Hurricanes and typhoons

Kwasi A. Boateng · Kristjan Solmann · Romain Sauvaget
November 29, 2021

2. Business understanding

Identifying your business goals

Background

Global warming is a phenomenon that affects the whole planet, no area will be spared. It is the greatest challenge facing our generation. Global warming implies a rise in air and water temperatures. Natural disasters will become more and more common and one of the most dangerous phenomena on our planet are hurricanes, cyclones and typhoons. These are exactly the same phenomena except that they have different names in the Atlantic, Indian and Pacific Oceans.

So we want to be able to study these phenomena, especially hurricanes and typhoons, to help our institute, the National Hurricane Center, which would like to know if the number of hurricanes in recent years has changed as a result of global warming and if they have become stronger than before.

The NHC is one of the World Meteorological Organisation's specialised regional weather centres responsible for forecasting and analysing tropical events in the North Atlantic and Northeast Pacific basins. It issues weather watches and warnings.

Business goal

This hurricane study aims to understand whether hurricane events are increasing and becoming stronger. Through this study, many countries could be better prepared for the arrival of such natural disasters on their soil and train their population to cope without falling into a general panic. We hope to publish this study before the next Cop 27 (Conference of the Parties) in Egypt 2022.

Business success criteria

The success of this study will be measured by its impact on those who are frequently affected by hurricanes and whether they take measures to protect their population. The effect of this report can be measured by looking at the evolution of each government's budget allocated to disaster prevention and whether it increases after the publication of this report.

Assessing your situation

Inventory of resources

The inventory of resources is the Kaggle records of the various hurricanes and typhoons from 1851 to 2014. Then there are the data miners, who will filter and clean all this data. Finally, the meteorologists and climatologists explain these phenomena and draw up an assessment.

Requirements, assumptions, and constraints

The data we have is freely available on Kaggle, so we don't have the problem of data security. The objective will be to have processed all the data within six months in order to give the experts as much time as possible to do their analysis and present the report in Egypt at Cop 27.

Terminology

Hurricane: cyclones of tropical origin with winds of at least 118 km per hour. In other words, a hurricane consists of powerful storm winds rotating around a relatively calm centre called the "eye". These storms are known as "typhoons" in the western Pacific, "cyclones" in the Indian Ocean and "baguios" in the Philippines. Each storm usually lasts several days.

Wind force: The strength of the wind is equal to the square of its speed. At 120 km/h, the wind is four times more destructive than at 60 km/h. At 60 km/h, wind exerts a thrust of 360 kg/cm². If the wind speed increases by 5%, the thrust increases by 44%.

Hurricane track: line of movement (propagation) of the eye in a given area.

Costs and benefits

Our institute is public. It exists because of the will of states and international organisations. Our study is requested by several states that have asked for our expertise and support to do our research.

Defining your data-mining goals

Data-mining goals

The deliverable will study the evolution of cyclones from 1851 until today with different curves showing the change during the years according to several parameters. (wind strength, duration etc.)

Data-mining success criteria

The success criterion of our extraction will be that there is an increase in the number of cyclones and that we find that they are more and more powerful over time.

3. Data understanding

Gathering data

We have a dataset of hurricane data from the Atlantic Ocean and typhoon data from the Pacific Ocean provided by Kaggle[1]. These are listed on the website as datasets in the Public Domain, not as a part of a competition. These datasets have documentation that is necessary to read and can be found in the same link provided. There is an additional dataset that contains information about El Niño wind patterns [3]. This particular dataset has no documentation.

Outline data requirements

We will be dealing with a .CSV document type. Below is the sample data with their data types.

Columns	ID	Name	Date	Time	Event	Status	Latitude	Longitude	Maximum Wind
Data type	Nominal	Nominal	Ordinal	Ordinal	Nominal	Nominal	Numeric	Numeric	Ratio

Verify data availability

The datasets can be found in [1] and [3].

Define selection criteria

There are 22 identical columns in both pacific and atlantic datasets; however, the columns of interest are listed in the table above.

Describing data

Currently, we have a dataset of hurricane data from the Atlantic Ocean and typhoon data from the Pacific Ocean. The ocean datasets are the main ones we are working with. The El Niño dataset is there if we want to check its effect on the hurricane data or refine a hurricane prediction model. These are not our main tasks.

The two Kaggle datasets are in the same format which is very helpful. They contain six-hourly information of every known hurricane and typhoon in the time frame covered. The Atlantic dataset contains information from 1851 to 2015, while the Pacific dataset contains info from 1949 to 2015. The Atlantic dataset contains 49105 cases and the Pacific dataset has 26137 cases. The six-hourly information contains the following features [2]:

- Hurricane ID

- Name (if given)
- Date
- Time - hourly
- Event (if given)
- Status - type of storm
- Latitude
- Longitude
- Maximum wind speed - in knots
- Minimum pressure - in millibars
- Low, Medium and High wind speeds for every ordinal direction.

This is the key data and perfect for us to mine data from. The main fIelds we expect to use are ID, Date, Latitude, Longitude and Maximum wind speed.

The El Niño dataset contains daily values for wind, humidity and temperature from 1980 to 1998. So it documents the wind itself and its temperature and humidity component. The features are:

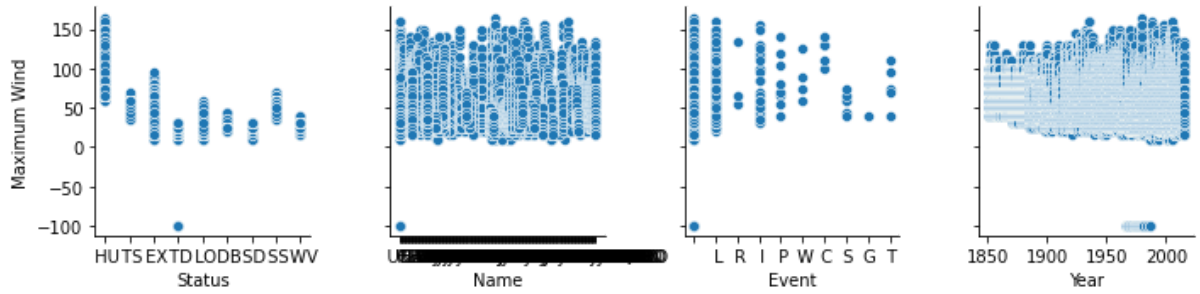
- Date
- Latitude
- Longitude
- Zonal Winds
- Meridional Winds
- Humidity
- Air Temp
- Sea Surface Temp

The fields we might use are: Date, Latitude, Longitude, Zonal Winds, Meridional Winds.

Exploring data

When looking at the data more closely the first thing that pops out is that the Low, Medium and High wind speeds are missing for most of the data, meaning they are represented in the data as -999. In the atlantic and pacific datasets, they start appearing from 2004 onwards. This means they are present for only about 10% of Atlantic cases and 20% of Pacific cases. This means these features are a lot less valuable since we would want to compare them historically.

The only anomaly in the numerical data that interferes with data mining is that in 1967 some cases had their Maximum wind speed set to -99 as a temporary patch in the data[2]. This could ruin numerical analysis results for this feature; therefore, it needs to be handled carefully.



A plot of the atlantic data shows that the Maximum wind relates to the following features; Status, Name, Event and Year.

We have done less exploration for the El Nino data as we are not confident it will be used within the given timeframe. The main problem with it is that the daily cases are only from one position, one that could be far away from any cyclones at the matching time, rendering the content of the features useless, as nothing relevant can be extracted from them. Furthermore, as a minor issue, humidity and wind data seems to be missing from some rows.

Verifying data quality

In summary, we have two datasets containing information about hurricanes and one dataset with information regarding the El Nino wind pattern.

The Atlantic and Pacific datasets are pretty big and contain enough features to study the evolution of cyclones. The quality issues they have are minor except for the values of ordinal direction wind speed being absent for most of the data, which significantly limits using these features for a historical analysis of cyclone evolution. However, ordinal direction wind speed is not one of our main features of interest, so we are not seeking additional data to cover this gap.

The El Nino dataset is minimal in terms of the time its data covers and is only from one location. We are not seeking out additional data to cover this since this is the best we could find. We are keeping this data since we could still use it to inform us on the possible effect of the wind pattern on cyclones.

4. Planning your project

Project outline

Each team member contributes 1-2 hours in weekly group meetings and 1-4 hours per task. The tasks are as follows:

- Clean up all Pacific and Atlantic data
- Interpreting the data
- Comparing datasets of atlantic and pacific to each other
- Measuring the frequency of cyclone occurrence each year in the Pacific and Atlantic Oceans
- To identify the evolutional strength of hurricanes
- Produce useful graphs to represent our data
- Observe how El Nino affects hurricane patterns. (optional task)
- Conduct statistical studies on the data to observe the evolution of cyclones since 1851. (This task is optional and can be done if we have time)

Tools we plan to use

- Jupyter Notebook for data cleaning, interpretation and visualization.
    - The installed python libraries include Numpy, Pandas, sklearn, seaborn, Matplotlib, etc.
- PowerPoint for the presentation of results

Appendix

[1] - https://www.kaggle.com/noaa/hurricane-database
[2] - more info in atlantic.pdf in our github
[3] - https://www.kaggle.com/uciml/el-nino-dataset

Github Link: https://github.com/new2me321/IDS_project_D23