Global Population Trends and Projections

Devki Desai, Paul Mello, Priyanka Moorthy, and Vamsi Chalamolu

San Jose State Univeristy

December 9, 2021

Abstract

The census has been the most crucial form of data collection for all of human history dating back as far as the Babylonians who used it to determine how much food they needed to survive. Today this data extends to demographics, geopolitical influence, and understanding the unnatural phenomenon that is climate change. Through exploration of the World Bank's census data we intend to develop a better understanding of the patterns and trends present in human life in an attempt to demonstrate relationships between countries and continents while also building models to offer predictions on future populations. We find that our regression models tend to predict future populations with statistically significant accuracy while classification by continent proves to be far trickier.

1 Introduction

Population studies are the most ancient methods of examining a populations' size, structure, and development over time. Researchers have applied statistical examinations on this type of data to explore mortality, fertility, and their associative factors such as poverty, employment, culture, migration, and religion. This often results in sweeping changes and action by giving people an understanding of where the modern population is trending towards. Many factors play an important role in developing these projections and predictions including climate change, health services, education, and future well-being. Having this information can better equip governments and individuals to prepare for the future.

The World Bank collects and interprets census data in a way that allows them to extrapolate and make predictions based on trends over the past 80 years. These predictions are often accurate outside of sudden or unseen global changes. The data set we will be using is broken down into many pieces consisting of continent information, country information, a column that describes what data the remaining row contains, and many more columns consistent with an interval of one year periods and each contains a subsequent numeric observations.

The focus of this paper aims to use the World Bank's population estimation data to comprehend global population trends by analyzing the data in unique and interesting ways, while also devising and developing models which can aid in furthering our understanding. We incorporated regression and classification based algorithms into our analysis in order to develop models which provided an additional complexity of understanding. This will be coupled with supplementary analysis, such as principal component analysis, in an attempt to extract any important years which may have altered the course of history within the past few decades. We hope that through our exploration we will see interesting patterns emerge, be able to make our own predictions and classifications, and shed some light on why these trends are occurring.

2 Methods

The World Bank provides an easy and accessible way to download historic, current, and future data projections directly from their main website. The data we worked with is broken down into three discrete categories consisting of Country, Series, and Time. The country section contains a list of all recognized countries on the planet as well as continental, age, and economic based data that is pooled together. The series section contains a list of all the relevant data information that can be extracted from each of the individual countries. Finally, the time section contains a yearly interval from 1960-2050. For our purposes we elected to download all the data available; However due to limitations in their servers, we were only able to download a grand total of 2.5 million cells for our analysis. As a result we downloaded every country and time, but chose to remove percentage based data from

the series section as that could be derived from the totals if needed.

The data we downloaded is laid out in the following manner. The first few columns consist of continents and country identification. Following these are a series description which elaborates on what the remaining observations in the row entail. These series consist of a variety of data points from age dependency, to net migration, and population totals. Each country contains an ordered list of 90 rows of the same series names which are followed by ground truth data from 1960-2020 and predicted observations from 2021-2050. The following figure demonstrates a small subsection of Afghanistan's data with many additional columns hidden after 1961.

Country_Code	Series_Name	Series_Code	[1960]	[1961]
AFG	Age dependency ratio (% of working-age populat...	SP.POP.DPND	81.617265593364	82.6886781269233
AFG	Age dependency ratio, old	SP.POP.DPND.OL	5.08221355458813	5.13013875077877

Figure 1: Inital Data

As there are many series and many countries, it is important to mention that low income and war torn countries often have significantly more missing data. As a result, there may be a bias in some of our models, depending on the data series we are working with, which appears to favor higher income and stable countries.

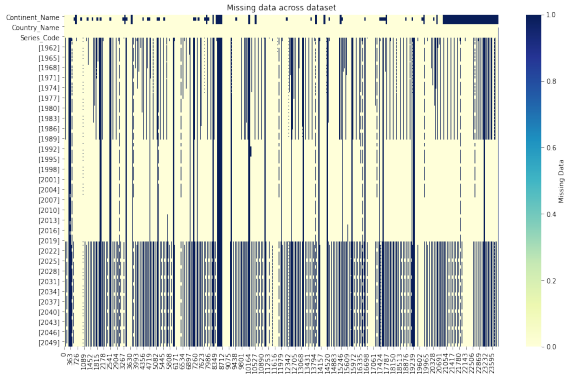

Figure 2: Missing Data

Our methods for analysis have been divided into many different components. Our first action was to collect and preprocess the data. Missing data was handled on a case by case basis where the missing observations may be pruned or cleaned as necessary. Following this preprocessing we subdivided the data into the appropriate data frames which will allow access to specific series, such as total population by country. As a result of this subdivision, visualization of the data was easily created and allowed for extracting further interesting information.

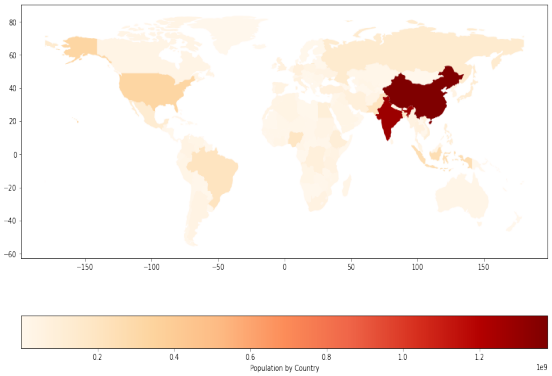

Figure 3: Heat Map of World Population Today

Once this visualization was completed we began conducting our methods of analysis to make predictions, categorize, and show relationships within the data. The two most notable methods we used consisted of regression and classification. Other methods such as clustering have been performed as well, but only in minor ways as applicability and accuracy are hard to determine. Each of these methods were used to interpret and understand if there were any underlying patterns inherent in the data we could not easily identify. Regression consists of an analysis which can help estimate relationships between dependent variables and one or more independent variables. This can help us model the trends within our data and make predictions about where a population is moving towards in the next few years. While classification is consistent with creating an understanding of relationships between groups by identifying and assigning categories to some dependent data. Both of these methods of analysis have provided key insights into making comparisons about the nature of the relationships.

Once these methods of analysis were compiled and completed we needed to create adequacy checks to demonstrate the accuracy of these respective models. We were able to create interesting plots which demonstrated our findings and our models. One method of analysis called principal component analysis (PCA) allowed us to see far deeper into the data than we previously could. PCA refers to reducing the dimensionality of a data set while retaining as much information within the data as possible. This is incredibly useful for our purposes because it allowed us to condense country and continent data down to a more manageable size. Additionally, the process paved the way for interpreting what years are contributing to our data's variation the most. This is an important step as it helped develop an understanding regarding important

years containing big shifts in the global populations trends.

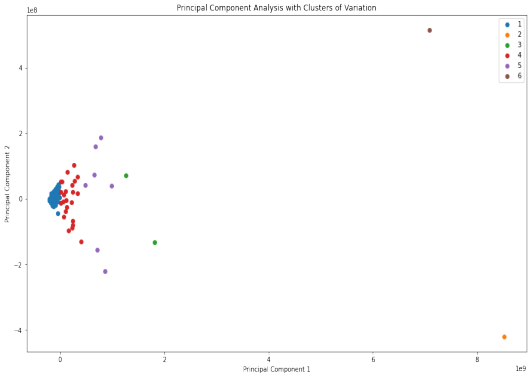

Figure 4: Conducting PCA on Population Totals

Here we demonstrate PCA by reducing our population totals data from 1960-2020 into a single two dimensional plot with their associated K-Means clusters as determined by our model. K-Means clustering is a method where we can partition the data into multiple categories based on their distance from a centroid which is iteratively updated until all data points are accounted for. We can see that most of the data is clustered into a tight spread which fans out with a few outliers forcing their own clusters. This provides a wealth of information as we can see how our population trends appear to be mostly normal. Diving further into the apparent outliers we were able to narrow them down to two specific countries, China and India, which have seen an immense population growth over the past century.

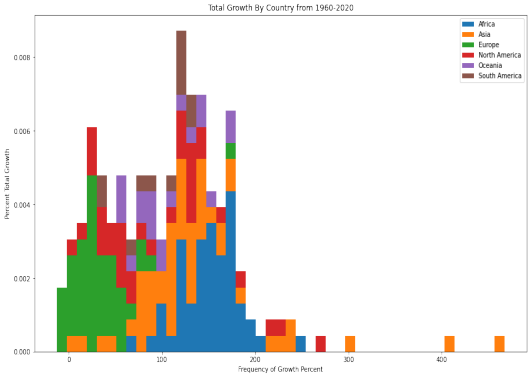

Figure 5: Continental Frequency Growth by Country

As demonstrated above these two countries have seen an explosion of growth likely due to their sudden and rapid economic expansion. In addition to this information we saw that, in nearly every clustering attempt, the data was best sectioned off into six distinct parts which mimic continent quantities.

Another method of analysis we conducted was regression. Regression can provide us with a way to model and predict future population trends. The following graph is an example of one of our regression models which fits the data almost to a

Due to the nature of this data set and our proposed work we expect to run into many different limitations and challenges as we deal with combining, separating, and cleaning these individually created data frames. However, we did not run into many challenges in our modeling because each data series we selected had a significant majority of data.

Finally, we conducted analysis and interpreted the results of our methods in a fashion that provided a clear and concise summary of our findings. In developing these methods we now have a greater understanding of important years and continents which impacted global trends in the population over the past 80 years.

3 Experiments & Analysis

In developing our experiments and analysis there were many models and visualizations that offered key and interesting insight about the series it represented. A few of these highlights can be found in the appendix section of this paper. These figures visualize and represent a broader range of modeling and help interpret our population data. However, for now in this section we will specifically highlight two examples which excited us.

Through our initial exploration of the data we discovered interesting trends around population totals. We found that, on average, western continents appear to have significantly less population growth within the past century, by in one case approximately 5x less.

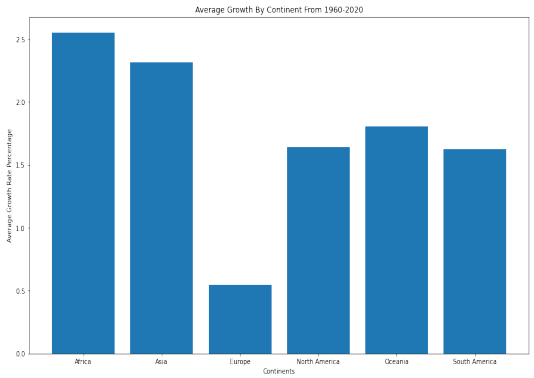

Figure 6: Growth Totals by Continent

As a result of this statistic we elected to dive

deeper into potential reasons for this discrepancy and found many potential justifications. According to the National Institute of Health, there are many factors which are contributing to these declines in western countries[1]. They cite increasing access to contraceptives, a rise in female education, and soaring housing prices as key factors to the decreasing fertility rate and thus decreasing population. They expand on this by explaining that unlike Europe, the Americas and Oceania have largely been unaffected so far due to the necessity of time to pass before generational equilibrium's standardize and appear in the data. For example, the Americas have not seen this slowdown in growth yet as evidently as seen in Europe, due, in large part, to most couples of child bearing age choosing to wait later in their life for economic stability before having kids. This process will repeat over a few more generations and inevitably result in the Americas of tomorrow appearing as modern Europe, in regards to growth trends.

Further research and analysis of this declining trend in fertility rates have been forecasted by the Lancet Project. They have demonstrated that by the year 2100 fertility rates will decrease from an average of 2.30 to 1.66[2].Resulting in one of the lowest fertility rates, for the global average, in human history. Specifically, we will see that some western countries, especially those in Europe, may see a 50 percent decline in their population totals[2]. Despite this drastic decline The United States will continue to gradually increase over the next century. From this new found understanding we began exploring other sections of the data in an attempt to pin down if there were any other contributing factors to this sudden change.

Through further data manipulation and visualization we came across an interesting statistic regarding net migration by continent. It seems that, due to economic prosperity in the west, migrants are moving in disproportionate numbers from Asia to Europe and North America.

On a yearly basis this difference alone does not make for a huge impact. However, overtime these migration trends can make a significant mark on the population at large. We suspect that the growing number of immigrants coming to western countries may result in a sizeable increase in population. We believe that the demographic of those making such a trip are primarily young and able bodied workers. Looking at global news reports about migrants on the borders of European countries only furthers this notion and demonstrates its relevance.

Evidence suggests that as global population continues to increase these migration patterns will persist. However, there will be a point where the world population will begin to decline and it is suspected that migration trends will reverse.

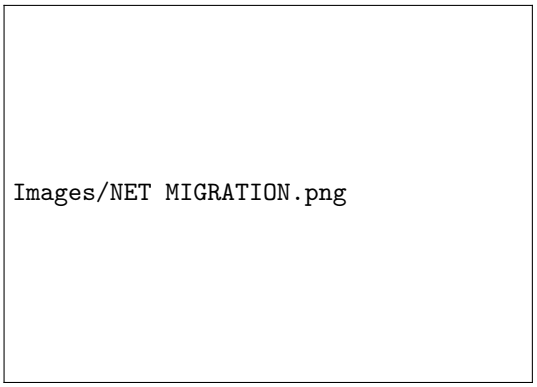

Figure 7: Net Migration by Continent

Through this analysis and understanding we proceeded with building our linear regression and classification models.

4 Comparisons

5 Conclusions

Population trends continue to increase on both a continental and global scale. We suspect that the global population is likely to peak somewhere around the middle of the century before a rapid decline due to low fertility rates. We have analyzed a significant amount data and created the relevant visualizations which demonstrate these defining feature to be unavoidable in our near future. While we were successful in developing statistically significant models to predict future population trends we can not be certain our analysis is accurate given the scope of this paper. We hope that countries will take these trends seriously as they prepare for the near future; not only for the sake of their own people, but for the common welfare of our planet.

References

[1] G. Nargund. Declining birth rate in developed countries: A radical policy re-think is required. *Facts, Views, Vision in OBGyn*, 2009.

[2] Prof Stein Emil Vollset, Emily Goren, Chun-Wei Yuan, Jackie Cao, Amanda E Smith, and Thomas Hsiao. Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the global burden of

disease study. *The Lancet*, 396(10258), July 2020.

World Bank Data:
https://databank.worldbank.org/reports.aspx?source=Health%20Nutrition%20and%20Population%20Statistics%3A%20Population%20estimates%20and%20projections#

World Bank Data Description:
https://data.worldbank.org/indicator/

National Institute of Health:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4255510/

The Lancet Project:
https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30677-2/fulltext

Our Repository:
https://github.com/devkisdesai/CMPE255-Team-6-Project-Fall-2021-