VRIJE
UNIVERSITEIT
BRUSSEL

Information visualization

DATASETS VISUALIZATION

Weather / Climate

Maximilien Romain, Steve Homer, Fabian Perez,
Abdelrahman Sanjekdar

Sciences and Bioengineering Sciences

Contents

1 Introduction

According to scientists, the earth has existed for about 5.55 billion years, while the first homo sapiens appeared about 200,000 years ago. Man is therefore only a speck of dust in relation to the total age of the planet.

The earth's environment is subject to a very large number of variables, the atmosphere, global temperature, sea level are all consequences of these different variables. It is impossible to see completely what is at the cause of what consequence. The earth has ice ages of different duration, meaning that the earth is always heated and cooled in turn.

It is often said that man has had a huge impact on his environment and the earth itself. It is quite reasonable to attribute many of its drivers of change to men, but can we attribute everything to them? Is global warming caused entirely by man, is it only helped by its pollution or does it have no impact on this factor?

2 Research Question

The aim of this visualization is to allow for visual exploration of historical data relating to climate change and a variety of different human-related variables that may hypothetically have a harmful or mitigating effect. Since it is difficult to understand long-term changes in a single dataset by inspection, let alone two, we are interested pairwise comparison of variables related to climate change. For instance, we might be interested in how an increase in population from 1900 to 1950 is correlated with an increase in average annual temperature during that same period.

Beyond that, we are interested in determining what regions are most affected by climate change and what regions are most responsible for it. Since the climate is a complex global system, it is very possible that regional differences will have a global effect, or that a global effect may be amplified, or absent, in a certain region.

Therefore, the different research questions we would like to answer are:

- What is the evolution of climate change and effective variables on Earth?
- What pairs of variables are correlated in this regard over time?
- What regions are most affecting and most affected over time?

2.1 Dataset Selection

We chose to work with multiple different datasets in our visualization because we wanted to examine different relationships between variables. Since it was impossible to find all the different variables we were interested in the same dataset, we utilized a set of them. Most of the datasets came from the Awesome Public

Datasets Github repository, though others were found elsewhere. These datasets were all found to have different formats, and varying granularity in both geographical region and time, so there was an amount of preprocessing done to synchronize the regions, times, and formatting of each dataset.

We began with datasets regarding average temperature, forest coverage, and CO2 emissions, which are meant to form the core of the analysis since temperature is the main measure of climate change, CO2 the main driver of that change, and forest coverage, one of the main inhibitors. Later, we added datasets regarding population and other emissions to enrich the exploration the user is able to accomplish. Besides the preprocessing step, the application is able to consume any sort of geographically-tied, numerical data which makes further expansion in the future simple.

3 Application Components

The application seeks to visualize pairs of variables to investigate trends and correlations. Since we are interested in pairwise comparison between variables, we will refer to a primary and secondary variable to distinguish them.

The application consists of three main visualization elements:

- **Choropleth map:** geographical overview of the primary variable.
- **Time series plot:** temporal overview of both variables.
- **Parallel bar chart:** ordinal overview of both variables.

It is also composed of two other functional elements used for the interaction:

- **Dataset Selectors:** selects the primary or secondary variable.
- **Time Range Slider:** selects the current time range.

3.0.1 Choropleth Map

A chloropleth map, Figure 1, is a visualization tool commonly used when investigating geographical data since it allows for intuitive qualitative overview of the given data through the use of color. Since we're interested in a global phenomenon, we employed a global map with a granularity at the country-level. The map is also interactive, in that the user is able to zoom and pan as they are accustomed on digital maps. In the lower left corner, a gradient color legend is shown to inform the user as to the range of values they are seeing. It should be noted that the map only summarizes the data for the primary dataset, leaving comparison with the secondary dataset to the other visualizations. The user is able to deselect and zoom out by clicking the globe in the upper left corner of the map.

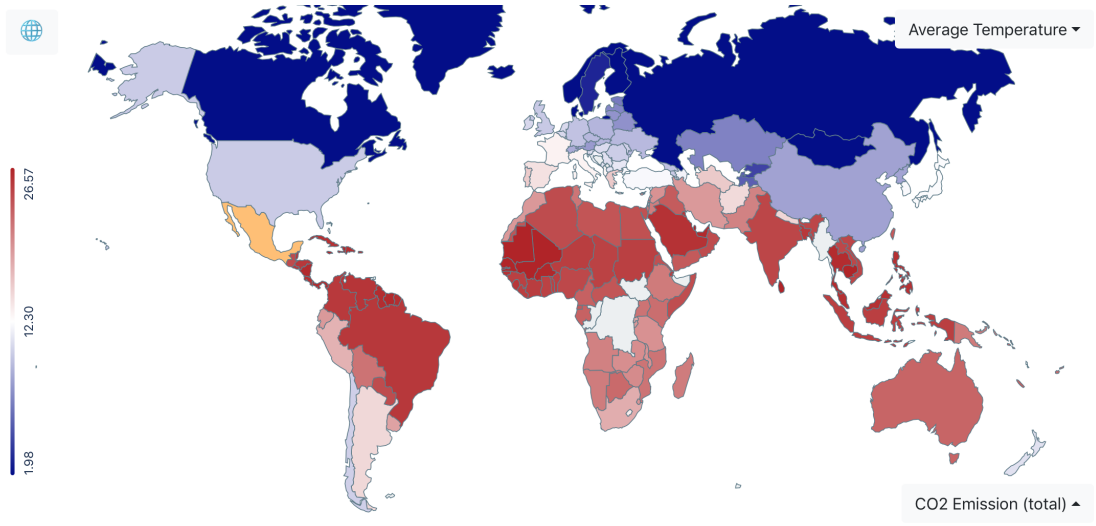

Figure 1: Map View

The user is able to hover over different regions, which will update the other visualizations described below, including the name of the region in the upper left corner of the application. The user is able to select regions for further investigation, and deselect by clicking again on a selected region.

Since different regions are color-coded according to whether they have high or low values for the given primary dataset, it allows the user to see at a glance, which regions are most affected or most affecting according to the current variable. If the user finds a specific region they are interested in, they can investigate it in more detail in the time series plot.

3.0.2 Time series plot

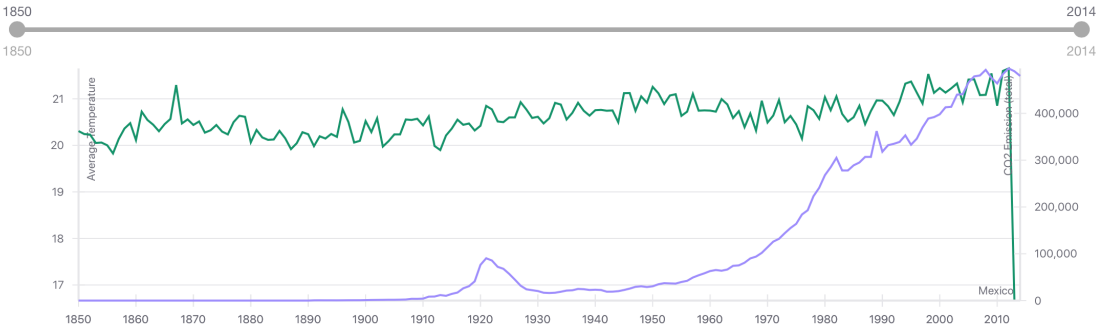

Figure 2: Timeline View

The time series plot, Figure 2, consists of two line charts of different scales. The left y-axis plots the primary dataset, whereas the right y-axis plots the secondary dataset, both over the same time range. The idea is to show different datasets in the same chart to examine any qualitative correlation between the two variables.

Each line represent a dataset that is being selected from the dataset selector buttons on the map, discussed below, and each one of them has a color to distinguish between them.

Since the chloropleth map gives a geographical overview of the data, the time-series plot gives a temporal overview. Examining this portion of the application allows the user to determine how the variables behave over a given time span. It aids in our analysis because it can illuminate distinct trends in single datasets while also giving qualitative trend information in the comparison of two.

3.0.3 Parallel Bar Chart

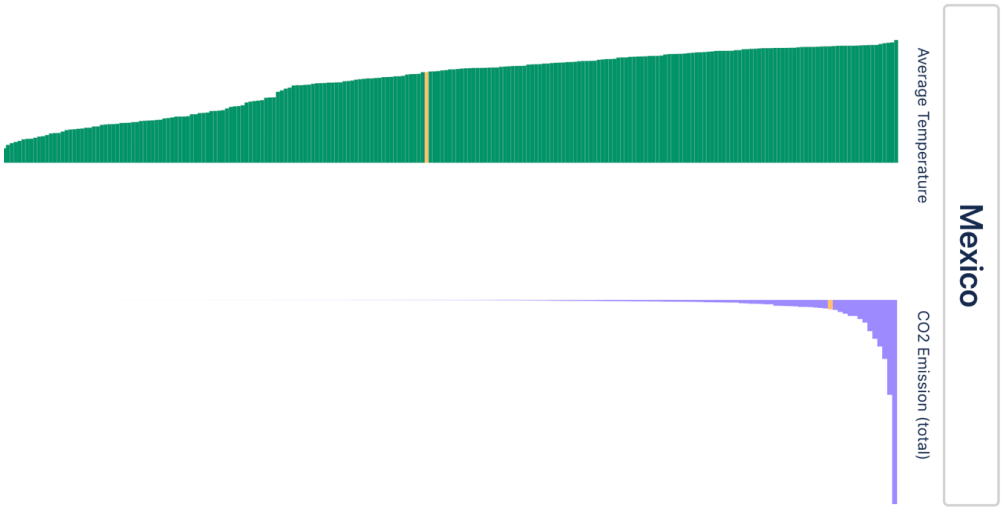

Figure 3: Parallel Bar Chart

The parallel bar charts, Figure 3, consists of two vertical bar charts: the left for the primary and the right for the secondary, to keep consistency with the time-series plot. Each bar chart is sorted in descending order of regions by their respective values. This visualization is used in conjunction with the chloropleth map to add another dimension to comparisons between regions. Since the bar charts are ordered, the user is able to examine how one country stacks up against another, and how top-heavy the distribution for the given dataset is, highlighting a disproportionality for certain regions.

Since both datasets are ordered in parallel, by hovering over one of the bar charts, the user is able to see where the hovered region falls in the other bar chart immediately. The goal is to be able to examine two datasets simultaneously, so the user needs to be able to see where a given region lies in relation to others, which is what the synchronized highlight accomplishes.

3.1 Dataset Selectors

The dataset selector is composed of two sub-menus used to allow users to choose the variables they want to compare using the different views of the application. Thus the user has the possibility to choose two types of data. The upper, primary dataset will be displayed on the choropleth map, and on the left side of the parallel bar chart and time-series plot. The lower, secondary dataset will only be displayed on the timeline and the parallel bar charts, in order to examine possible correlations with the primary dataset. Users are free to choose the dataset they want to display in order to complete their analyses.

3.2 Time Range Slider

The time range slider is a multi-node slider that is located above the time-series plot and allows the user to choose the desired time period for their analyses. By allowing the user to focus on a period of time that they find more interesting, they eliminate superfluous data that is unnecessary for their analyses. The views are updated instantaneously when the time range on the slider is changed. The chloropleth map and parallel bar chart will summarize the data by averaging over the selected time range, whereas the time-series plot will display the data as is.

The slider has two ends that indicate the beginning and end of the time range, which can be shortened or lengthened to the user's needs. This time range can also be scrubbed over the full range, updating the other views and allow the user to see how a given variable behaves dynamically in the chloropleth map or parallel bar chart. This aids in the analysis of how different variables affect different regions over time, as well as examining their distributions.

4 Application Functionality

4.1 Main functionalities

The application consists of thee main functionalities:

- **Region Selection:** select a region on the map.
- **Region Extension:** investigate the summary of a region with more detail in the time-series plot.
- **Region Comparison:** compare a region's performance to other regions.

4.1.1 Selection

Though the choropleth map is used to visualize the dataset at a glance, it also serves as a region selector. When a region is selected, it will reflect on the other components:

- Selected country will be highlighted.
- Time series plot will show more details of the selected country.
- The name of the selected country will be displayed.
- The selected country will be highlighted in the parallel bar charts.

4.1.2 Extension

The time series plot expands on the details that is being shown in the map. In the map we are showing the mean value of the given time range, but when we select or hover on a region in the map, we can see the progression of the data over the years. In addition, we are showing the two datasets selected in the same chart using different lines to allow for comparison and correlation analysis.

4.1.3 Comparison

The parallel bar chart present all the countries in an descending order by their values. We can hover the mouse over a bar element in the chart so we could see which country it is, displayed above in the upper left. It will also highlight in the map and display in time series.

4.2 Layout and Color Scheme

4.2.1 Layout

The application is split into three panels, a panel for each element of the application.

<table>
<tr><td rowspan="2">parallel bar charts</td><td>choropleth map</td></tr>
<tr><td>time series plot</td></tr>
</table>

Table 1: Application layout

The choropleth map takes the majority of the view because it is the main visualization where you can select a country and check the mean value of the dataset selected for all the countries in a glance. The time series plot is located beneath the choropleth map with a relatively big size of the application view. The parallel bar charts are located on the left side of application's view.

This layout was chosen to not only work with the necessary form factors of the chosen visualization – e.g. the time-series naturally wants to be wide, but not tall – but to take advantage of how the user's eye travels across the screen from top-left to bottom-right. Instead of placing the menus in the prime top-left real

estate, we placed them in neutral colors over the map where they are unintrusive, but easily available.

In addition, the legend and time sliders serve as natural boundaries between the different panes, allowing to save on the data-ink ratio while maintaining the semantic separation between the visualizations.

4.2.2 Color scheme

First, the time-series plot shows two different datasets. To distinguish between them, color theory indicates that for categorical data, a difference in hue properties should be used. Since because the two different datasets have no natural ordering, they are categorized using dark green for the first dataset and purple for the second, as can be seen below in Figure 4. These were chosen according for being not only colorblind-friendly, but also being "optimally distinct" in terms of color perception.

Figure 4: Time series palette

Second, in the same way as the parallel bar chart, the time series plot uses hue properties to select the colors for the two datasets that are shown. This component use the same colors that in Figure 4 for the bars to represent each of the two datasets. A tertiary color is employed in hover and selection that is distinct from the primary and secondary colors, while still having a sense of being a highlight, shown in Figure 5

The choropleth also employs the same orange color shown in Figure 5 to select and hover region in the map to maintain consistency with the highlighting in the parallel bar chart.

Figure 5: Parallel bar charts palette

Third, the choropleth map employs a different palette of colors depending on the datasets that the application is currently visualizing. This difference in color palette highlights the differences between the datasets, and gives a visual cue as to what the data is. For instance, forest coverage uses green and CO2 emissions uses red, since they are very different quantities. In this component, the data that is going to be displayed has order and is continuous, meaning varying the hue

is improper. For each one of the datasets and explanation will be given for the chosen colorblind-friendly palette:

- **Temperature:** In this dataset, diverging color schemes were used, signifying two kinds of characteristics to be displayed: hot and cold. These correspond to two different hues, red and blue, as the Figure 6 shows. Then, for the temperature levels the more extreme values have a greater saturation and less extreme values have less saturation the way that sequential color schemes indicates. Red and blue were chosen because in western cultures these colors represent hot and cold respectively.

Figure 6: Map temperature palette

- **Forest Coverage:** Sequential color schemes were used, because the data has an order and is just one category. The color that was chosen here was the green, because green in western cultures represents life and nature. For high forest coverage the green has a greater saturation and for lower levels the color has less saturation, as is show in Figure 7.

Figure 7: Green sequential color schemes

- **CO2 emission:** For similar reasons as the previous dataset, sequential color schemes were used. The color that is used to represent this dataset the red, because CO2 is a colorless gas, but many times is associated with danger, which is related to red. In Figure 8 it can be seen that for highest values of CO2 emissions a greater saturation is used, and less saturation for the lower values.

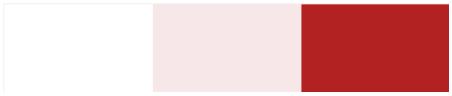

Figure 8: Red sequential color schemes

- **Sulfur emission:** In the same manner here, sequential color schemes were employed. Because sulfur is yellow, this color was chosen to represent it, and as previously seen, greater saturation for high values of sulfur emission and for lowest values less saturation, as is show in Figure 9.

Figure 9: Map Sulfur emission palette

- **Urban population:** Also uses sequential color schemes, in this case with the grey color, because urbanism is associated with concrete, and concrete has a grey color, for high percentages of urban population is used a greater saturation, as it can be see in Figure 10.

Figure 10: Grey sequential color schemes

- **Population:** For the population dataset the green color was selected, because of green signifying life, here also sequential color schemes are used, with high values of population with more saturation as Figure 7 shows.
- **Coal, Oil, and Gas Consumption:** This dataset also uses sequential color schemes, the color for this one is yellow, because energy in western societies is represented by it, for higher coal consumption values greater saturation of yellow is used, as it can be seen in Figure 9.

4.3 Interaction

Interaction has been used here for the two reasons that the theory indicates, for exploring and managing information, there are many ways in which you can interact with the app, that are explained below.

4.3.1 Connecting

In order to create a cohesive, integrated visualization application, the different components should interact and reinforce the information garnered from the others. One way to accomplish this is to use brushing and linking to indicate corresponding quantities on different visualizations. Selecting is obviously important to allow the user to investigate further something of interest. Filtering allows the user to focus in some particular portion of data. In this application, each component is dynamically changed according to the other components, so that all three visualizations are indicating the same quantity, but in different modalities.

These kinds of interaction are important, because it makes it easier for the user to interact with the different graphs and provides more detailed information about a region or a range of dates.

Three methods were employed in linking the components:

- **Brushing:** The brushing can be used in the choropleth map and in the parallel bar charts components. By brushing a region or a bar, the corresponding element in the other is highlighted. This allows the user to simultaneously see where a given region is located geographically and stacks up against other regions. If there is no region currently selected, hovering will also update the time-series plot. This allows the user to quickly scan through different regions.

- **Selecting:** Selecting is similar in functionality to brushing, only it maintains the highlight of the bar chart and map, as well as the time-series plot. This is necessary for further investigation in the time series plot, as solely brushing would not allow. It should be noted that if no region is currently selected on the map, the global mean of the given dataset is defined as selected.

- **Filtering:** Filtering is used in information visualization to selectively hide data which is not useful for certain moments. To filter the datasets with respect of time, the time slider allows the user to filter the data for different ranges. The user can lengthen or shorten the time range from either end, or scrub a smaller time range over the full range, allowing for a dynamic view of the data since it updates instantaneously.

4.3.2 Zoom and Pan

In order to explore with these visualizations, certain affordances are expected of them. Specifically, digital maps are not only assumed to be clickable, and for that click to result in a highlight, but that the user should be able to zoom and pan on the map. These were implemented so that the user can investigate more closely some geographical regions versus others in a more "natural" manner.

5 Conclusion

We have implemented an application that gives the ability to the use to observe and analyze data related to climate and humanity. This is achieved through several integrated visualizations that make up the application and a number of datasets that the user has the ability to choose between. The choropleth map, which is the main view of the application, helps the user to take a quick glance of a selected dataset on all the countries. The map also serves a country selector so we could give more information in the time-series plot where we expand the data that is being shown in the map over the years available. Finally, the parallel bar charts that serves as an auxiliary view show the rank of a country with respect to a different country.

5.1 Future Work

The main purpose of the application was to explore a pair of datasets related to climate change in tandem. We hoped that through this analysis, the user would be able to investigate interesting correlations discovered between datasets. However, it might not be the case that all the datasets we've chosen are interesting in this regard. Therefore, the addition of more interesting datasets would increase the value of this integrated visualization.

Crucially, the application is agnostic about what sort of data it is seeing, as long as it is in the proper format, meaning that dataset extension is somewhat trivial for the application. Some datasets we would have liked to explore would be about weather anomalies and catastrophes for each region over time. This would yield interesting insights into climate change with the use of this solution.

5.2 Demonstration

A video demonstration of the main components and functionality in action can be seen at :

https://www.youtube.com/watch?v=gYr9FYqgkFIfeature=youtu.be