SUBJECT 4 - Big Data and Data Mining techniques applied to the characterization of atmospheric tornadoes

Project's context:

Our approach to carry out this project is divided into multiple steps.

The first step is a bibliographic search on tornadoes in order to grasp the physical mechanisms involved and have a better understanding of the discrepancies between various characterisations. This study enables us to determine the type of data needed to classify tornadoes. We also find out that a large quantity of mathematical modeling were to use to describe tornadoes. Thus, on the advice of our supervisor, Joël Chaskalovic, we will focus on the equation of Navier-Stokes which is the most global one. Moreover, this equation can be shortened and therefore simplified according to the type of fluid flow : turbulent, laminar, viscous, compressible… In fact, not only the knowledge of these parameters simplifies the study of tornadoes' mechanisms, but it also alleviates the numerical simulations.

In a second step, we are trying to contact organizations holding data (see a model of the mails in the organization part). Knowing that more than 1500 tornadoes roam the soil of the United States of America, we try to contact in priority American organizations such as NOAA, UCAR, AMS, or the University of Oklahoma. To spare no track we will contact French organizations such as Keraunos and others.

Then, in a third step, we will have to process and analyse the data collected. Several cases can occur. Ideally, we will have access to real-time tornado data of all types (Force 1 to 5 on the Fujita scale). The information will therefore be real-time spatial data of pressure, temperature, wind speed, shear rate, viscosity, Mach number, tornado's geometry, moisture content and Reynolds number.
In another possible case, the collected data might be raw (ie such as sensor data from satellites for example). Such data is not processed, it is not directly exploitable. For example, some meteorological satellites have sensors measuring the albedo (reflecting power of a surface) indirectly giving information on pressures. This type of data will have to be processed upstream to create more significant and exploitable characteristics, especially for the study of Navier-Stokes equation.
The worst scenario would be that we were unable to collect data from the various organizations contacted. So, we will have to be content with data already found on the internet being quite poor in information (ie we only have for each tornado: the trajectory, the most extreme wind speeds, the duration of life, the force, and the damage caused). This data might not be useful for characterizing a tornado according to the Navier-Stokes equation.

The rest of the project will therefore depend on the type of data obtained. However we have already planned our project in the ideal case. Our first part will be based on the study of the classification of tornadoes according to the Fujita scale from classification algorithms on a supervised learning. We will try to draw the notable characteristics of this classification. For the second part we will try to create groups in a non-supervised way thanks to Data Mining techniques. This will lead us to compare the groups obtained, to the different models of simplification selected, thanks to the algorithms, by the study of the equation of Navier-Stokes.

Organisation:

To carry out this project we are a team of four students: Siloée Montreuil, Nicolas Playe, Gabriel Ducasse and Gufran Erol.
At first, we barely knew each other, so the allocation of the tasks was not clearly given.
But after a first contact between us four and a second meeting with our attributed professor, the doctor Joel Chaskalovic, we eventually had a more precise approach of the work that must be done and a better idea of the tasks that will suit each student according to its abilities.
That's how we made the following distribution:
We all had to search a maximum information on the physical phenomenon and the possible linked models. In fact, in order to be effective for the upcoming tasks, we needed to have a minimum culture on tornadoes and on the institutions and laboratories working and collecting data on this phenomenon. Then we decided that:

- Nicolas Playe will deepen the former researches on the institutions to find intern contacts or intermediaries for each institution culled which are:

- *spc.feedback@noaa.gov*

- *amsinfo@ametsoc.org*

- *patrick.marsh@noaa.gov*

- *http://meteorology.ou.edu/contact/metit@ou.edu*

-*http://weatherstationew0157.net/weather28/index.php?p=incContact&lang=en#data-area*

- *roger.edwards@noaa.gov*

- *https://www.ucar.edu/contact*

- Gufran Erol and Siloée Montreuil will write standard mails according to the institution such as the following example:

  Dear Sir or Madam,
  We are a group of 4 students completing a master's degree in Fundamental Physics at the Sorbonne University. We are contacting you because we are working on a scientific project toward Tornadoes and Big data. We browsed your website which provided us many information. But this project consists in a statistical study and a characterisation of this phenomenon using Big Data and Data Mining technics. Thus, we would like to know if you could provide us some numerical data such as velocity, temperature, pression, geometrical characteristics and any complementary information that could help us to deepen our study. We could use processed data, but also raw data linked to its documentation. Alternatively, may you provide us some contacts who could help us in that way.

  Best regards.

- Gabriel Ducasse will send all the mails to their recipient and oversee the exchanges with the laboratories.

For now, the results are thin because we only had answers from two laboratories for ten laboratories contacted.