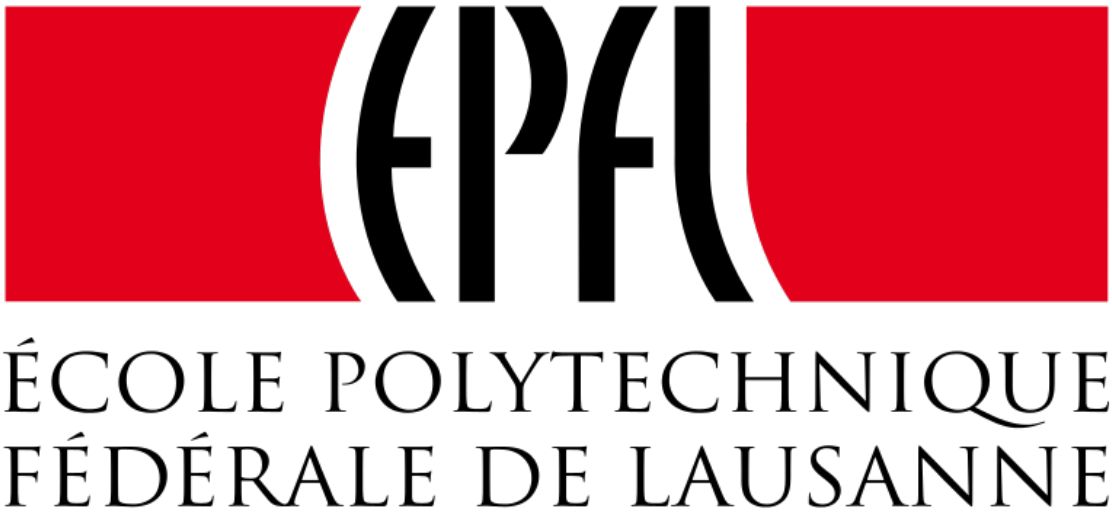
Food Production in 2050

2050 : A new society for a new food production system



COM-480 Data Visualization

Process Book

JOULOT Eliott PELLETIER Kevin PETITPIERRE Maxence

Introduction

This process book will detail the visualization projects from the brainstorming to the technical implementation. It will then cover our different merged ideas, and the thoughts behind each step following the good practices and implementation linked to the Data Visualization. Indeed, this project is a realization aiming the course Data Visualization (COM-480) at EPFL during the 2018 Fall Semester in collaboration with the Stanford University and especially the Natural Capital project. This team is developing practical tools and approaches to account for nature's contributions to society. Their studies will then help important actors within companies, countries and organizations to make smarter decision for a more sustainable future.

Overview

We are currently, in 2018, more than 7.5 billions humans on earth. In 2050, this population number will increase up to more than 10 billions. It will then mean to feed this increased population. The Natural Capital team analyzed and produced data and prediction regarding the food production environment for 2050. Hence, this data presents many attributes, as the calories produced, the temperature, the production and others, each described in the world map with precise point map with the associated attributes. This study followed 5 different behavioral model following the global society organization linked to a greenhouse gas concentration value. This allow the people to see the possible food production results for a sustainable implicated society with a low greenhouse gas concentration, compared to a nationalist society, possibly leading to wars and less concerned about the ecology. Our project will then explore, analyze and visualize this data in a spatial manner. Currently, 40% of the world terrestrial lands are used for agriculture. This could then be interesting to see how the world map could evolve according to these attributes.

Motivation

The subject concerning predictions about food production in 2050 possess obviously an ecological aspect. First, this part of the project really interested us, as we are sensitive to the ecology. Indeed, in another EPFL course named Applied Data Analysis we're doing a project analyzing and predicting ecofriendly trends using the Amazon dataset. Moreover, the fact that our work could help to understand the different tradeoffs and possibilities is a motivational aspect. We wished to learn about the type of prediction scenarios related to the food production, and to apply our knowledge into a map visualization. This was an opportunity to work on a concrete and impactful research.

Target Audience

Our target audience is anyone interested or involved in the ecological aspects. We implemented our visualization to allow non-specific users, as we were, to understand the different scenarios which are properly explained, and the different values shown in the map. We hope to give a better understanding of the scenarios involved, and maybe a self-thought about the behavior of each person to choose the best possible future in 2050.

Dataset

Our given data corresponds to 20 csv files of approximately 1 million samples described by 20 features each. Indeed, there is 5 SSP (Shared Socio Economic Pathways), which are then split in 4 different RCPs (Representative Concentration Pathways). To resume the SSP models will integrate how the society will evolve and behave (Sustainability, Middle of the Road, Regional Rivalry, Inequality and Fossil Fuel Development are the 5 different pathways). The RCP in each model will give the results using a possible value of the concentrated gas emission / radiative forcing (for example if we get to +4.5 W/m^2).

Exploratory Data Analysis - Processing

First, we visualized the data using Python and especially the pandas/numpy modules. We provided in the github our Jupyter Notebook allowing the loading of a data csv, and the total process of cleaning and extraction. First, we focused our efforts to handle the SSP1cc model data. We then visualized the data using some description methods to see the ranges, means, max/min of the values and how the points were organized. For the organization the points are ordered by the latitude feature, and in each discrete latitude values, the corresponding longitude values are ordered.

To have a first overview of the map, we decided to extract and process the "calories" feature with the associated position (latitude/longitude). We then cleaned the new dataframe by removing all the invalid values, and "standardized" the data, by taking the log and a coefficient of the calories, and then round the values in order to have a reduced length of string. Indeed we will then convert it to geojson so the string values have to be optimized.

As it was not really efficient to plot all the data points in a map (1 million), we decided to create data compression. For this we implemented an algorithm that will iterate through the different discrete values of latitude (north to south) and for each latitude value, iterate through the longitude west to east. Then we will cluster every 8 (parameter) points of following longitude to compute the mean of each feature and store it in a single point, if and only if the distance between each point is less than a threshold (parameter). This avoid then to compute the mean of some points starting in the east cost of the USA with some points in the west cost of Europe, leading to an averaged point a bit meaningless and with a position in the ocean. This will then allow having a data size divided by approximately 8 (parameter) with a really small loss of precision. We then tried different values for the cluster size and the distance threshold, to find an appropriate reduced data. Indeed there was a tradeoff between the precision of the data and the size of the data.

Finally we decided to transform the dataframe into a geojson file, because the format is well implemented by our map API (cf Design) and which reduce the size memory.

Using this parameterizable python script we were then be able to load, clean and extract all the different csv models that we have, and select the wanted feature (calories, production, population etc..).

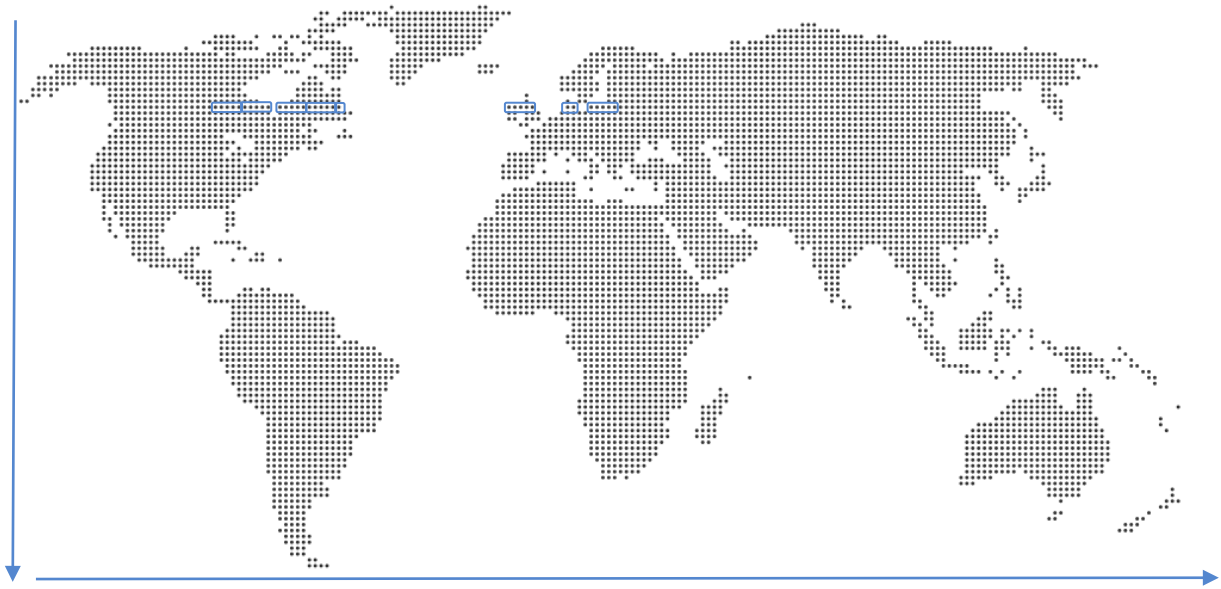


Figure: Data clustering schema: Iterate through the latitude values, then the associated longitude values, and aggregate by computing the mean of the meaningful nearest points

Designs

Visualization:

In this section we provide more insight in the design process. We include the sketches and elaborate on the evolution of the visualization from the idea, to the high level concept on paper to a prototype in code.

Different steps have motivated us from basic comparison visualization to a final approach. As we focus on spatial map comparison through several climate models, we though about visualizing two maps on the screen. The scenarios can be selected from a dashboard menu.

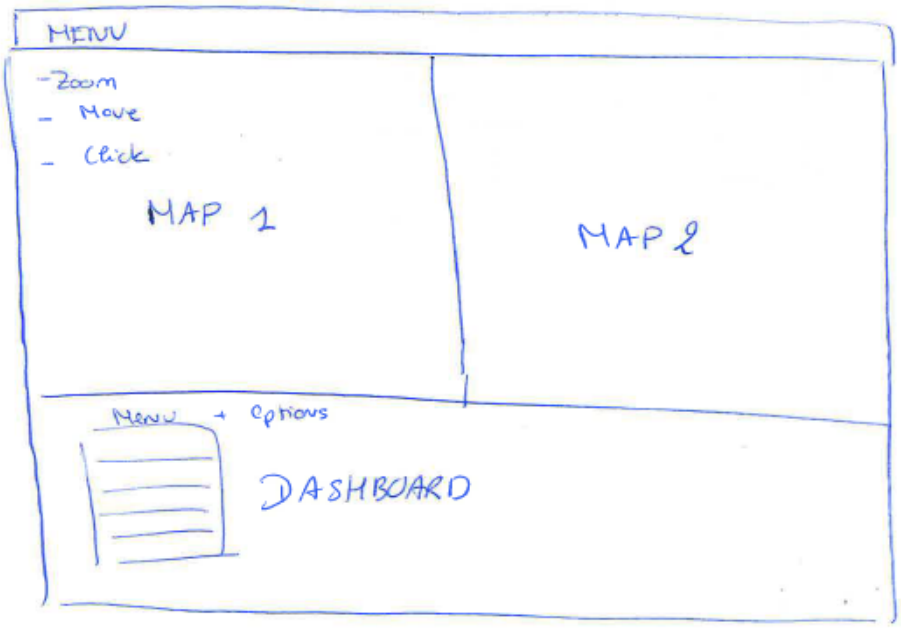


Figure: Initial high level concept

After discussion and reviews (in group and with professors), we deviate from that initial proposal and concede that two maps are not efficient at all and too heavy for a user experience. It is not user-friendly for an audience and is not really relevant. The comparison has to be made by our visualization and not by the user who has to make (or guess) the difference between two regions on two different maps. Thus we will implement a graphical user interface as a menu selection. We will discuss about it later.

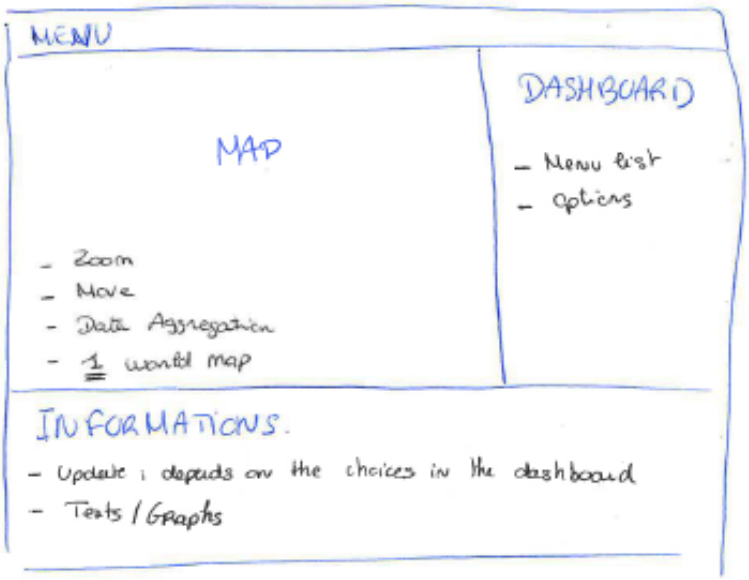


Figure: Final high level concept

Then comes the visualization itself, there are a lot of things to define:

1. Which data to visualize
2. How to visualize the data
3. How the user can interact with it
4. Color scales for the data
5. User guide

1. What to visualize

At first, we selected for all scenarios, four parameters to visualize, such as: calorie production in calories/year, cropland in % of hectare, yields in cal/ha and population count per pixel. Thus we had in total 20 possible map visualization.

When discussing potential use-cases for this visualization, it came to us that this is far away too much. So we reduced them to keep the most differents scenarios (substainability, inequality and fossil development) with only three parameters for each: calorie production, yields and population.

Then we can easily see and compare where we produce food with or without high yields vs where the population is mostly located.

2. How to visualize the data

As explained in the pre processing data, some csv files provided us. Our data consists of one millions points that we reduced by the data processing, and comes the idea of displaying it on a map. At first we query a world JSON file online and display the data on it as points. The result is shown below.

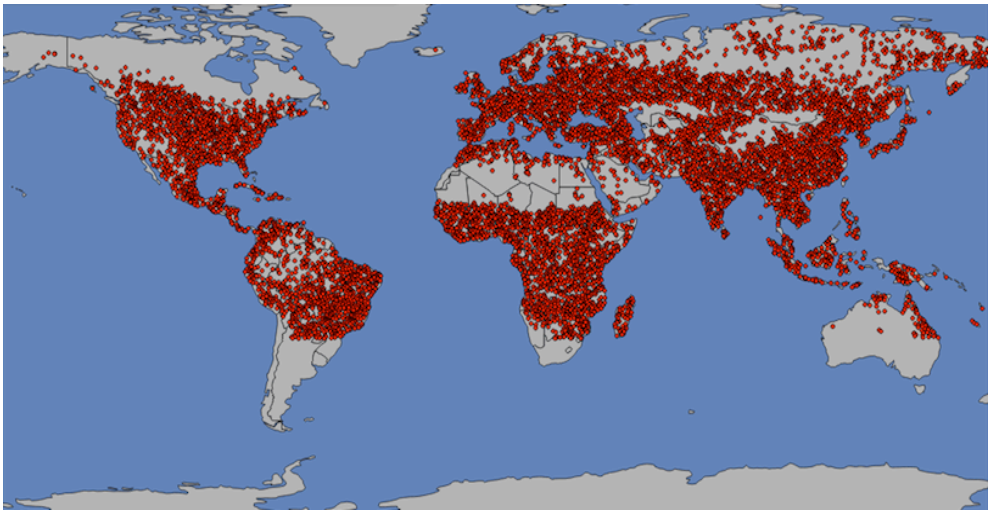


Figure: Initial visualization

The data consists of future climate scenario over the year 2050, a high resolution is not that mandatory as the data might change and variate in the future. So, instead of a high precision at every coordinate point, a global overview by region/country can be more relevant. Thus we decided to go though a heatmap.

Heatmaps are good for showing variance across multiple variables, revealing any patterns, displaying whether any variables are similar to each other, and for detecting if any correlations exist in-between them.

In that case, we are using Mapbox api (free until 50 000 visits/month) with Mercator map. The Mercator map is not most efficient because it distorts the size of objects as the latitude increases from the Equator to the poles but is relevant for worldwide overview.

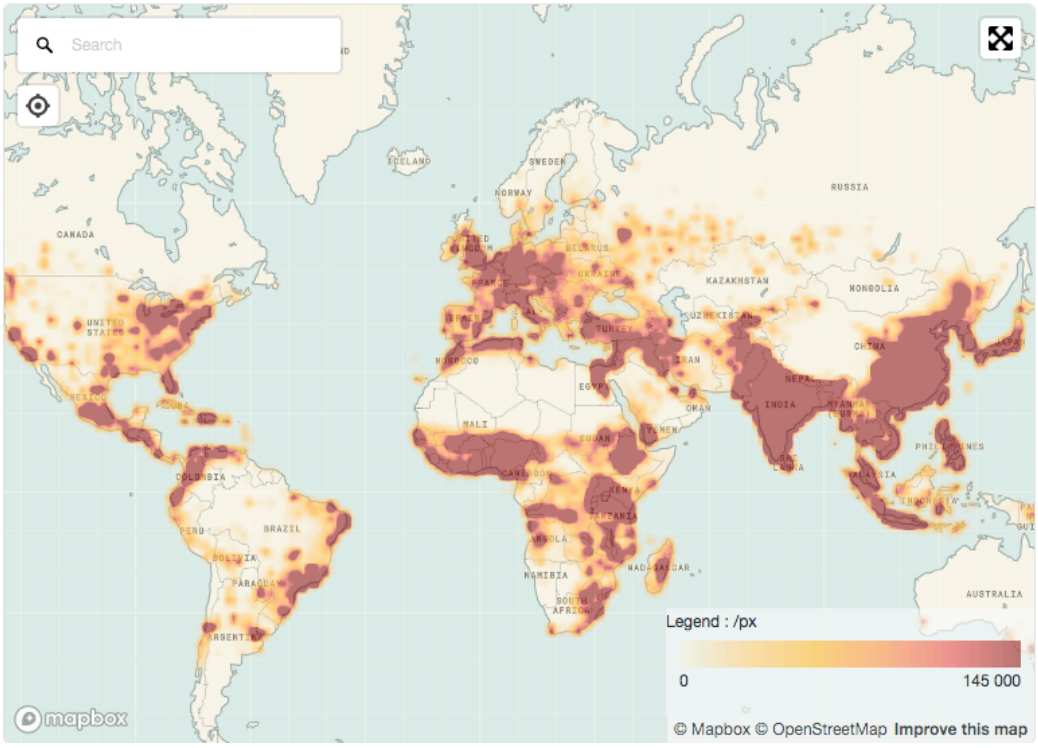


Figure: Final heatmap visualization for population counts

Heatmaps are powerful in giving a really global overview by region but also has some bias. From this point of view, if we take a look at Europe, it seems completely covered by green. But then if we zoom at it, this is what we obtained:

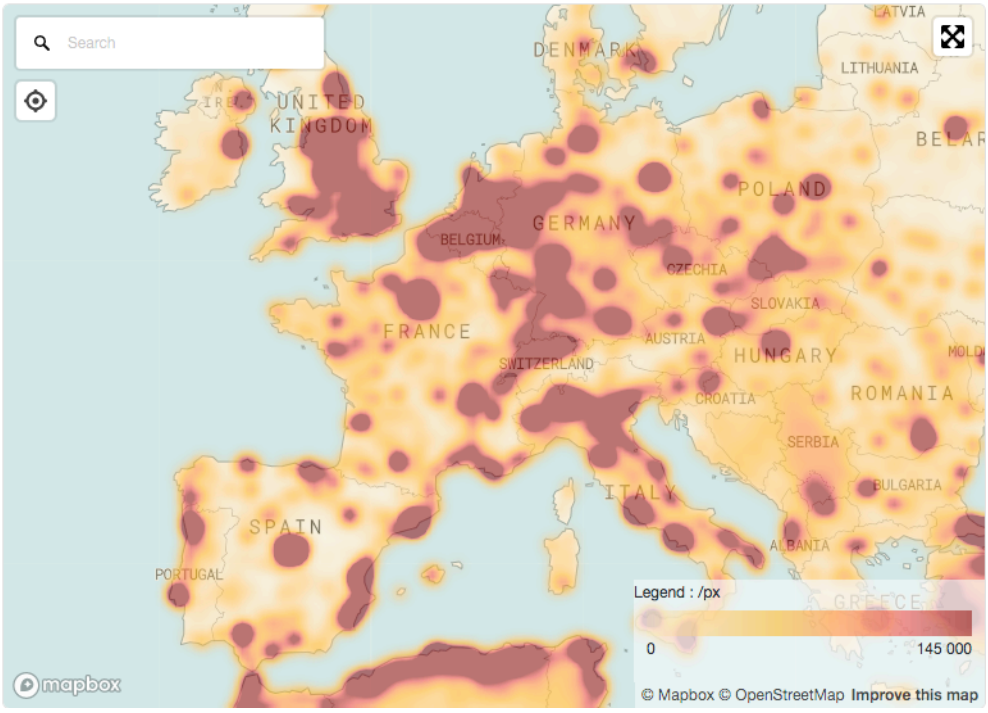


Figure: Europe zoom on population count

Indeed users can have trouble interpreting the map. This is why we implemented a third level of zoom: data points. The goal is to allow user to zoom deeper and then interact with the data.

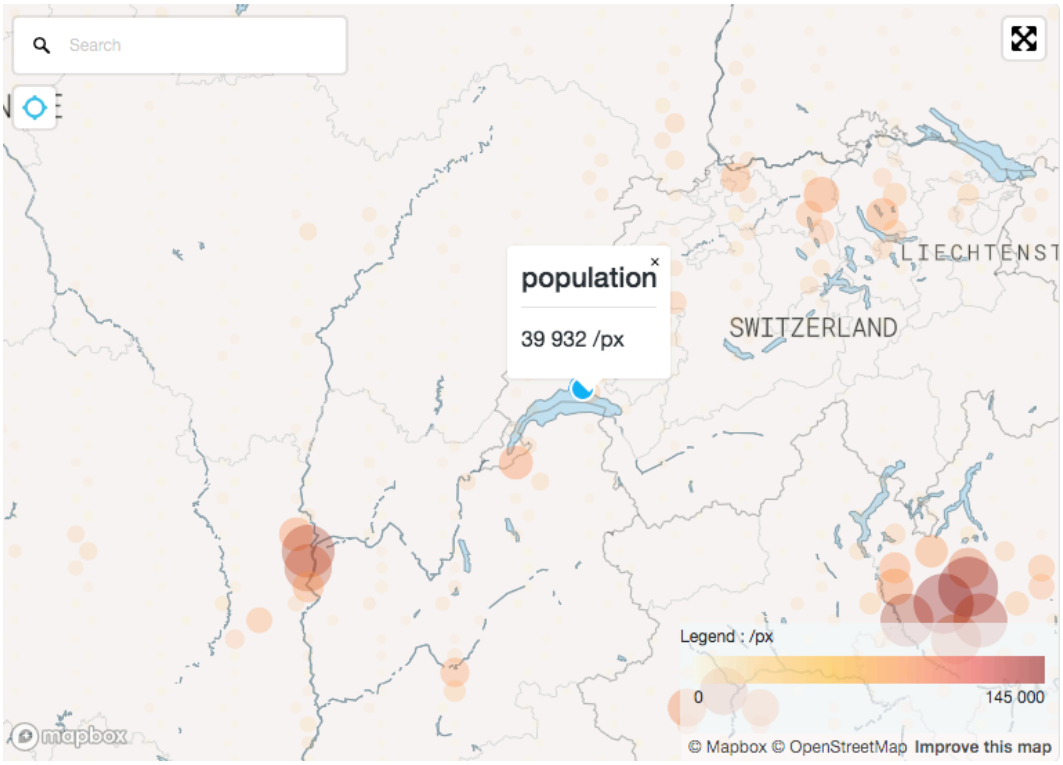


Figure: Lausanne population count

3. User interactions

Mapbox API provides several tools, some of them are fancy but can be useful (search for location, full screen mode, ..). Actually the user can move and zoom on the map. The zoom limit (or level) will be restricted to our precision kept from the data reduction we made and to keep regional overviews. Indeed our data contains predicted scenario not ground truth, this is why we can set a precision threshold.

As we presented earlier, there is a possibility, from a certain zoom level, to click on the circles that appear in order to get the value.

On the right side, there is user interface for scenario and parameters selection. It is not apart from the visualization itself but contributes to the user experience. Initially, with the prototype with five scenarios, the menu selection was implemented like a drop-down menu. By reducing the number to three, therefore we changed it and use a more visual selection menu.

It is made of circle sorted by the increase order of challenges that the scenario has to face in development challenges. The user can select one of them by simply clicking on the matching circle.

Below the map will be placed a container/placeholder in order to display more information about the scenario that the user has chosen. We do not forget that our target audiance can be either researchers (Stanford users for example) or common users that do not have strong knowledge in the data visualized.

4. Color scales for the data

The color scale is defined by the mapbox's heatmap API. This API automatically normalized the data from 0 to 1 and then we just have to specify the range and colors for each level. As we are treating calorie production and yields, the evidence color scale seems to be in green. But for population count, green is not what is mainly use and we moved to red.

The challenge to face was to adapt the color scale to the data distribution. Indeed, mapbox API provides a linear color scale for heatmaps while our data do not follow it. At the present type the data follow any type of specific or known distribution.

5. Popup: How to use

In order to let the users clearly understand the several parameters and visualization, we decided to create a “How to use” pop-up which will appear once per session to guide the user. Additionally, this help is accessible at anytime from the website, by clicking the dedicated ‘Help’ button. This popup contains explanations of the goals of the visualization and introducing the different models and features.

After reading this paragraphs, we let the user the choice to read a bit further about the models and the features in our dedicated page, or to proceed with the visualization. Finally, to help the user with the implementation of our maps, we provided some quick steps to guide him through the different parameter’s choices, and the possibilities with the map. In addition to the written explained steps, we provided two animated images to have an overview of the parameters, and the map results.

Implementation

In this section we will further discuss implementational details, based on the final implementation model in the previous section.
We will cover technical details: libraries, the functionalities and reasoning behind it, as well as the overview and uses of the visualization.

Technical details

To create the map interface, we used the mapbox API. We already talked about it a little earlier, but we will now go into a little more detail to understand the features of this API we used, and how we adapted our data to it.

First of all, it is actually a free API that allows to integrate a map very quickly to any website. Its first characteristic is the great possibility of customization that it offers, by allowing to choose each of the elements to be displayed, the colors, etc. We have chosen a very simple and minimalist theme, so that our map is not overloaded with information, and the focus is on the data.

With this map, mapbox allows us to add some interesting features. We therefore added those that we thought were coherent and useful for the visualization we wanted. We have thus added the possibility to view the map in full screen, and to search for a specific location on the map, or finally to locate yourself to directly see the data associated with your geographical position.

Then we started to add the visualization of our data to the map. To do this, we have linked a geojson file containing our data. It will be loaded when the map is created, and directly interpreted by the library to place the points in our map.

Mapbox organizes its view into layers. Thus, adding a data visualization is actually like adding a layer. We started by adding all the points, represented by circles of which the color varies according to the value represented.

For each point, our data file contains several pieces of information: population, production, etc. For each of them, we can then create a different layer that is displayed or hidden depending on the model selected by the user.

This leads to one of our biggest problem: the size of our data files. Indeed, even after data processing and reduction, each file is around 40MB. Each time the user changes the scenario, we have the remove the previous source and bind the new file as the new map's source, and it takes time. The solution would have been to store those data in the user's session storage. But we can not do that since the user session or local storage are limited to 5MB.

This is one of the main drawbacks of our visualization. We did not found a proper solution to solve that. We though about reducing even more the number of data points and thus the size of the file, but if we compute the mean of too much points, there are no more differences between scenarios and consequently no more comparison.
Another way around to proceed was to compute the differences between scenarios and parameters. But this would not actually reduce the amount of data.

Asynchronous changes

There are several asynchronous tasks that we had to handle for the visualization. The first one is manage by mapbox API that corresponds to all user interactions with the heatmap and datapoints.

The others are the scenario and parameters selection. Because changing the scenario makes the map's source update while changing parameters only changes layers. The graph below sums up the handler events.

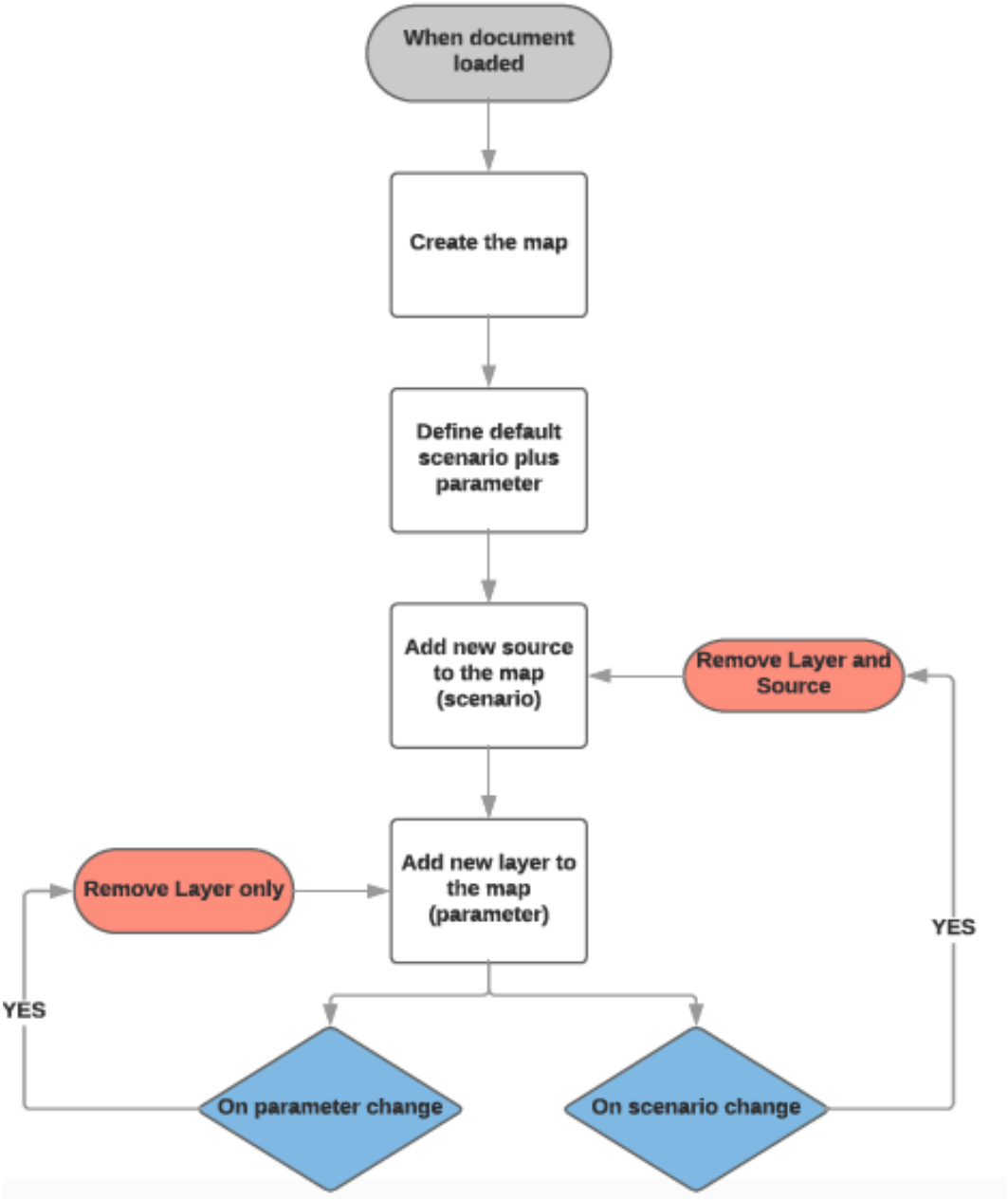


Figure: State diagram of the map update

The scenario selection is handle by d3.js when the user clicks on circles while the parameter is handle by an event handler when a new checkbox is checked.

Evaluation

In this section we discuss the results we have obtained, the data insights we have achieved, pitfalls to look for as well as the prospect for the future improvements and work.

More details about practical aspects are in the code and comments to point out to more practical details and insights of implementation in Javascript or D3.

Data insights

We started this project without having strong knowledge in the field of shared socio economic pathway but we were very interested in sustainability development.

We had some difficulties to go deep in the data and really understand them. As a matter of fact there are a lot of scenarios, with a lot of features, not all of them are accessible and seem very specific for this domain and restricted to a limited audience. But that also allowed us to easily choose among them which ones to visualize.

The pre processing step was one of the main points of the data visualization's insights. After some exchanges with Stanford coordinator, we learned that those data are one of the main innovative points of this study because it addresses this question on a global scale, at a very fine resolution. And so it would be much better to show the production results at the highest possible resolution.

Consequently our problematic quickly became how to hand this high resolution at a world scale level. And we faced the tradeoff of resolution vs global scale due to the size of the data.

As we can see on the visualization, the comparison between scenarios is almost impossible from a global worldview and the user has to go to deeper in details to get and observes differences.

Finally a non-scientist user can experiences the visualization with only one scenario and explore the differences of repartition between calorie production, yields and population without trouble between scenario loading.

Whereas a scientist/researcher in this fields can go for some analysis insights between scenarios but will have its user experience reduced by the loading timeouts.

We also learned by the end of the project that the population data was containing some errors and that the units were not totally corrects. However we still decided to keep it to visualize because it gives a quite good global insight even if the unit per pixel has no real meaning.

Future work

One the point that can be really improved in the future is the global vs local scale. The choice to keep global view with fine resolution can be difficult to realize.

For future work, we recommend a new strategy which is to do as we did but on a country scale. Thus the data can be more compressed, less big and the compromise between global scale and fine resolution can be more concret.

Some improvements could be done on the visualization side, improving the heatmap color scale to each features' distributions.

With further insights from experts and scientists, we could possibly create an even better and insightful visualization that would be not only visually interesting to nonexperts, but potentially provide deeper and practical insight into the importance of those future measurement and how to really apprehend them.

Peer assessment

In its extensive overview, the group went very well over the project. Every one was willful is progress and improve the project. The team meetings

often ended on agreement for the next steps. All ideas were always taken into consideration until we discarded them.

The contribution was equally spread, when one of the members was working on data processing, the others were carrying out on the visualization design and implementation, or in the other way around.

There was a positive and respectful atmosphere and behaviors between every member from the beginning to the end of the project.