Visualizing Multidimensional Data

Anton Arbring

Abstract— This is a paper on information visualization. A data set containing 260 countries and 43 different dimensions was visualized using parallel coordinates, a choropleth map, scatter plot and a data mining algorithm called principal component analysis.

Index Terms—Principal Component Analysis, Parallel coordinates, Scatter plot, Choropleth map

1 Introduction

A commonly occurring issue in information visualization is how to present multi dimensional data in 2 or 3 dimensions. This paper will discuss a few different measures to manage that.

A web application has been developed using a JavaScript library called d3 [2] and Bootstrap [1] to visualize . The data is presented using scatter plots, parallel coordinates and a choropleth map.

This is a paper describing a project performed in the course ”Information Visualization” - Linkp̈ing University, Spring 2014

2 Background

Some time was spent looking on different data sets. This project was performed during the second half of a course, following a lab series on information visualization focusing on web visualizations.

3 Data

The data set contains information about 260 different countries in 43 different dimensions. The data contains a wide range of categories such as geographical data (area, population), energy usage data (electricity, oil) and economic data (Gross Domestic Product, Inflation Rate) etc.

The data is not complete (i.e. some dimensions have a lot of missing data). This causes problems when handling the data by calculating mean values, variance and performing other kinds of analysis.

Each row of the data set represents a country and each column a data dimension. All values following the country name (which are strings) are doubles, in varying size.

4 Method

To visualize this data we used three main views. The first one contains a choropleth map and parallel coordinates. The second view is the scatter plot view and the last one is the result of a principal component analysis, also demonstrated as a scatter plot.

4.1 Parallel Coordinates and Choropleth Map

When handling geographical data such as countries, a map is a very intuitive way to show data. The user is given eight drop down menus to choose what dimensions to plot in the parallel coordinates. The first menu will determine the coloring of the countries.

There are nine possible colors, fading from a pale yellow to dark red. To make sure that outliers would not make all the other countries have the same color; there are always four colors representing the countries below the mean value and four values to represent the ones above.

It is possible to filter the parallel coordinates and a selected line gets highlighted in the map and vice versa.

4.2 Scatter Plot

Another part of the application was the scatter plot. There are a few different ways to be able to show more than 2 out of 43 dimensions at once. In this application, one has the ability to choose what dimension to decide radius and color, where the color is plotted similarly to the drawing of the map and a bigger radius represents a higher value.

In case of missing data, there are a few different measures to take. To minimize confusion in the scatter plot; it was decided not to draw the dot if the y, x or radius dimensions data values were missing, but the color would turn gray if that data is missing. The missing data fields could have been replaced with some average of the rest of the data but that would imply a clustering around the mean value which might not be the actual case.

4.3 Principal Component Analysis

Principal component analysis or PCA for short was a very useful algorithm when one wanted to visualize data in many dimensions. Its main objective was to find two orthogonal vectors which together could capture as much of the total variance as possible.

The first step in the PCA was to standardize the data, which basically sets every dimensions mean value to zero, to be able to get relevant results.

The covariance of each pair of dimensions was then determined resulting in a covariance matrix. This matrix had eigenvectors representing the possible principal components and eigenvalues which was increasing with the amount of variance obtained from each of those components. [3]

5 Implementation

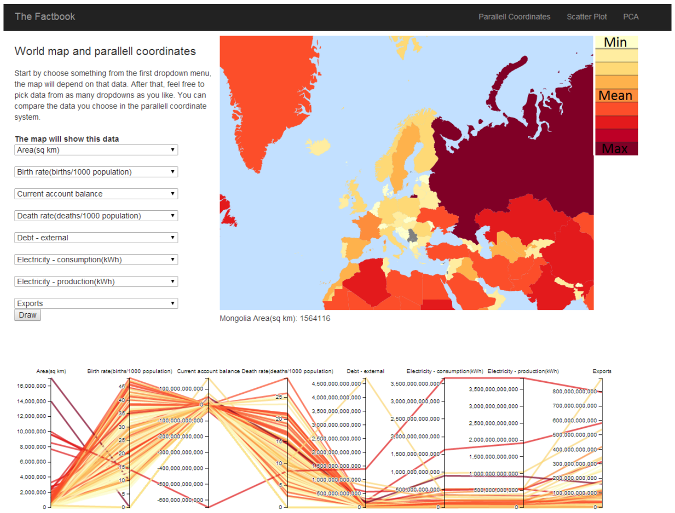
The algorithms were built and run on average laptops and the application could be run on almost any computer with an Internet connection and a browser that supports JavaScript.

For the HTML, CSS and JavaScript code, Sublime Text editor was used and this paper was written using a LaTeX environment called TeXstudio.

6 Results

Fig. 1 shows the first view in the application. The linking between the components make the application quite intuitive and a map is very good if one wants to locate a specific data fast. The value for the country is shown as a tooltip when the mouse hoovers over it.

- *Anton Arbring is with Linköping University, Sweden, e-mail: anton@arbring.se.*



World map and parallell coordinates

Start by choose something from the first dropdown menu, the map will depend on that data. After that, feel free to pick data from as many dropdowns as you like. You can compare the data you choose in the parallell coordinate system.

The map will show this data

Area(sq km) ▾
Birth rate(births/1000 population) ▾
Current account balance ▾
Death rate(deaths/1000 population) ▾
Debt - external ▾
Electricity - consumption(kWh) ▾
Electricity - production(kWh) ▾
Exports ▾

Mongolia Area(sq km): 1564116

Fig. 1. A view of the web application

After a PCA was performed on data from the 43 dimensions, a 2D scatter plot is found in Fig. 2. This algorithm turned out to be a very good match for the chosen data set where approximately 99 percent of the variance could be shown in a two dimensional space. Outlying countries such as the United States, China and Russia does however make the rest of the countries look fairly similar. The fact that a missing value was replaced with the mean of the rest of the data might also have made countries with a lot of missing data seem very mediocre.

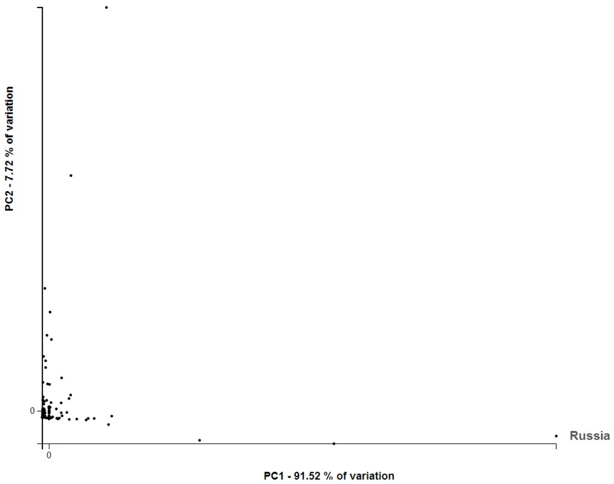


Fig. 2. Principal Component Analysis plot

The scatter plot was used because it is very easy to understand, even for people not used to graphs and diagrams and it is effective when showing the relation between two dimensions. In Fig. 3 the relation between GDP and number of Internet users is easy to comprehend

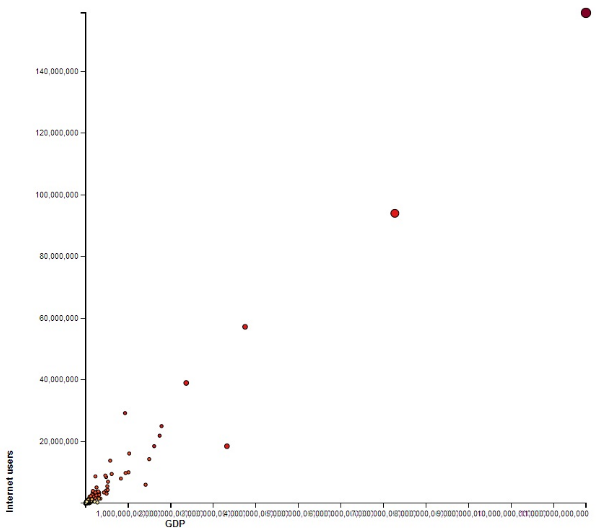


Fig. 3. Scatter plot

7 Conclusions and Future Work

The team was satisfied with the resulting application considering the limited time that was given to the project. Some features could have been improved such as the ability to chose a subset of countries or dimensions to create a PCA based plot.

A user that is not familiar with the different plots and figures might have a hard time understanding the parallel coordinates and the PCA plot. The filtering possibilities in the parallel coordinates are for example far from obvious. This could be explained in a tooltip or maybe in another view called ”How to make use of this application”.

This application could definitely benefit from a usability study. It is easy to miss obvious design flaws as a developer.

References

[1] Bootstrap.

[2] d3.js - data-driven documents.

[3] D. H. Jeong, C. Ziemkiewicz, W. Ribarsky, and R. Chang. Understanding principal component analysis using a visual analytics tool.