



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

BIG DATA

Project Report

---

## Weather Dashboard

---

*Authors:*

Aubry CHOLLETON  
Jonathan DUSS  
Anders HENNUM  
Alexis KESSEL  
Quentin MAZARS-SIMON  
Cédric ROLLAND  
Orianne ROLLIER  
David SANDOZ  
Amato VAN GEYT

*Professor:*

Christoph KOCH

*Supervisor:*

Amir SHAIKHHA

May 20, 2014

# Contents

<b>1</b>	<b>Data Retrieval</b>	<b>2</b>
1.1	Looking for data . . . . .	2
1.2	The <i>Integrated Surface Database</i> . . . . .	2
1.3	Downloading the data . . . . .	2
<b>2</b>	<b>Nearest Neighbors</b>	<b>3</b>
2.1	Why? . . . . .	3
2.2	Implementation . . . . .	4
2.3	Results . . . . .	4

# 1 Data Retrieval

## 1.1 Looking for data

Initially, the project idea was to get Alps weather information from social networks using natural language processing. Hence we first looked how to mine Twitter, Facebook, Instagram or Campocamp. We also searched for other datasources such as the *Institute for Snow and Avalanche Research* (SLF), MeteoSuisse or ski resorts. After discussion and clarification with our TA, the scope and goal of our project changed in order to get closer to a *Big Data* project. Hence we needed to process a big dataset and the *Integrated Surface Database* from NOAA seems to be the one.

## 1.2 The *Integrated Surface Database*

The *National Oceanic and Atmospheric Administration* (NOAA) is the US agency responsible for the weather surveillance and forecast. They have a *National Climatic Data Center* (NCDC), “responsible for preserving, monitoring, assessing, and providing public access to the Nation’s treasure of climate and historical weather data and information”. One of their dataset is the *Integrated Surface Database* which has data not only for the USA territories, but for the whole world (20,000 stations). We can access this dataset under the following condition, which we fulfill.

“The following data and products may have conditions placed on their international commercial use. They can be used within the U.S. or **for non-commercial international activities without restriction**. The non-U.S. data cannot be re-distributed for commercial purposes. Re-distribution of these data by others must provide this same notification.”

The data is accessible by FTP at `ftp://ftp.ncdc.noaa.gov/pub/data/noaa/` using anonymous login. There is one directory per year and within each directory one file per weather station. Within a file, one line correspond to one data record of the station at a specific time.

All stations are listed in the file `ish-history.txt` with, among other attributes, their coordinates and the country in which they are located.

Every station can have two IDs: the *Air Force Datsav3 station number* (USAF) (6 figures) or the *NCDC WBAN number* (WBAN) (5 figures). Most stations have only one of these two IDs (the other being set to 9s). Hence we defined the ID of a station as being the combination of the two IDs, which guarantee uniqueness.

It is important to note that despite the fact that each stations has precise coordinates, each line entry in the dataset give the coordinates again because a station might be slightly moved.

## 1.3 Downloading the data

The dataset is over 500GB large when uncompressed. As we only have 300GB on our server, we couldn’t download the whole dataset. In order to have a sample dataset, we downloaded 10 years (2004-2013) of data for Switzerland, which is around 2GB large. Then we downloaded 39 years (1975-2013) of data for the USA, which is around 200GB large. We can notice that these 39 years of data for the USA are a big part of the whole dataset. We can therefore deduce that they richer (more frequent sampling).

To proceed the download from the NOAA FTP server to our server, we wrote a Shell script taking in argument the `filename` of a file containing a list of the stations we want and the `year` for which we want the data. As all the stations are not active, there will not necessary be a

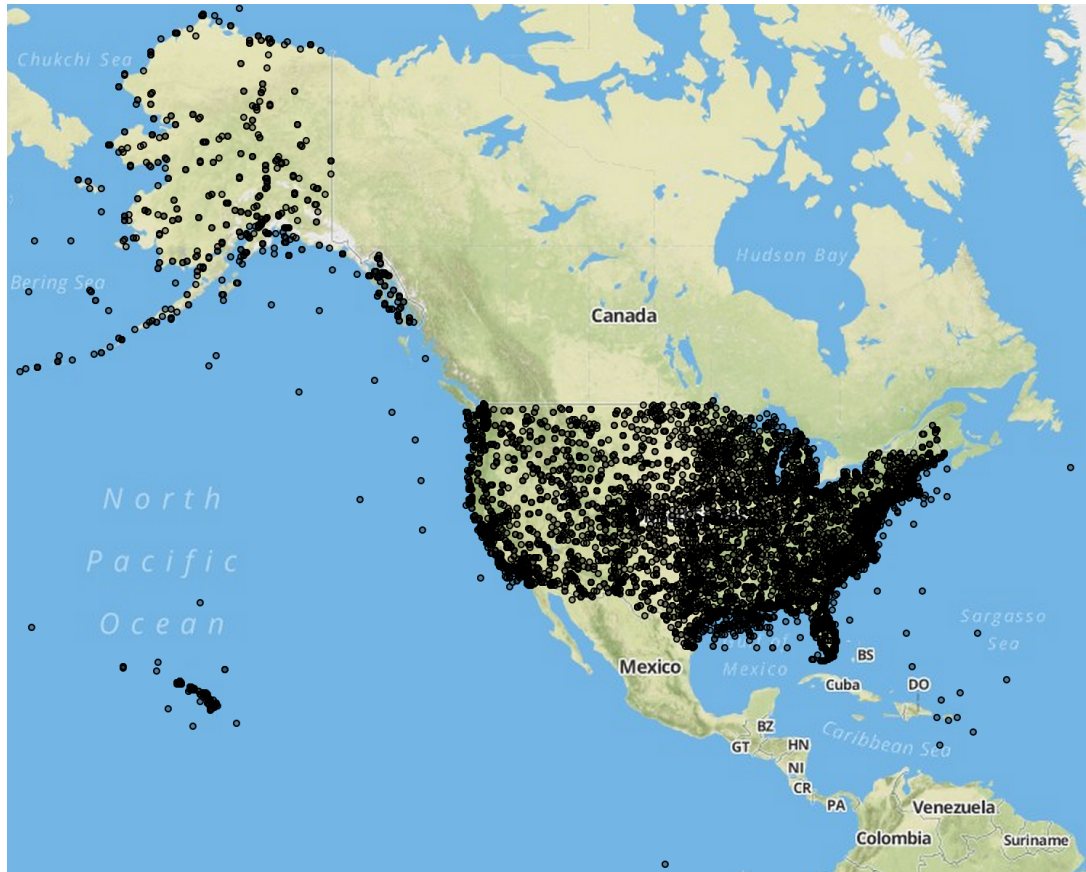


Figure 1: Stations categorized as being in the USA

file for each station in our list. Hence, before downloading, the script compute the intersection between the stations in our list and the files available on the FTP server. This will improve the downloading as it will not ask for files which doesn't exist.

## 2 Nearest Neighbors

### 2.1 Why?

The main purpose of implementing nearest neighbor, is to find similar weather periods in the history and possibly find patterns in the climate. This is useful in terms of understanding the climate and to forecast specific events.

As an example, here in Switzerland, it could be useful in terms of avalanche danger. If most of the nearest neighbors of the current winter so far, led to many and massive avalanches, it is reason to believe that also the current winter can become a such a winter. By knowing this, it would be possible to take action before the events occur. The same method can be used to forecast events like floods or drought.

It can also be used to find patterns in the climate such as cycles and periodic structures.

TODO: Rewrite to better language

## 2.2 Implementation

The k-nearest neighbor is a fairly simple algorithm. You choose a metric, calculate the distance from the reference node to all the other's, and pick the  $k$  nearest neighbors. This has been done more or less straight forward.

In this case a node is a period of time. We fixed the period to one month, such that you can pick one month for a specific year, and then find the most similar months for the other years. Often climate data is summarized by month, so this makes it easy to find data to compare with.

Next question is to determine how many intervals the period of time should be divided into. Should we compare averages for an hour, a day, a week, or just one value for the whole month? For a small region (a few stations) and a short period of time (a day), then comparing hour by hour could make sense and give good results. But for a large region (thousands of stations) and for a longer period (a month), this will mostly just return noise. Here, the method is implemented with a period option so one can choose how many periods a month should be divided into. The final results is run with this set to 1, meaning that averages for the whole month is used.

TODO: add figure shows how this was implemented with map reduce on Hadoop. Write some more and clean language.

## 2.3 Results

TODO: Present some results with some figures to show that it works.