EarthQuakes Analysis
Visual Analytics
Sapienza University of Rome

Federico Bucci 1649197, Michael Capponi 1711759

2020-03-12

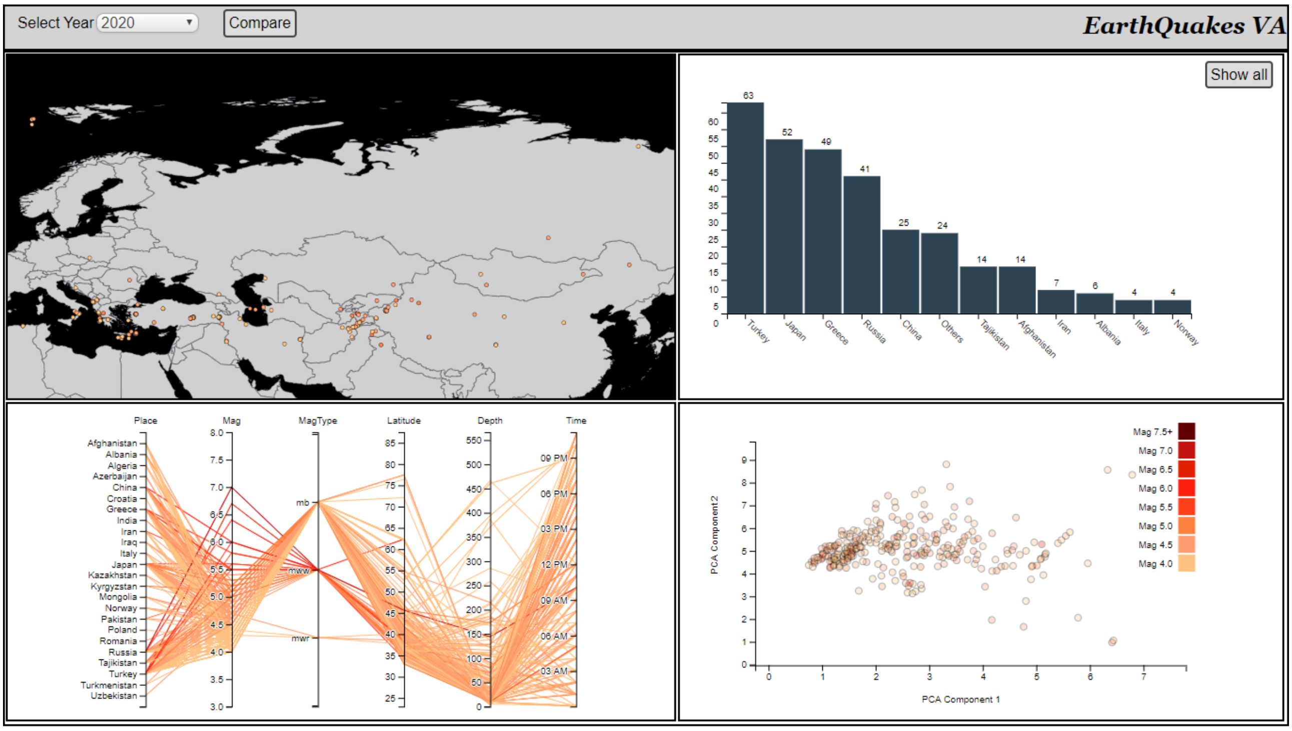


Figure 1: Project overview: earthquake distribution of current year (2020).

Abstract — Over the time, we have very often heard (and sometimes unfortunately experienced) the terrible consequences of earthquakes around the world. Especially for those that occur in large cities or near the coast, which always cause great problems for the local population.Even today, experts still study the causes of these earthquakes trying to predict where a new one will arise in the future and which are the places and countries where earthquakes are most frequent. Not only the frequency of earthquakes, but also their magnitude and many other aspects provide information about the disasters that an earthquake could cause in different locations.

This project aims to visualize earthquake data from 2000 to 2020 to help scientists and seismologists easily discover the correlation between earthquakes through many data visualization graphs and analytics techniques.

1 Introduction

Earthquakes have always been considered a natural disaster, bringing destruction and devastation in cities and countries.

Especially in the last century, where people grown exponentially in number and cities become ever more populated, an earthquake could cause a very high number of victims, without counting the damages to the buildings. For these reasons, the research of many subjects focused on trying to predict where and when an earthquake will arise. So, from engineering to science, from seismology to economy, the study of the earthquakes is conduced starting from the data analysis and the study of the past earthquakes.

Our project purpose is to help our users to better understand the correlation between earthquakes that arises in the last 20 years aiming to provide an easy

visualization of the data.
Of particular interest could be the monitoring of the various earthquakes of a specific country during the years or the comparison of two adjacent countries to see how, also between near territory the earthquake occurrences changes.

2 Dataset

The dataset is taken from the official site of the USGS, the United States Geological Survey. The USGS monitors and reports on earthquakes, assesses earthquake impacts and hazards, and conducts targeted research on the causes and effects of earthquakes.
USGS site allows to select a rectangular region of the world map and also to choose a starting date and an ending one. We selected the countries the european and the asian region for two reasons: representing the entire world map was too expensive in terms of computation resources and it tilted our systems causing an overloading. It also was difficult to have a clear idea of all countries showing them in a map because their representation would have been too small and qualitatively poor.
So we come with the choose of select european and asian regions because they are part of an unique plate, the Eurasia plate, that causes the current continent disposition and still today is cause of the majority of earthquakes in this zone. Finally, as USGS requires, we chose a magnitude of 4 or greater, considering a lower magnitude very low significant in terms of impact and effects over the territory.

2.1 Preprocessing

Original dataset is made up of many columns, but some of them are removed during the preprocessing part since not so much significant for our goal.
Moreover, the column relative to the Country, named "place", is processed to adapt its values to match with the same name provided in the map we used for the project. In the final part of the preprocessing, PCA algorithm is performed in order to calculate two PCA components and store them as two new columns of the dataset. PCA is performed on different features of the dataset (not all) that have been already standardized by removing the mean and scaled to unit variance. To avoid to overload the computation cost of the project, PCA is just computed once before the start and they will not change anymore then. So, at the end of preprocessing we have aroung 30 thousands tuples with the following columns:

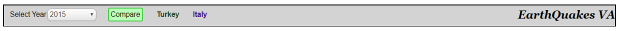
- **id**: specifies an integer number to identify an earthquake.
- **latitude**: specifies the latitude at which the epicenter of the earthquake is.
- **longitude**: specifies the longitude at which the epicenter of the earthquake is.
- **mag**: specifies the magnitude level of the earthquake.
- **magType**: specifies the algorithm used to evaluate the earthquake magnitude.
- **depth**: specify the depth in km of the ipocenter of the earthquake.
- **nst**: the number of seismic stations used to determine earthquake location.
- **gap**: the largest azimuthal gap between azimuthally adjacent stations, the smaller this number, the more reliable is the calculated position of the earthquake.
- **rms**: the root-mean-square travel time residual, calculated in seconds. It provides a measure of the fit of the observed arrival times to thepredicted arrival times for this location.
- **place**: the country where the earthquake happened.
- **type**: the seismic event type, could be "earthquake" or "quarry".
- **status**: could be automatic or reviewed based on the process that found the earthquake event.
- **PCA_Component1**: the first component computed by PCA algorithm, the one with the highest variance.
- **PCA_Component2**: the second component computed by PCA algorithm, the one with the highest variance that is orthogonal to the first, so it is linearly uncorrelated with it.

3 Visualization

3.1 Year Overview

3.1.1 Header

In the header of the page we can see the command line used from the user to change the values of the dataset. Through this the user can filter the dataset by year (from 2000 to 2020). When a country is selected its name is displayed in the header.Then there is the compare checkbox which allows for countries comparison and will be specified later in this paper.



Select Year 2015

Figure 2: Header filtering.

3.1.2 Geographic Map

The geographic map is the method used to visualize the position of all earthquakes in the world map. In the dataset, for each earthquake there latitude and longitude coordinates. These values are used to position the circles (the earthquakes) into the map. The map represenst only the Eurasia continent (Europe and Asia).

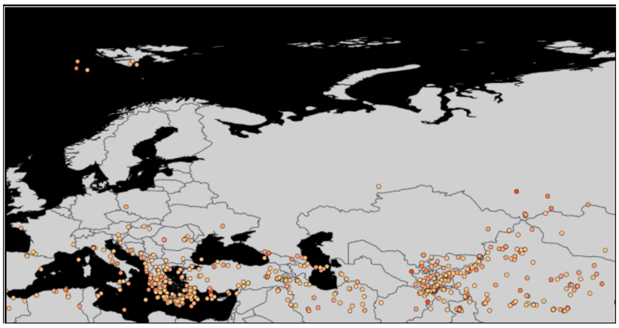


Figure 3: Geographic map.

There are different types of circles, each circle has a color that represents the magnitude level. The user can select a country and the system will allow him to zoom it and see the position of the earthquakes.

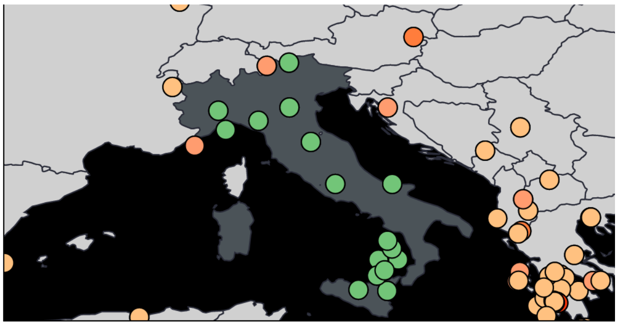


Figure 4: Country selection: Italy.

When the user selects a country the interaction with the system is the filtering of the dataset, all the graphs change their visualization for a single selected country and the page switches graphs from the scatter-plot and the bar-chart to other three graphs: the line-plot, the radar-chart and two box-plots. In this way the user can observe deeper information about a single country (Figure 9).

3.1.3 Parallel Coordinates Graph

The parallel coordinates graph is the graph that coordinates the other graphs. Through this graph the user can modify the values of the bar chart, scatter-plot and the other graphs. The user can observe in this graphs different attributes of the dataset, in particular, for each year can observe the values of the magnitude of all earthquakes in the Europe and Asia , the depth, the time of start, the latitude and the magnitude type.

The user can interact with the graph and change the filters of visualization on the dataset. He can select a number of lines of the graph and through this (brush selections on each axis), he changes also the visualization of all the other elements of the page. The lines of the graph represent the earthquakes while the colors represent the level of the magnitude for each country.

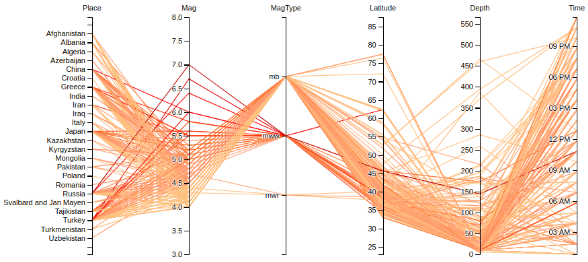


Figure 5: Parallel coordinates chart.

3.1.4 ScatterPlot

A scatterplot is used to display elements of the all countries corresponding to the selected year. Each point of the scatterplot represents an earthquake and its position is given by the first two components of the dimensionality reduction produced by PCA algorithm. What emerges is that points are quite all clustered and only some of them are far from the main cluster representing the so called outliers. So the scatterplot helps the user to compare earthquakes observing the distance from each other in the graph, to find out the outliers and it also offers the possibility to brush an area to select only some points of interest to point them out in the map and the parallel coordinates chart.

Finally, a color encoding is given to each point to indicate the earthquake magnitude.

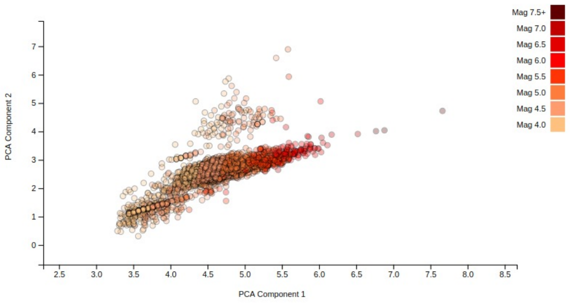


Figure 6: Scatterplot.

3.1.5 BarChart

To show the frequency with which an earthquake is perceived in a country for the selected year, a bar chart is created.

As just said it shows the number of earthquakes for each country, that is indicated by the y axis, relative to the frequency that is also specified on the top of each bar. To avoid confusion and make the barchart difficult to read, we grouped all those countries which have been hit from an earthquake less than four times that year in a single bar named ”Others”. However, if an user is interested in those countries, he can check the apposite checkbox to display them modifying the chart.

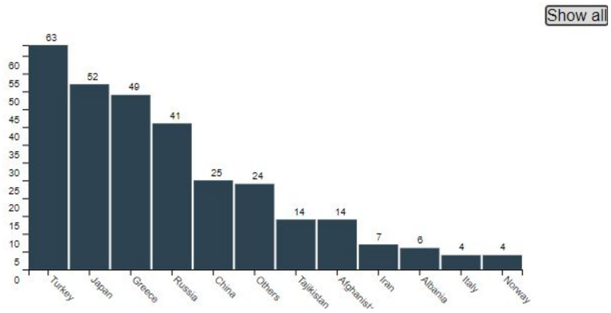


Figure 7: Bar chart.

Finally, we offer the user the possibility to click on each bar to show new inner bars displaying the number of earthquakes for the different magnitude values: the colors used for the magnitude are the same of the ones used in the scatterplot and also a legend is shown to better help user understand level of magnitude.

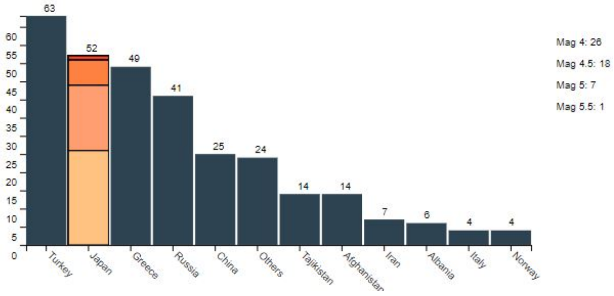


Figure 8: Selected bar: magnitude levels percentages.

Finally, using the parallel coordinates chart, the barChart could be filtered displaying only the selected rows of the dataset.

3.2 Country Overview

The page (Figure 9) displays now the selected country on the map, zooming and pointing out on it, the radar chart, the box plot (depth and magnitude), the parallel graph for a single country and the line plot. When a country is selected, to differentiate and better evidentiate it from the others, the colors used to represent it in the various graphs is a scale of green (violet if it is the second country selected for comparison).

3.2.1 LinePlot

When a place is selected, the entire page shown changes.

This line plot is perfect to show the evolution of the frequency of earthquakes, for the selected place and year, during the months thanks to the path tha evidentiate in each month the level of the frequency, fixed with a circle.

Again the interaction is obtained through the parallel filtering of the earthquakes for the country, pointing out only some data of interest.

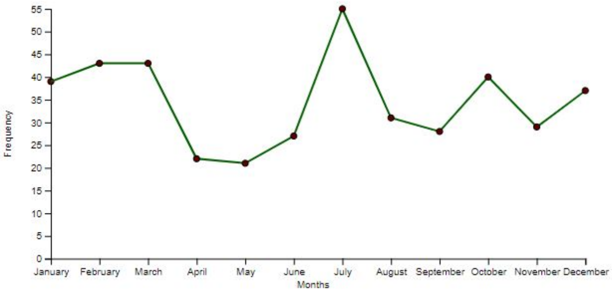


Figure 10: Line plot.

3.2.2 BoxPlot

For some distributions/datasets, you will find that you need more information than the measures of central tendency (median, mean, and mode). So we used the box-plot. A box plot is a graph that gives you a good indication of how the values in the data are spread out.

We used two box-plots, one for the magnitude distribution and the second for the depth distribution of the earthquakes.

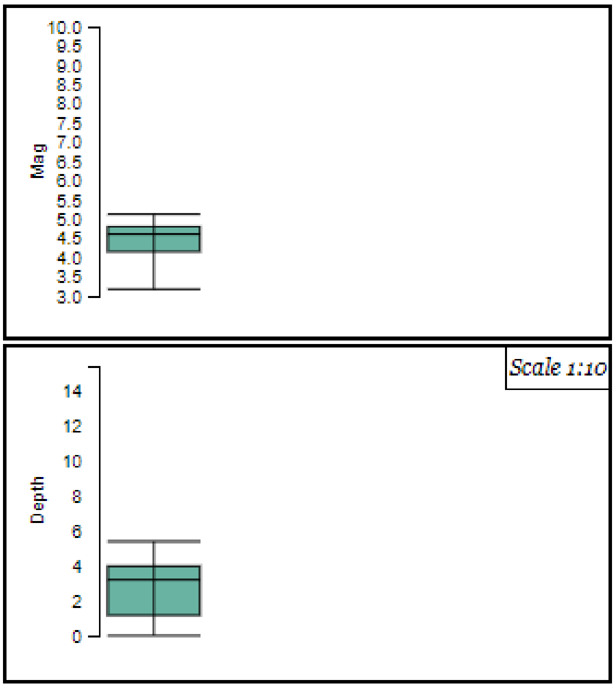


Figure 11: Magnitude and Depth Box plots.

3.2.3 RadarChart

A radar chart is used to display multivariate in two-dimensional. The quantitative variables used are five and each of them is represented with a axes starting from the point. The variables used are: nst, magnitude, depth, gap, rms. The graph is useful to better make directly visible the correlation between these variables in the sense that the area covered from the obtained graph for the selected year and country indicate the maximum level of each feature.

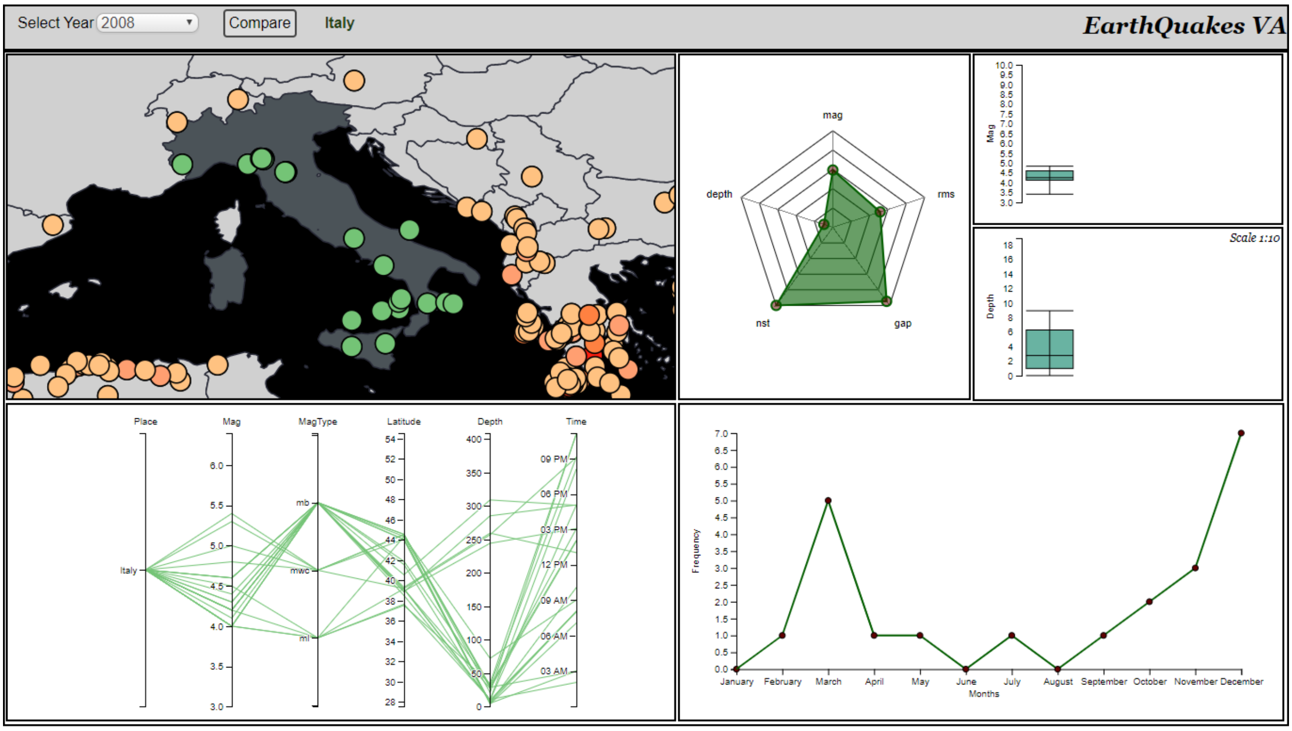


Figure 9: This picture represents a country selection (Italy).

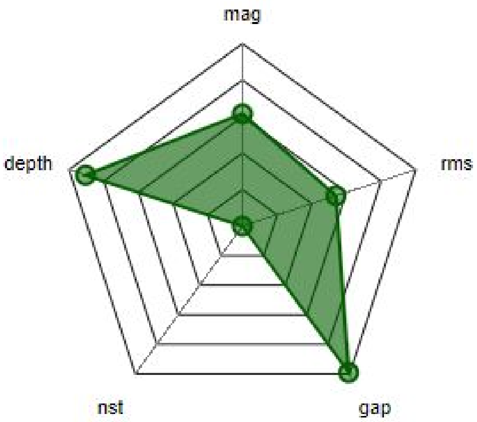


Figure 12: Radar chart.

As we can see in this image, the chart shows how for a high value of the depth the gap value is also high, indicating how for earthquake happened in deeper points, the accuracy of the measure is lower.

3.3 Country comparison

The country comparison (Figure 13) is a tool used to compare two countries and in particular the values of their earthquakes (by year).
To do this the user can select the comparison button in the header and then select two countries on the map. The parallel graphs, the radar chart, the box plot and line plot will show a comparison between them (the two country). In the graphs the first country will be represented with the green color and the second country with the violet color. What is interesting is the comparison between adjacent countries or near ones that better evidentiates correlation and differences between the two.

4 Analytics

4.1 Interaction

User interactions with our project are facilitated through the use of interactive graphics and with the presence of an header that allows to filter the dataset according to an year. These interactions lead to deep anaysis of the data since the user can select, filter, brush and focus on specific type of earthquakes. So, let's see more in detail how user can interact with graphs and header:

- **Header**: it provides a select tool for choosing an year to display all earthquakes arised in that year. It also has a checkbox where the users can activate the "comparison" mode and have a feedback of which country has been chosen.
- **Parallel coordinates**: it has many axes, each one representing a feature, that allow the user to brush them to select an interval of data whose he is interested in.
  The data selected will be filtered also in any other graph of the page, so it could be considere as the control center of the interaction with the other parts of the project.
- **Map**: the map allows to pan and zoom for better analyze countries and earthquakes.
  Moreover, clicking on a country, the user can deeply analyze data of that country.

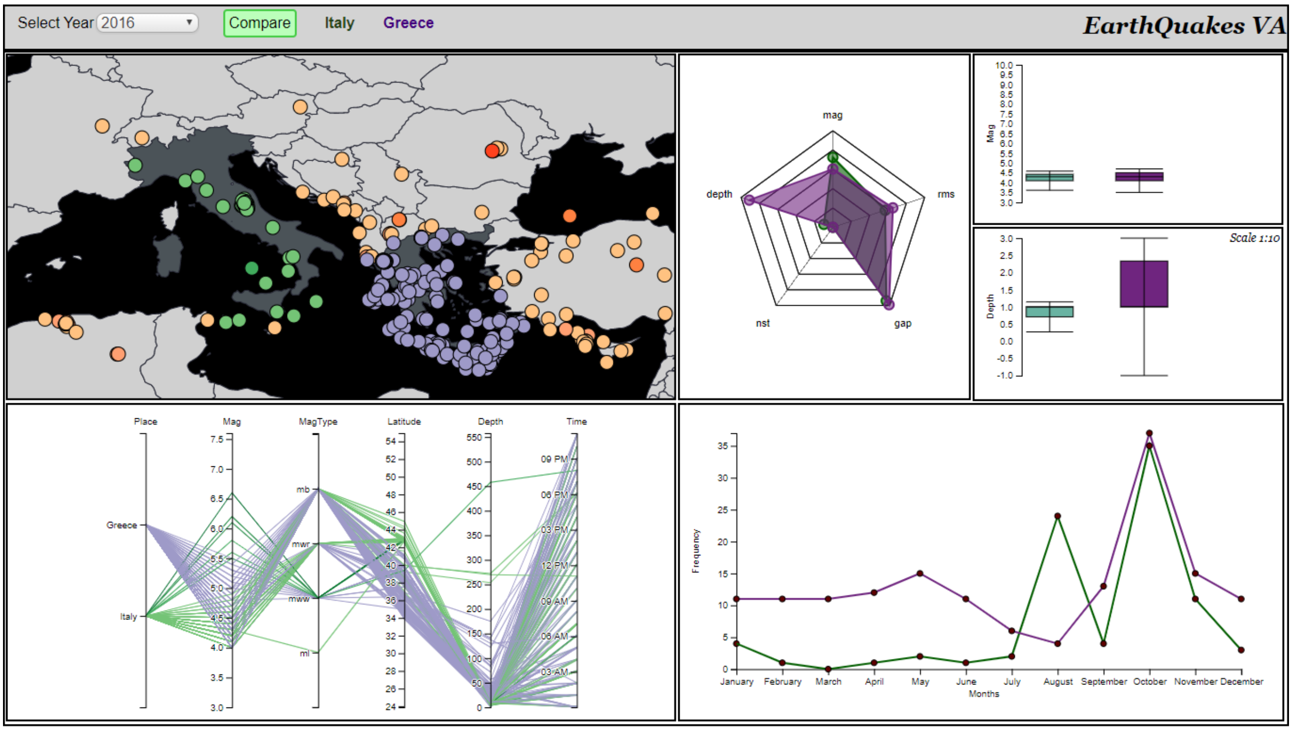


Figure 13: Comparison mode.

- **Scatter Plot**: here the user can again brush an area in which, the data containes, are pointed out and this is also reflected in the parallel chart, where the correspondent paths are highlighted (and so it is created a bidirectional interaction), and in the map, where the correspondent points changes their color.
- **Bar Chart**: the bar offer the possibility to click on it to analyze different levels of magnitude frequency for the relative country

Other graphs not directly influence the others, but are only influenced by the filtering or computation of other graphs and so they are intented to make the information and data easier to read and to analayze them.

4.2 Analysis

Analytics part involves some of our graphs and also the computation of other informations. Let's see them in a list:

- **Frequency**: both for barchart and lineplot are computed different type of frequencies. In the lineplot is computed the frequency as the number of earthquakes that for that year and that country will arise in each month. For the barchart there are computed both the frequency of the earthquakes in each country for that year and both the frequencies of each magnitude level in each country for that year.
- **Mean**: in the radarChart, for each of the represented feature of the dataset, it is computed the mean for the selected country in that year.
- **Scale**: many scale adaptation are used to adapt data to graphs, for example, to display data on the depth boxplot, values are normalized according to te min and the max value of the depth, also, for each boxplot in the project, quantiles are computed every time the selected country changes.
- **Other computation**: for example we had to recompute, each time the selected year or country changes, all the views, the size and the range of the axis.

5 Discussion of our approach

Through this visualization we allow the user to have a good analysis of the earthquakes in Europe and Asian. We create a system based on graphs to display the attributes for each earthquake in a clear way.
The system is composed of eight graphs and the main goal is to compare earthquakes for different country and to determine the greatest concentration of these in the various territories.
The analysis of these events has been divided into ranges of years and for each year it is given the possibility to filter the dataset to understand each possible variations in terms of: magnitude, latitude, longitude, depth, nst, gap, rms, place, type and status.
In the end the main method used to filter the dataset is the creation of a parallel graph and scatter plot which interact with all graphs in the system.

6 Conclusion

This system represents a good method to the analysis and visualization of earthquakes, it offers many interactions to change the data and correspondent visualization.
In the future an additional goal that can be implemented is the possibility to have a feedback from the users and the possibility to add information in real time to improve the system of visualization of earthquakes and their attributes.
It would be perfect for users to understand also the gravity of each earthquake in terms of human losses and economic for each country.
At the moment only two countries can be compared, this for a good visualization. In the future it would be perfect to find an approach to compare the attributes of earthquakes for more than two countries without worsening their visualization.

7 Bibliography

References

[1] ”Course’s material”

[2] ”d3.js Data Driven Document” `https://d3js.org/`.

[3] ”d3.js Graph gallery” `https://www.d3-graph-gallery.com/`.

[4] ”USGS Earthquakes” `https://earthquake.usgs.gov/earthquakes/search/`.