# Correlation of Population Data

Divyank Garg, Priyadharsheni Balasubramanian, and Swati Ramachandran

October 3, 2018

**Abstract**

The task is find the correlation information of a country's population with that of the other countries. Population data of 258 countries between 1960 and 1999 are used and the correlation information model derived from this data is used to predict the population correlation of data between 2000 and 2016. The results are visualized as a network graph using Gephi and are interpreted using various network visualization features.

## 1 INTRODUCTION

Linear Regression is a statistical tool that is used to model relationship between selected features and the outcome. But in certain models, the linear regression when applied still doesn't provide the optimum results. To get the optimum results regularization can be done and there are two ways to achieve this - LASSO regression and ridge regression.

The following paper deals with the implementation of a regularization technique to predict a model that best reflects the behaviour of a country's population depending on other countries.

## 2 PROBLEM STATEMENT

The dataset used has the population of various countries in the world between the years 1960 and 2016. The task is to find 5 countries that closely represent the population behaviour of a country. The data between years 1960 to 1999 is used to train a model and this trained model is used to test the prediction over the data for years 2000 to 2016.

## 3 IMPLEMENTATION

Both Ridge and LASSO are regularization techniques that can be used in predicting the model by suppressing weak features and accentuating features that matter.

In the ridge regression method, an alpha value is chosen which makes the coefficients of certain features reach zero and not become absolutely zero to

minimize the least squares error of the predicted model. In LASSO regression method, the alpha value chosen changes the values of the coefficients of the features and brings it down to absolute zero for those features whose effects are considered insignificant in making the least squares error value become minimum. For the final implementation, we chose the Lasso model, with a range of possible alpha values.

# 4 EXPERIMENTS

The first step in computing the coefficients was to ensure the data set was appropriately formed. The data was cleaned up by dropping unwanted columns. With Ridge regularization, it was seen that many countries are weak contributors, i.e. coefficient values close to 0. As a result, we decided to work with Lasso, to ensure we would not have to manually tackle the non-zero coefficients. The final step was to choose a range of alpha values such that for every country, we can determine 5 contributors for as little error as possible. Figure 1 lists the various experiments performed along with the time taken for computation and the number of countries with properly predicted coefficients.

| Alpha range | Number of countries with 5 coefficients | Computation time (s) |
|---|---|---|
| 10**np.linspace(10,-2,5)*0.5 | 127 | 58s |
| 10**np.linspace(11,-2,200)*0.5 | 258 | 290s |
| **10**np.linspace(10,1,100)*0.5** | **258** | **125s** |

Figure 1: Table of experimentation with alpha values for LASSO regression

The mean square error was approximately same for the bottom two rows. Based on this, we chose the last row, because it was faster in terms of computation.

# 5 RESULTS

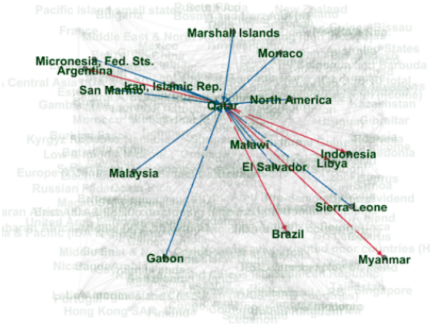
With the experiments performed above, we used Gephi for visualizing the results.

Figure 2 shows the connected graph, with each node representing a country and edges representing the correlation between each country, based on the coefficients calculated.

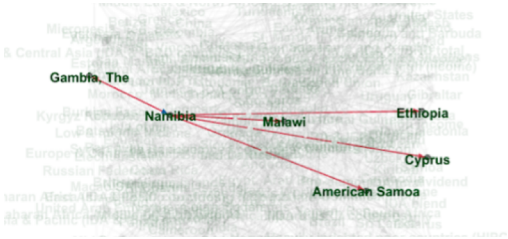
Figure 3(a) represents the connectivity for an example country (Qatar). It can be seen that the red edges (outgoing) represent 5 countries that this country is correlated to. The blue edges represent connections from other countries where Qatar has a correlation. It can be noted that every country will have

Figure 2: Connected graph representing all countries

exactly 5 red outgoing edges and any number of blue incoming edges. This is illustrated in Figure 3(b). For Namibia, there are exactly 5 outgoing red edges and 0 incoming blue edges. This means no country was found to have a direct correlation with Namibia.




(a) Connections for Qatar




(b) Connections for Namibia

Figure 3: Connections

Additionally, it is also observed that the coefficient value is taken as a weight for the edges. Figure 4 illustrates connections from Yemen Rep. to Turks and Caicos Islands as a thicker edge. Figure 5 has a table that lists the coefficient values from Yemen Rep. to all correlated countries.

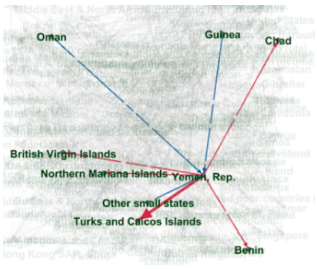



Figure 4: Connectivity from Yemen Rep

| Target | Weight |
|---|---|
| Benin | 1.039 |
| Northern Mariana Islands | 10.572 |
| **Turks and Caicos Islands** | **330.058** |
| Chad | 0.116 |
| British Virgin Islands | 22.315 |

Figure 5: Coefficients for Yemen Rep

# 6 CONCLUSION

From deeper network analysis, it was found that population data for any given country can be estimated using historical data from other countries. However, this exhibits some mean squared error, which is expected, because other factors like economic, political, geographical etc. are not taken into consideration. Lasso is an effective technique, especially for feature extraction and can provide reasonable correlation in spite of diversity and redundancy in the datasets. Future work can target other factors, along with a country's own historical data to give an even more accurate representation of a connected world.

# 7 REFERENCES

[1] An Introduction to Statistical Learning by Trevor Hastie, Robert Tibshirani, Gareth James, Daniela Witten, June 24, 2013
[2] Gephi Tutorial Quick Start- version 0.7alpha2