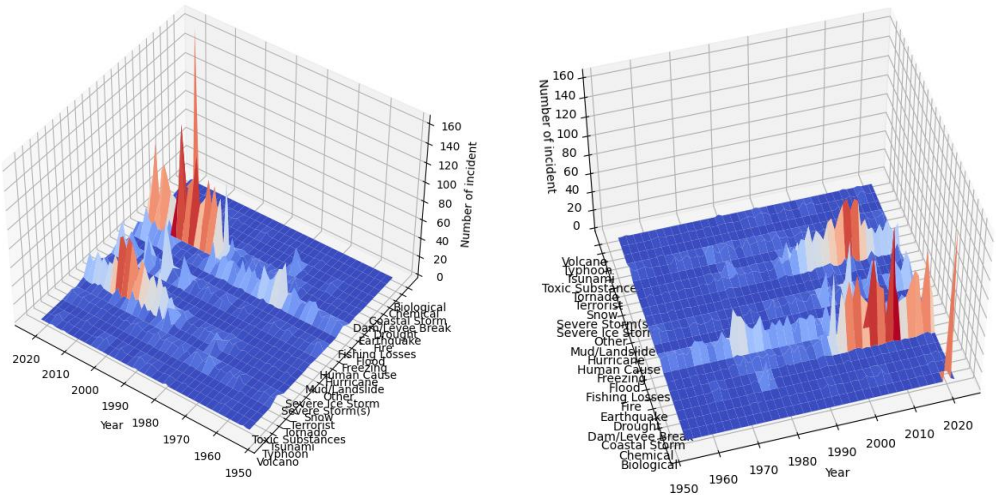
There is no further suspect correlation between our dataset’s columns, which is why no **dimensional reduction** will be conducted.
Finally, each part will have its own data cleaning, to use precisely the data we want to work with.

Work distribution

As of work distribution, Mathilde did the 2nd half of part II (code *Part II-2)*, Lauryne did part III (code *Part III*), Apolline did the 1st half of part 2 (code *Part II-1*), and Solal did the 1st part (code *Part I*).
The work distribution was therefore overall balanced.

I –TYPE AND LOCATION OF THE DISASTERS

An overall view



You will find above a **3D graphic** of the **number of disasters depending on the year and on their types**. We can clearly see an evolution in the global number of disasters this last decades: these evolutions will be studied in detail in the second part. The peak that can be spotted in biological disasters in 2020 (bottom-right corner on the right picture) is the Covid-19 pandemic. One can easily see that its heigh (here related to the number of disasters reported) clearly eclipses the others in numbers. It is therefore crucial to take the **imbalance** caused by this phenomenon into account so that it does not distort the analysis.

We can remark another curious fact, namely the differentiation between “human cause” disasters and “terrorist”. We will see below that since these two kinds of disasters are quite uncommon (2 different entries in total), we will study it. Still, it is a curious fact to point out.

In this first part, we will be ignoring the dependencies of the disasters in time, i.e., we will be taking the number overall from 1953 to 2022. Indeed, this will allow us to highlight the general trends of disasters in the USA.

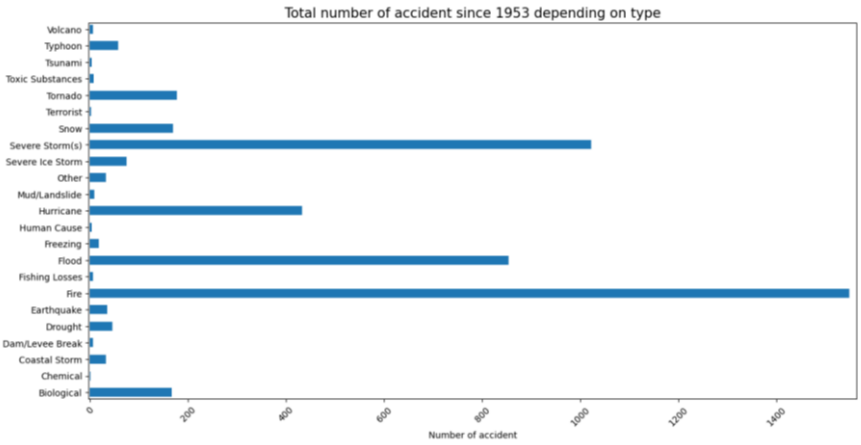
We will start by analysing the types of the disasters in the dataset and propose a concrete categorization of disasters. We will then focus on the location of these disasters, and we will develop a clustering model to identify the high-risk regions of the USA. Finally, we will emphasise the specificities of the Covid-19 crisis in this dataset, and more generally as a biological disaster.

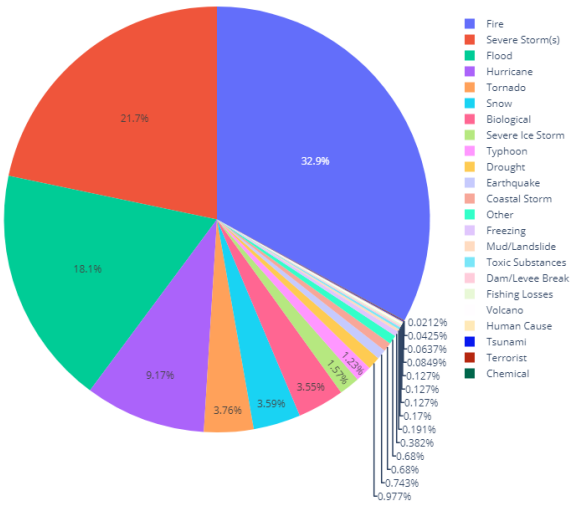
Thereafter, we will consider the evolution in time of the disaster, and we will try to construct a model for future natural disasters, by using our general knowledge on the climatic crisis. Finally, we will shortly put the emphasis on the efficiency of the interventions of the government depending on several criteria.

Type of disasters

We will firstly use different types of graphics to scrutinize the repartition of the disasters and try to establish a classification.

Thanks to this **bar chart**, we can observe the total number of disasters that happened since 1953 (69-year span) in the USA. We can therefore witness that some are extremely common, as the fires, hurricane, and floods, while several are particularly rare, such as the tsunamis (3 in total), chemical disasters (1 in total) or terrorist attacks (2 in total).





The **pie chart** on the left clearly shows the predominance of four main phenomenon, namely fires, severe storms, floods and hurricanes, which add up to 81,87% (3855 disaster out of 4709) of the total quantity of disasters (it could be discussed if hurricanes are to be taken into account in this category, as it constitutes less than 10% (471 disasters), whereas the other account for more than 18% (848 disasters)).

Moreover, we can identify a second group of disasters, in which each type constitutes between 3.5% and 4% of the global number of accidents. This type of disasters are the tornadoes, snows, and biological disasters.

However, as seen before, putting the biological disasters in this category might not be wise, as all biological disasters are linked to the Covid-19 pandemic, and thus take place in a 3 year-span starting in 2020.The third category is composed of all the disasters that account for less than 2% and more than 0.5% and are quite uncommon as an event of this type happens on average less than twice in a year. Finally, the four category regroups the phenomenon that have no more than 0.4%, i.e., that happen on average less than once every three years. Those four categories are summarized in the afterward table.

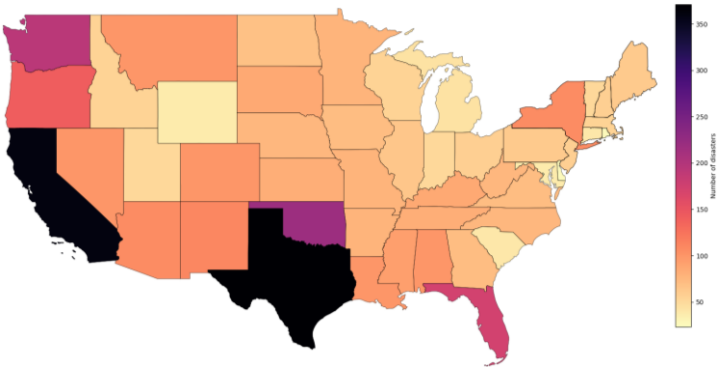
Category of disasters	Percentage interval	Corresponding number of observation range	Which disasters
Type I (very common, on average more than 6/yr)	**9%<x**	**424<y**	*fires, severe storms, floods, hurricanes*
Type II (common, on average at least 2/yr)	**2%<x<9%**	**141<y<424**	*tornadoes, snows, biological disasters*
Type III (uncommon, on average less than 2/yr)	**0.4%<x<2%**	**19<y<141**	*Severe ice storms, typhoons, droughts, earthquakes, coastal storms*
Type IV (very uncommon, on average less than 0.3/yr)	**x<0.4%**	**y<19**	*All other accidents*

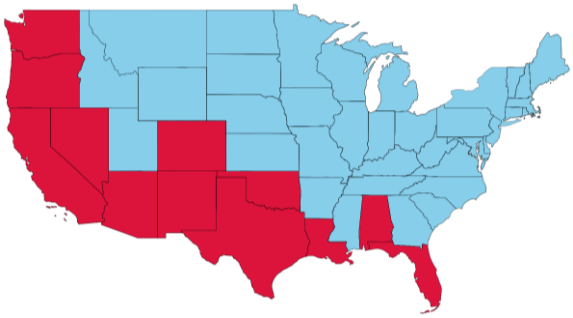
The above categories are useful to understand and improve the management of the relief forces and the efficiency of the interventions of the government. Indeed, the relief forces are often well-prepared to face common disasters and know how to act as they have already experienced of such accidents. In addition, the allotment of the government forces and funds depends on the probability of an event to happens. Therefore, we might counsel to allows more funds to firefighter forces than to services that aim at fighting type IV disasters.

Finally, this categorization can be linked to the evolution of the climate: as we said, the global warming will make some uncommon disasters usual, therefore leading to changes in those categories. This will be discussed later in our analysis.

Location of the disasters

Now that we have a general idea of the distribution of the accidents, and a coherent classification of the disasters depending on their occurrences, we want to understand *how the phenomena are distributed over the USA*. Once again, we will be **ignoring the dependency in time to have an overall view**. We use the *geopandas* library.

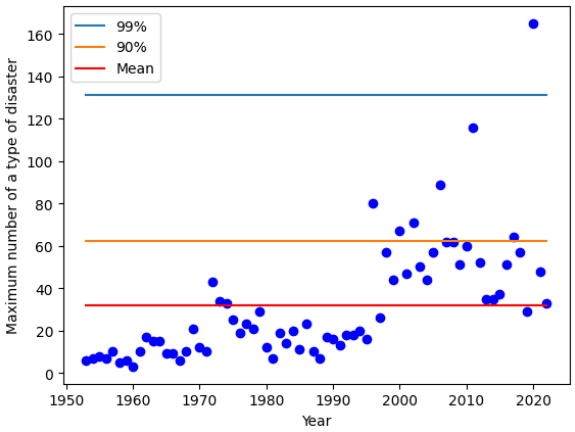




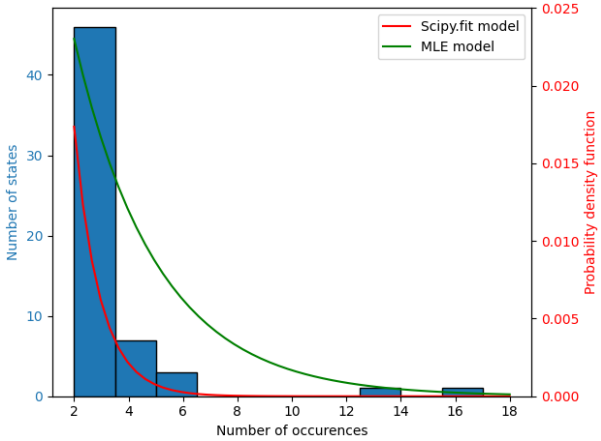
We first created a map showing the number of disasters depending on the state, but it was *not efficient to observe general tendencies* of the distribution of disasters over the whole USA. We therefore built a **gradient of colours** which depends on the total number of accidents. We then used this gradient to create a **colormap**. We subsequently built our **own algorithm** for **geographic clustering**, using the same method of categorization than the previous table, but with only 2 categories.

We **balanced** those categories, using the *75% percentile*, and plotted the map above. We can observe in red *the part* of the USA *most affected by disasters* with more than *96 disasters* declared since 1953.
Finally, with those maps, we can efficiently detect the tendency of the **western and southern states** to be **more impacted by disasters** than the rest of the country, and therefore where to emphasize the means available to the rescue forces.

Analysis of the Covid-19 pandemic



We first tried to prove that the biological disasters are **imbalanced** regarding the data, by computing the *annual number of the disaster that has the most occurrences at a given year,* for each year since 1953. The graph allows us to clearly observe the peak of *165 disasters* provoked by the pandemic in 2020, weighting more than **99%** of the other entries. We therefore **decide not to include the biological disasters** in the **Type-II category**.
Subsequently, we analyse the pandemic *geographically*. We first make the hypothesis H_0 that the biological disasters are *uniformly distributed all over the USA*, as it seems coherent to consider the effects of a pandemic approximately uniform over a country. We can reformulate H_0 as "the distribution of biological disaster over the US follows a **normal distribution**". However, using the *Shapiro-Wilk test*, we obtain $p_{value} = 1{,}8.10^{-13}$. We therefore **reject the null hypothesis** with a confidence of 95%.
Notwithstanding, the form of the histogram of the number of disasters by states prompts us to try an **exponential model**. We try two solutions to build our model: fitting the exponential curve to our data and using the **maximum likelihood estimator** θ_{MLE}. Nevertheless, using the *Kolmogorov-Smirnov* test, we find $p_{value,fit} = 3{,}7.10^{-10}$ and $p_{value,MLE} = 3{,}2.10^{-5}$. Therefore – although the MLE model is a significant improvement from the other models – we must **reject H_1** with a confidence of 95%.



Finally, although we failed to build a distribution model, this study allowed us to assert that *most of the states handled the pandemic the same way*, limiting their declaration of disaster under 6; whereas *two states faced complications*, namely Oklahoma and New Mexico, which declared respectively *17 and 13 disasters*, while the **mean is 3.05 disasters per state**. We can therefore recommend the government to pay a particular attention to the emergency services in charge of biological disasters in those states.

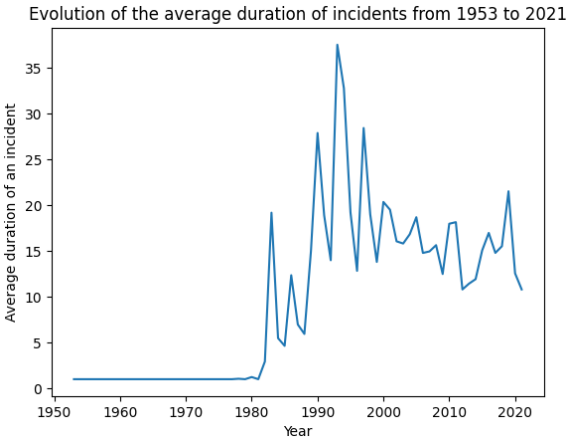
II – TIME EVOLUTION OF DISASTERS

Specific data cleaning

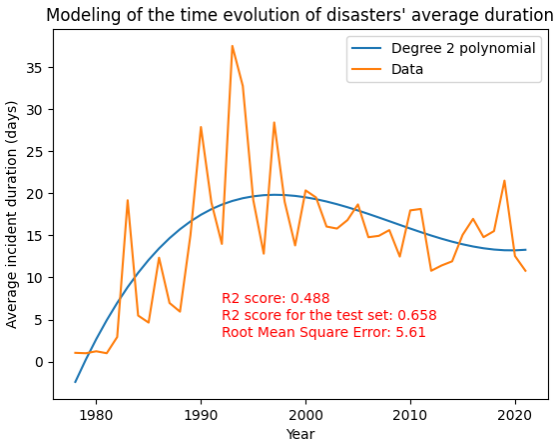
This part of the project deals with the link between the recorded incidents and climate change. Consequently, we do not take into account the incidents whose types are classified as *'biological'*, *'human cause'* or *'terrorist'*.
Moreover, we compute the duration of disasters, among other things. We thus drop all rows having a null value in the **incident_begin_date** or **incident_end_date** columns. As the dataset was last updated in October 2022, this year does not have records for every month, and will therefore be dropped as well.
Finally, the disasters that impacted several geographical locations appear several times in the data, once for each place. In this part, we only keep one occurrence of each disaster. *4,185 entries* remain after the cleaning process.

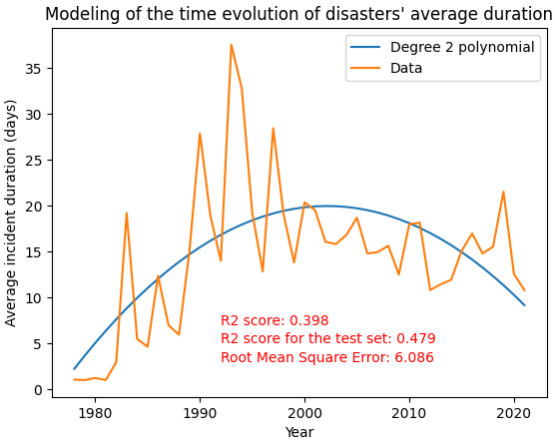
Duration of the disasters

We firstly focus on the *average duration of incidents per year*. The durations, expressed in days, were computed using the begin and end dates of each disaster, and were **capped at 365 days**. This decision was due to the observation that some disasters had lasted several years, and thus **distorted the data**.
The resulting graph shows a leap in the average duration of incidents around the year 1980. This particular observation might not only be due to climate change, but also the result of a more systematic recording of incidents in more recent years. However, the average duration also shows a sharp increase from the 1980s to the year 2000, and a steady evolution since then.



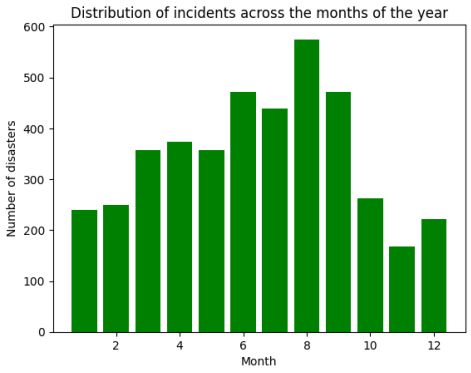
Given that practically all the incidents recorded before 1978 lasted for only 1 day, the following regressions only consider the years between 1978 and 2021. The modelling was conducted for polynomials of degrees 2 and 3, using an **OLS regression**.



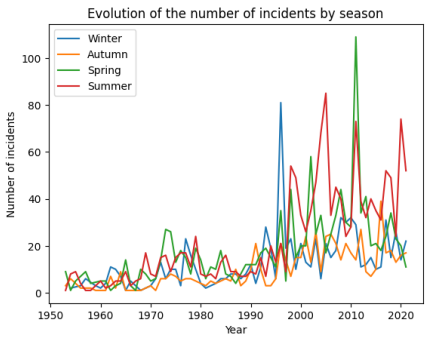


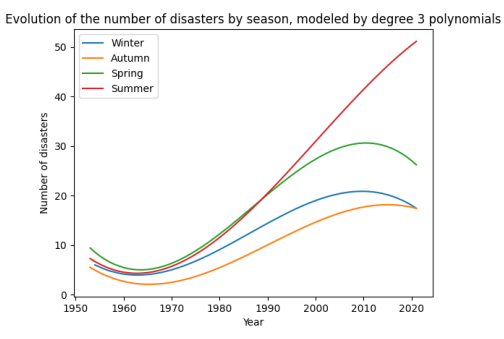
In order to fit the curve, the data was split into a train set and a test set. As shown by the evaluation metrics, a third-degree polynomial function is best suitable to model the evolution of disasters' duration. Firstly, the RMSE is lower than for a second-degree polynomial; secondly, the model is able to predict the results satisfyingly, given that the context does not require great precision. The resulting graph gives a good account of the rise in the duration of incidents after 1980, and of its stabilization at a significant number of days until 2021, with a potential increase in the next years that can be linked to global warming.

Seasonal dependencies



According to research led by climatologists **[7]**, climate change strongly affects the *timing and nature of seasonal events*. An interesting approach to the evolution of the number of natural disasters might thus be to test this hypothesis by **differentiating the incidents by season**.
This first graph simply aims at helping us visualize the general distribution of disasters across the months of the year. The **peak during the hottest summer months** seems in accordance with the number of fires, floods and tornadoes recorded.

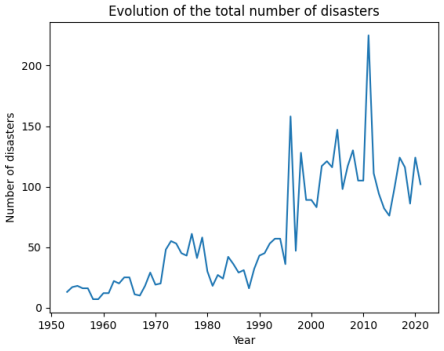




The first graph on the left being rather hard to read, we visualize the evolution using an **OLS regression** in the second graph.
The third-degree polynomial regression underlines the **sharp increase in the number of disasters occurring in the summer**.

In summer, the number of fires, hurricanes, floods and tornadoes can be directly linked to the effects of climate change. This effect is all the more striking as the rise has been so steadfast since the beginning of the 1970s. However, when it comes to the other seasons, the increasing number of natural disasters does not seem to have such a significant impact. Nevertheless, the **overall evolution** of natural disaster follows – as will be observed – a **sharply ascending curve**. The strong increase of the summer disasters must therefore cause a general increase of the annual natural disasters.

Annual evolution of the number of natural disasters

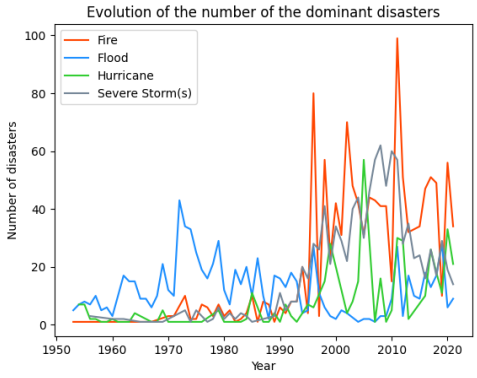


This section will concentrate on the evolution of the number of natural disasters over the years. Firstly, let us visualise the evolution of the number of natural disasters over time.

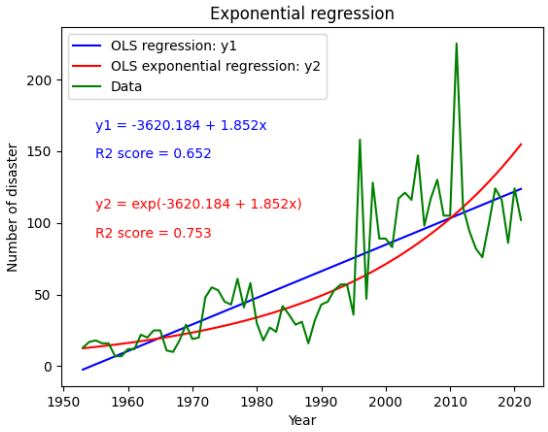
As we can observe, the increase is quite moderate from 1953 to 1996. Then, a sharp growth in the number of disasters occurs, leading to an extreme plateau which is still ongoing. It seems consistent with the global warming curve **[1]**, the global temperature starts to rise and exceeds 0°C in 1980. This might therefore be the trigger for the increase in disasters in the USA, especially in fires.

The category of **dominant disaster** has been identified in part I as being composed of 4 major disasters (**Type I disasters**). All four of them can be correlated to global warming and its effects.

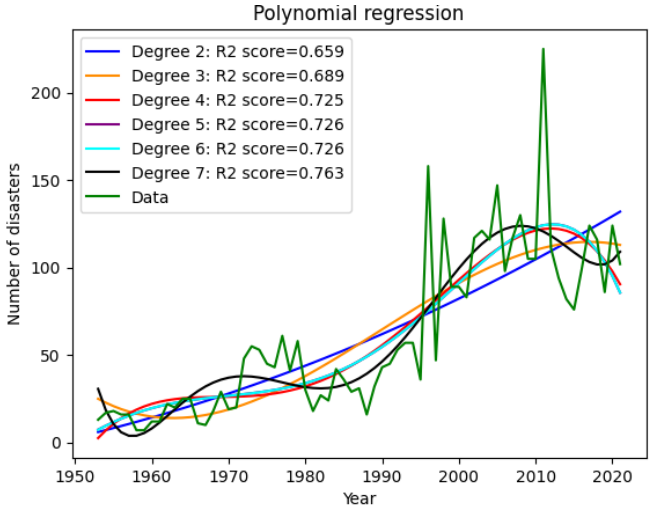
In 1996, we detect a sudden increase for three of the type-I disasters (fires, hurricanes, and severe storms), while the number of floods seems to remain rather steady. Having done these first visualisations, we will now build several models to predict the evolution of the number of natural disasters over time. We will choose the most coherent model regarding the information in our possession, the bibliography and the numbers we obtain. Aside, let us remind that we cannot apply these models for more than a few years in the future, as an unforeseen disaster could disrupt the model.



In a first time, we try a **linear model** and an **exponential model**. As we can observe, both displays the same behaviour with a sharp increase. However, the first R^2 score is too low to consider the linear model a success. The exponential model fits undoubtedly better: its R^2 score is strictly greater than the one of the OLS linear model. However, we will try building models with even better R^2 scores.



These curves emphasize on the fact that if we do nothing against global warming, disasters will most probably continue to multiply. Moreover, these curves have the advantage **not to be overfitted**, such as some of the models we will consider thereafter.

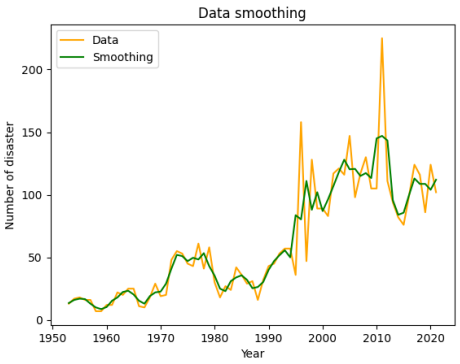


To remedy the low R^2 scores of the linear model without using an exponential curve – which has the inconvenient to grow exponentially with time, and therefore not being very subtle – we are going to work with polynomials. We will first fit the curves, and then carefully select the degree of the polynomial. We chose to compute polynomial models with degrees lower than 7, to avoid **overfitting** (i.e., a high sensitivity to small variations in the number of disasters, and thus an unreliable model).

As expected, the R^2 score is overall higher than previously. However, it is still lower than the exponential model for degree strictly lower than 7.

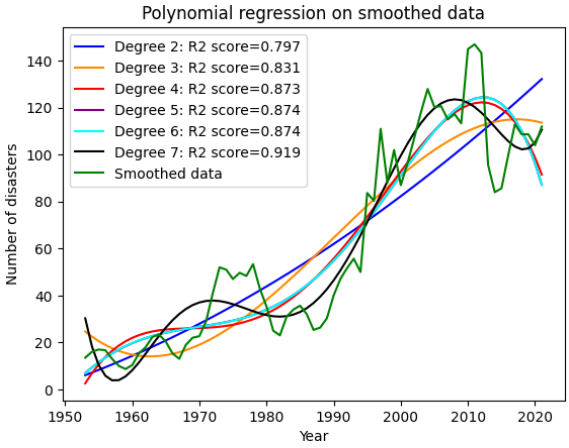
To select the degree of our model, we try to **compromise** the R^2, the behaviour in the last decade, and the need to avoid overfitting.

We first notice that the degree 4,5 and 6 polynomials reach a (local) maximum in 2020, and starts to decrease drastically, whereas – on the contrary – the number of natural disasters is expected to grow in the next decades. The degree 7 polynomial has a higher R^2 score than the exponential model, but clearly overfits the data. We therefore reject the aforementioned models.
The 2nd degree polynomial has an acceptable behaviour, a slightly better R^2 than the linear regression and is free from overfitting. On the other hand, the 3rd degree polynomial reaches a local maximum around 2015 and starts to decrease slowly. Even though this model is not fundamentally unacceptable, the fact that the exponential regression has a greater R^2 score and a more consistent behaviour make us reject this curve.
We finally accept the exponential model along with the 2nd degree polynomial regression to model our data. Nevertheless, we are going to try another method to obtain a more reliable model.



Data smoothing

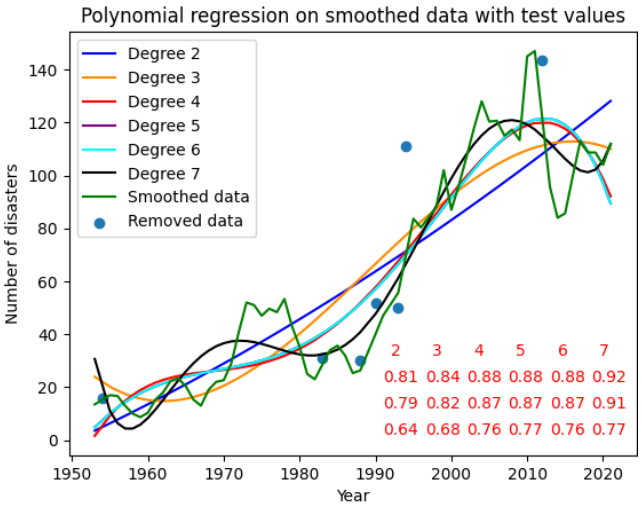
To make the data less sensitive to abrupt variations, and therefore increase the R^2 score, we are going to **smooth the data** and thus eliminate the most imbalanced peaks. For this, we will use the **moving average method** with the coefficient $n-1$ and $n+1$. By using a few values, we can keep the global aspect of the curve while removing the strongest variations. Using the smoothed data, we try once again to construct polynomial models.



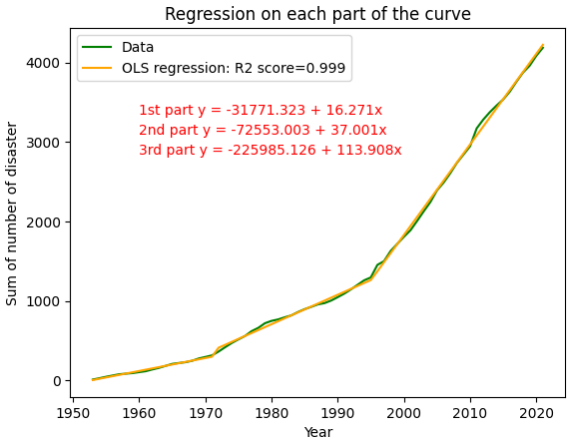
As we can observe below, the R^2 score have strongly increased, and the models have undoubtedly improved, even though their behaviour are quite similar. Notwithstanding, we still have the issue of sharp decrease in the last decade for degree 4,5 and 6 polynomials.
Moreover, the 7th degree polynomial curve still overfits the data. As for the 3rd degree polynomial model, we apply the same reasoning than previously.
We will therefore choose the last **2nd degree polynomial regression** to model the data. Indeed, this function has some major advantages: firstly, it is based on the smoothed data, and therefore less sensitive to abrupt variation. Secondly, it is a degree 2 polynomial, thus extremely easy to compute an analyse. Finally, its growth is not exponential, leading to a more reliable model for long-term predictions.

Test and regression on the smoothed data

In this part, we try to **train and test our regression** to evaluate the fitting of new models. For this, we remove 10% of the data randomly (7 points), then we compare these values to the new polynomials.
The table on the bottom-right corner of our plot shows the three characteristic R^2 **scores for each polynomial**: one for **all data**, one for the **model** and one for the **test data**.
These new polynomials are quite close to the previous. However, by observing the third row, we can observe that – for this particular set of random points – the higher degree polynomials do not give the best results. We can therefore conclude that not selecting these functions was a good choice regarding the reliability of our general model (less complexity, less overfitting, and a low standard deviation).



Cumulative method



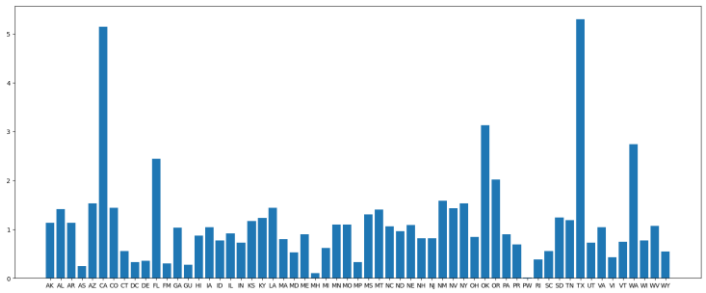
We finally visualise the data using a new mathematical tool: the **cumulative method**. This technique allows to monitor variations while smoothing the curve, without altering the data. The equation grounding this method is the following: $\boldsymbol{f(year)} = \sum_{\boldsymbol{i=1953}}^{\boldsymbol{year}} \boldsymbol{Nb_{disaster\ in\ year\ i}}$.

By plotting the curve, we identified **3 different trends** which could not be predicted with the standard visualisation. To conclude, under the hypothesis of a **similar trend** in the near future, we can assert that this model is even better that the previous ones, as the overfitting is low (linear model) and the R^2 is extremely high.

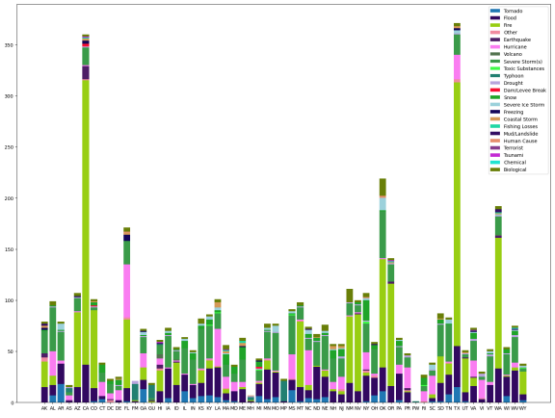
III – STUDY OF RELIEF FORCE MOBILIZATION BY STATE

We will now focus on finding the proper way for the government to intervene and distribute the relief forces. This brings us to study the repartition of the disasters.

Breakdown of disasters by states

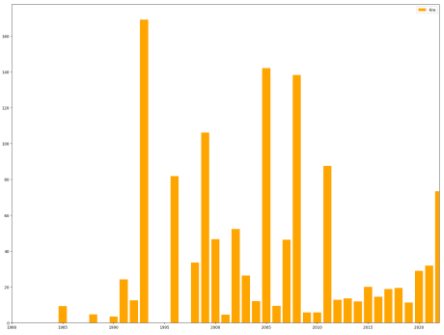


As explained earlier, a disaster that affected several counties appears several times in the database. Here we only count each disaster once for each state it affected, even in the case of multiple counties being concerned. 4,708 rows remain after the removal of the duplicates. The **bar chart** of the **average annual number of disasters per state** shows that Texas and California are the most affected by disasters, with an average of 5 incidents per year. By cumulating the information of this chart with the ones of the 1st part, we can advise the government on how to deploy its relief forces. The next four states with the most disasters per year, after Texas and California, are Oklahoma (OK), Washington (WA), Florida (FL), and Oregon (OR). Only Oregon is among the ten largest states in America. We can hence argue that a state's sources of vulnerability cannot be reduced to its size, but also include other factors such as its climate and geographical situation. Consequently, it is important not to plan to simply dedicate more relief forces to the largest or the most populated states, but to adapt the distribution to each state's needs.



This **cumulative bar chart** of **disaster types by state** indicates that the most common types of disasters are fires, severe storms, floods and hurricanes. These observations are consistent with the results of part I and allow us to establish an order of priority as to the training of the relief forces. Furthermore, the great diversity of disasters compels to adapt responses to each state, making prevention more complex. Therefore, providing adapted infrastructures and responses to all disasters that may occur appears as a necessity.

In the case of a fire, the duration of the disaster might reflect the intervention efficiency of the relief forces. Given that fires are additionally the most common natural disasters, we decided to further analyse this particular type of incident. This study should provide some insight on the type of decision that the government could make, based on the analysis.



In terms of the **data cleaning** specific to this part, the multiple occurrences of one disaster having affected several counties were all taken into account, since the duration of the fire may vary depending on the location. This final bar chart displays the **evolution of the average duration of fires** in the USA since 1953, expressed in days. After a decrease in fires' duration around 2010, it is now increasing again with an average of 80 days for the year 2022. With summer heatwaves growing in number, we can predict an intensification and a multiplication of fires. It thus seems necessary to strengthen the relief forces for this type of incident, and to do so by concentrating on arid, fire-prone regions such as California.

KNN algorithm and confidence

We now attempt to predict future disasters in a different manner, using classification thanks to the **KNN algorithm** ($k = 69$). The aim is to predict the type of incident most likely to happen in a given state, on a given date. After converting the data into numerical values and normalizing the results, we obtain the classification report on the right.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	15
1	0.45	0.16	0.24	88
2	0.56	0.87	0.68	164
3	0.00	0.00	0.00	6
4	0.00	0.00	0.00	2
5	0.59	0.48	0.53	48
6	0.00	0.00	0.00	1
7	0.39	0.45	0.42	91
8	0.00	0.00	0.00	1
9	0.00	0.00	0.00	4
10	0.00	0.00	0.00	4
11	0.00	0.00	0.00	1
12	1.00	0.05	0.09	22
13	0.00	0.00	0.00	5
14	0.00	0.00	0.00	2
15	0.00	0.00	0.00	2
18	0.00	0.00	0.00	1
20	0.00	0.00	0.00	1
22	0.32	0.92	0.47	13
accuracy			0.50	471
macro avg	0.17	0.15	0.13	471
weighted avg	0.47	0.50	0.43	471

Disaster types that were never predicted have their accuracy and F-score set to zero by default. We note that the only disasters predicted are on rows 1, 2, 5, 7 and 22. The first three correspond to floods, fires, hurricanes and severe storms, which were shown to be the most frequent incidents. The last type of disaster to be predicted is the biological one, whose importance in the dataset starting from 2020 is due to Covid-19. Most of the values for accuracy, precision and recall are inferior to 0.7. We could probably improve the accuracy of the results by adjusting on the size of the train set, which was 0.1 here. However, considering the limited information provided in the dataset (we can only base predictions on a place and a month), classification does not seem the most effective method to predict future events. For instance, a system of probabilities as above might be better suited. Running the algorithm with additional information on the weather conditions, such as temperature or pressure, might also improve the results.

We also created a **confidence indicator** (summarized on this last table) which for a given state and month will give the probability for each type of disaster that incident currently happening is of that type. I.e., the confidence indicator tells, knowing that there is a disaster in a state in a month of the year, the probability that it is of a given type.

	state	month	incident_type	confidence indicator
0	AK	1	Earthquake	0.333333
1	AK	1	Flood	0.333333
2	AK	1	Severe Storm(s)	0.333333
3	AK	2	Mud/Landslide	0.250000
4	AK	2	Severe Storm(s)	0.750000
...	...	...	...	...
1573	WY	8	Tornado	0.142857
1574	WY	9	Fire	0.833333
1575	WY	9	Toxic Substances	0.166667
1576	WY	11	Fire	1.000000
1577	WY	12	Severe Storm(s)	1.000000

CONCLUSION

Having analysed the FEMA dataset, we are now able to answer to the questions we asked ourselves at the beginning of this report.

Firstly, we have seen that some locations in the USA could be considered as particularly dangerous, in the sense that they are especially affected by disasters in regard to the rest of the country. Namely, the western states and the southern states are the most affected by disasters.

Secondly, we have constructed numerous statistical models that can be used to improve the management and allotment of the relief forces, as they can be more prepared to face certain types of disasters depending on the state they are in and the current period. Furthermore, the analysis might help the government to prepare nationwide by knowing where to allow more budget, as they are states where disasters occur more often.

Thirdly, we observed the effects of climate change on the number, size and duration of natural disasters in the US, and we built models aiming at predicting the number of accidents in the near future.

Finally, we have understood how the Covid-19 was considered in this dataset. Indeed, as the size of this pandemic was unprecedented, the number of reports clearly distorted the analysis, and thus a specific study has been made for this phenomenon.

BIBLIOGRAPHY

[1] https://www.ecologie.gouv.fr/impacts-du-changement-climatique-atmosphere-temperatures-et-precipitations

[2] https://www.usgs.gov/faqs/how-can-climate-change-affect-natural-disasters

[3] US Natural Disaster Declarations | Kaggle.

[4] https://www.statista.com/topics/1714/natural-disasters/#topicHeader__wrapper

[5] https://jcutrer.com/python/learn-geopandas-plotting-usmaps

[6] https://www.epa.gov/climate-indicators/seasonality-and-climate-change