Global Population Trends and Projections

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

San Jose State Univeristy

December 10, 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 and life expectancy. We find that most of our developed models tend to predict future trends with statistically significant accuracy, while classification and clustering by continent proved to be significantly more difficult than we anticipated.

1 Introduction

Population studies are among the most ancient method 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 countries and government an understanding of where their population is trending towards, and their primary demographics. 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 their future survival.

The World Bank collects and interprets census data in a way that allows them to extrapolate and make predictions based on trends from historic information. 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, country, and yearly information.

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. We devise and develop models which can aid in furthering the understanding of global trends. We incorporated regression and classification based algorithms into our analysis in order to develop models which provided an additional layer complexity. This will be coupled with supplementary analysis, such as principal component analysis, and external research, in an attempt to extract information and extrapolate analysis. 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 Data

The World Bank provides an easy and accessible way to download historic, current, and future data projections directly from their main website. Due to limitations in their servers we were only able to download a smaller, but still significant, portion of the data.

The data we downloaded is laid out in the following manner. The first few columns consist of country identification. Following these are a series of descriptions which elaborate what the remaining observations in the row refer to. 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 series types 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 annual columns extending to 2050.

We analysed global population trends such as

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: Example Data Subsection

world averages, country and continent specific information, population growth rates, net migration and life expectancy, children deaths under 5 and a few more.

2.1 Data Pre-processing

As there are many series and many countries, it is imperative 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 appear to favor higher income and stable countries.

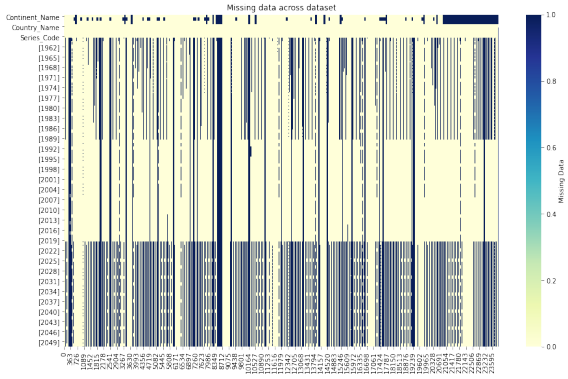

Figure 2: Missing Data

From Figure 2, it can be observed that a significant amount of data is missing from the years 1960-1989 and also from the estimated projections from 2020-2050 when compared to the years 1990 to 2019. We can postulate that between 1960-1989 many governments, or third-party census trackers, were attempting to figure out how to collect this data from remote or dangerous places. We can however assert that estimated projections are very difficult and can have serious consequences if calculated improperly. As a result, the data from 2020-2050 is far sparser.

Estimating missing values for population analysis can result in unreliable data as population is subject to many external factors. Consequently, within our own project, missing data was handled on a case by case basis where the missing observations may be pruned or cleaned as necessary. Following this cleaning we subdivided the data into the appropriate data frames which allowed access to specific series, such as total population by country. Through this subdivision, visualization of the data was simple.

2.1.1 Data Integration

The World bank data does not come with direct information regarding the continent a country is located in. Another data set has been incorporated which adds continents and their abbreviations to our main data. This will provide an easy way to model and predict by continent. Apart from the continent integration, we have also added the ISO 3166 international standard Alpha-3 codes to the data. This will help in the analysis and generating choropleth maps.

3 Methods

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. Through this subdivision, visualization of the data was successfully achieved. Below we can see a heat map of the world, which demonstrates current population totals of the most recent year.

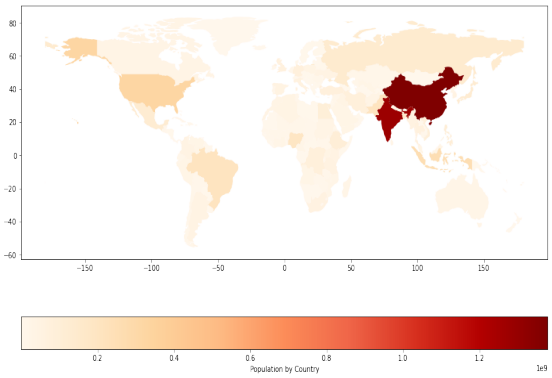

Figure 3: Current World Population

Another example of these visuals demonstrates total world population and growth rate. This gave us significant amount of insight into the totality of our data in a way that allowed us to consider many different options for approaching our series.

Once these visualizations were created and completed we began conducting our methods of analysis to make predictions, categorize, and show relationships within the data. The two

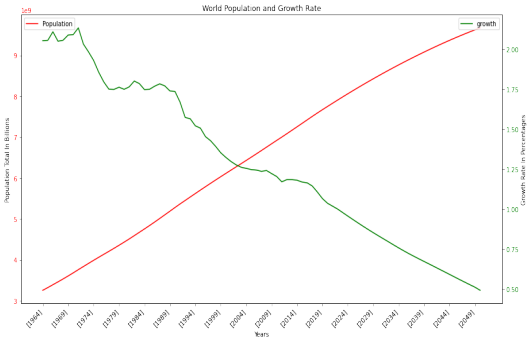

Figure 4: Population Totals and Growth Rates

most notable methods we used consisted of regression and classification; However other methods such as clustering have been attempted. 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 trending in the immediate future. 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 between countries, not only by continent, but by global comparison as well.

Our modeling parameters were hyper-dependent on what series we desired to extract and our goals for prediction or classification. These series included population totals, net migration, life expectancy, and deaths under 5 years old. For nearly every series we used, our data consisted of each observation from 1960-2020, and then depending on specific analysis goals we would add 2021-2050 as projected data. Having projected data was pivotal in being able to test our models. The strength of these parameters are apparent in that they provide consistent and unflinching data regarding population trends on a country by country basis. If a country has some data, they are overwhelmingly likely to have all the data. However, the opposite can be stated as well such that, any country which tends to have missing data often has a significant portion of data missing.

Once these methods of analysis were compiled and completed we needed to consider adequacy checks to demonstrate the accuracy of these respective models. Some examples of these include R-squared, confusion matrices, and mean squared error. 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 about the data and its higher dimensions as possible, consider it as compressing the data. This is incredibly useful for our purposes because it allowed us to condense country and continent data while demonstrating variation.

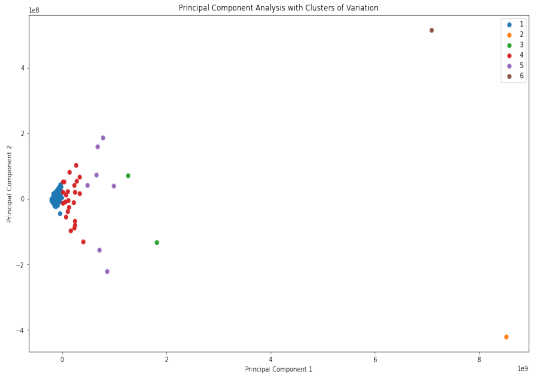

Figure 5: Conducting PCA on Population Totals

In figure 5 we demonstrate PCA by reducing our population observations from 1960-2020 into a single two dimensional plot with their associated K-Means clusters as determined by our modeling. 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 updates until all data points fall in their appropriate cluster. 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 the same. However, certain countries appear to have extensive variation from the mean.

We were able to narrow these outliers down to two specific countries, China and India. These countries have seen an immense population growth over the past century when compared to global an continental trends.

As demonstrated in figure 6, 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 mimics continent quantities. Further investigation demonstrated this to be a coincidence rather than a correlation. The scree plot in figure 7 helps demonstrate that our selected clus-

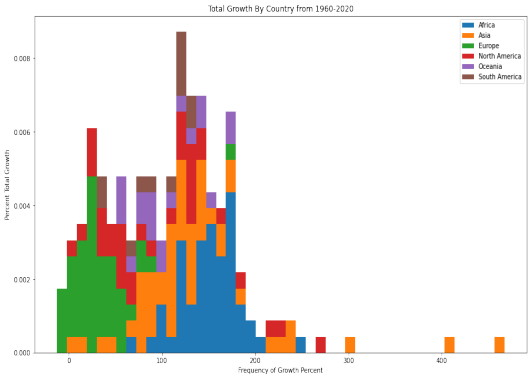

Figure 6: Average Growth Rate Frequency

tering determined six to be the optimal cluster count.

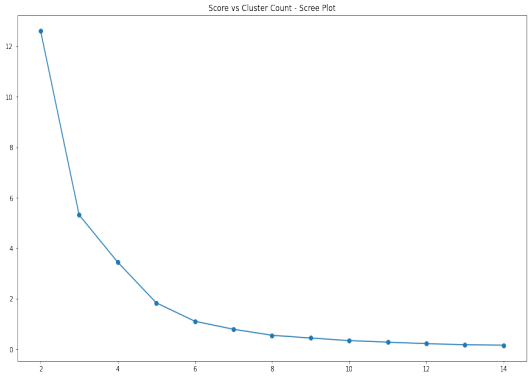

Figure 7: Deciding Country Based Cluster Counts

Another method of analysis we conducted was regression. Regression can provide us with a way to model and predict future population trends. As we will soon see regression proved to be far more difficult to implement accurately than we expected. This is in large part due to regression necessitating that each column be independent of each other. This was however, not the case for our data set given its many complexities. We only realized these limitation once we conducted our modeling. The same can be said about classification due to difficulty in developing accurate classifiers for our data given the sparsity of countries in certain continents.

Finally, we conducted analysis and interpreted the results of our methods in a fashion that provided a clear and concise summary of our findings, even despite the inherent flaws we would later find in our models. This was done in part through extensive research and analysis of similar studies. We will cover this experimentation and analysis in the following section.

4 Experiments & Analysis

In developing our experiments and analysis there were many models and visualizations that offered key and interesting insight about the series it described. 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 we will specifically highlight two examples which excited us population totals and net migration.

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 from the beginning of the data collection, by in one case approximately 5x. Figure 8 demonstrates this difference in growth rate over the past 60 years.

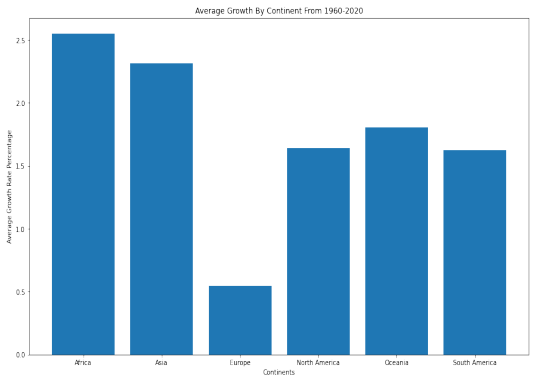

Figure 8: Growth Averages Statistics by Continent

As a result of this visual 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 growth rate and thus decreasing population. They expand on this by explaining that unlike Europe, the Americas and Oceania have largely been unaffected due to a lack of time passing.

We believe that once generational equilibriums standardize we will expect to see a gradual decline. For example, the Americas have not seen this slowdown in growth like Europe has, 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 North America of to-

morrow appearing as modern Europe, in regards to growth trends. Hence we felt it necessary to build a model for predicting European population trends as a means of visualizing a possible future for North America.

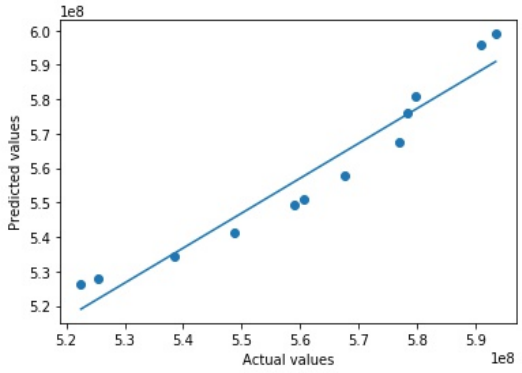

Figure 9: European Growth Trends Model

We found that modeling the population trends for Europe was far more difficult than any other continent. We discovered that modeling Europe's growth trends produced an R-Squared value of 92 percent, while models for all other continents could explain their variation by at least 96 percent if not greater. As a result we can not say that our European model was statistically significant, but we can attempt understand why this is happening.

Continent	R-squared	MAE
S. America	0.99	3414753
Africa	0.96	45820472
Oceania	0.99	559575
Europe	0.91	5838186
Asia	0.99	27991317
N. America	0.9994	2141151

Analysis of this declining trend in fertility rates has 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 global growth rates in modern human history. Specifically, we will see that some western countries, especially those in Europe, may see a 50 percent decline in their population totals over the next half century[2]. Despite this drastic decline The United States will continue to gradually increase in population 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 relevant factors contributing to this increase despite growth rates decreasing.

We began to analyze other series in the data, but found no evidence as to why these trends will occur. Through further visualization we were now under the impression that this may be a result of net migration trends by continent. Unfortunately, our data does not allow us to see where a country or continents population is migrating to or from. However, the data suggests that a primarily Asian population will be moving to Europe and North America in the coming decades.

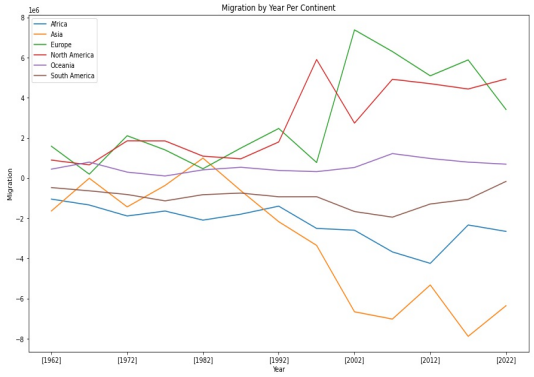

Figure 10: Net Migration Trends

Annually this difference does not appear to make a huge impact. However, overtime these migration trends can compound to be significant. We suspect that the growing number of immigrants coming to western countries may be climate or resource refugees. The demographic of migrants making such a trip will primarily be young and able bodied individuals.

We attempted to model this data but we soon ran into issues. We had naively assumed that our net migration data was some how independent between continents. Upon attempting to model migration we became aware that such an assumption was deeply flawed when our mean squared error was greater than our population totals.

We thus returned to figure 4 in hopes we could find another interesting notion to model. We noticed a gradual but evident decrease in the total population year after year. Meaning there may be an inflection point where the population will begin to decrease as a direct result of declining growth rates.

We now understood that average world growth rates are projected to fall below the crucial level of "replacement births" around 2020 with some continents being affected significantly more than others.

This meant that less kids will be born when compared to fertile adults. Secondly, population totals will begin to gradually slow down around the start of the 2030's. Consequently, we can extrapolate that within a few decades after 2050 we will likely see an inflection point where the

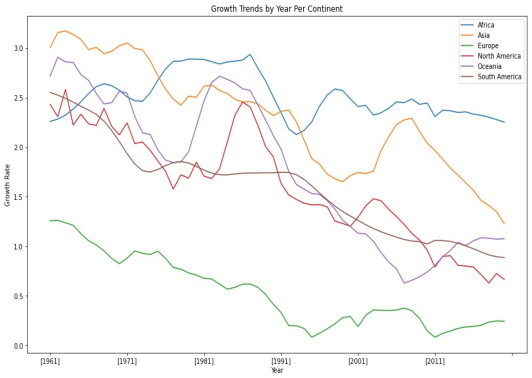

Figure 11: Declining Growth Rates by Continent

population will peak before a significant decline.

We suspected that such a large increase in life expectancy will cause far more problems as generations continue to age. As we illustrate in figure 12 the average life time has increased from the early 1960s in all continents significantly.

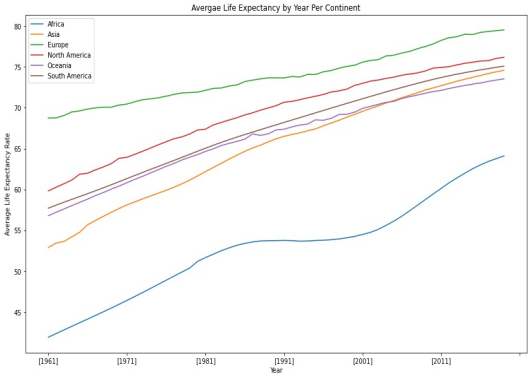

Figure 12: Annual Average Life Expectancy

This will have a significant effect on the future as the population becomes more strained in a large multitude of ways. Diminishing resources, increased pollution, strained social programs, and, for the first time, a population above 80 years old which will be double the population under 5 years old[2]. This extreme shift in human life will have a profound impact for future generations.

This sudden and emerging population of the elderly will put an extensive strain on the already depleted resources available. As a result it is necessary to prepare for these situations by building out systems which can alleviate the burden when our current population inevitably ages.

We thus attempted to model this change in human life by attempting to model and predict life expectancy in the coming decades by country.

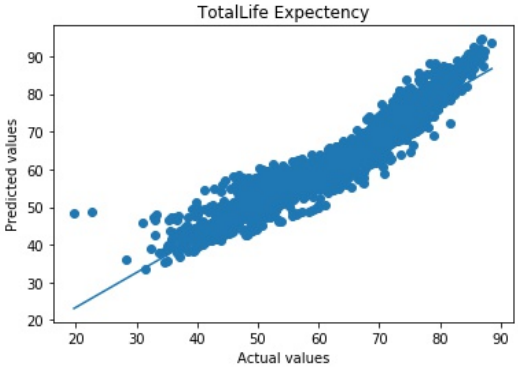

Figure 13: Average Life Regression Model

Performance of Multi-regression model to predict life expectancy of a country, at a given year.

Series Type	R-squared	MAE	RMSE
Total	0.90	2.49	12.02
Male	0.90	2.43	11.74
Female	0.91	2.60	12.93

We found that life expectancy will continue to increase, however, not for every country nor by the same rates. Due to economic prosperity and an increase in the overall quality of life different countries will experience different life expectancy's.

We face an incredibly grim reality, but through these types of models and predictions we can prepare for tomorrow.

5 Comparisons

We used Linear Regression to model our data. It is one of the simplest models to build. However, we ran into some challenges since our data set was limited in nature. Similar research conducted found that some modeling techniques demonstrated far better results than our attempts. Modeling life expectancy through a generalized learning model, which adjusts weights based on closeness of prediction may offer additional predictive accuracy. Models such as random forest, and Artificial Neural Networks (ANN) have extreme potential for massive performance gains. The performance of the model would be evaluated using the accuracy of predicting the correct category of the Population Growth Rate. Future studies can focus on using various ML techniques and performance metrics to evaluate the performance of predicting the population trends.

6 Challenges and Considerations

Due to the nature of this data set and our proposed work we expected to run into limitations and challenges, however they were far more extensive than we had expected. Due to the importance of the census data not much needed to be done to pre-process and prepare the data so long as we worked with the proper series. Most of the effort and challenges came from developing proper data frames and have a thorough understanding of the extensive data. We did however struggle significantly with the modeling of our data as our assumptions about the data proved to be problematic. In the future working with such complex and intertwined data will necessitate understanding the implications and meaning of each individual series.

Additionally, our approach to the data is flawed due to the complex nature of population data. Much of the initial planning we developed had to be pivoted or removed as we noticed we could not accurately or statistically prove that our assumptions about the data were correct. One example of this short coming was a lack of data for North America, simply put there are few countries in North America and thus few opportunities to train and classify the data properly. This resulted in a lot of time being sunk into efforts which proved futile. If we were to conduct this analysis again we would like to attempt to incorporate many more training styles such as K-Fold cross validation or finding another data set with more direct data.

7 Conclusions

In conclusion, 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 this century before a rapid decline due to low fertility and growth rates. We have analyzed and modeled a significant amount data and created relevant visualizations which demonstrate these futures to be unavoidable. As a result, we can confidently say we have shed some light on the data and its subsequent meaning. While we were successful in developing some statistically significant models to predict future population trends by continent, we can not be certain our analysis is accurate given the complexity of our data, our naive assumptions, and at times sparse data series. Despite this, 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 survival of the population.

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-

Related Papers: https://academicjournals.org/article/article1379677191_Folorunso%20et%20al.pdf

8 Appendix: Interesting Visualizations

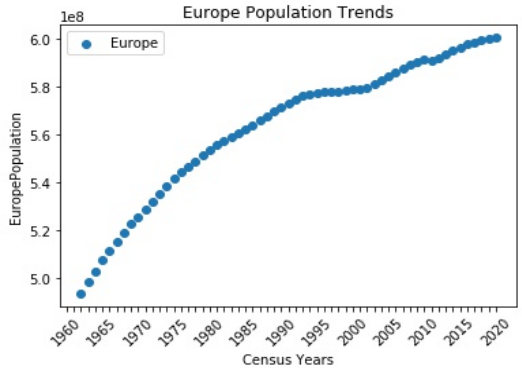

Figure 14: European Population Trends

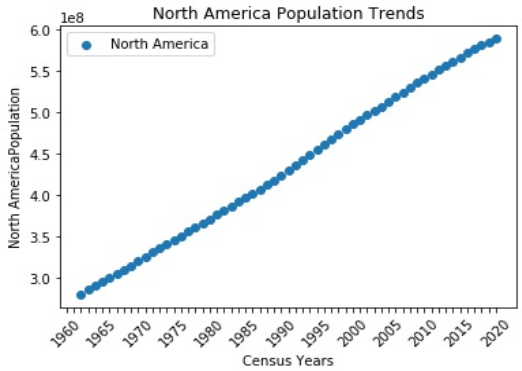

Figure 15: North America Population Trends

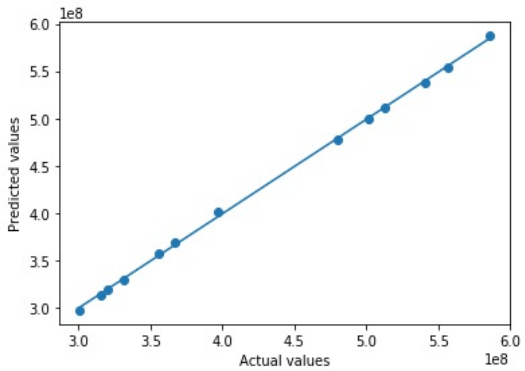

Figure 16: North American Population Regression Trends