Correlates of War Lab

Matthew Boundy and Jasper Lemberg

***Abstract*— Using the National Material Capabilities Dataset from the CorrelatesofWar.org, we were able to analyze historical data on countries throughout the world and understand more about them. Using this data we did a variety of tasks including scaling all of the countries data and figuring out if we can measure how similar one country is to another.**

I. INTRODUCTION

This lab used Python and the Correlates of War data set in order to uncover key insights and information about various countries and their history of war after the post-Napoleonic era.

The Correlates of War Project was started in 1963 by J. David Singer at the University of Michigan in order to collect data on the history of wars and the conflict among states in the post-napoleonic era. Academic resources are used to publish datasets as professors and scholars publish datasources to the COW project. Some of the people involved with this project are Jacob Singer, Stuart Bremer, Scott Bennett, Glen Palmer, and Zeev Maoz who all served as directors of the project. Some of the universities involved in the project are the University of Michigan, Penn State, UC Davis, University of Arizona, and Michigan State to name a few. Some of the Datasets that are in this COW Project (As there are 15 total) are Militarized Interstate Disputes (records all instances of when one state threatened, displayed, or used force against another), Formal Alliances (records all formal alliances among states between 1816 and 2012, including mutual defense pacts, non-aggression treaties, and ententes), Direct Contiguity (registers the land and sea borders of all states since the Congress of Vienna, and covers 1816-2016), Diplomatic Exchange (tracks diplomatic representation at the level of chargé d'affaires, minister, and ambassador between states from 1817-2005), and Trade (tracks total national trade and bilateral trade flows between states from 1870-2014).

II. DATA

Looking at the Data we have the following Variables in the main dataset: stateabb, this is the three letter country abbreviation, so for example the United States is USA. ccode, this is the COW Country Code. year, this is the year of observation. irst, this is the iron and steel production in the thousands of tons. milex, this is the military expenditure, from 1816-1913 in the thousands of british pounds, from 1914+ in the thousands of US Dollars. milper, this is the military personel in the thousands. energy, this is the energy consumption in thousands of coal-tons equivalent. tpop, this is the total population in thousands. upop, this is the urban population living in cities larger than 100,000 in the thousands. cinc, Composite Index of National Capability score. version, this is the verision number of the dataset. In the supplementary version of the dataset there are sources and notes on each of the variable being measured and quality of the measurements given. There is also a new variable called pec which is the PEC score for the given country at the given time.

A. Q2

Version 5.0 is not the most recent version of National Material Capabilities dataset, the most recent version of the dataset is version 6.0. Version 5.0 expanded the data into 2012 from 2007 and also added additional documentation sources. Looking at the Variables we have the following in the dataset: stateabb, this is the three letter country abbreviation, so for example the United States is USA ccode, this is the COW Country Code year, this is the year of observation irst, this is the iron and steel production in the thousands of tons milex, this is the military expenditure, from 1816-1913 in the thousands of british pounds, from 1914+ in the thousands of US Dollars milper, this is the military personel in the thousands energy, this is the energy consumption in thousands of coal-tons equivalent tpop, this is the total population in thousands upop, this is the urban population living in cities larger than 100,000 in the thousands cinc, Composite Index of National Capability score version, this is the verision number of the dataset In the supplementary version of the dataset there are sources and notes on each of the variable being measured and quality of the measurements given. There is also a new variable called pec which is the PEC score for the given country at the given time. The data was collected from academic journals such as the Journal of Peace Research. Causal and Predictive Data are both possible, but Causal is a bit harder to figure out since some of the data is from 1816, which makes it much harder to find accurate data of all measures. Predicitive could be used to figure out certain values given other values are happening.¶

B. Q5: Heatmap

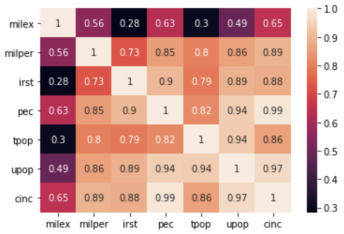

Fig. 1. Correlation Heatmap

C. Q5: Continued

"milper" and "pec", "tpop" "upop", and "cinc" as well as "irst" and "pec", "upop", and "cinc" and "pec" and "tpop", "upop", and "cinc" are highly correlated. If they are included as explanatory variables in the same predictive model, then there would be a stronger trend with the rest of the data. Negative correlation would cause a negative trendline versus a positive one.

III. RESULTS

A. Q3

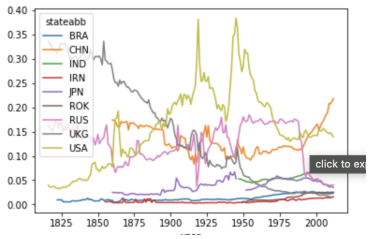

Fig. 2. Line Graph

Looking at this graph the USA did better compared to everyone else between 1900 and 1975, when it was usurped by Russia and then China. USA has sharp increases around 1860, 1920, and 1945. It decreases during the Great Depression and decreases from 1945 onward.

B. Q4

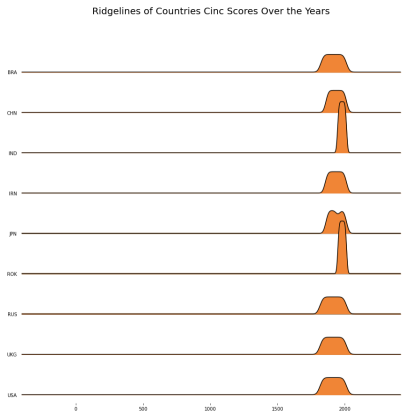

Fig. 3. Rigeline Plot

Most of the countries above are normally distributed, with India and South Korea being most normally distributed. Japan is the only country that is not normally distributed because it has two peaks. Besides that, most of the plots have a fairly large plateau at the top and do not go immediately down after going up.

C. Q6

The ten smallest euclidean pairs are the following:

- 'NAU –> TUV': 3.820809191230833e-07
- 'MSI –> SKN': 2.2438762566501622e-06
- 'TON –> KIR': 3.0013776638230176e-06
- 'LIE –> MNC': 3.1273620535521213e-06
- 'SVG –> GRN': 3.7921268585288873e-06
- 'LIE –> SNM': 4.770612484460835e-06
- 'WSM –> SLU': 6.243581795223367e-06
- 'NAU –> PAL': 8.304173492666257e-06
- 'AND –> DMA': 1.13276079853209e-05
- 'GRN –> SEY': 1.841901735609515e-05

The ten smallest manhattan pairs are the following:

- 'NAU –> TUV': 3.820809191230833e-07
- 'MSI –> SKN': 2.2438762566501622e-06
- 'TON –> KIR': 3.0013776638230176e-06
- 'LIE –> MNC': 3.1273620535521213e-06
- 'SVG –> GRN': 3.7921268585288873e-06
- 'LIE –> SNM': 4.770612484460835e-06
- 'WSM –> SLU': 6.243581795223367e-06
- 'NAU –> PAL': 8.304173492666257e-06
- 'MSI –> DMA': 1.13276079853209e-05
- 'DMA –> AND': 1.841901735609515e-05

These results make sense because all of these countries are small countries that do not have large militaries.

D. Q7

Although the map did not fully load, one can assume that countries with a lower cinc value have more imbalance, whereas countries with a higher cinc value have less imbalance. Conflict will most likely happen in areas with higher imbalance (i.e. not North America and Western Europe), which I believe is fairly accurate.