UNIVERSITY OF STUDY MILANO-BICOCCA

Masters Degree Program in Data Science

Report project Data Management

THE EFFECT THAT RISING TEMPERATURES, DUE TO GLOBAL WARMING, HAS ON GLACIERS.

REPORT OF:

BARBERA VALENTINA N.856780

LUCA SINANAJ N.844540

INDEX:

FONTI DATI :

https://cds.climate.copernicus.eu/cdsapp#!/dataset/insitu-glaciers-elevation-mass?tab=overview

https://data.world/data-society/global-climate-change-data

INTRODUCTION

The goal of this study is to show the impact that rising temperatures have on our planet's glaciers.

Global warming is an atmospheric phenomenon that has been present for several decades now and is characterized by the increase in the average global temperature and the continuous change in atmospheric phenomena, caused by the emission of greenhouse gases into the atmosphere and the continuous exploitation of available resources, factors attributable to human activity. This atmospheric phenomenon over the years has had a strong impact on the fauna and flora of our planet, of significant importance turns out to be the melting of glaciers, which, in addition to causing concern in the scientific community, because of the large impact it can have on the entire ecosystem, has recently begun to capture public attention.

The data collected and used for analysis come from different sources, the first of these is the Copernicus platform, from which we obtained three datasets, the first containing data on elevation change per individual glacier ,the second containing data regarding annual balance, also per individual glacier and finally a third dataset including latitude and longitude for each ice boulder. The second source we used is Data world, is an Austin, Texas-based knowledge management company founded by Jon Loyens, Brett Hurt, Bryon Jacob and Matt Laessig, where you can search, edit and download data, like Kaggle platform ; from the latest source we obtained the average temperature readings by individual country.

Copernicus is an initiative of the European Space Agency (ESA) and the European Commission created in 2001, aimed at providing information by 2021 for the European Union in order to be able to act autonomously in the field of security and environment through satellite surveys. Copernicus is based on 4 pillars, the space component, airborne and ground based measurements, data harmonization, and user service.

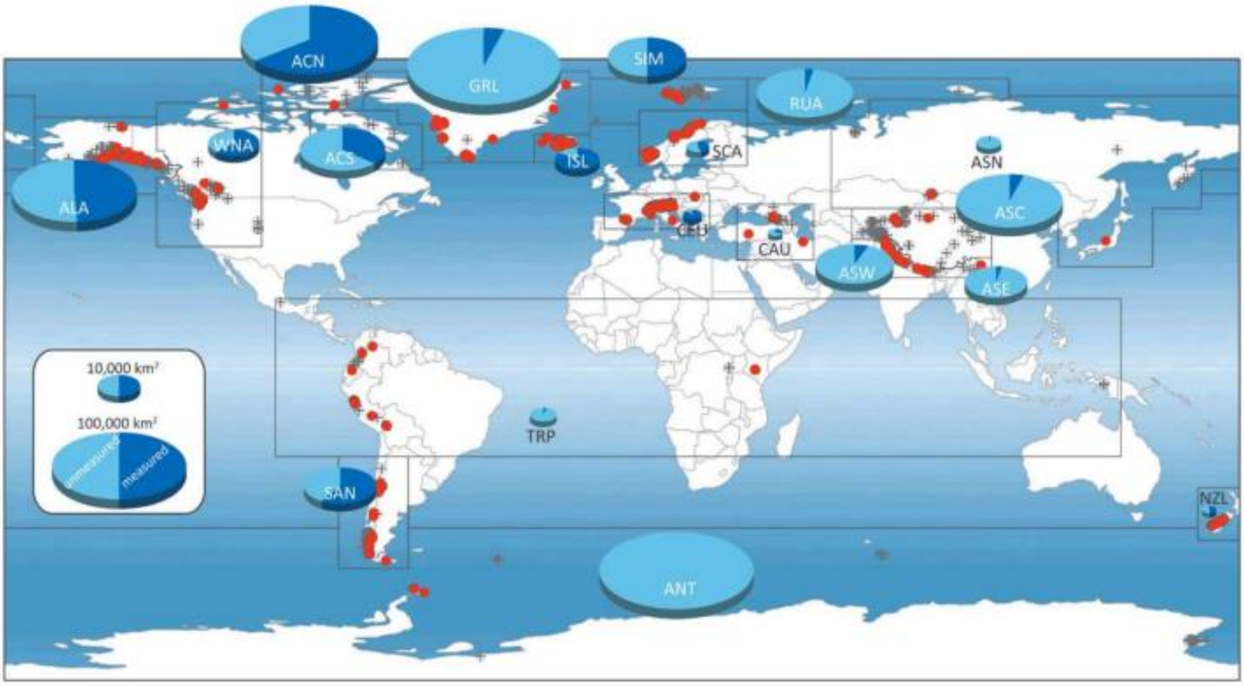


Figure 1: Distribution of glacier area and fluctuation records in 19 regions. The pie charts show the regional glacier area (excluding the ice sheets in Greenland and Antarctica) and the fraction covered by available observations of changes in glacier length, volume and mass. The dots show the location of continued (red) and interrupted (black cross) series. (https://datastore.copernicus-climate.eu/documents/insitu-glaciers-elevation-mass/C3S_312b_Lot4.D3.GL.8-v3.0_Product_User_Guide_Specification_Change_i1.0.pdf)

The following figure shows the distribution of glaciers present in both datasets (elevation change and annual balance), it can be seen that for some the surveys have been discontinued precisely because of this, in some cases, measurements for the most recent years are not available.

Even for the dataset obtained from Data World, on temperatures, there are no Average Temperature values for the most recent years, but we still decided to keep it because it allowed us to understand whether, over the years, there has indeed been a general increase in temperatures that has resulted in a contraction of elevation and glacier mass.

METODOLOGY:

DATA ACQUISITION

As already mentioned, the first three datasets were obtained through the API of the Copernicus platform, which provides the following data in tabular format. It is, in addition, possible to choose from possible versions of the following datasets, we decided to select the most recent ones available.

The elevation change dataset contains elevation change measurements (in millimeters) per individual glacier, while the dataset on annual balance contains mass change measurements per individual glacier, also in millimeters.

Both datasets have some similar variables :

- WGMS_ID : Glacier identification ID
- SURVEY_DATE: Date of survey
- AREA_CHANGE : Variation of area (in millimeters)
- AREA_SURVEY_YEAR
- NAME: Glacier name
- PU: Political Unit

The dataset on elevation change contains about 111.600 records, while the dataset on annual balance contains 7.186 records.

The third dataset obtained from the Copernicus platform contains latitude and longitude values for each glacier.

With the respective variables:

- PU
- WGMS_ID
- NAME
- LATITUDE
- LONGITUDE
- GLIMS_ID
- RG60_ID
- RG50_ID
- WGI_ID
- GEN_LOCATION
- SPEC_LOCATION
- GLACIER_REGION_CODE
- GLACIER_SUBREGION_CODE
- PARENT_GLACIER
- LATLONG
- STATE

The last dataset, was obtained from the Data World platform via download (composted from about 239.000 records) and contains the following variables:

- Dt: date
- Average temperature : in Celsius
- Average temperature uncertainty
- City
- Country
- Latitude
- Longitude

DATA INTEGRATION

After acquiring the data we noticed that for the dataset on annual balance the variable "AREA CHANGE" (change in area of the glacier compared to the previous survey) was not present , so we decided to create a new column taking as a reference value the variable "AREA", i.e., the value of the area for each individual survey and to subtract from each value that of the previous survey, so as to obtain a column with a type of values of the same nature as those of the variable "AREA CHANGE" present in the dataset on elevation. In this way a comparison can be made between the change in area due to a decrease in elevation and a change in area due to a decrease in glacier mass.

The dataset from which we started is the one on elevation to which we decided to integrate the values on annual balance. The number of records in the second dataset compared to the first one is inferior because glaciers for which the survey has not been interrupted are taken into account and we have measurements for the most recent years as well. Despite this, we still decided to supplement the elevation measurements with the respective measurements on annual balance, where possible, because the glaciers in the second dataset are the most globally relevant and consequently the main ones on which to perform the analyses. In the elevation dataset there is the variable "SURVEY ID," containing a unique value for each survey, so we decided to create a new column, for the mass dataset, containing the unique values for each survey in order to perform integration through a unique key ("SURVEY_ID").

Next to the dataset containing the information for elevation and mass we integrated the dataset containing latitudes and longitudes using the variable 'WGMS_ID', so that we could create a useful map to show the distribution of the considered glaciers in the World.

Finally to perform a join between the following integrated dataset and the dataset with temperatures we defined a function able to define the Country for Latitude and Longitude of the integrated dataset, through the library 'geopy' of Python, because in the dataset on measures we do not have the variable "COUNTRU". The join was performed with the variables “COUNTRY”.

As a last thing, we decided to make a new column for the temperature table called 'TEMP_ID', containing unique values for each record, so that we have a univocal key.

DATA STORAGE

The type of database we chose to save the data is the RDBMS, MySQL, for the following reasons:

- The data are already presented in table format.
- Consistency is assured, possibility of having all data at the same time.
- Availability is ensured, availability of the data to respond to all requests received.
- Since there is no real time data, it does not make sense to use distributed db.

In addition, MySQL, as an RDBMS, takes advantage of ACID properties:

- ATOMICITY: the operation is atomic, happens in full or returns an error and the database returns to it’s initial form. For example, either all data is updated, or none possible to update only part of it .
- CONSISTENCY: New data entries reflect the predetermined pattern.

- ISOLATION: Single operations do not affect other operations, this is because the database builds a process execution sequence so that the state of the database does not change during the execution of a request.
- DURABILITY: persistence of the file always guaranteed, even if the System crashes.

DATA QUALITY

Regarding data quality, after data acquisition we identified the presence of null values for the variables "AREA_CHANGE" and "ELEVATION_CHANGE_UNC" (the uncertainty of elevation change) of the elevation dataset and for the variables "ANNUAL_BALANCE_UNC" and "AREA_CHANGE" for the mass dataset.

We decided to impute the null values by construction of a linear regression model, and using the MICE package found in the RStudio language.

The mice package implements a method to deal with missing data. The package creates multiple imputations (replacement values) for multivariate missing data. The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data. The method chosen and included in the topic ‘methos’ of MICE is the Predictive Mean Matching (PMM), that involves selecting a datapoint from the original, non-missing data which has a predicted value close to the predicted value of the missing sample.

Through this methodology it was possible to impute the missing values correctly, but it is necessary to make a small note: the "AREA_CHANGE" Column for the elevation dataset had a high number of missing values, it is necessary to specify that the values imputed through PMM method were based on few original values present, consequently, for this Column the accuracy is not high.

Regarding timeliness the most recent year for glacier measurements is 2019, we can consider ourselves satisfied in that because the surveys are not necessarily done every year and time is needed to integrate new results.

For the latitude and longitude dataset, no missing values were found, and after entering the information on Tableau to create a map and compare it with one in the documentation provided by Copernicus, the glacier positions were found to be correct.

Overall, we can consider the collected data reliable since the main source turns out to be Copernicus, a European agency.

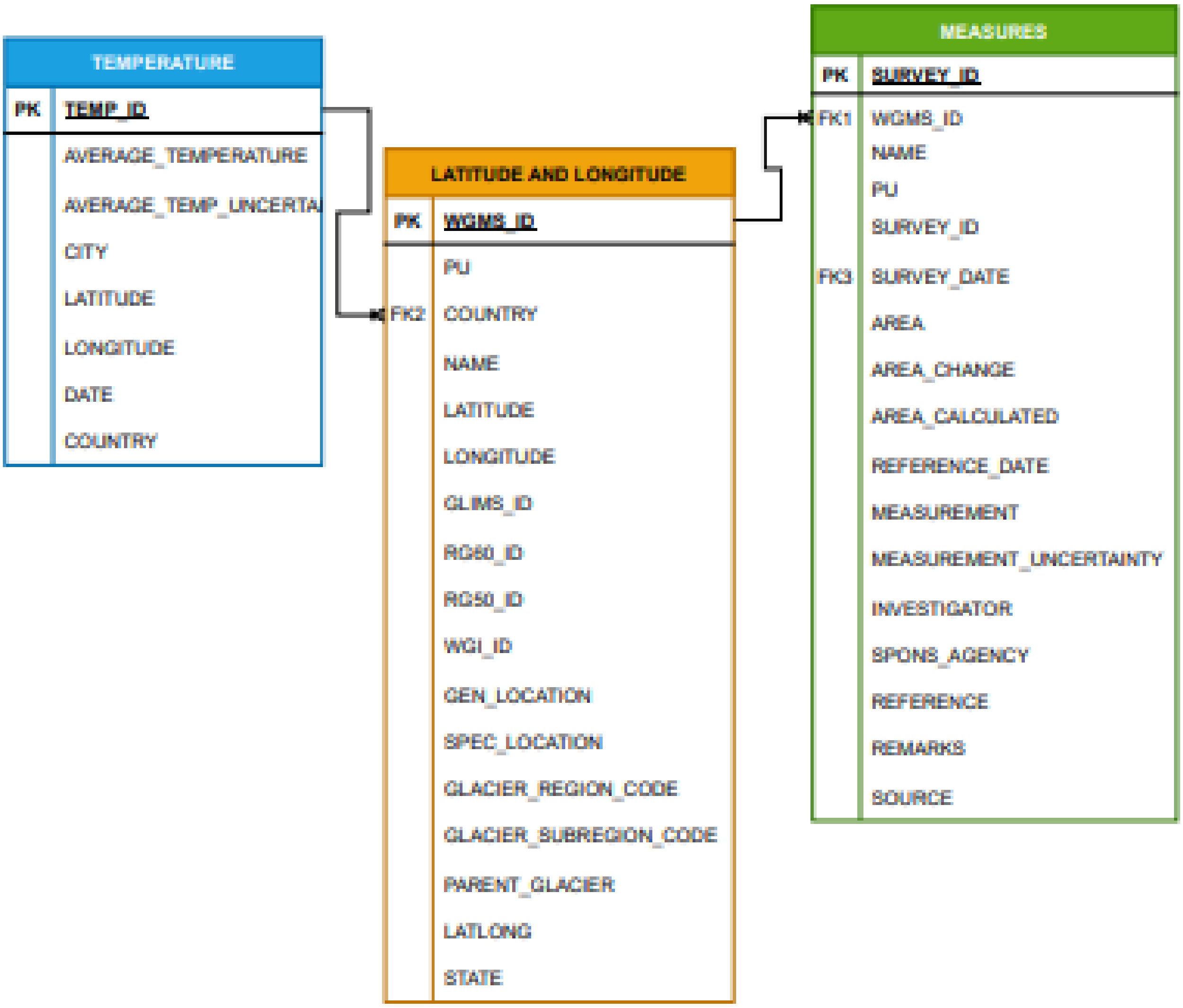
The temperature dataset obtained from the Data World platform contained some null values for the variable "AVERAGE_TEMPERATURE", we decided to delete these records, as they are a small number, less than 2% of the total information.

Temperature measurements cover a fairly broad range of time, from 1749 to 2013, sufficient to detect an overall increase in temperatures as a result of global warming. The failure to collect temperatures for more recent years has not limited the analysis.

The temperature information is of relevant importance, as after graphing some glacier elevation and annual balance trends over the years, we detected an increasing decrease in mass, elevation, and area values, attributing this phenomenon to glacier melting and thus to rising temperatures. Thanks to the following dataset we were able to confirm our hypothesis, because it is possible to detect a general increase in temperatures around the world.

Since these are annual or monthly surveys for each glacier, in all datasets, no redundancies were detected.

CREATED DB SCHEME



TEMPERATURE	
PK	TEMP_ID
	AVERAGE_TEMPERATURE
	AVERAGE_TEMP_UNCERTAI
	CITY
	LATITUDE
	LONGITUDE
	DATE
	COUNTRY

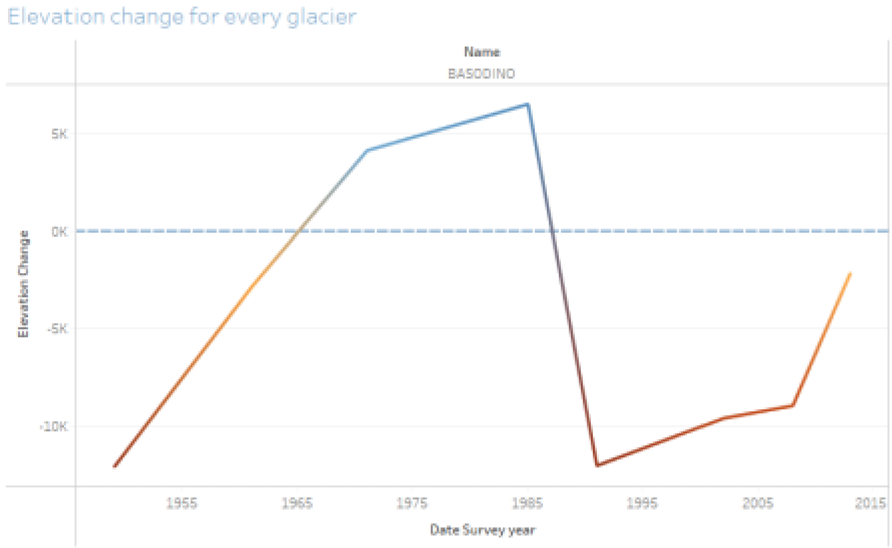
LATITUDE AND LONGITUDE	
PK	WGMS_ID
	PU
FK2	COUNTRY
	NAME
	LATITUDE
	LONGITUDE
	GLIMS_ID
	RG60_ID
	RG50_ID
	WGI_ID
	GEN_LOCATION
	SPEC_LOCATION
	GLACIER_REGION_CODE
	GLACIER_SUBREGION_CODE
	PARENT_GLACIER
	LATLONG
	STATE

MEASURES	
PK	SURVEY_ID
FK1	WGMS_ID
	NAME
	PU
	SURVEY_ID
FK3	SURVEY_DATE
	AREA
	AREA_CHANGE
	AREA_CALCULATED
	REFERENCE_DATE
	MEASUREMENT
	MEASUREMENT_UNCERTAINTY
	INVESTIGATOR
	SPONS_AGENCY
	REFERENCE
	REMARKS
	SOURCE

COMPUTED ANALYSES

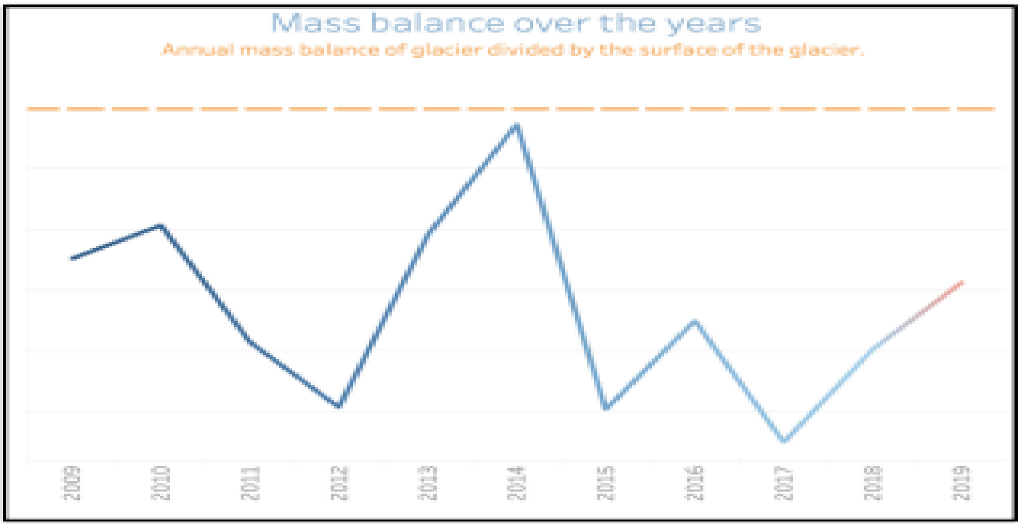
Following are some examples of graphs made:

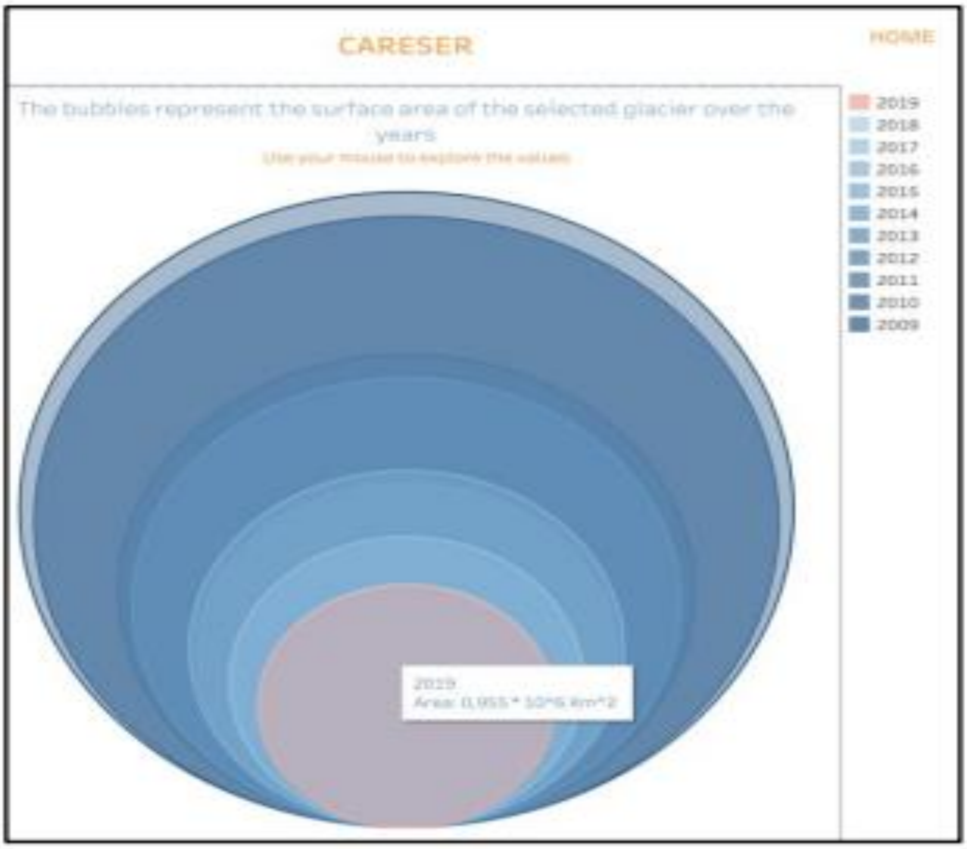
Ex. Glacier Basodino



The following graph shows the trend of the different surveys for Basodino Glacier, located in Switzerland. Over the years the elevation change has not been uniform, it can be seen that from 1965 to 1985 the surveys report a positive value, as it is above the zero threshold. From 1991 to 2005, elevation change values are negative, this indicates glacier melting in those years. For the year 2015, the graph line turns out to have an upward trend, because compared to the previous survey, which took place in 2008, elevation change is still negative but with a lower value than the previous point. For the following glacier, the last survey is for the year 2015, but it can be assumed that over the years, as temperature increases, further elevation declines will occur.

Ex. Glacier Careser





CARESER

HOME

The bubbles represent the surface area of the selected glacier over the years

Use your mouse to explore the values

2019
Area: 0,955 * 10^6 km^2

The following graph considers the Careser glacier, located in Switzerland, where it is possible to see a sharp contraction in mass between the year 2011 and 2012, again it is plausible to consider this change a result of rising temperatures. From 2009 to 2019, the last year of the present survey, the bubble appears to have decreased by a great amount, consequently the mass of the following glacier over the years has particularly decreased.