Visual Analytics for data quality

Alessandro Migliore
Daniel George Gavrila

Introduction

The goal of this project is to illustrate the data quality of the ETER (European Tertiary Education Register) database in which are present lots of data relative to European Institutions that provides Tertiary Education.

More deeply, we want to allow the user to make comparison between Institutions or Countries simply analyzing their parameters and mostly their quality in order to let find partners in research, institutions in which we can invest and many other stuffs based on the parameter contained in the dataset.

The software can be also a starting point for ETER to improve the data contained in the Database.

Outline

Dataset
Technologies
Visualizations
Case Study
Related Works

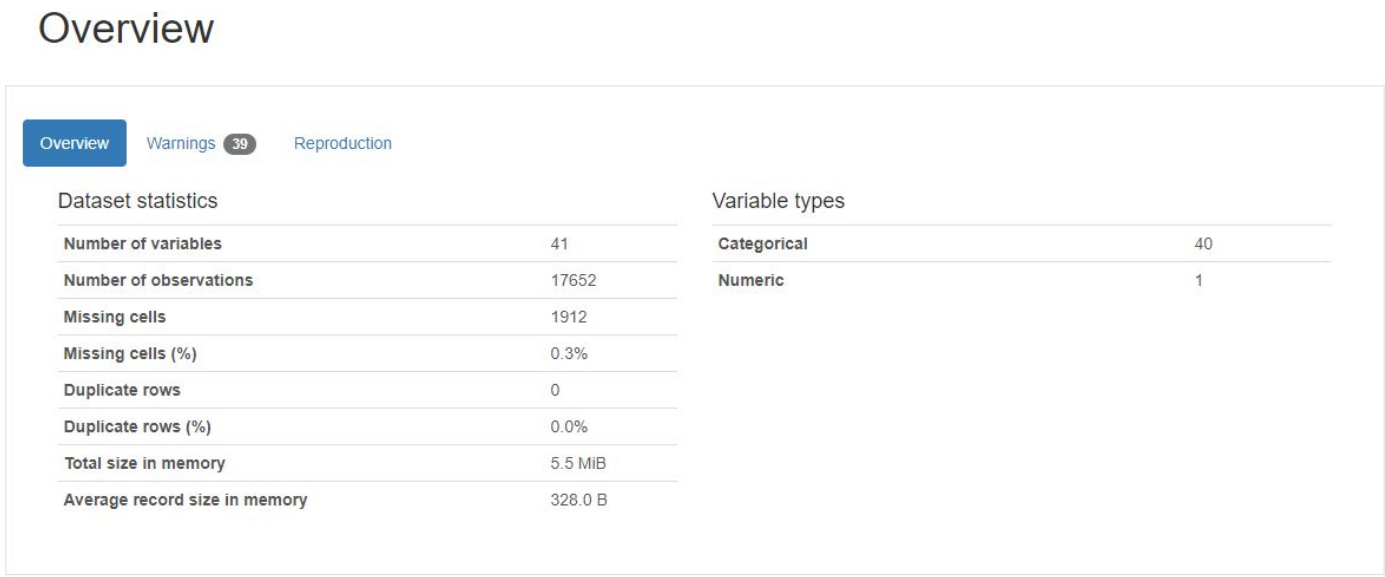
Dataset - Starting point

European Tertiary Education Register (ETER) database

We have selected the core of the dataset that contains: basic institutional descriptors , data about education , research , finance and others of interest.

- **Format :** xls
- **Dimension :** 5.1 Mb
- **Years :** 2011 - 2017
- **AS rule :** #tuples * # dimensions = 723.732

Overview of dataset with Pandas-Profiling

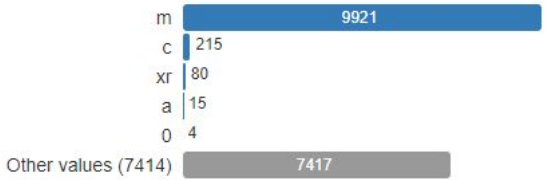


Overview of a parameter of interest

total_current_revenues(P…
Categorical

HIGH CARDINALITY

Distinct	7419
Distinct (%)	42.0%
Missing	0
Missing (%)	0.0%
Memory size	138.0 KiB



Toggle details

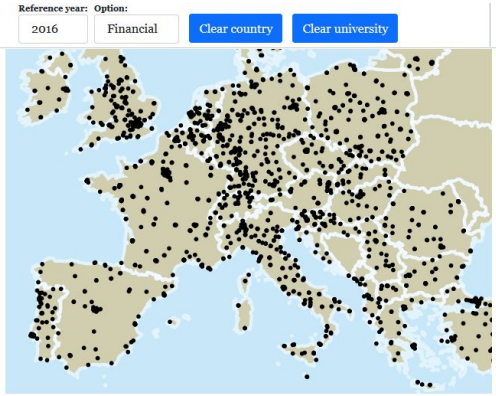
Overview | Categories

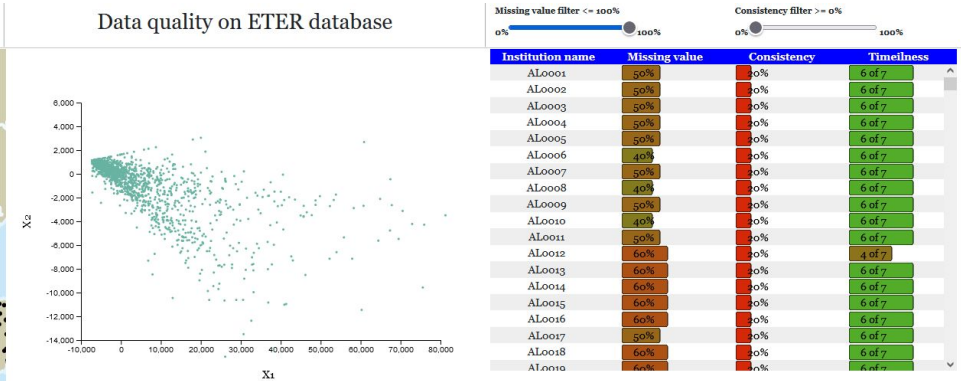
Value	Count	Frequency (%)
m	9921	56.2%
c	215	1.2%
xr	80	0.5%
a	15	0.1%
0	4	< 0.1%

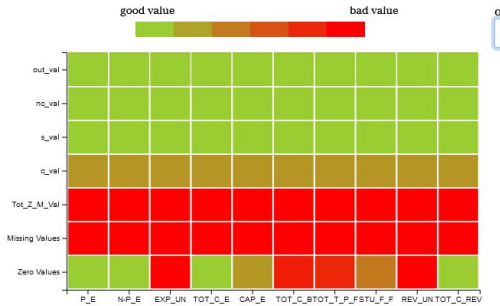
Technologies

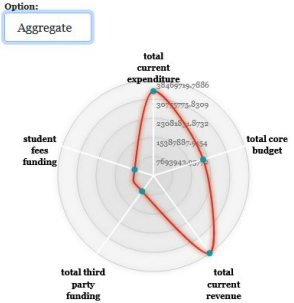
- **Python**
  - Data manipulation
- **D3JS + JS + Bootstrap**
  - Visualization

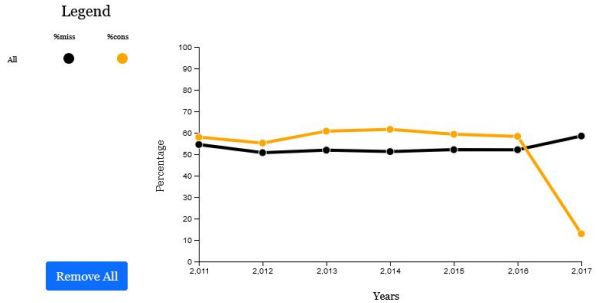
Quick Overview











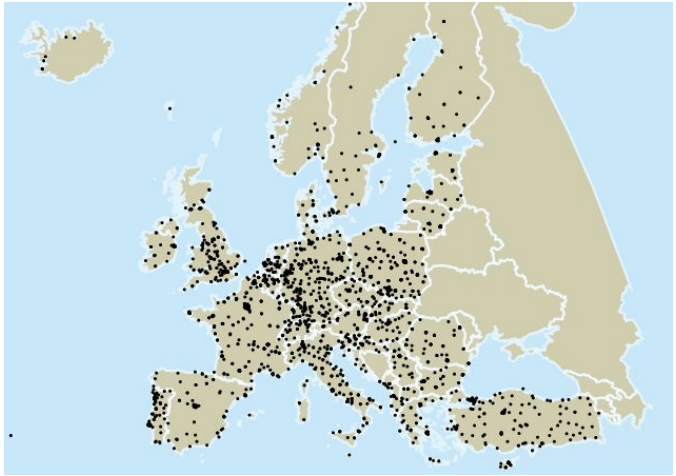
How it was conceived

The entire system was thought mainly to make comparison between institutions or country. Brief description of the purpose of every view:

- **Map -** to allow user to select institution or country of interest
- **PCA -** reduce each tuple taking the most important values to 2-dimensions
- **Table lens -** shows data quality parameter in tabular form
- **Quality metrics matrix -** display a more detailed data quality of the dataset.
- **Radar-chart -** the graph shows the pure values of the selected parameters and institution.
- **Linear time chart -** trend of data quality parameter per year

Visualization - Map

The first visualization is a simple european map that shows the geographical position of the institutions in relation to their country.



Visualization - Map

The map was meant as the view used for selection, in order to allow the user to select country or university of interest and to study their data quality and their pure data.

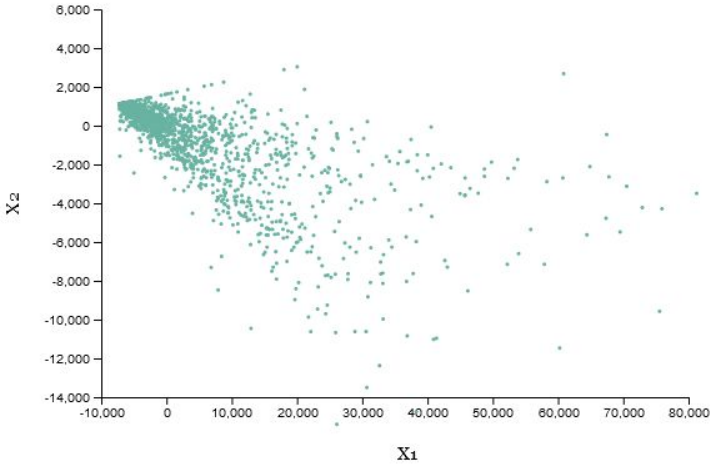
It’s also thought to visualize the institution for which there are data in the DB per year.

It’s completely zoomable and interactive.

Visualization - PCA

The initial idea was to cluster the institutions with the Kmean, but given the nature of the database, it did not report satisfactory results, so we simply opted to use the PCA and reduce each tuple taking the most important values to 2 dimensions and display it.

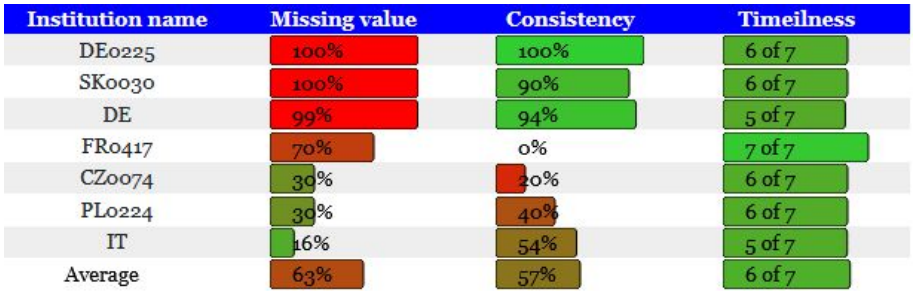
In particular, passing the mouse over the institutions in the PCA it is possible to see the trend of the whole nation.



Visualization - Table lens

Table lens allow the user to visualize data quality parameter in tabular form for each institution or country. Moreover it allows to order the lines for each column and to have an overview of the trend of the selected value.

The last row represent always the average of the selection.



Visualization - Table lens

Every line has a bar that goes from green that represents good value to red that represents bad value. This choice was made because in the human perception green is always associate with good and red with bad values.

Moreover it allows to order the lines for each column and to have an overview of the trend of the selected value.

The last row represent always the average of the selection.

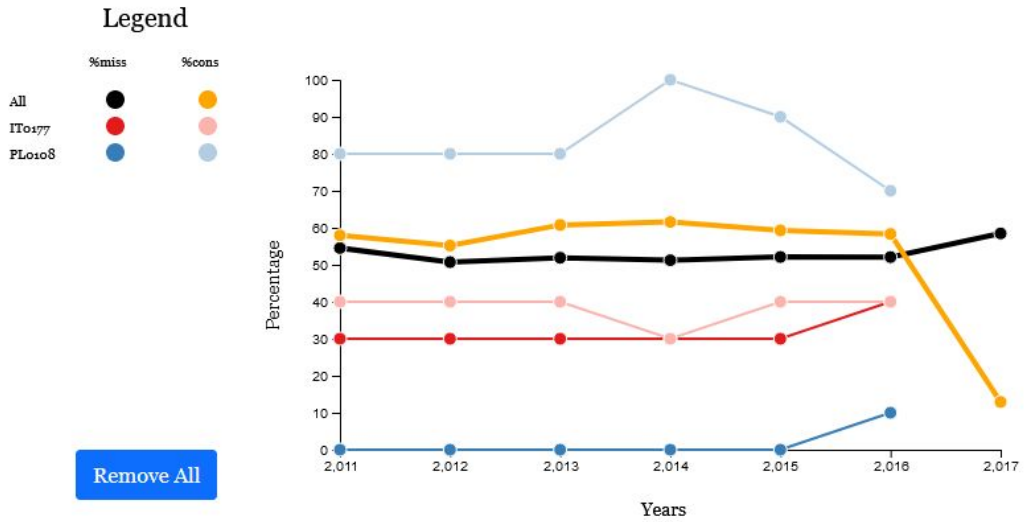
Visualization - Linear time chart

Axis:

X: Years (2011 - 2017)

Y: Percentage of the value

Every line represent the annual trend for the selected institution, one is made for the consistency and one for the missing value percentage.



Visualization - Linear time chart

The time line chart was designed to compare the annual trend based on what has been selected.

There is a dynamic legend that creates an association between what has been selected and the colors in the graph to allow better usability.

And a button to remove the trend of the entire dataset if it is not of interest.

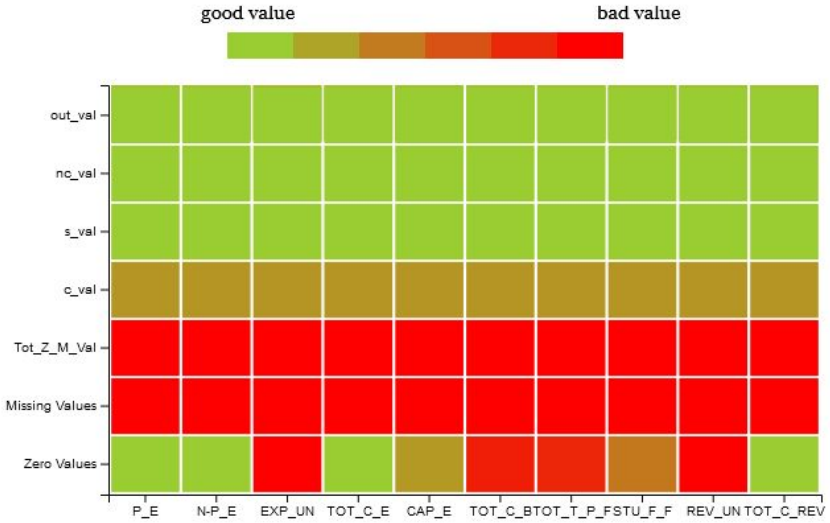
Visualization - Quality metrics matrix

Axis:

X: parameters selected by filter (Financial, Educational, Other)

Y: parameters of interest on data quality

This graph allows us to visualize the quality of the data on the entire dataset. As you can see from the legend, each value is linked to a color ranging from green, if it is the given quality of the data is excellent, red instead , if the data quality is not good.



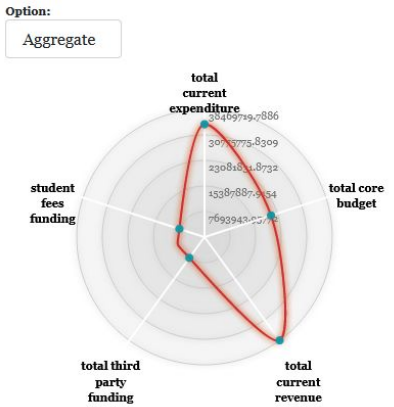
Visualization - Radar Chart

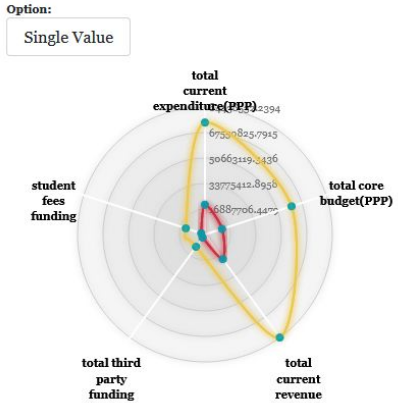
Axis: parameters selected by filter (Financial, Educational, Other).

The user can select two options :

Aggregate, the graph shows the pure values of the parameters of interest and a total average of the selected parameters, going to perform a more in-depth analysis on data quality. In fact, by selecting a state or a single university we can compare the trend of the data with the total average.

Single value the analysis is aimed at a comparison between individual universities and between states.





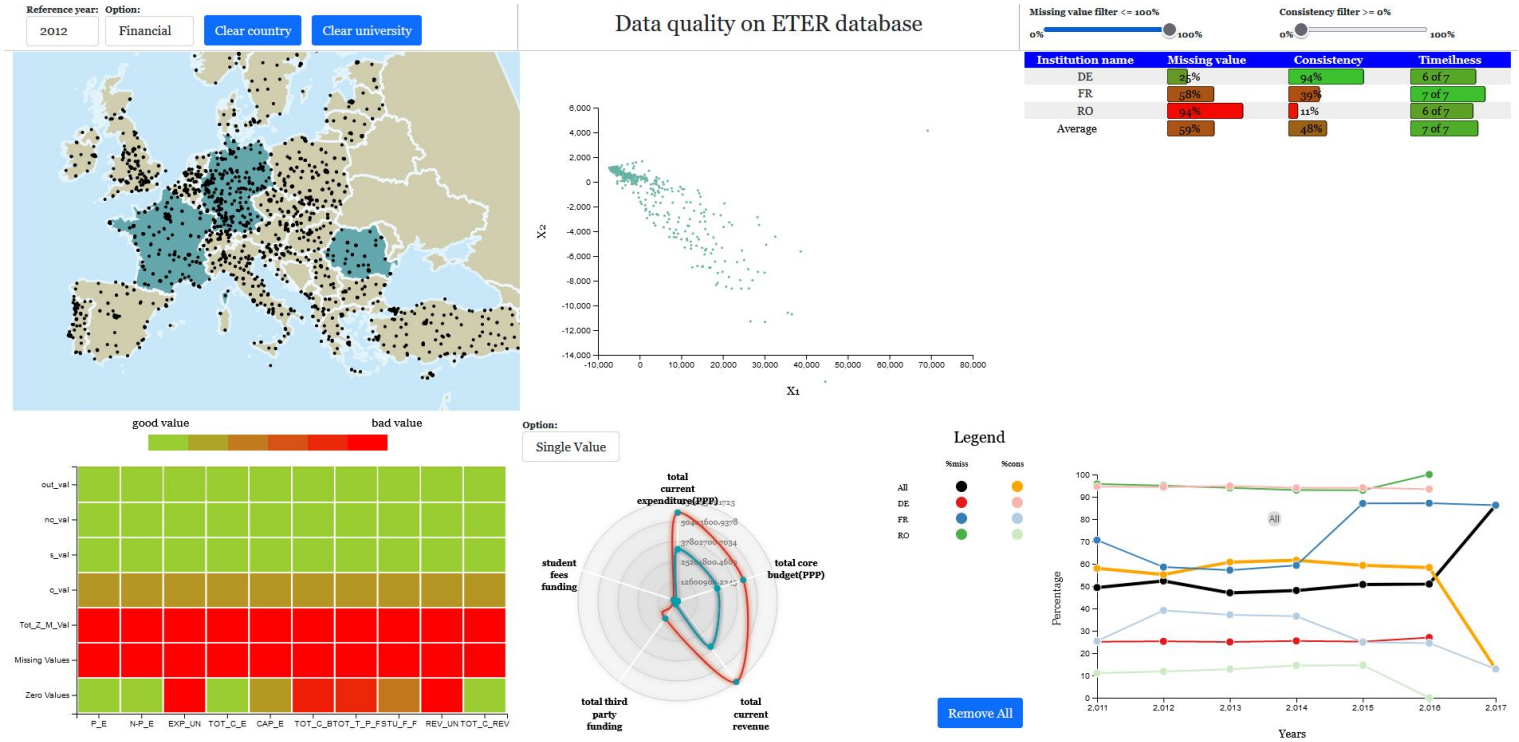
Filters

With the top bar the user can filter on:

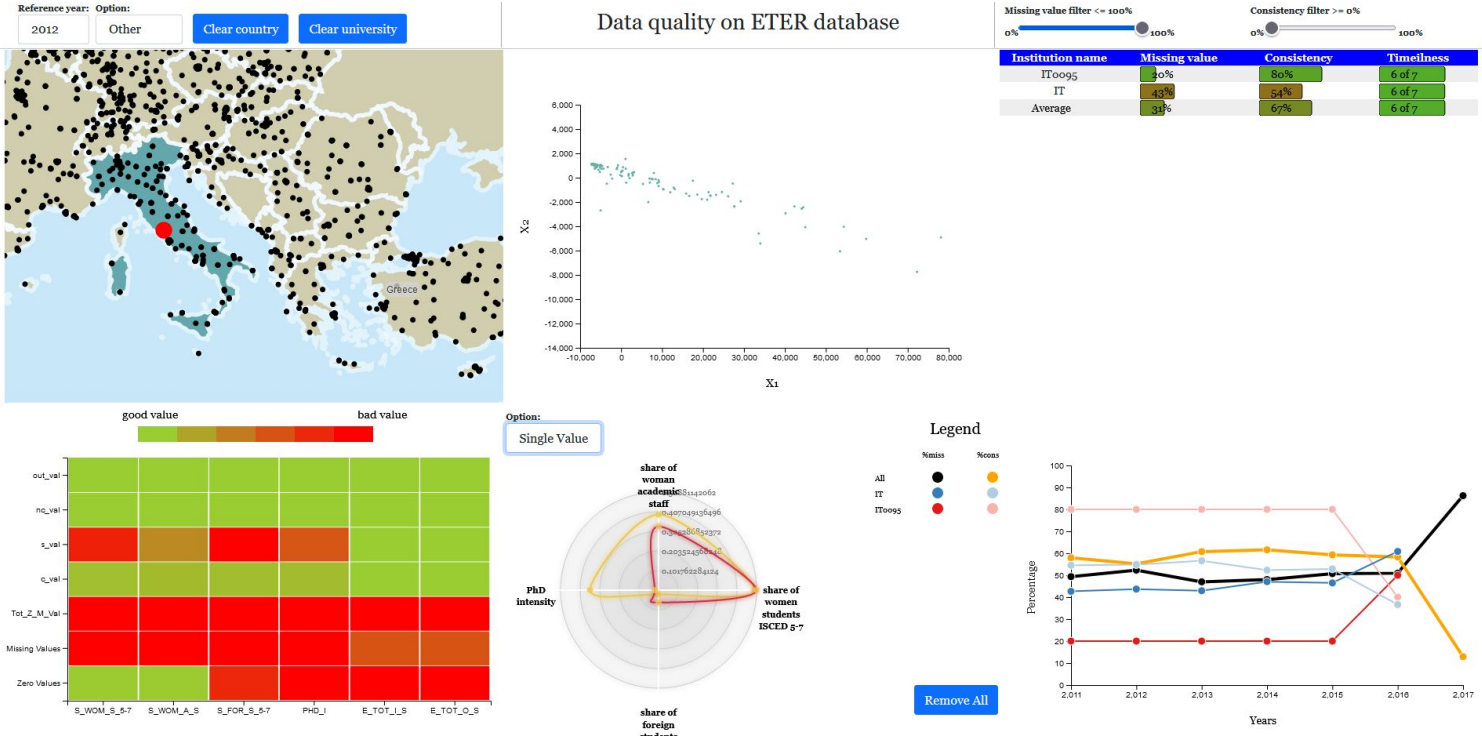
- year of interest
- data of interest (financial, education, other)
- on missing value %
- consistency %

There are also two buttons to clear the selected country and university.

1. Case study - Where should I invest as a company?



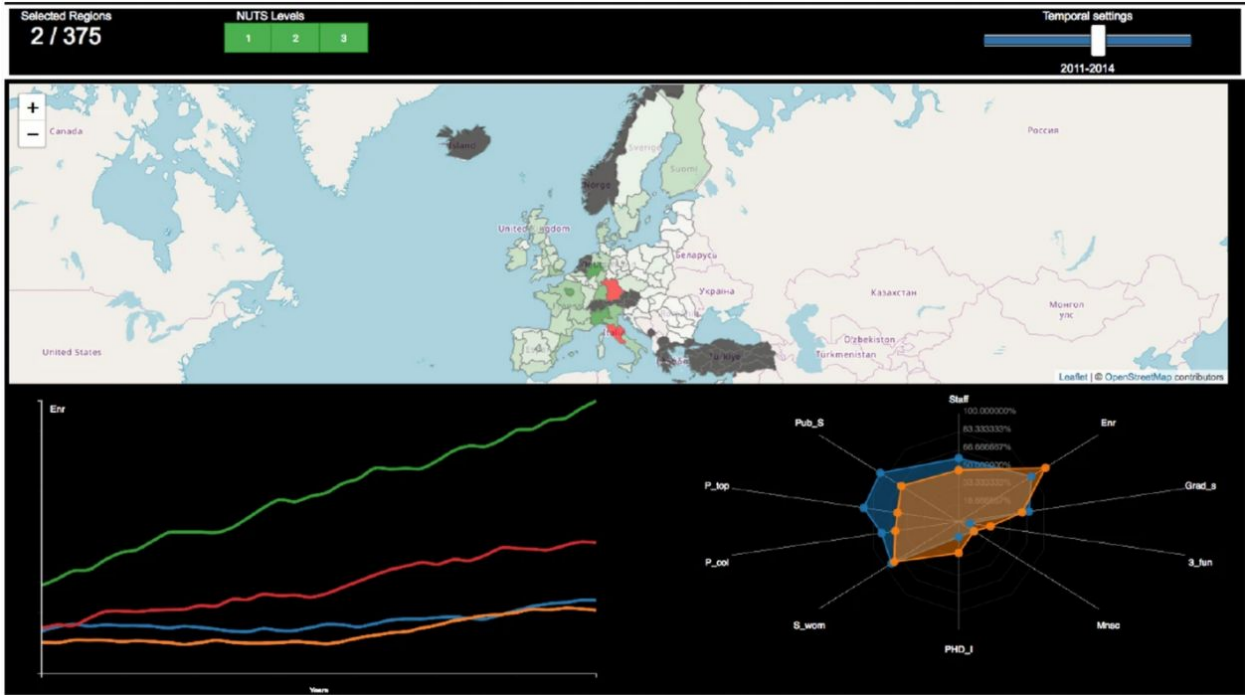
2. Case study - Where should I be a researcher in Italy?



Related works - ETER data quality

In the literature there are many scientific articles related to data quality, with different types of analysis and methodology, in particular the closest to our work is "A tailor-made Data Quality Approach for Higher Educational Data", of which however there is no visual analysis system.

Related works - Visualization



Thanks for the attention