Visualising World Temperatures

Quentin Laville, Paul Nicolet

Data Visualisation (COM-480)

EPFL

INTRODUCTION

This project aims to leverage the knowledge we acquired in the class to extract visual insights from a temperature dataset with interactive and clear visualisations.

Global warming is a quite popular topic nowadays and more and more people are getting conscious about this phenomena. However, it is not always easy to realise what it implies on our world, and we need to trust media in order to see global warming action. But, with the growing popularity of data in general, it is possible to find some datasets gathering lots of temperatures samples across the world, starting many decades in the past. The goal is then to leverage this data, with the help of data science and data visualisation techniques, and provide to anyone interested in exploring earth temperatures a simple way to extract meaningful insights about climate change from interactive visualisations.

The ideal goal of this project is to be able to visualise real change reflecting global warming. With old enough data, we think that it should be possible to design interesting interactive visualisation able to show the following key points:

- True evolution of temperatures
- Evolution of geographical distribution of temperatures
- Climate change depending on countries and geography
- Correlation between temperatures, population and growth domestic product

Even though we expect to extract these points from the data, we have originally no clue if the dataset is able to provide base content for this facts. Then we need to design the visualisation accordingly and draw conclusions at the end of the process.

DATASETS

We use three datasets in the project: a main one gathering temperatures, and two auxiliary ones to correlate the first with population and growth domestic product data.

Temperature dataset

The dataset is accessible following this link.
This dataset provides many different temperature files (by country, by city, by continent and by area). We used temperatures by country to build the time series and bubble chart, whereas the map is built on top of temperatures by city. Both files theoretically start around 1750 and go until 2013, but most of the data is missing before 1850. Note finally that both contain data for the 1st day of each month.
The map and time series start in 1850, as most of the country were represented. For the bubble, we take only from 1950, because the two following datasets only start in 1950.

Population dataset

The dataset is accessible following this link.
After intense investigation, we couldn't find a correct dataset (even a poor one in fact) with data before 1950. We suppose that gathering population statistics started a short time after the second world war.
Then, the chosen dataset gathers population data for every country. from 1950 to 2015. The document contains multiple columns (estimate, low and high variance, …). We picked the "estimate" column as it represents the data needed for the bubble chart. One should not that the file is really fancy and awful to parse.

Growth Domestic Product dataset

The dataset is accessible following this link.
Once more, it was hard to find old data for growth domestic product, then, exactly like the previous case, the data gathers growth domestic product starting from 1950 and is complete until 2015. The file presents multiple features as well: 2005 international dollars, 2005 USD, etc.... We picked quite arbitrarily the first one, i.e. "2005 international dollars", as its behaviour suited our needs, since it is often used to be able to compare the GDP of multiple countries (one can find the exact definition following this link).

PRE-PROCESSING

Exploratory data analysis

The first step of the project has been to make sure that we were trying to show something real. We then focused on the temperature dataset in order to quickly extract insights about the temperature

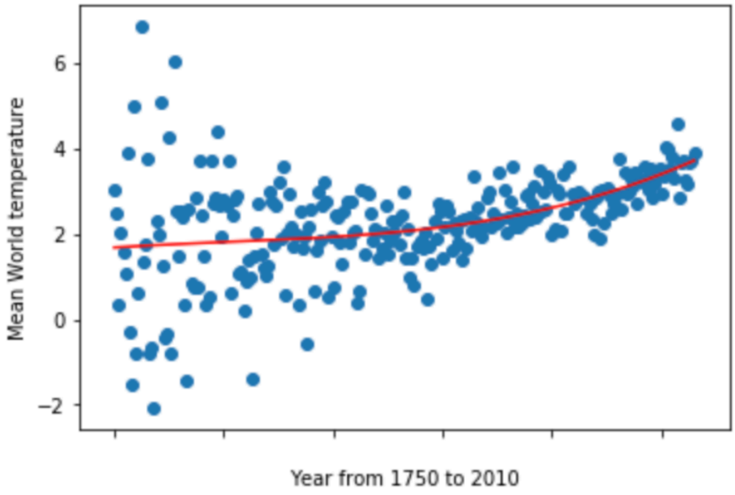


Figure 1: World temperature regression

evolution and see if global warming was reflected by the data we collected. Using regression and basic plots (see Figure 1), we were able to see that the world temperatures are actually getting higher from 1750 to 2010. This step gave us the possible to securely continue and develop our main visualisation, which will be able to extract the same information in different forms and more details. You can find this work in `/preprocessing/exploratory.ipybn`.

Data preparation

Data preparation has also been a major part of the work. It consists in extracting essential information needed for the visualisation to work, from the raw datasets.
Preparation for both the time series and map has been quite straightforward, since it is relatively easy, with adapted tools as `pandas` to filter a dataset and format it in the desired way. You can find this work in the `/preprocessing/map.ipynb` and `/preprocessing/time-series.ipynb` notebooks.

The hard part however has been to build the data for the bubble chart. Given three heterogeneous datasets representing temperatures, population and GDP, we needed to build one only, gathering each of these points for each country, for each year. The challenging part has been to match country names, as different datasets do not necessarily have the exact same representation for a given country. Some of this matching has been done manually. Additionally, some of the datasets were presenting missing data where some others weren't, which has lead multiple times to incomplete formatted datasets, making the visualisation crash. We thus had to be particularly careful and fine tune our data merging process. The corresponding work can be found in `/preprocessing/pop-gdp.ipynb` and `/preprocessing/Merge.java`.

We previously said that the temperature dataset contains the temperature for the first day of each month. At first, we were extracting our data taking into account a particular day and month, namely January 1st for each year. However, we realised there can be high variations from one year another

given a particular month, and this does not accurately represent the overall temperature of the given year. Then we decided to average the temperatures of all the months of a year in order to get a unique value, smoothed over the months.

At the end, we are able to produce three final `json` files in the `/app/data` folder.

DESIGN

Considered visualisations

The first kind of visualisation which came to mind at design time was obviously the time series chart, which is clearly the simplest way to represent time dependent data. We were impressed by the time series following this link for example, which finally appeared to be overkilled for our project.
It also made sense to represent the dataset spatially, and as we are using geolocated data, a world map was an obvious choice. However, adequately representing temperature on this map has been quite a struggle. Our first thought consisted of a heat-map, as we had samples for single points and not areas. It turned out to not be adapted since areas around data points were appearing colder with growing distance, which is clearly not reflecting reality. We finally decided to use the simplest representation consisting of a colored point on each city with a scale from blue to red to show the temperature.
We considered implementing a clustered chart grouping countries suffering from the same temperature changes. By looking at related work, we found an interesting visualisation supported by d3 using zoom techniques with bubbles (link). By discretising temperature variations and putting each country in different bubbles following their temperature variation between 1900 and 2013 we could be able to create a hierarchy between countries. This visualisation type quickly got abandoned though. Indeed we didn't see how to make different levels of bubbles to zoom without creating several dozen of subsets of variation just for the sake of the visualisation. In a first place we thought of doing the first level by continent, then by country to finally separate cities in a given country by variation. The interest of such a hierarchy is however limited. Another idea was to use external data. Thus we gathered GDP and population for each country in order to build another kind of hierarchy. This marks our last trial for this visualisation.
However, the population and GDP datasets led us to our last plot, the bubble chart, inspired by a chart discovered on the d3 website. It allowed us to represent 5 types of data (the change of temperature, GDP, population over several years as well as a temperature variation) and still be understandable by the any user, as somebody who does not want to understand how the data evolute through years can simply look at the color and see that the bubbles go toward the same direction in a pretty intuitive way.

Design decisions

During the whole process, we followed the Tufte's integrity principle shown in class. We were able to achieve the two first one, but we had problem with the third one "*strive for clear, detailed and accurate labels and scale*" for the second and third visualisations. We followed the two core principles of the mark and channels as well, thus we do not use visualisation for non-existing data (expressiveness) and we gave to the most important attribute (the temperature here) the most noticeable channels (effectiveness), which is the color.

We go through the three charts one after the other in the following paragraphs.

For the time series, we remind that we had more than 100 countries. Thus we allow the user to brush and filter the temperature for a country as desired. This way the plot is light in its base state, and the user can itself add and remove complexity at wish with an incremental search allowed by the autocomplete input. Finally the ultimate goal was to show that whatever users choices, using "0 mean" button would show a similar growth for every country based on the similarity concept of perception.

There was no need for any specialised map for our map chart, thus we use the natural projection. The color scale is picked to follow the quantitative representation of temperature. Unfortunately here we are failing at the third principle of Tufte's integrity. The chart was supposed to show a difference in temperature, hence that there is a big variation around 2000 in comparison to 1850. We couldn't express this phenomena without introducing a lie factor, for example by making sure that all dots are blue in 1850 and red in 2000, which we choose to not do. Additionally, to avoid text blobs all over the map, users can hover a point to get precise insight about a specific location, using the brushing concept.

With the bubble chart, we wish to associate each available channel to ensure maximum selectivity over the year and get a clear difference between 1950 and 2010 and an easy associativity between country with high temperature variation. By moving the slider between these two dates, you should directly understand that the color is used as a pre-attentive feature. Nevertheless, we fail at correctly rendering the population scale. In fact, to be able to see most of the country, we found ourselves with a lie factor: most of the countries have the same radius which absolutely doesn't mean they have the same population.

IMPLEMENTATION

Time series

The time series allows the user to add any number of desired country or continent. However by picking a hot country along with a cold one, it is hard to easily get interesting information as the graph will present a large temperature domain on the Y axis (try to add Somalia and Denmark for example). To solve this problem, we add the "0-mean" button helping users to visualise the temperature issue as it enables comparing regression curves. It is also possible to hover on a data

marker or a regression line to get more details. Hovering the legend will highlight a specific data, and clicking on legend elements allow users to hide or show specific countries or continents.

Map

The goal of the temperature map is to geographically show the temperature distribution in the world, and try to highlight the climate change depending on location. It is possible to drag the slider directly to go through the timeline, or just tap the “Play” button and let the slider animate the visualisation automatically year to year every half second. More details about the temperature and city name is available by hovering a particular dot on the map.

Bubble chart

The bubble plot tries to show the relation between the temperature augmentation and the GDP/ population values. The timeline can be manipulated the same way than for the map, i.e. by dragging the cursor or with the “Play” button. One bubble represents one country. Hovering a bubble gives more details about that country for the particular year selected by the slider, in the table above the chart. This values are represented the following way:

- The population is used as radius of the bubble: higher populated countries have a bigger radius.
- The GDP is represented by the X position of the bubble: higher GDP countries are on the right.
- The temperature is represented by the Y position of the bubble: warmer countries are on the top.
- The color is the variation intensity compared to 1950, which is the first available year, from green to yellow to red, red being a high variation (~2.5°).

The final visualisations reflect the original plan, as we essentially kept the same components, except minor variations.
The autocomplete input was not in the original design as we were planning to have a large checkbox area where you could add or remove countries. We currently have more than 100 countries, thus we can easily imagine the problem that occurs with 100 checkboxes and user experience.
The bubble chart and the map were supposed to be visible on a single screen, and share a single slider. First of all, we had viable data from 1850 for the map, but only from 1950 for the bubble chart. Next, the visualisation were too big to fit on a single screen, and the amount of data to display was too big for each step to be handled smoothly by the browser as we were experiencing lags. That is why the two visualisations are finally independent from one another.

A preview of each visualisation is available under the Annex section.

EVALUATION

As a conclusion we didn't learn much from the data, as it's a common fact that world temperature is quickly increasing, especially this last few years. We can do the same comment for the GDP and population, we knew beforehand looking at them that they were growing.
However we can propose a few remarks about the temperature dataset. When we plot the temperature for cities, we only have data for Europe and the East Coast of the US. Even in 1850, there are severe inconsistency in the data for South America and Africa from year to year, since data is very sparse. Some could question the veracity of temperature data before 1950, and indeed the variance observed in Figure 1 is extremely high before 1850. This leads to our decision to remove every data before this date.

If our question was simply: is global warming a real thing? The answer will be given by the first plot and be positive. Can we see a relation between this growth and other data like GDP and population? We could actually. However we didn't process the data to use statistic on them, but only to get a visually credible plot, hence we do not affirm the causality between the behaviour of the different datasets.

In the next few paragraph, we evaluate our visualisations themselves.
The time series gives exactly what was expected. The base state with the world data has a really nice slope with the regression showing the temperature increase. For those who could argue that we lack data before the beginning of the slope, and that the entire system is biased, they can easily add any number of country and check the regression of all these countries to see that the assumption that temperature rose for the world since 1900 is in fact true for every country in the world.

On the other hand, it appeared around the end of the development that the map wasn't as interesting as thought in the first place. Instead of showing an augmentation, it simply shows the temperature distribution over the world. The main problem here is that we didn't implemented a "0 mean" button, thus the range of color is following the domain of possible temperature. With this implementation, a slight difference in temperature as shown in the first plot (around 1° in 60 years) can't be seen on the map. Note that this "0 mean" would have been extremely interesting to show precisely if some locations were more affected than other. Unfortunately, the fact that we come up late with this idea and the amount of data for the map being important (more than 70Mb for 150 years), didn't allow us to add this feature and keep a pleasant experience for the user.

Finally, the bubble chart is visually catchy and gives us the expected result, but in a "weak" way. The color and the movement on the X axis could be linked as we wanted, at the end almost all the bubbles have progressed on the right side of the graph and most of the color are from yellow to red. However, the movement on the Y axis is again penalised by the big range of temperature. We

thought of implementing a “zoom” button which would allow the user to zoom on the Y axis between 20° and 25°. In this state, moving the slider would lead to serious change of bubble on y axis, clearly demonstrating a rise in temperature. The size of the bubbles isn’t as satisfying as we hoped as well, and again it’s due to the great disparity among the country (China vs Qatar for example). This leads to two problems. The main one being that most of the countries look like they have the same size, but it’s just the minimum size. The second one is that we don’t see the growth in the population, except for already big countries (China, India or US), as small ones do not increase enough to reach a significant radius change. There is a tradeoff between using a linear scale (which is the current implementation) and having a lot of small bubbles, and using a log space scale and having a lot of big bubbles, which does not represent better reality and gives a less visually pleasant chart.

PEER EVALUATION

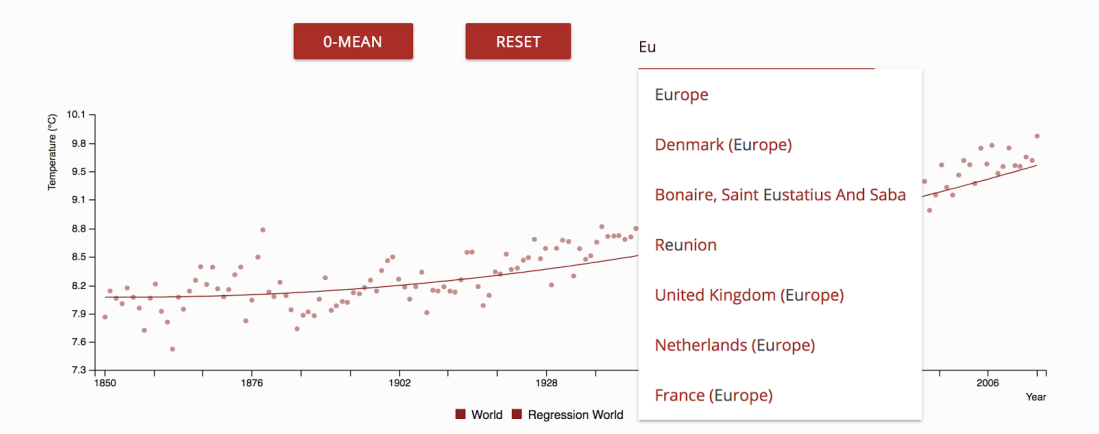
Everything went very smoothly and we both agree to say that we produced the same amount of work. This was a great team project.

CONCLUSION

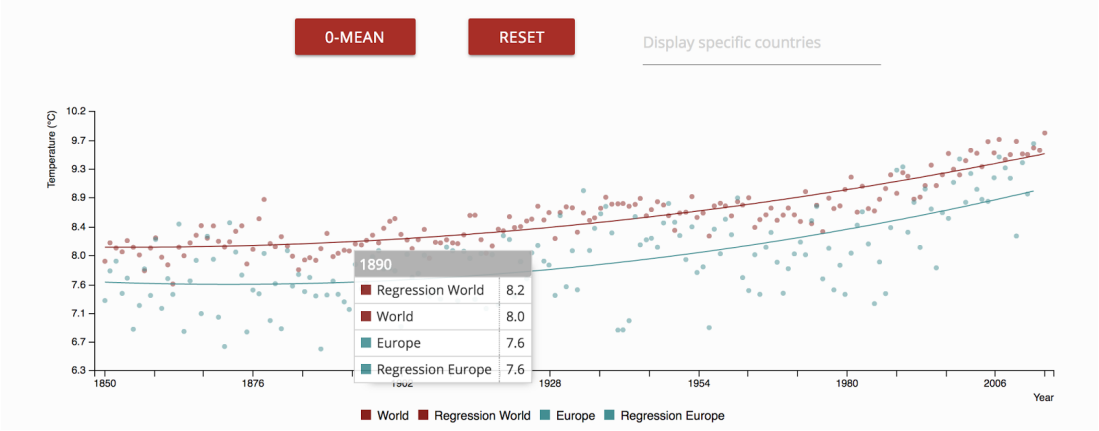
To conclude, we can say that we implemented three nice visualisations which are appealing for users to play with. Even though we do not always succeed to clearly highlight climate change, it is possible to see temperatures changes. But most importantly, this project allowed us to manipulate data in a visual way for the first time, get a first experience with popular frameworks like `d3.js`, and understand the main challenges that one could face when trying to show crucial information using raw data.

ANNEX

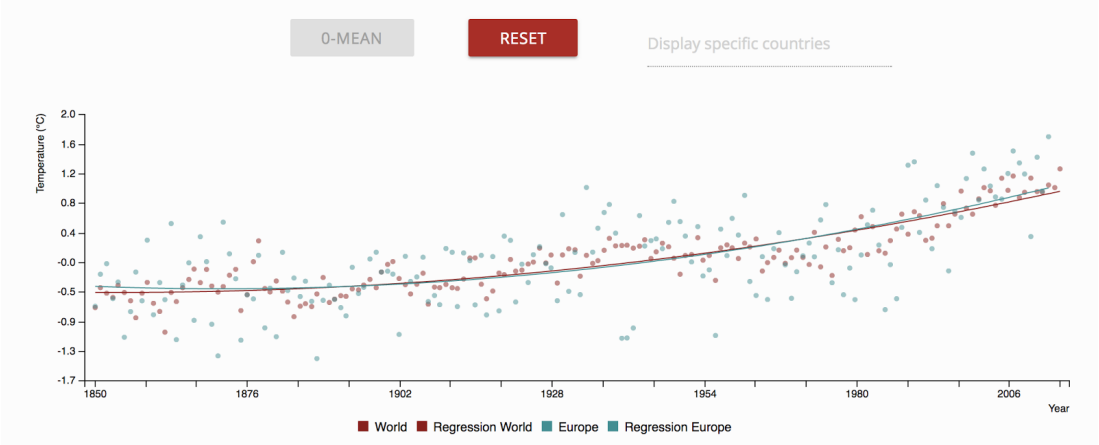
Time series



Annex 1: Time series autocomplete

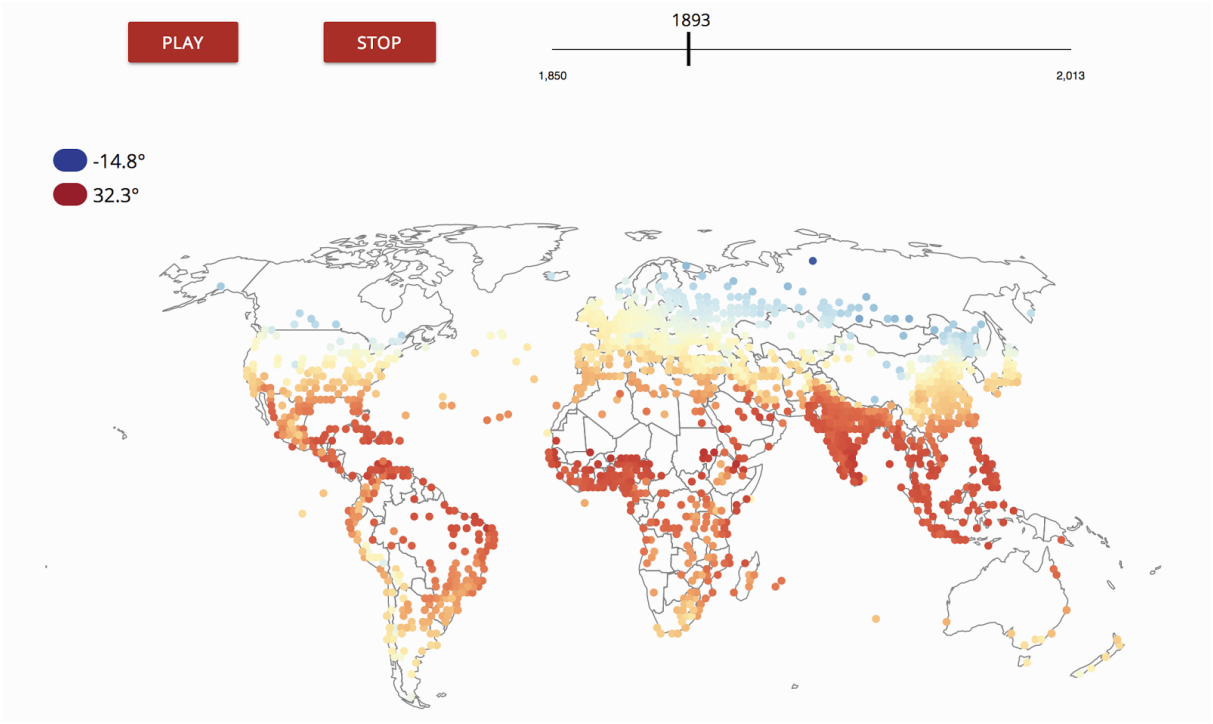


Annex 2: Time series details

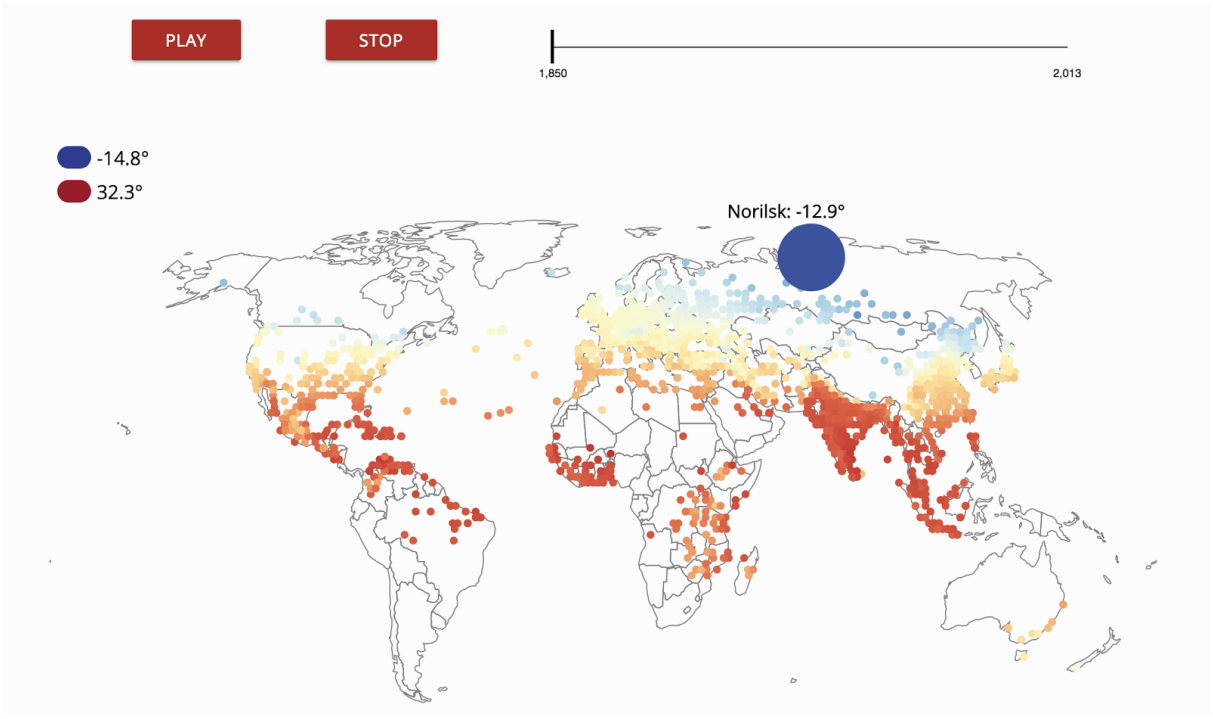


Annex 3: Time series 0-mean

Map

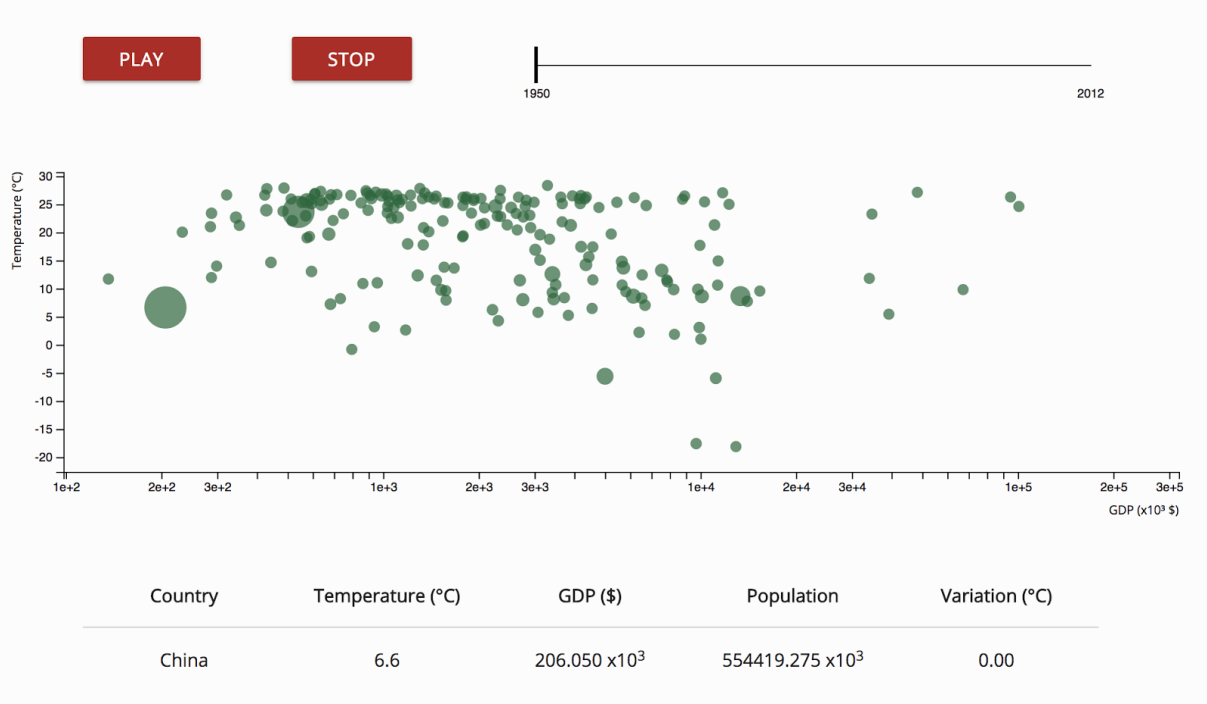


Annex 4: Map



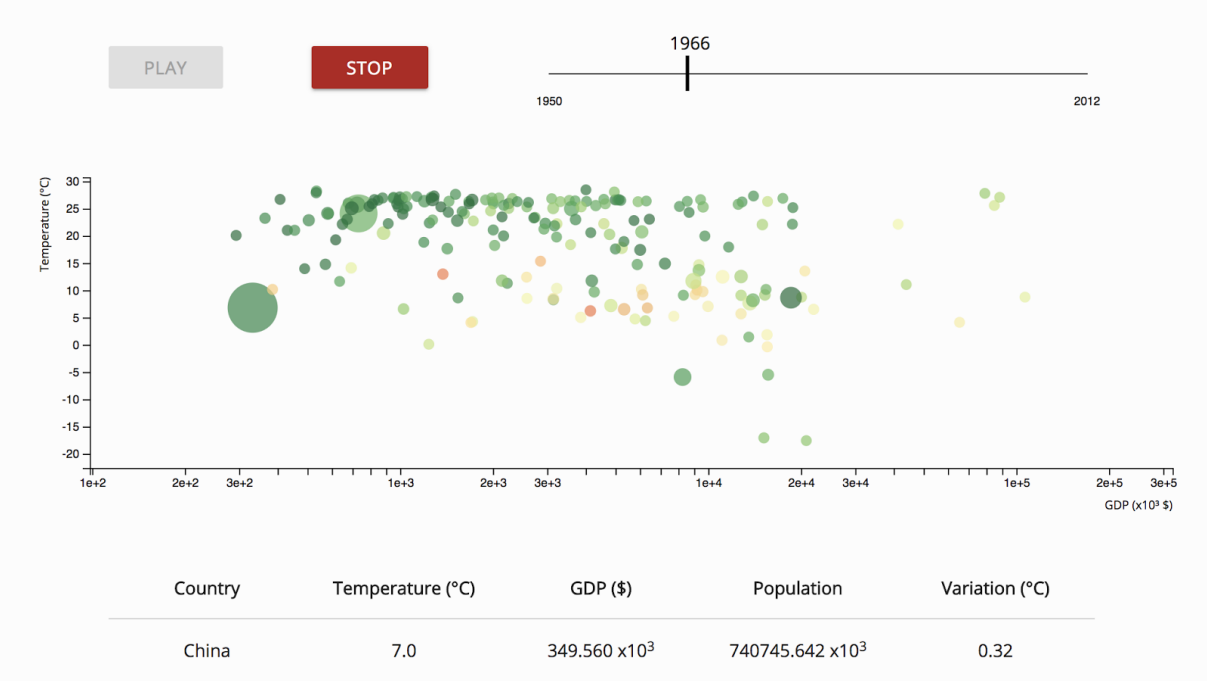
Annex 5: Map detail

Bubble chart



Country	Temperature (°C)	GDP ($)	Population	Variation (°C)
China	6.6	206.050 x10^3	554419.275 x10^3	0.00

Annex 6: Bubble chart



Country	Temperature (°C)	GDP ($)	Population	Variation (°C)
China	7.0	349.560 x10^3	740745.642 x10^3	0.32

Annex 7: Bubble chart variation