**INF 229 – Data visualization**

# Project report

# Visualizing the “Countries” dataset

* Caldas Felipe
* Huang Yingpo
* Huy Dinh
* Petel Cyril
* Truong Thai-Chau

# Plan

**I) About our project: data and potential users**

I.1 Data

a) The dataset “Countries.txt”

b) Issues about the dataset

I.2 Our potential users

1. Users experience with the data
2. What does the user want to understand?

**II) Our visualization**

II.1 Ideas that guided us during the project

II.2 The cartogram

II.3 Parallel coordinates and table lenses

II.4 2D correlation scatterplot

II.5 An interface to switch between views.

**III) Critique of our visualization**

References

**I) About our project: data and potential users**

**I.1 Data**

1. The dataset « Countries.txt »

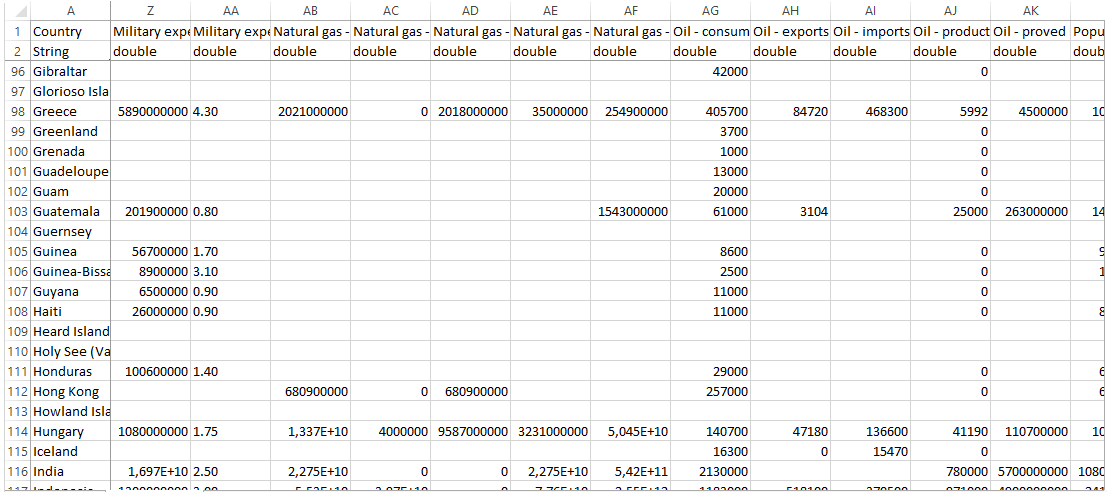
We hesitated a bit before choosing this dataset, but this option soon became obvious since we are an international team: four nationalities are represented among us (Brazilian, Chinese, French, and Vietnamese)

The dataset is composed of 264 countries, that is to say almost every countries in the world. Some of them are small islands (ex: Tokelau, pop. 1400 ) and might not be really interesting for our user. The dataset provides 45 different criteria to have information about these countries. These criteria can be classed in different domains: economic (for example GDP), geographic (ex: area); health (ex: AIDS prevalence adult rate)… It gives general information about countries as well as detailed information for some fields (for instance for oil: exports, imports, production and consumption). It enables to understand many things about geography, geopolitics… All the data are displayed with double numbers: **data cases are quantitative.**

1. Issues about the dataset

When studying the dataset, we realized that, while it was really interesting to work on a huge number of data cases, the dataset was not so easy to handle.

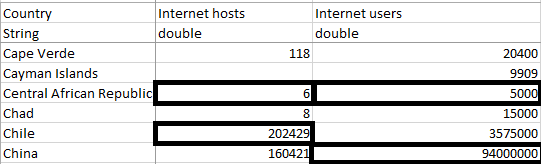
First problem: for some countries, data is not available for some (or all the) criteria: this had been an issue for our project. In theory, we were supposed to display 11 880 cases, in fact it is less due to these missing information on countries. Most of the time, we lacked figures on small countries like small Islands in Oceania, but sometimes more important figures were missing (for instance Japan’s external debt).



*A part of our dataset with missing information*

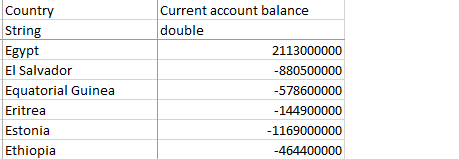
Given the time we had to complete the assignment, we could not deal with this missing value issue.

What took more time to discover was the variety of data. Since the dataset lists all the country in the world from the tiniest to the biggest, for a given criterion, it deals with figures that can vary up to (ex: pop. of Tokelau / pop. Of China). Then it would be really hard to compare figures for some countries. One could say that micro-countries are not relevant for the user, but sometimes such differences could be seen in big countries. The variance of data case was really impressive in some cases, and would be problematic for some classical representations. We will explain in the visualization part how we tried to tackle the issue to some extent.



*An example among many of the huge variance in the data*

Negative values were a similar problem: for some criteria, there were negative values, due to the nature of the measured quantity which could be negative (for example Current Account Balance).



*Negative values for Current account Balance*

**I.2 Our potential users**

1. Users experience with the data and visualization tools

Geography is taught from middle school to university, and everyone needs to get information about countries from time to time. Everybody has some notions of geography and knows richest countries of the world, or biggest… To sum up, it is clear that our user has experience with at least some parts of our dataset. But he does not know the figures by heart, and he ignores where all the countries of the world are.

Our potential users could be anyone: curious users who want to know more about economic figures in a country, students working on the AIDS plague.

We do not expect our users to be information visualization professionals. They are basic IT users, who instinctively understand how to manipulate a software if it is not too complicated.

1. **What does the user wants to understand from the data?**

What was more interesting than our users’ background was to think about what the user wants to do with our visualization. Here are some ideas we had: (for syntax purposes, the user is referred as “he”)

* First, for a given criterion, he might want to see at a glance what the most prominent countries are, while having an actual vision of the geographic positions of the country on his mind.
* Related to this first request, he might also want to know the biggest and lowest values (and corresponding countries) for each criteria
* He wants to compare countries with one another. For example, he could be curious to compare G7 countries on their GDP, unemployment rate, debt… This might be particularly interesting for geopolitics purpose: …
* The user may want to see if there is any correlation between economic and social factors, depending on countries and continents; or more generally between any couple of criteria. For example, that is what we’ve been through during our project: we spent much time trying to study correlation between things out of curiosity.

**II) Our visualization**

II.1 Ideas that guided us during the project

First, we wanted a multi-function visualization, which could be useful for different purposes depending on what the user wants to do. We had found three main requests the user may have (map to compare all the countries, compare several countries on several criteria, and prove correlation between criteria), and we truly thought that these three requests were legitimate and important. Obviously, different purposes would lead to a more complex visualization, but we were ready to make this clear choice.

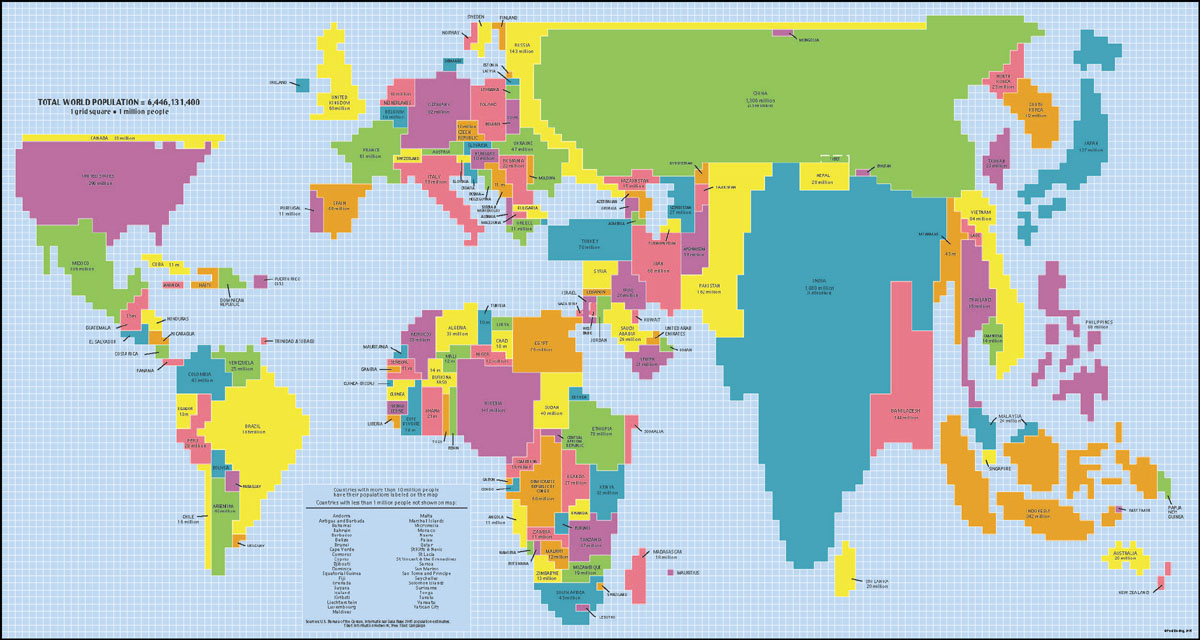
The design document was a critical step: we had to reduce all our ideas to only three. The three we selected were an anamorphosis map (from the French “carte par anamorphose”) or cartogram, a bunch of multivariate graphs, and a 2D scatterplot to prove correlation between criteria. These three ways of representing the data were radically different and answered different questions the user may have, so we decided to keep them and to integrate each one of them in an interface to navigate between different representations.

We will start by describing each one of the different visualizations, and then focus on how the user can navigate from one to another.

II.2 The cartogram

The most original and exciting idea we had was to make a cartogram. We wanted to show a map of the world, in which country would be deformed depending on their figures for one criterion. Obviously, with this kind of map, the user can be disappointed in the beginning because he is used to the countries’ shape. But once the concept is understood, it is a really practical and clear way to visualize data.

The first implementation we wanted to make was a rectangular cartogram. The country would only keep their position on the map, but their shape would disappear, replaced by only squares of different sizes depending on the criterion.



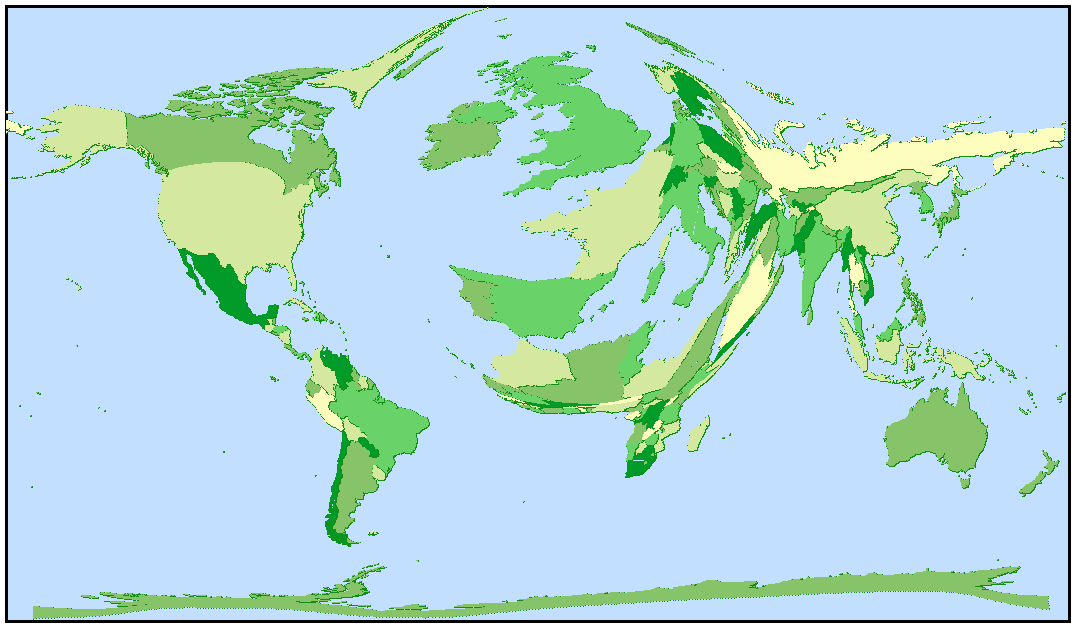
*A rectangular cartogram found on the Internet*

Although it was to us the best visualization we could think about, it seemed difficult to implement from the beginning. To do that kind of map, we would need to put manually the relative position of all the countries in the world, and then test it for all the criteria to see if that works. There were also problems of disappearing countries, or countries which became too big for some criterion. We were forced to think of other implementations.

Another kind of cartogram is one obtained by computing some deformation algorithms on an actual map. Even if it was difficult as well, we tried to develop this idea.

First, we had to load a map of the world compatible with our processing environment. We found a map of the world with a SVG format that could be handled with our program: it gave all the coordinates of all countries of in the world in numerical values. (The map can be downloaded here: <http://ruig.grid.unep.ch/wp-content/uploads/2006/11/country_shapefile.zip> ) We then had to match this SVG file with the countries we had on our dataset, dealing with different names for the same countries in the datasets. Once the country was associated with its position on the map and the corresponding shape, we applied an algorithm to deform the map.

Different algorithms are available on the net, with various level of mathematical difficulties. We chose one we can understand, based on the ScapeToad project as well as the Gastner/Newman diffusion-based algorithm. (see references)

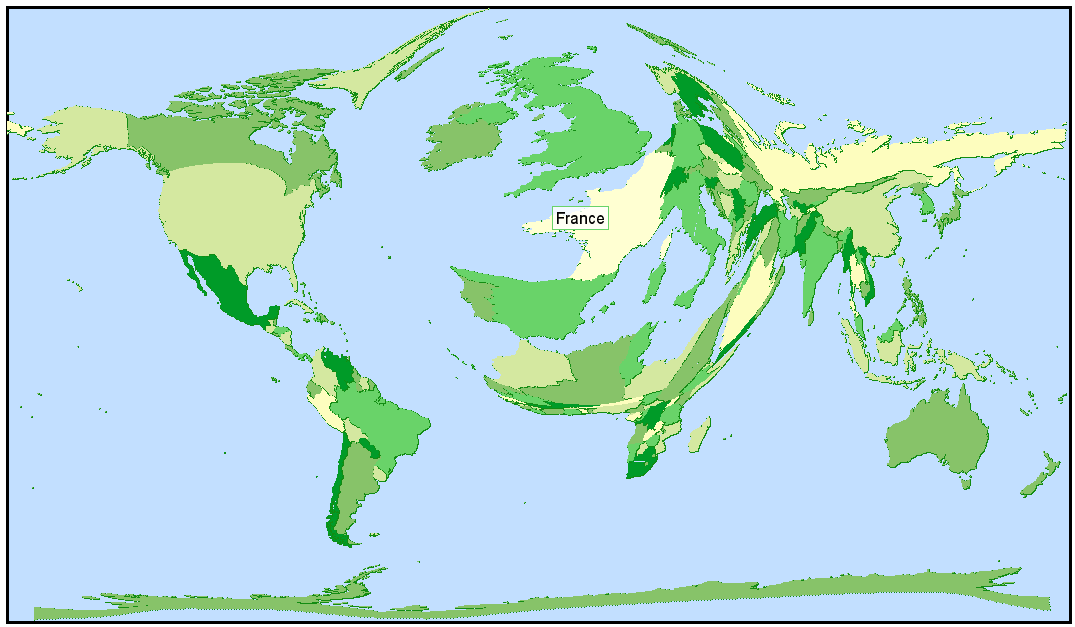


*Cartogram for external debt*

The example provided above is for the criterion “External debt”. At a glance, we understand the tendencies: Western European countries have a huge debt, American countries average, and Asian countries low debt. We will cover on II.5 how to select criteria.

This vision also has some drawbacks: it is hard to have a correct vision of the debt for a given country. Moreover, some missing data make it somehow false: one might think Japan has not been deformed enough on its map regarding its really huge well-known amount of debt, but it is simply because the figure is not available in the dataset.

If the user puts the cursor on a country, the country is highlighted and he can see its name.



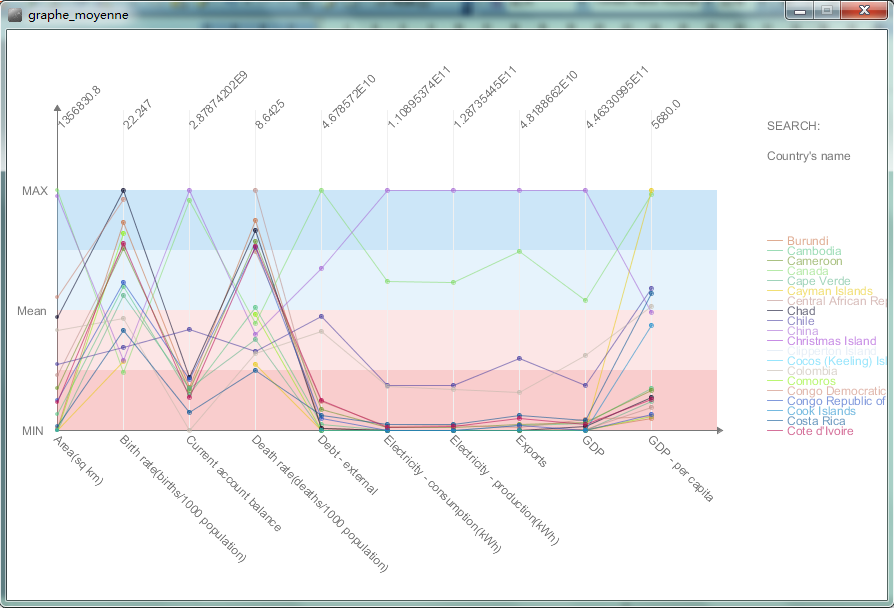
*Here France is selected*

II.3 Parallel coordinates and table lenses

Although the cartogram is an original and interesting idea, it is not practical if the user wants to compare some figures for some given countries. To do so, we thought that the most efficient way was the classical way for hypervariate data: parallel coordinates and table lens. This representation is useful because it can work for an important number of criteria as well as for a big number of countries, so it is really relevant for our dataset.

1. Parallel coordinates

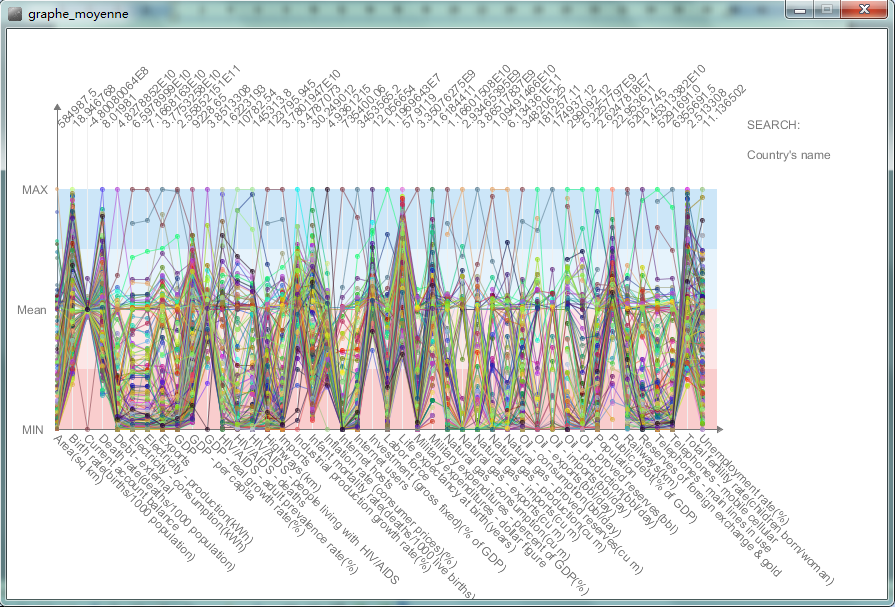
On the x axis are displayed criteria the user wants to know about. Each line represents a country, and we decided to color the lines to make it easier for the user to recognize countries. The size of the y axis is constant and values are displayed depending on their relative position with the maximum and minimum values for all countries selected. On the right of the screen appear the countries names with their respective color.



*Parallel coordinates for 20 countries (chosen by alphabetical order) and 10 criteria*

On this first example, the user can see some features we have added to make him understand the figures. For each criterion, we have put the minimum, maximum and average value. For instance, for exports, we can see that Comoros have the lowest value and China have the biggest, and they are displayed respectively at the bottom and at the very top of the graph. We can also see which countries have a value above average, and which countries have a value below: if the point is on the red area, then the value for this country and this criterion is below average (for the selected set of countries). The mean value has been displayed on the top of the graph: even if it does not seem pretty, it gives the user a first clue on how to interpret the data.

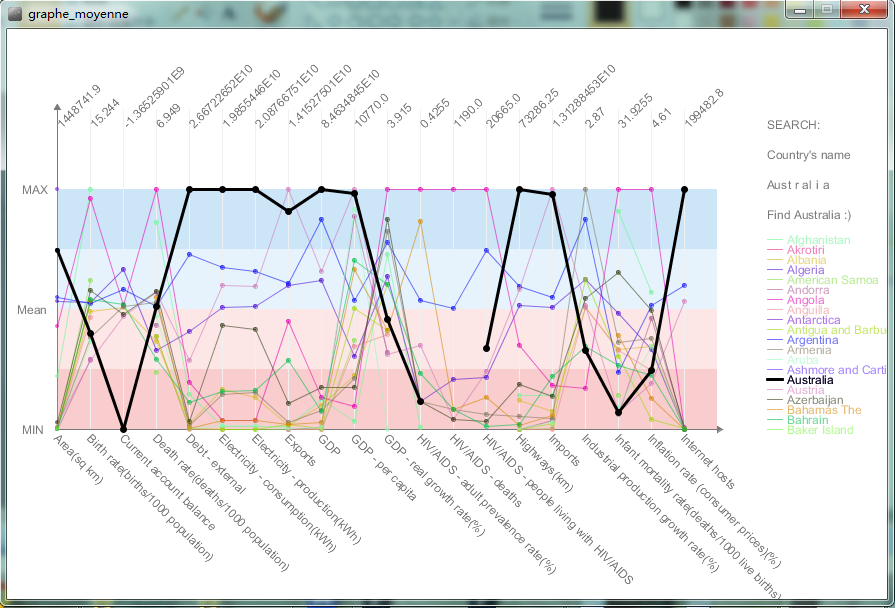
With the interface we will cover later on, the user can choose which countries and which criteria he wants to display, with no limit.



*Parallel coordinates for 263 countries and 44 criteria*

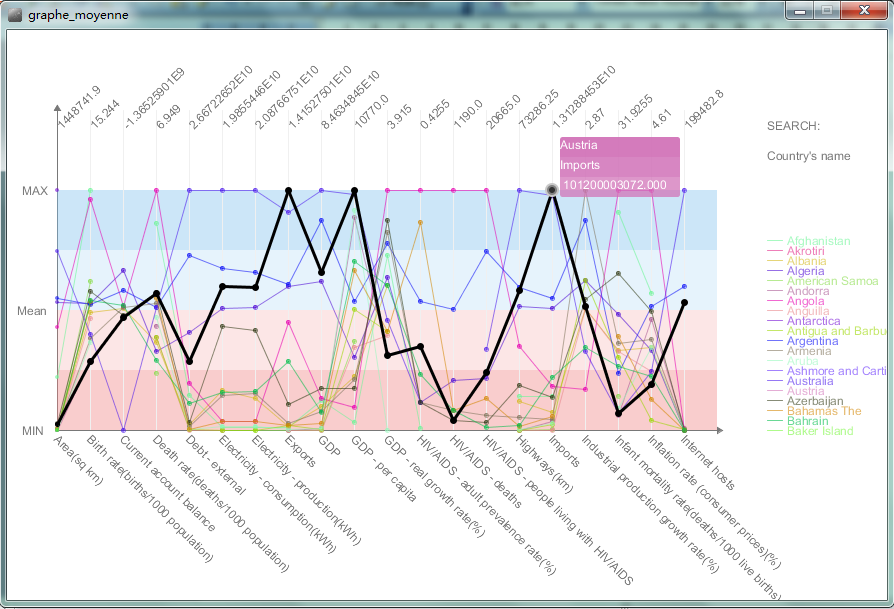
Even if this graph is hardly readable, it can give some trends for the statistical repartition of countries for each criterion.

To get figures he wants about one country, the user just has to put the cursor on the country’s name. Then the country line is highlighted in black, to enable the user to see all the country’s features for all the selected criteria.

f

*Australia highlighted among 10 countries and for 20 criteria (HIV/deaths was not available)*

If the mouse is moved on a special point (let’s say Austrian imports), the country is highlighted, plus the figures for the selected criterion and country appear on a box.

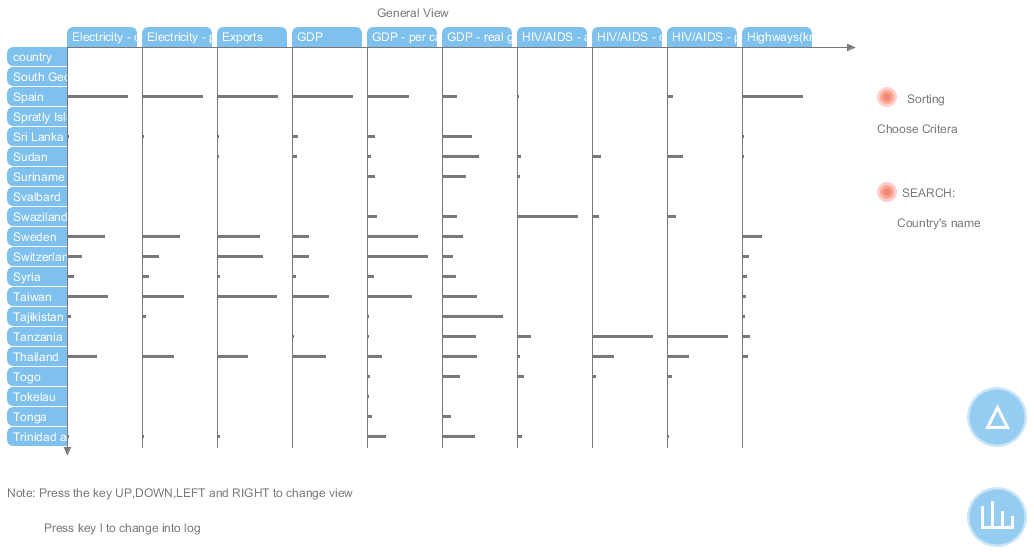


*Austrian imports with 20 countries and 20 criteria*

If the user wants to have all the information about the country for the criteria he has chosen, he just has to type the name of the country on the keyboard, and then press enter. If the country is on the list, let’s say Spain, a little message “Found: Spain” will appear and the lines and points for Spain will be highlighted as well.

1. Table lens

The table lens can be used for the same purpose than parallel coordinates (with a huge number of countries on various criteria) but can be more practical because it enables to classify, find minimum and maximum values easily.

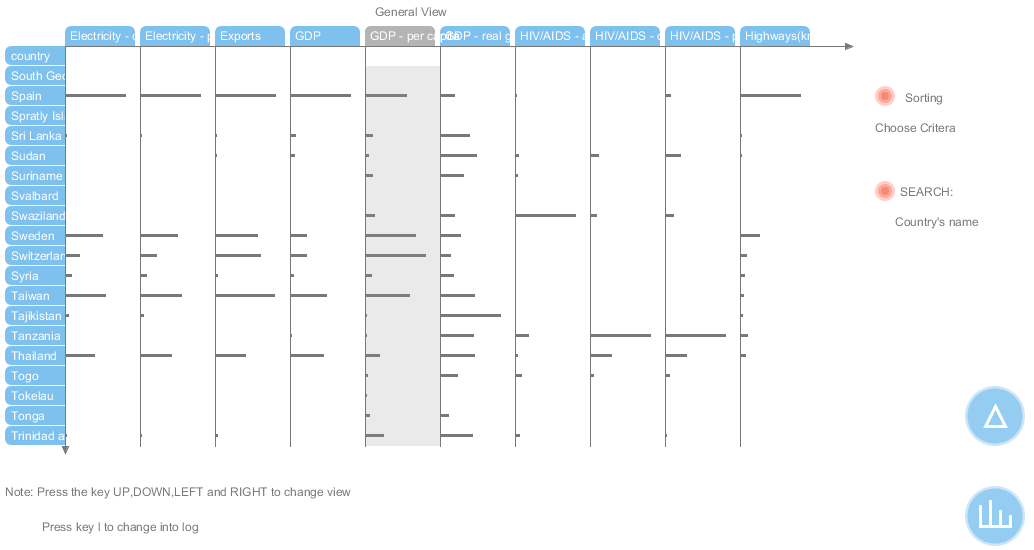


*How the table is displayed*

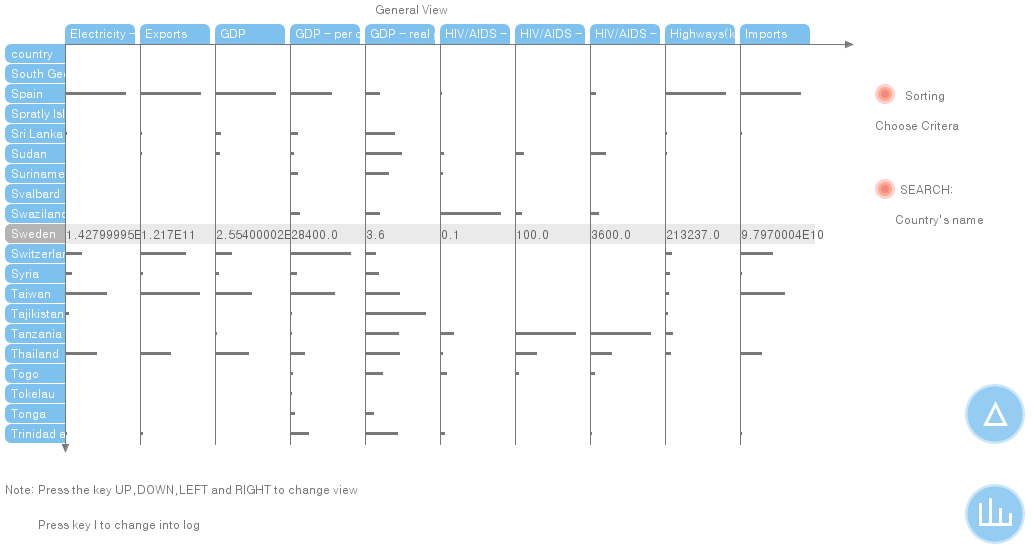
First we can see that there are two blue buttons. They also appear in the parallel coordinates graph, even if it was not there on the images above. They enable to switch from one design to another:

 is to go to parallel coordinates,  is to go to table,  is to go to starplot.

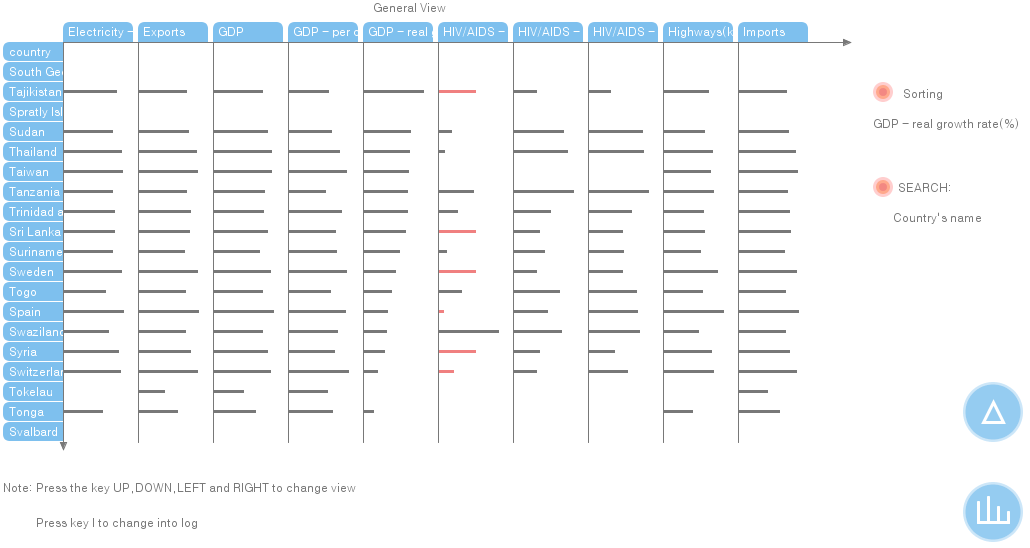
Many interactions are available with the table: they enable users to sort, find maximum and minimum values, find a specific values, have logarithm plots, or directly find a specific country.



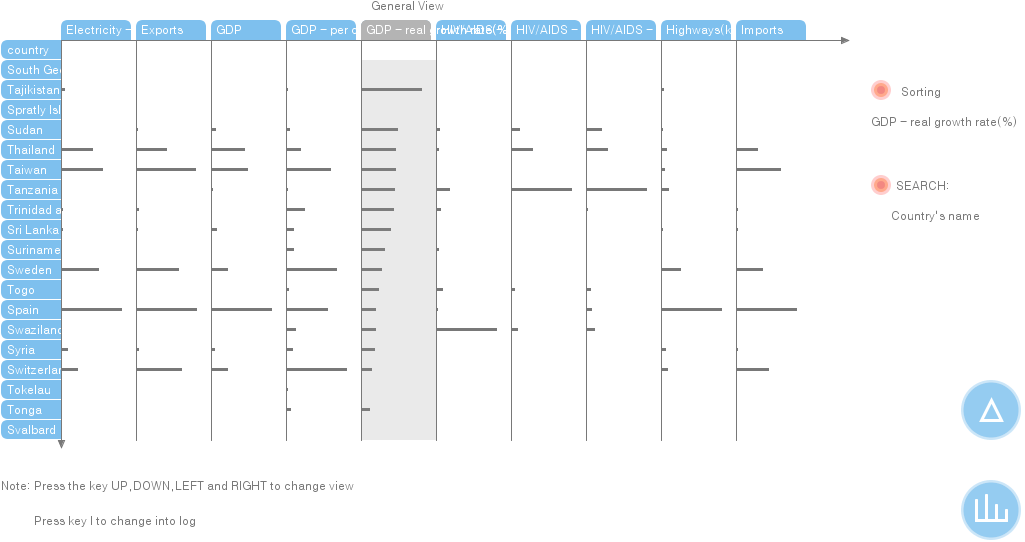
*Cursor on the label GDP per capita: it becomes highlighted*



*Clicking on a label: numerical values are displayed (here for a country label: Sweden)*



*Pressing l: puts all the values in logarithm form*



*Pressing s with the cursor on a criteria: sort all criteria*

There are also two other kind of interactions (we don’t have enough space to illustrate them here). First, if the user puts the cursor on a case in the table, the value will be displayed. Moreover, if the user presses ctrl and then the country’s name, the country will be highlighted in the same was as label are highlighted above.

1. Star plot

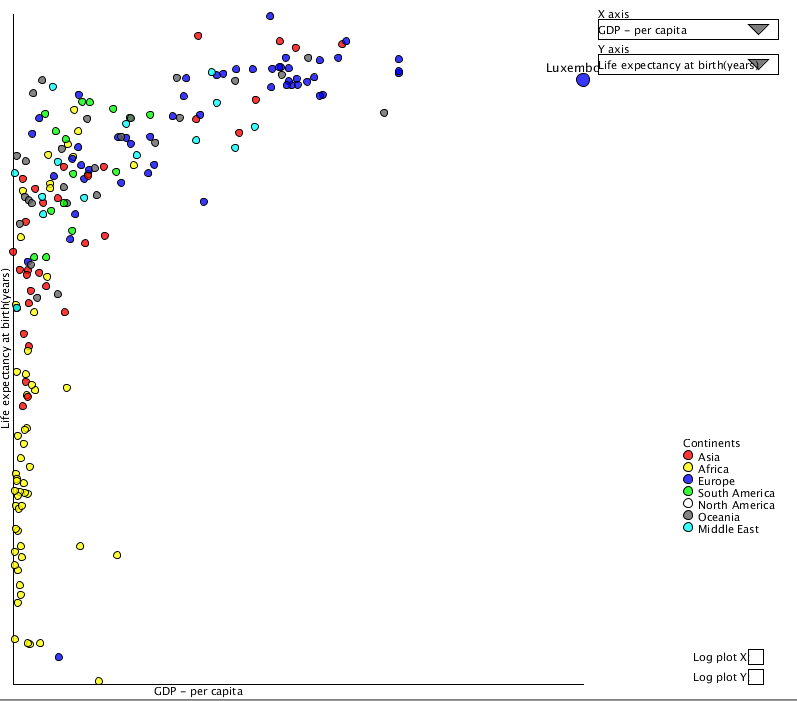
We also made a star plot visualization, which is really related to the parallel coordinates one. Except for a different shape (star instead of a regular graph), interactions and purposes are exactly the same.

II.4 2D correlation scatterplot

The third way to visualize our dataset is a 2D correlation scatterplot. This visualization was directly designed to answer questions the user might have on the possible links between criteria. It also enables the viewer to see which country has the biggest (or the lowest) figure for a couple of criteria.

The implementation and tests for this visualization went well and we did not face difficult problems. Each graph is scaled depending on the minimal and maximal value for each criterion,

The user can choose two criteria to analyze thanks to a list on top-right corner. Of course if he puts both axis with the same criterion, he will obtain a line.



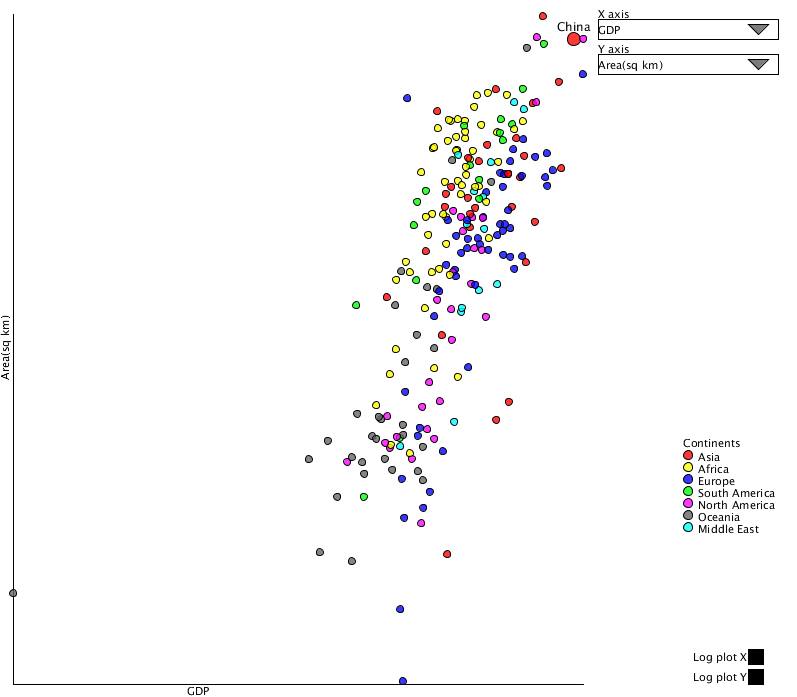
*Correlation between imports and GDP for all the countries in the world*

As an example, we can see above that the user has chosen to display Life expectancy at birth on the Y axis and GDP per capita on the X axis. A non-linear correlation is then demonstrated.

We decided to add manually a new variable for countries: the continent they belong in. It enables us to categorize countries and to see trends for each continent. For instance, we can see here that African countries have lower life expectancy at birth and lower GDP, while European have countries the biggest. This kind of visualization is really interesting because it can in some cases illustrate some topical issues: here the migration from Africa to Europe can be a consequence to the figures this graph shows.

The big issue here was negative values and high-variance values. The presence of outliers in the dataset for almost every criteria made most of the scatterplots unreadable: a majority of the points were aggregated on the left-right corner of the graph, and a couple of outliers were located really far away.

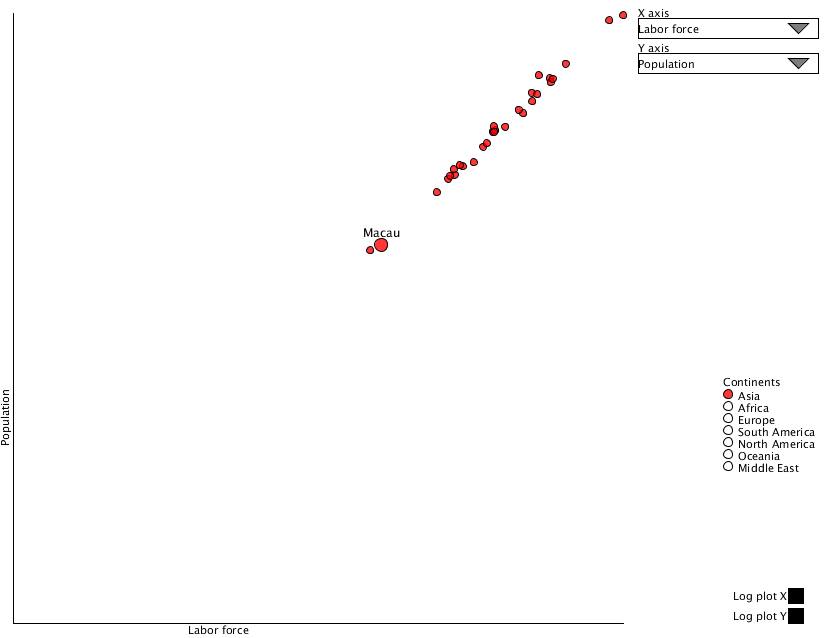
The best solution we’ve encountered to cope with that was to use a logarithm plot. Values for x and y are not linear, but displayed in logarithm, to reduce space between circles and getting rid of the negative values problem while keeping the idea of correlation. To remain loyal with graphical integrity, we did not impose this choice: the user can still choose if he wants to use our log plot for x and/or y or not if he clicks on the squares at the bottom left corner of the screen.



*Area and GDP for all the countries in the world with log plots for x and y*

Here we see that area and GDP are correlated, and we can easily see each points thanks to the log plot. The user has to keep in mind that spaces between points are not representative of the reality: using logarithm gives only an idea for correlation, and does not enable to compare a country with one another.

We also added a possibility of interactivity: if the user scroll on a point, he can see the country’s name. There is also another way the user can interact with the graph: he can choose which continent he wants to study, by clicking on the circles near to the continents on the right.



*Population and Labor force only for Asian countries*

This option provides a more readable graph, and can be more convenient for the user if he only wants to focus on some continents only. It is also practical because for some features (for instance AIDS-related data), the figures for one continent (in this example Africa) are really different from those of other continents.

We could have tried to implement linear regressions, and putting a line to represent it. We chose not to do it for two reasons:

* For a significant number of couples of criteria, there is no linear correlation
* It was surprisingly difficult to implement.

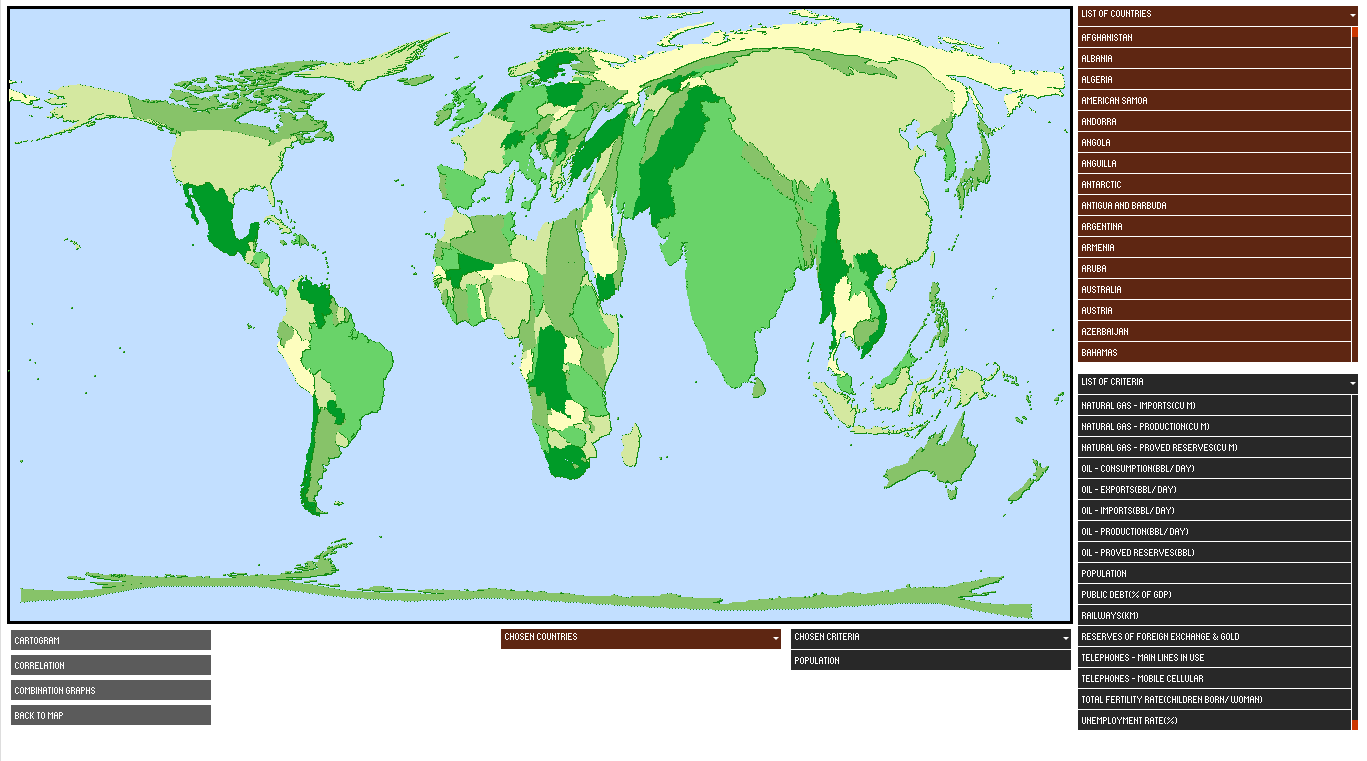
II.5 An interface to switch between views.

So far we have explained how our three kind of visualization work independently. To conclude, we will shortly explain how we implemented an interface to switch between views.

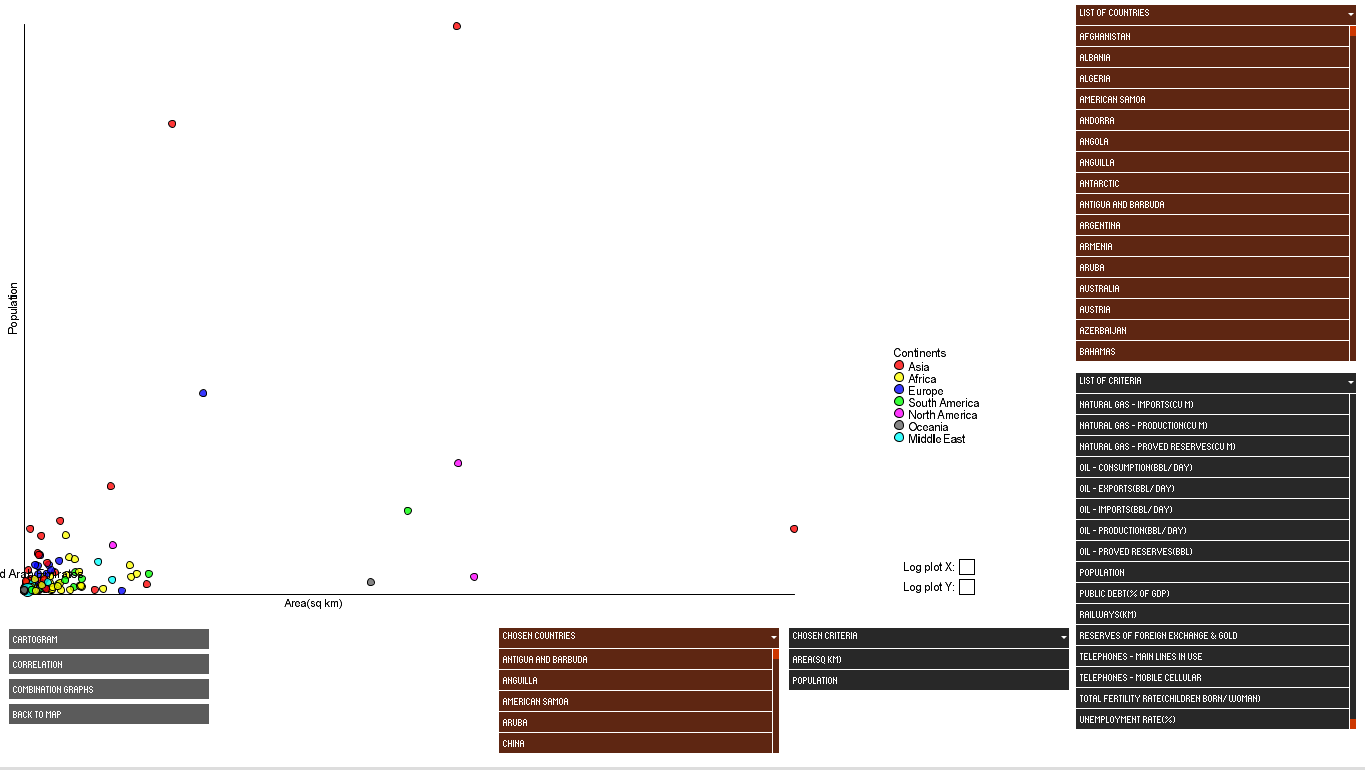
We decided to have a common frame for the three different visualizations: choosing criteria and countries and switching from one visualization to another is done similarly, regardless the visualization “you are in”. We used a library called “ControlP5” to create the implementation (see reference).

To choose a criterion, the user has to click on it, and same goes for countries. Click on an added criterion on the below list to delete.

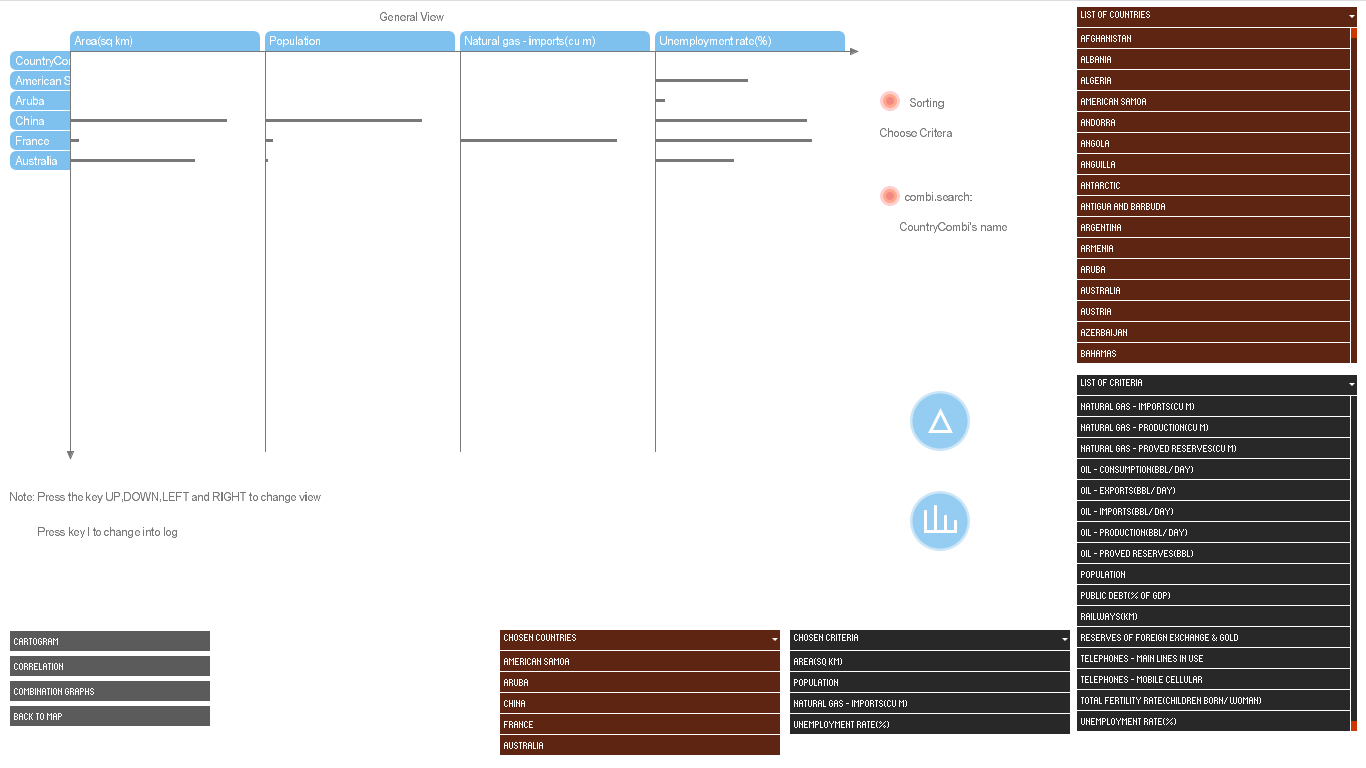
For the map, he can only choose one criterion. If he has chosen more than one, the first criterion on the list will be selected. Then the user just has to click on “cartogram” to get the image.



If he wants to switch to correlation scatterplot, he just clicks on the button. Then he can add countries and criteria by clicking on them on the right, and click again on “correlation”. To remove items, he just has to click on them where they are displayed below “chosen countries” or “chosen criteria”.



If he wants to switch to table lens and parallel coordinates, he clicks on combination graphs. To switch between both (or even to have star plots), he can click on the little buttons that are explained in II.3.



**III) Critique of our visualization**

III.1) What (we think) it does good

Our visualization is original because it enables to have different insights on data: with only one criterion (see cartogram), with a couple of criteria to compare (see 2D correlation scatterplot), or with as many criteria as the user wants (table lens and parallel coordinates). In that way, they do provide an answer to the user’s needs as expressed in the first part of this report: comparing countries, having an idea at a glance, proving correlation between criteria is doable: many things can be processed

III.2) What our visualization could do better

As any work, our project have some drawbacks that may have been fixed with more time/competencies/ideas:

* First, we have to admit it: the project includes the three really distinct ideas we had for the first report, so it looks quite like a superposition of different visualizations, even if they are complementary. None of these visualizations is at its best, and interactivity (to sort, or to have more information) could be improved for some features.
* We did not manage to have a visualizing tool mixing overview and detail at the same time on the screen: cartograms and correlation scatterplots can be used for overview and parallel coordinates for detail but we can’t see everything at once (except when we put all the data in parallel coordinates, but it can’t be read easily).
* Even if we tried our best, our interface is not that easy to use, and it might take some time for a beginner to understand how to handle it.
* The cartogram does not work as well as we expected. This is due to missing values, a complex dataset, but also to a complex and radically new implementation.
* We cannot prove correlation between more than criteria. We would need a 3D scatterplot or other tools we did not have time (or courage) to implement.

Thank you for your reading, hope you enjoyed it.

**References:**

For the svg map: <http://ruig.grid.unep.ch/wp-content/uploads/2006/11/country_shapefile.zip>

What helped us to implement the algorithm:

The original paper: <http://www.pnas.org/content/101/20/7499.full>

A project really close to our own: <http://scapetoad.choros.ch/index.php>

For the interface: <http://www.sojamo.de/libraries/controlP5/>

For the data: <https://www.dropbox.com/sh/im7ywtt9uela54s/AAB2q1ADKd0YSWJ6C8sMl-Mva?dl=0>

(Copy to the folder of the project)