A Socio-Economic Assessment of Climate Vulnerability

"The Risk Is Real"

Abhinandan Dubey
Stony Brook University
New York, USA

Teja Madiraju
Stony Brook University
New York, USA

April 2017

Abstract

Climate change poses real and present danger to the future of the world. Also at stake is practically every major life form's existence that resides on this planet. Primarily, addressing climate change requires developing awareness in peoples' minds. We devised a time-series model on the socio-economic factors contributing to climate change, and produced several visualizations which establish relations between several socio-economic factors. More generally, we have created an educative interface which provides visualizations to trigger climate-aware thought-processes and provides some useful insights through basic statistical and machine learning methods.

1 Background

It is estimated that climate vulenreability will displace 250 million people by 2050 [5]. A more shocking fact is that if everyone in the world lived the way people do in the U.S, it would take five Earths to provide enough resources for everyone [3]. Additionally, thirty seven percent of Americans believe that global warming is a hoax, and 64 percent don't believe that climate change will seriously affect their way of life. [4] The greater problem is how nations choose to tackle this problem of climate change. There have been around 2,950,000 publications on climate change according to Google Scholar. Although most of them produce interesting results, a majority of them fail to attract attention of the general populace through simple media like interactive visualizations. We intend to achieve a part of this function through the present experiment.

2 Introduction

The project aims at performing a thorough analysis of Climate Change and Socio-economic datasets made available by the World Bank OpenData Bank. We found many relations between factors affecting climate change, and produce thought-provoking visualizations to increase awareness. Also, through the power of learning theory, **We devised a time-series model that presents the socio-economic factors contributing to climate change, and predicts future conditions if human actions proceed as they do today. We rendered many interesting visualizations providing vivid representations and run-throughs of data that has been collected over many years by different sources.** We produced interactive plots that show different factors against each other in a time-lapse charts using D3js.

2.1 Why This is an Important Problem

- Most of the research on climate change has been limited to papers and mathematical figures. It fails to provide interactive visualizations for people to see.
- Possibly, there are interesting relationships between factors leading to or affecting climate change. It would be interesting to find out these relations.
- Machine Learning methods can allow us to model the conditions and predict the future. This has been done before, but what makes our project different is the insights and visualizations that we have generated from this model.

3 The Datasets

We have used various datasets from the World Bank's OpenData bank [1] and Berkeley Earth[2]. Some of these include:

- CO_2 emissions (metric tons per capita)
- Electric power consumption (kWh per capita)
- Energy use (kg of oil equivalent per capita)
- Forest Area
- Improved sanitation facilities (% of population with access)
- Mortality rate, under-5 (per 1,000 live births)
- Population growth (annual %) and Population, total
- Renewable electricity output (% of total electricity output)
- Literacy - School enrollment, primary and secondary (gross)
- Urban population (% of total)
- Average Annual Temperatures of all Land Area across countries - Berkeley Earth

4 Implementation & Approach

Primarily, our aim iss to find out correlations between Population, CO_2 emissions, forest area, energy use, urbanization and literacy. For achieving this, we divide this problem into two subproblems which deal with two kinds of factors.

4.1 Correlating Factors

These are factors that indicate the effect of climate change. Some consequential indicators include CO_2 emissions, Energy use per units of GDP, or per capita, average temperatures, urbanization and literacy. For these features, we have real datasets from World Bank which we can use to find correlations and build predictive models.

4.2 Causal Factors

They are factors that are the real cause for climate change, such as deforestation, agriculture, mining. These factors can probably account for any "shifts" or differences in expected and predicted values from our model.

Our project was divided into three stages. First stage involved running correlation tests such as the Dickey-Fuller Test and finding out if there are relations across the time series of different features. This stage also involved data cleaning like using interpolating values for missing values etc.

The second stage involved performing MDS for visualizing the level of similarity of countries with respect to some features, and treating various factors such as literacy rate, population growth and CO_2 emissions, we also trained a Machine Learning model to predict the average temperatures across the land over time across the countries.

The third stage will involved visualizing results and answering questions such as how much the Earth will heat up, or how does the literacy rate affect the climate change, or how does the urbanization, agriculture affect the climate change. We also produced several visualization using D3js, ThreeJs, and WebGL. We chose the "Gapminder's Visualization", *inspired by Hans Rosling's memorable 2006 TED talk*. This visualization shows the dynamic fluctuation in values across time for multiple features in a simple way. One can mouseover the year to move forward and backwards through time

Backend: The backend involves using Python, **TensorFlow, Keras**, scikit-learn, Pandas, NumPy, SciPy.

Frontend: The frontend involves HTML, CSS, Javascript and **3.js (three.js), D3.js, WebGL** for visualization.

5 Implementation & Approach Description

5.1 Data Preprocessing

This stage involves assimilating all the datasets from the World Bank's OpenData and Berkeley Earth. We also used the MLEDoze's Countries Dataset as a common database for all countries. We identified a country using it's `CCA3` **coding scheme**. The code is available here. The preprocessing involved the steps in Figure 3.

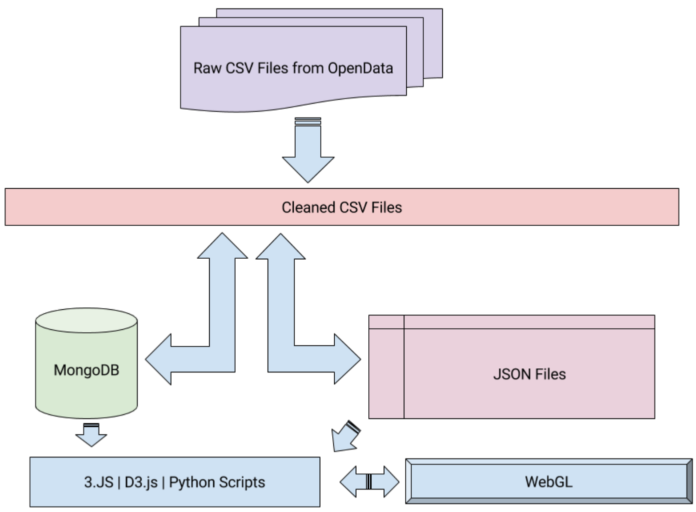

Figure 1: Preprocessing Backend

- The raw data is fetched from OpenData using `Thunder/wget` and cleaned for required information using `genCleanCSV` function. This function takes care of encoding (Unicode to ASCII), missing values (average values, or declines entirely in case of large number of missing values)
- The cleaned data is saved into CSV and loaded into the MongoDB database for faster access in python applications.

- The cleaned data is also read into JSON files for WebGL and 3.js to be able to read it.

5.2 Data Analysis

This stage involved a lot of parts

5.2.1 Feature Set Description

We were able to perform MinMax scaling on the dataset for rendering visualizations, and also for tasks like Time Series Prediction.

5.3 JSON Encoder

The ThreeJS library expects data to be encoded in a very specific data format. We wrote python scripts which clean the data, and encode them in the JSON format required by ThreeJS.

5.3.1 LSTM For Time Series Prediction

RNNs have been extensively been used for Time-series prediction. Our problem involves a time series, and this makes it difficult to form predictive models. Unlike regression predictive modeling, the additional attribute of time adds the complexity of a sequence dependence in the features that are trained upon.

Neural networks have been widely used for prediction tasks lately. A special kind of neural networks are designed to handle sequence dependence. This class of neural network is called recurrent neural networks. LSTMs (Long Short-Term Memory) networks are a type of recurrent neural network used for time series prediction using deep learning. Since our project also involves handling time-series sequences, we thought of using recurrent neural networks for time-series prediction. So, we trained a Long Short-Term Memory network (LSTM network), which is a type of recurrent neural network for deep learning. We used **Keras** on top of **TensorFlow** to code our model. Our LSTM Network has 4 blocks (pairs of Neurons and Memories). We framed the prediction problem as a regression problem for predicting CO_2 emissions five years down the line. Our model was able to achieve an RMSE Error of 0.60 on training and 0.78 on testing, which proves that our model is able to predict the carbon emissions with reasonable amount of accuracy. The same can be visualized in Figure 3 (for the United States). The same has been done for all the countries which had no missing values.

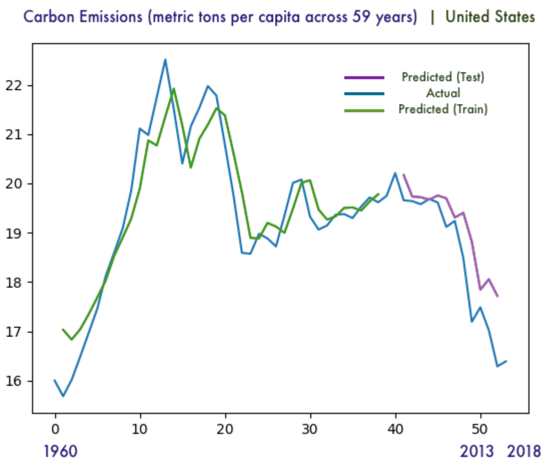

Figure 2: Predictions Using LSTM

5.4 Clustering for finding countries with similar socio-economic trends

To determine the countries following the same trends in terms of socio-economic factors, we performed clustering over our dataset. We used the K-Means clustering algorithm for this task, and use the **elbow-method** to determine the optimal value of k, which came out to be 5.

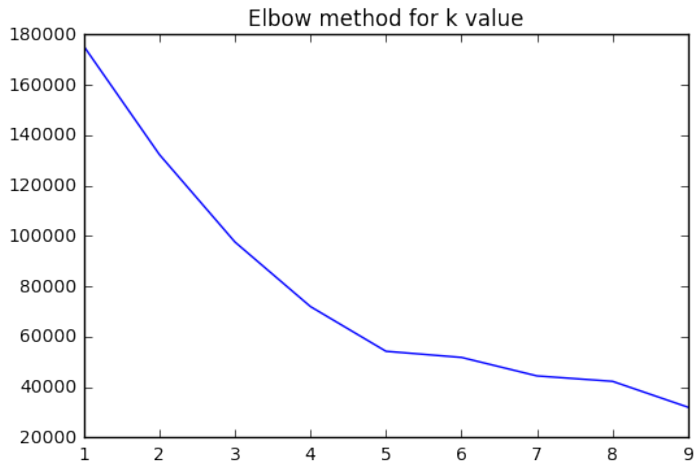

Figure 3: Elbow method for finding optimal value of k

5.5 MDS for visualizing similarity

Multi-dimensional Scaling is used when there are multiple features in a dataset and it is difficult to find similarities in different rows of the dataset. Normally, MDS is employed where we need a visual representation of a complex set of correlations. Since graphs are usually two-dimensional objects, MDS translates to finding an optimal configuration of points in a 2-dimensional space. In our context, this optimal configuration implies a set of contries distributed by their similarities.

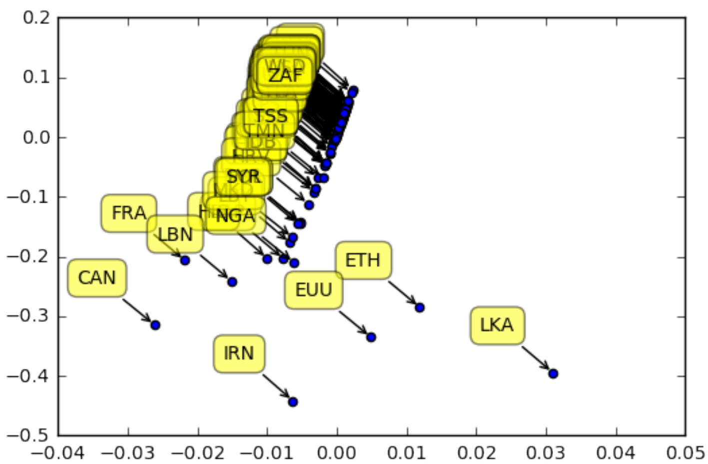

Figure 4: Multi-dimensional Scaling for Similarity detection

5.6 Axis re-ordering for parallel coordinates

Visualization of dataset with multiple dimensions is a challenging task because of non-uniformity in the data. To visualize data for a better analysis, we used parallel coordinate. Furthermore, we resorted to a correlation based axis ordering of different features for obtaining a visually appealing parallel coordinates graph. We obtained very insightful results from this, which we will discuss in observations section.

5.7 Data Presentation & Visualization

This is a dashboard project which aims at creating an educative platform which captures attention of the people unaware about the climate change, through insightful and interesting visualizations.

5.8 Carbon Emissions (in metric-tons per capita)

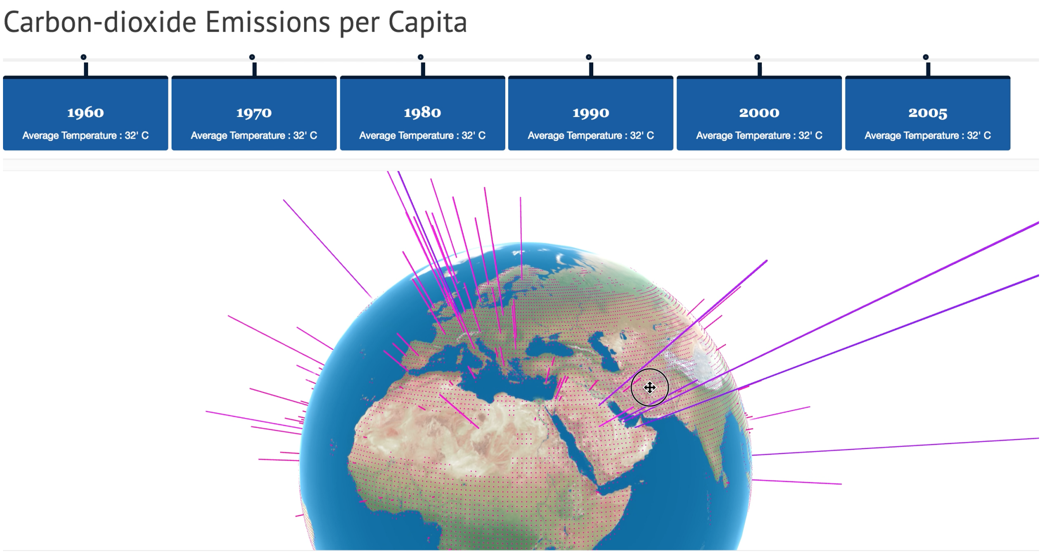

Figure 5: Carbon Emissions (in metric-tons per capita) - 3D Visualization Using WebGL

We used the JSON from the cleaning and preprocessing to generate a 3D Globe Visualization for Carbon Dioxide emissions across the globe across several years. This beautiful visualization is rendered using WebGL. The user can rotate and interact with the visualization. The visualization is displayed below timeline over which the user can hover to know the current value above the country on the globe itself. The height of the line indicates the amount of carbon emissions in metric tons per capita. The higher the number, the darker is the color's intensity. The timeline above the globe adds value because we can compare the values just by hovering over the timeline.

5.8.1 What makes this visualization useful

The ability to see the carbon emissions of a country right over it on the globe helps a lot in visualizing. The interactions of dragging, panning involve the user more, and thus stimulate perception. The users can also compare different countries right on the globe.

5.9 Dashboard

Our project has a very intuitive dashboard which has two flipcards with interesting questions which act as a click-bait for capturing users' attention. Once clicked, the card flips and shows interesting visualizations such as Forest Area vs Carbon emissions, or carbon emissions per capita of a country etc.

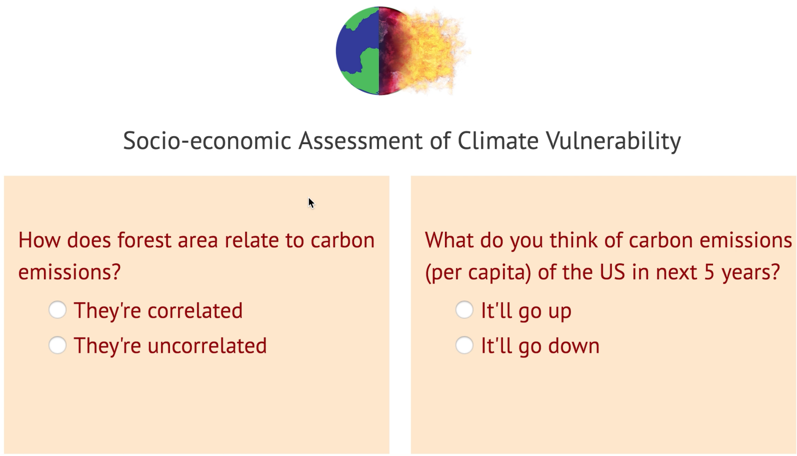

Figure 6: Main Dashboard

5.9.1 What makes this visualization useful

Our project was primarily focused on increasing awareness about climate change. It is crucial for our dashboard to capture users' attention in the first glance. These inquisitive questions help us in achieving these tasks.

5.10 How does urbanization impact carbon emissions ?

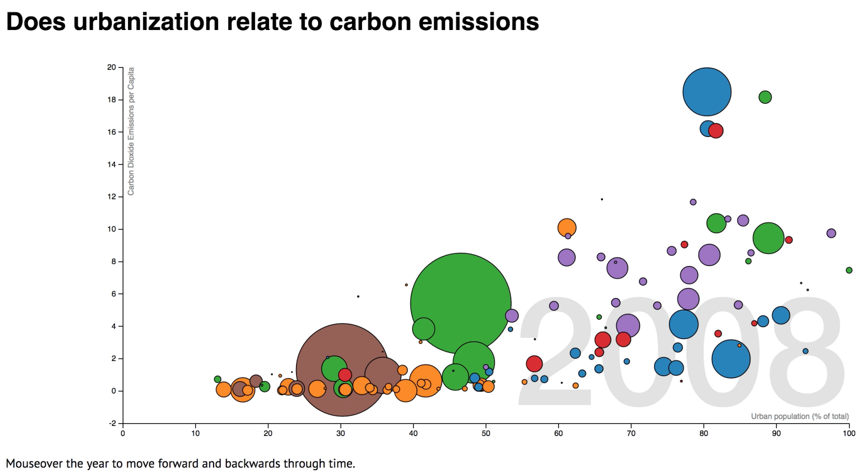

Figure 7: Urbanization vs Carbon Emissions, Population across years. Bubble size indicates population of a country. Color denotes region.

Urbanization plays a major role in determining consequent factors such as number of vehicles, factories, commercial buildings, electrical energy consumption etc. Our timegraph captures the trends in urbanization, population, and carbon emissions of different countries nicely in one graph.

5.11 Parallel Coordinates - which factors impact the most?

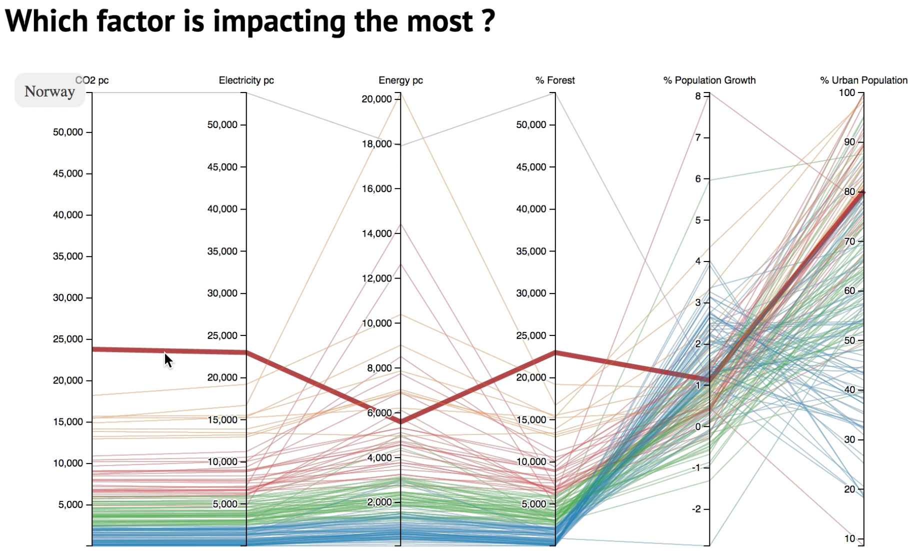

Figure 8: Parallel coordinates using different features

To find out which factors impact the most, we used the parallel coordinates visualization. Since the order of the axes is very important for visualizing the relation between features and finding trends, we needed a mechanism for the axis ordering. We performed clustering to cluster countries with similar features and color coded them for displaying on the parallel coordinates.

Let $x_i = (x_{i1}, ...x_{i6})$ be the set of our features described above, and the centroids of the clusters are $m_1, m_2, ...m_5$ (since optimal $k = 5$), with partitions $p_1, p_2, ...p_5$, then we try to converge to a *local* minimum of

$$\sum_{k=1}^{5} \sum_{i \in p_k} ||x_i - m_k||^2 \qquad \text{Euclidean distance}$$

(within cluster sum of squares)

We also ordered the axes using a correlation based axis ordering. MDS was also used to help achieve these results. Furthermore, we found out that the carbon emissions is indeed a deciding factor which has a direct impact on climate change. We found many more useful insights from this visualization which we have discussed in the next section.

5.11.1 What makes this visualization useful

It is important to view all the features and their impact as a whole. Parallel coordinates is specifically helpful for multidimensional data.

5.12 Average Annual Temperature vs Carbon Emissions per capita

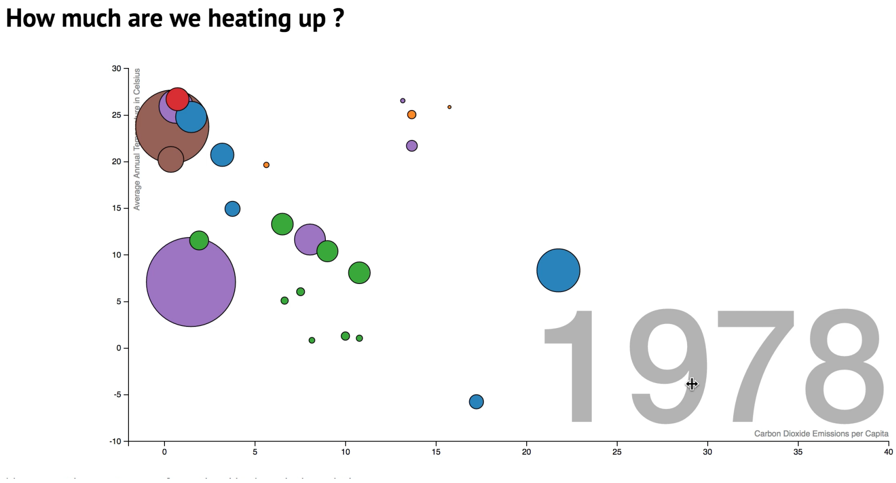

Figure 9: Average annual temp. vs carbon emissions in metric tons per capita. Bubble size indicates population of a country. Color denotes region.

The visualization plots the average annual temperatures of different countries, color coded by their regions (eg; North America, Sub Saharan Africe etc.) across their carbon emissions. We observe that the countries closer to equator have high annual temperatures and lower carbon emissions, while countries far from the equator like the United States have lower annual temperature but much higher carbon emissions.

5.13 Heatmap - Who's burning fast?

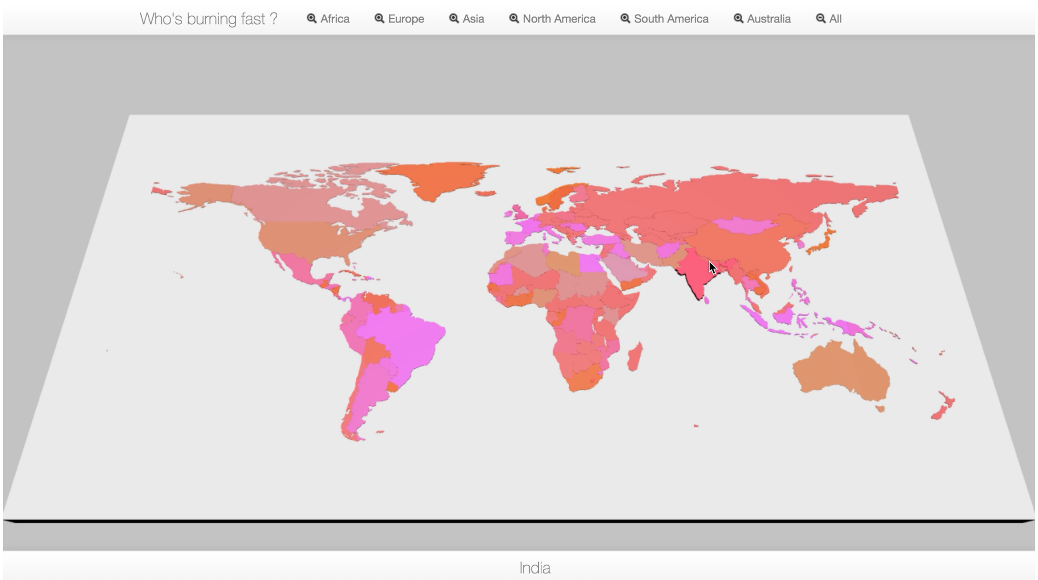

Figure 10: Heatmap visualization which shows countries colored by their rate of emissions (More Pink is less emissions, more orange is more emissions)[6]

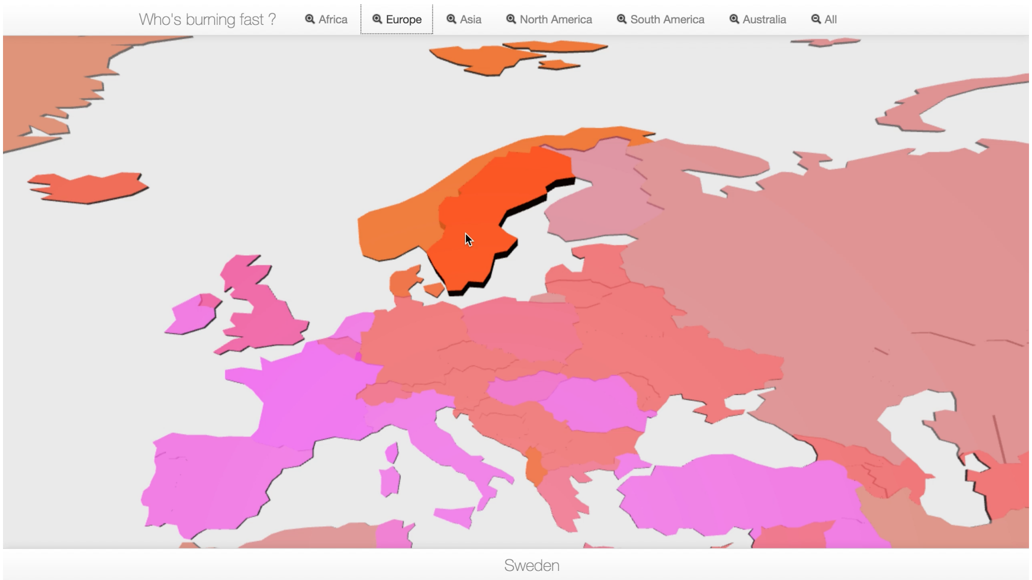

Figure 11: User can zoom to a continent

5.13.1 What makes this visualization useful

This visualization allows one to observe the rates of different countries and compare them at the same time. The interactions involve zooming, panning and clicking which help the user visualize interactively.

6 Observations & Conclusions

we can observe that there is a sharp rise in the CO_2 emission and urbanization rates of certain Middle Eastern countries in the early 1960s. This is because of the formation of the OPEC (Oil and Petroleum Exporting Countries) that allowed these members to trade oil as foreign exchange in the global markets. Consequent liberalization of the oil business increased both the GDP and co2 emissions. GDP in turn improved the levels of urbanization of these countries.

We can also notice that Kuwait had a sharp rise in its CO_2 emissions rate because of its massive nationalization of oil drilling. Interestingly, there is a total drop of CO_2 emissions in 1991, which was the year of the Gulf War where Iraq invaded Kuwait and shut down all of their oil wells.

We also observed that several developed countries have a higher rate of carbon emissions than other developing nations. Also, countries with higher carbon emissions were the ones having high urban populations. The population growth was found to be higher in countries with lower carbon emissions, which is quite surprising. Thus, this kind of interaction offers a fantastic opportunity to use interdependent factors as sources for real world narratives.

References

[1] The World Bank - OpenData "climate change". `http://data.worldbank.org/topic/climate-change`. [Accessed: 2017-04-02].

[2] Berkeley Earth Climate Science - time series data. `hhttp://berkeleyearth.org/data/`. [Accessed: 2017-04-04].

[3] Fast Facts About Climate Change - National Wildlife Federation. https://www.nwf.org/Eco-Schools-USA/Become-an-Eco-School/Pathways/Climate-Change/Facts.aspx. [Accessed: 2017-04-04].

[4] The Guardian - 'A tipping point': record number of Americans see global warming as threat. hhttps://www.theguardian.com/environment/2016/mar/18/climate-change-record-concern-us-global-warming-poll. [Accessed: 2015-04-10].

[5] UNHCR - The UN Refugee Agency. http://www.unhcr.org/en-us/news/latest/2008/12/493e9bd94/top-unhcr-official-warns-displacement-climate-change.html. [Accessed: 2017-04-10].

[6] Worldmap - clientside web engineering at university of applied sciences salzburg. http://torghele.at/worldmap/. [Accessed: 2017-05-06].