Global population from 1960 to 2016

Process Book

Contents

Yanyan Zhang
e-mail: u1127804@utah.edu
UID: u1127804

Lin Yan
e-mail: lin.yan@utah.edu
UID: u1140293

The project repository is here.

1 Introduction

1.1 Motivation:

World population has changed over years in many respects such as "Age dependency ratio (% of working-age population)", "Age dependency ratio, young", "Fertility rate, total (births per woman)", "Life expectancy at birth, female (years)" and "Female population 00-04". Users may need to study these changes to predict future population so that they can plan ahead to provide good education, health care and other services in the future. Different aspects of the world population data are closely correlated to different types of services. For example, studying "Fertility rate, total (births per woman)" and "Life expectancy at birth, female (years)" may tell users how much health services for childbearing will be needed in future years.

There were also interesting changes in values of some attributes throughout years. For example, in the US "Age dependency ratio, young" changed from ~51% in 1960 to <29% in 2016, and "Fertility rate, total (births per woman)" dropped from ~3.6 to ~1.8. In China "Life expectancy at birth, female (years)" increased from ~45 in 1960 to ~78 in 2016. We are interested in visualizing all these changes across countries and years. We also would like to find out explanations (e.g. from economical and political points of view) for these changes if possible. Studying the potential explanation that causes these trends could help users predict the global population in the future and make plans that could meet the current or future needs of people according to different distribution of population.

1.2 Data:

Downloaded from a Kaggle link.

This database presents population and other demographic estimates and projections from 1960 to 2050. They are disaggregated by age-group and gender and cover approximately 200 economies. We can extract all the information we are interested in from this data set,

such as “Population Growth(annual %)”, “Population, Female(% of total)”, “Population Total”, “Work-age Population”, “Birth Rate”, “Death Rate”, “Fertility Rate”, and “Life Expectancy”. All of them were stored in one csv file, in the size of 44.2 MB.

1.3 Task:

1.3.1 The primary questions we are trying to answer

1. How do values of certain attributes of global population change over the years?
2. How do we compare changes of certain attributes over the years among selected countries?
3. How to visualize the population data of all or selected countries in a fixed year?
4. How to visualize the population data of a certain country in all or selected years between 1960 and 2016? Are there any correlations among attributes that can be observed from our visualization charts?

1.3.2 What we would like to learn and accomplish

1. Give users a global view of population change. Provide users tools to track attributes they may be interested in from a global point of view.
2. If we can find similarities in the change trends of population among certain countries, users could try to figure out explanations that may lead to these similarities.
3. Users may be interested in certain attributes of certain countries that are correlated to the demand of certain services. Studying the changes in these attributes over the years will provide quantitative information that may help them determine their future plan.
4. If we could match economical or political changes of countries with their population changes in certain years, our visualization could help us explain how economical or political changes could affect global or specific regional population.

1.3.3 Solution overview

· **World map**

Give users a global view of population information according to the attribute and the year users select.

· **Cluster**

Group countries with similar patterns and properties in a fixed time period (10 years) according to the specific attribute users are interested in. Users can explore potential reasons for similar patterns within a cluster group.

· **Line chart**

Users can select multiple countries they are interested in by clicking them on the world map, and make a comparison among them according to the selected attribute. They can also remove lines that are no longer interesting by double clicking them one by one.

- **Parallel Coordinates**

Display all the attributes of the selected country from 1960 to 2015. Users can observe normal or abnormal trends in the attributes from this view. To see the detail discussion about our task and solutions, please refer to section 4 of this document.

1.4 Users:

The target audience could be the general public, the ones who are interested in global population change and specific information. It also could be city planners who would like to rearrange resource based on population in their plan regions. Researchers and educationist in related field may get inspirations from this visualization as well.

1.5 Related Work:

Both members in our group come from China, one of the countries with the largest population in the world. Population is always a hot topic in our country. We believe that developing simple demographic profiles to understand past and current trends of global population is a realistic and interesting topic. Moreover, during the visualization class, we have been exposed to many interesting views and methods for visualizing data, such as table, map, parallel coordinates. By choosing this data set, we can implement some visualization methods we are interested in.

2 Process

2.1 Data Processing

There are several problems we solved during data processing.

Firstly, there was missing data. We tried to interpolate data to fill up small holes in data (during drawing world map) or drop a small amount of lines in which there are too many missing attribute values (during data analysis).

Secondly, the data set is relatively large compared to data sets we used in our previous homework assignments. In order to improve the speed of visualization, we have tried to pre-process the data using Python during data analysis. For this project, since we focused on a few attributes, we used Python to extract data of those attributes from the original data.

Thirdly, we had decided what attributes we want to visualize. Although there are many attributes we are interested in, we cannot visualize all of them due to the limit of time and space. For now, we selected “Population Total”, “Birth Rate”, “Death Rate” and “Life expectancy” for the world map.

Moreover, for clustering, we need a computational method which involves topological, machine learning and dimensional reduction ideas. We decided to pre-process our data in Python, and use Python results as JavaScript input for visualization.

Finally, because there are several different views in our interface, which require different formats of data. And because we do different part separately, we need convert the data to the desired format that can be easily used across the charts.

Commented [yz1]: Did we do this?

2.2 Design Evolution

All attributes we are interested in are options in a selection bar. For different attributes, we show data in a world map with each country colored accordingly. A color bar is displayed beside the world map. There is also a year chart in which users can select a specific year, and a year slide bar with which users can see transition throughout years.

Users can select one country by clicking it on the map. A line chart of year vs. the selected attribute for the selected country will show up next to the world map. Clicking another country will add another line to the line chart, and so on. This line chart shows comparison among selected countries. If users would like to show comparison in another attribute among the same countries, they can keep the selections of countries and select another attribute in the selection bar. By hovering over a line, a tooltip of the country name will pop up. Names of the countries shown in the chart are also listed above it. Clicking a name in the list highlights the country across the world map, the list and the line chart. Double clicking removes highlights. Lines for different countries have different colors. Users can remove a line they are not interested in any longer by simply double clicking the line.

To visualize the population data of a certain country in all or selected years between 1960 and 2016 and find whether there are any correlations among attributes that can be observed from our visualization charts, we implement clusters to group countries with similar trends and properties in a fixed period. In order to visualize countries with 10 years' data, we use PCA to do dimensional reduction, and colorize clusters with the value of the first principal component. We also use Mapper, which is based on the idea of partial clustering of the data guided by a set of functions defined on the data. We want to use this topological method for the analysis of high dimensional data sets, and display the subgroups in the data sets which could get insights from. These view provides users with a computational method for global population analysis.

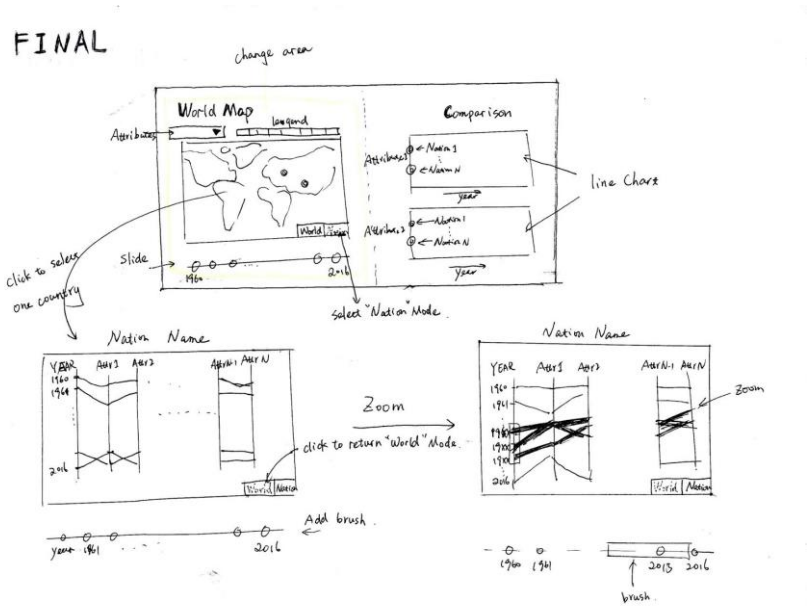
To show multiple attributes in one graph, we use parallel coordinates. For a selected year (or a selected country), we plot countries (or years) and attributes as parallel coordinates. By this chart, we can observe correlation between attributes. Each coordinate can be sorted and be moved to be reordered. Use brush to select part of a coordinate and to highlight paths that go through the selected part of the coordinate.

We can also use table to show data for a selected year. The columns of the table are countries and attributes. The table can be sorted by each column. Countries from each continental can collapse to one group.

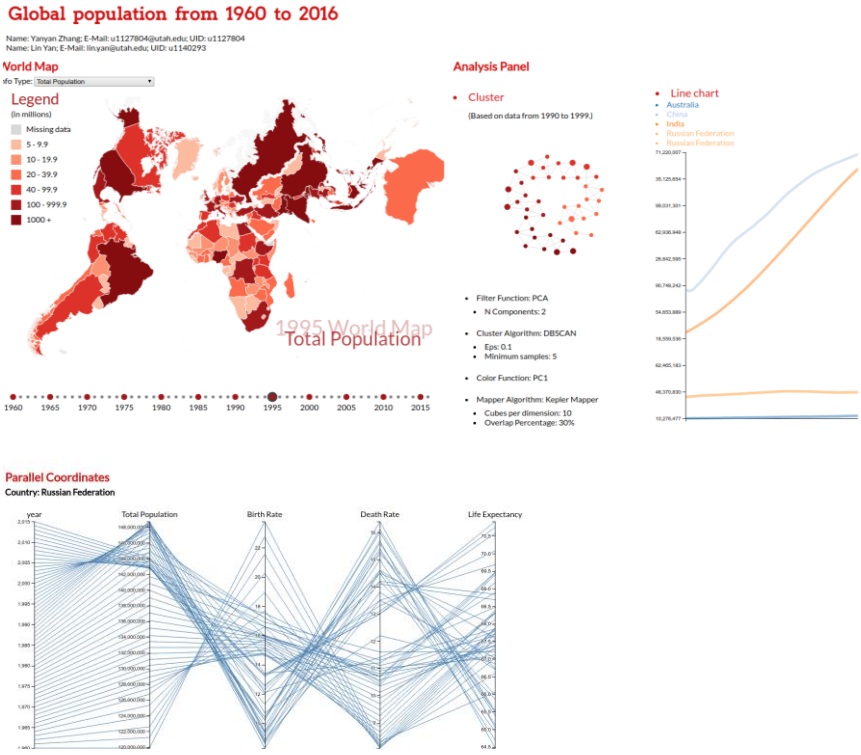
3 Implementation

3.1 Overview

Up to now, we have finished selection bar, world map, year chart, line chart, cluster and parallel coordinates. See here for our project. Following are our hand-written design and images of the part we have done.



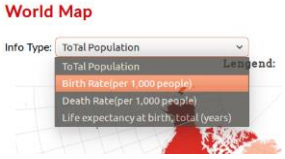
Hand-written Design



Overview of Implementation

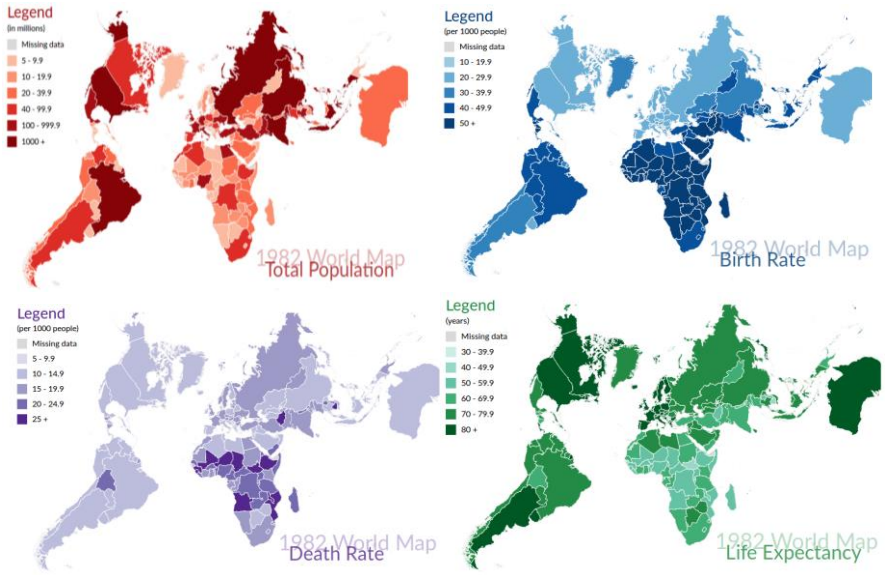
3.2 Selection bar

In the selection bar, there are 4 attributes from which one attribute can be picked. The values of the attribute in a specific year can be visualized in the world map.



3.3 World Map

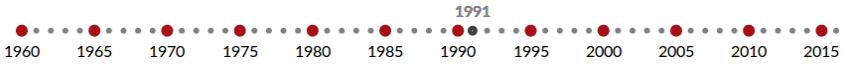
The world map can be updated according to the year and the attribute selected by users. It uses different color scales to represent different attributes.



The World Map Updated Based on Different Attributes

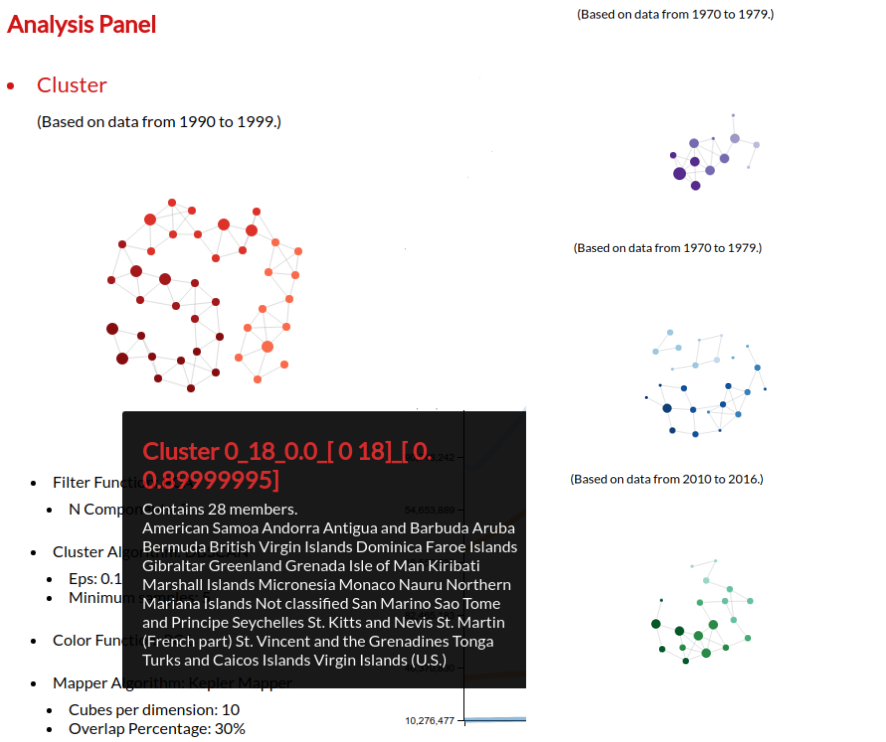
3.4 Year Chart:

The year chart displays every five years users can select. When user's mouse hovers over the smaller circle, the year will display on the top of this circle, and will disappear when user moves the mouse away.



3.5 Cluster

In this view, we implement a method, which was called Mapper, to reduce high dimensional data set (10 years' data) into simplicial complexes with far fewer points which can capture topological and geometric information at a specified resolution.



Clusters from different time periods based on Total population, Death rate, Birth rate and Life expectancy respectively from left to right and then from top to down

Filter selection transforms the similarity information into a visual representation, a 2D rendering of a 3D graph. The first two principal component scores (PC1 and PC2) of the variance-normalized data were used as filters, such that two participants were judged to be similar if both their PC1 and PC2 scores were sufficiently similar. We colorized each cluster according to the value of PC1.

· **Methods:**
Filter Function: PCA (N Components: 2)
Cluster Algorithm: DBSCAN (Eps: 0.1, Minimum samples: 5)
Color Function: PC1

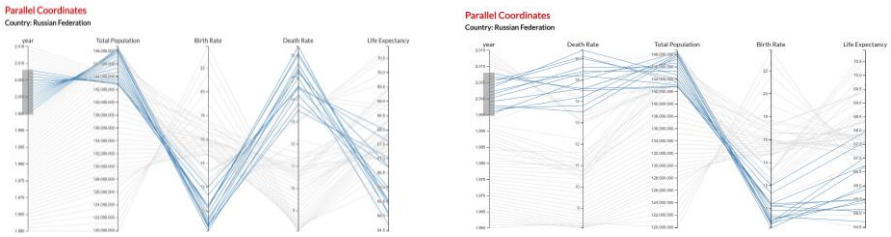
Mapper Algorithm: Kepler Mapper (Cubes per dimension: 10, Overlap Percentage: 30%)

3.6 Line Chart

Clicking a country on the world map adds the country to the line chart. The line chart shows all values of the current attribute of selected countries. There is a list of countries that the lines represent. Clicking a country name highlights its line in the line chart and its path in the world map. Double clicking a name or clicking another name removes highlighting on the previous country. The country name shows up when its line is hovered over. Double clicking a line removes the line from the line chart and the name list. When the attribute in the selection bar changes, the y axis of the line chart changes correspondingly. But the previously selected countries remain the same.

3.7 Parallel Coordinates

A test dataset extracted from the original dataset was used for the parallel coordinates for now. This test dataset contains the values of two attributes of the USA from 1960 to 2015. There are three axes in the chart: ‘year’, ‘Total Population’, ‘Birth Rate’, ‘Death Rate’ and ‘Life Expectancy’. Users can use brush to select a range of values on each axis. The paths going through all the brushed areas are highlighted. Users can also grab an axis name and drag the axis to a different position to reorder the axes.



The unordered and reordered parallel coordinates for Russia

4 Evaluation

Our visualization was designed to answer proposed questions:

Q1. How to visualize the population data globally in a fixed year?

Q2. How to visualize the global population data in country groups in years from 1960 to 2016? What similarities do the countries in the same group share?

Q3. How do values of a certain attribute of selected countries change over the years? How to compare change trends among multiple countries?

Q4. For a selected country, how to visualize the change trends of all attributes over years from 1960 to 2015? Can we identify normal or abnormal patterns? Are there stories behind them?

In this section, we first discuss how these questions were answered using our visualization (section 4.1). We also found some interesting phenomena using our visualization. We will specify some examples in section 4.2.

4.1 Evaluation

Q1: In order to answer Q1, we designed the world map. Users can select the attribute they are interested in to display its data on the map by selecting the attribute in the selection bar. To make our world map clear, we used the legend with different hues for different attributes, and set countries with missing data to gray.

Advantage: It has good visual effects. The accuracy does not need to be high for this global view.

Disadvantage: Small countries are hard to see or to choose on the map.

Q2: In order to find similarities of attributes that are shared by a group of countries, we designed cluster chart. We could also drag the cluster on our visualization to arrange a good distribution.

Advantage: The information we can get from this visualization is highly topological and geometric, and is very useful for data analysis.

Disadvantage: It is hard to understand for people without related knowledge. Parameters cannot be changed by users.

Q3: In the line chart, we show values of the selected attribute of the selected countries throughout all years. Each line represents a country. By comparing the changes in the lines, the difference among those countries can be easily seen.

Advantage: By using this chart, comparison of a certain attribute among countries becomes easy.

Disadvantage: The chart can become crowded if more than 10 countries are selected to compare.

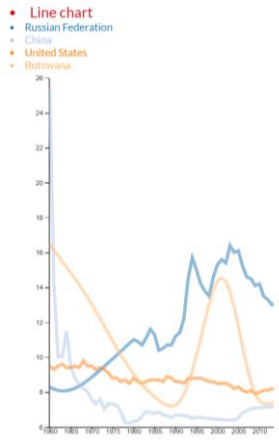
Q4: For a fixed country, the parallel coordinates can show all the information at the same time. Each path shows all attribute values in one year.

Advantage: Patterns and change trends can be easily observed in this chart.

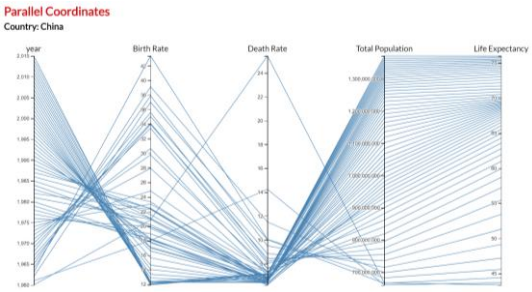
Disadvantage: It takes some space. Showing the charts for multiple countries on the same page might not be practical.

4.2 Examples

It was observed that a normal trend for most countries from 1960 to 2016 is that as time went on, the total population increased, the birth rate and the death rate dropped, and the life expectancy went up. Let up look at the line chart of the death rate in the following. The US has a relatively normal trend. But the high death rate before 1975 could be related to the Vietnam War. China had extremely high death rate in 1960 and 1961. This was mainly caused by The Three Years of Great Chinese Famine from 1958 to 1961. Russia had high mortality rate since 1990s. Reasons behind this are more complicated, that could include economic and social instability, tobacco and alcohol consumption and bad health care system. What happened in Botswana? Users can find the story behind the curve. In addition to this, we show the parallel coordinates for China. From this chart, we can also observe the high death rate in 1960 and 1961. Besides, the total population has doubled in the last 56 years. The life expectancy has increased ~ 75%. The birth rate has decreased more than 70%. Many of users probably know the One Child Policy in China. From this chart, users can identify when this policy started.



The line chart of the death rate



The parallel coordinates for China

5 Conclusions

Up to now, we have finished several tasks related to this visualization:

1. (World map) Give users a global view of population change.
2. (Cluster chart) Group countries with similar properties in a fixed time period (10 years) according to the attribue users are interested in.
3. (Parallel coordinates) Users may be interested in certain attributes of certain countries that are correlated to the demand of certain services. Studying the changes in these attributes over the years will provide quantitative information that may help them determine their future plans.

What should we further improve:

1. We can show more information via the world map. We could add tool-tip and zoom on it.
2. The year chart is hard to select on. We could add the left/right keys/buttons for small changes of the year selection.
3. We could use another world map projection so that a country's size is proportional to its value of the attribute so that its value can be visible on the map even if the country is very small.
4. In the cluster chart, the information in tool-tip is too professional. We could change the title of the tool-tip which could be understood by people without mathematical background.