# Correlates of War Lab Report

Xuanyu Shen, Yushu Lyu

***Abstract*— The Correlates of War (COW) project is a study about the international relationships and especially focuses on the history of wars. The idea behind of the project is to use accurate and reliable data in scientific studies to extrapolate the correlates of wars from different factors of different countries over the world**

## I. INTRODUCTION

The history of the COW project traces back to the year of 1963, found by J.David Singer, from the University of Michigan and thus has its nature as an academic resources and strictly not for any commercial uses. The dataset are published by the directors as well as in cooperation with advisory board with both being responsible for keeping track of the data sets. There are many data sets with unique features that makes up the project. This includes "State System Membership" that marks the change of composition over time; "COW War Data" that serves as the datalist that contains a list of wars since 1816; The "Militarized Interstate Disputes" dataset that keeps on track when one state threatens another; and most importantly, the "National Material Capabilities" which is used in this report that contains important annual values espeically the six indicators;and many more datasets such as religion data, alliances, territory changes and so on.

## II. DATA

The "v. 5.0 of National Material Capabilities dataset" is currently not the latest version as there is another v.6 version avaliable. As of the v.5 version, the main updates comparing to the v.4 version is the update at the year of 2017, by extending the data for the six components(NMC) to 2012. In addition, urban population data was changed significantly as now is the data provided by the UN, and the procedure in calculation of the Composite Index of National Capabilities(CINC). Starting with the “stateabb” code which is a 3 letter abbreviation for country names; comes to the important measures section with “irst” being the iron and steel production and unit is "thousands tons"; “milex” is the measure of military expenditures; “milper” is the military personnel number being recorded in unit of thousands; “pec” simply being the energy consumption with its unit in thousands of coal-ton and equivalent amounts; “tpop” stands for the total population where as “upop” is the urban population; lastly it is the “cinc”, the composite index of national capability score being less than 1. The supplementary data file for v.5 contains all core variables included in the v.5 file of NMC dataset while having additional information. Supplementary information is added for the six main variables including "milex", "milper", "irst", "pec", "tpop" and "upop". While "milex" and "milper" both have a source and note file; "irst", "pec", and "tpop" are supplemented by another set of code information, quality code and the anomaly code which are also named by addition of strings; lastly, "upop" has all the additional materials listed above as well as a "growth rate" variable with its source. Most sources comes from recognizable sources: ”U.S. Department of Commerce and Bureau of the Census","Historical Statistics of the United States", or "UN Energy Statistics Database". However, one thing might cause problems in causal inference is the consistency of the sources. Even for the same country, the sources varies a lot as the time goes and the accuracy may not be as high if one single source is used for the entire time.

## III. ANALYSIS

### A. Descriptive Analysis

Based on the observation from the graph (Fig. 1), it can be tell that even with top 10 countries in the ’cinc’ measure, they tends to vary a lot, especially looking back through the history. Before in-depth analysis can be made, according to the Anscombe’s Quartett, the graph illustrates that only five countries are recorded all the way back from the year of 1816; three countries started around the year of 1860 being CHN, JPN and IRN; and two countries doesn’t have any data until the year around 1950. Starting from the modern time, at the year of 2012, CHN and USA are on the top half(in CINC value) as other countries remain in the bottom half of the group. RUS had a high value between the year of 1950 till the late 80’s and UKG was started as the highest, around 0.35, but was decreasing for the entire 200 years time. Rest countries were basically stayed at the lower half only with occasional rises.

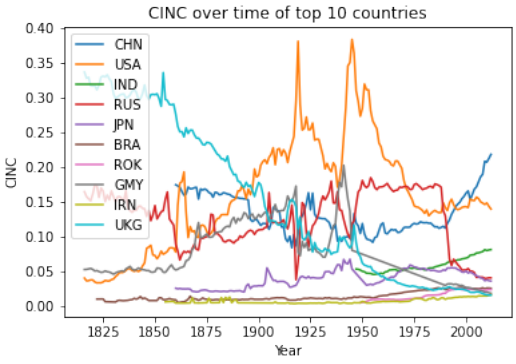


Fig. 1. CINC TOP10

The ridge line plot (Fig. 2) provides the view of change in distribution of certain measure. From the graph above,

it can be tell that countries "BRA", "IRN", and "ROK" all have a crowded cinc value while "JPN" and "IND" have a little spreader distribution, "CHN", "GMY" and "RUS" have a more spreading distribution that covers a larger range and lastly, "USA" and "UKG" have the spreadest value that covers all the way from range 0.0 to about 0.4. The distribition of the CINC data indicates in the past 200 years, or 100 years for some countries, the cinc values that they've had. Distributions that look not normal include "IRN", "USA", "UKG" where IRN has all it's data distributed at one point, and the other two data are perfectly spreaded accorss the range and even not that visible on the plot.

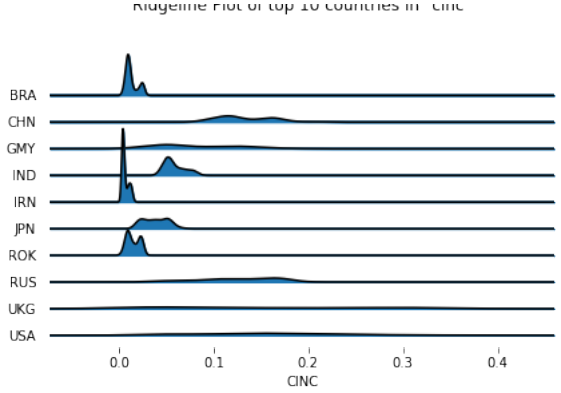


Fig. 2. Ridgeline for CINC Top10

## B. Causal Inference

According to the correlation heatmap (Fig. 3), there are several variables that might be naturally correlated. Firstly is the "milex" and the "milper" where when a country spend more fund in the area, more people must be recruit in order for the system to operate, whether it's researcher, armed force in the front line, back line, or officers. "irst" and "pec" can also have the strong positive correlation as one countnry produce more iron and steel, energy are consumed by producing, transporting and usage of these production. "pec" also has the positive correlation with both "tpop" and "upop" because based on theory, the more poeple there are, especially for those in the urbanized area, more energy will be consumed. Last but not least, "milper" might have a correlation with "tpop" because more people the avaliable, the more armed force there could be. Taking it back to the study itself, these variables that are potentially correlated with each other by nature will actually decrease the actual correlation than what is being displayed from the data or the graph. Due to the fact they are already correlated, only a small positive correlation makes a big correlation on the heatmap, especially the strong correlated factors.

On the other hand, variables with potential negative correlation by nature will strengthen the outcome of the study. As if the natural correlation is defined to be around "-0.5", then a positive "0.5" displayed on a heatmap might potentially indicating a positive "1.0" correlation. This moves the level of correlation from low to high, dramatically changes the study

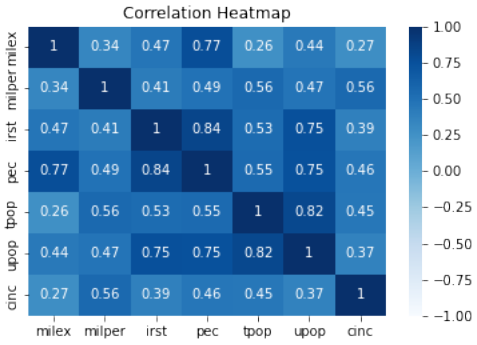


Fig. 3. Correlation Heatmap

outcome and the idea behind it. Next, we utilized Euclidean distances and Manhattan to calculate the similarities between each country. The top 10 smallest Euclidean distance country pairs are: 'TUV NAU', 'SKN MSI', 'KIR TON', 'MNC LIE', 'GRN SVG', 'SVG TON', 'LIE SNM', 'GRN TON', 'MNC SNM', 'SLU WSM'. The top 10 smallest Manhattan distance pairs are: 'TUV NAU', 'SKN MSI', 'KIR TON', 'MNC LIE', 'GRN SVG', 'GRN TON', 'SVG TON', 'MNC SNM', 'LIE SNM', 'SLU WSM'. We can see all these pairs are small countries. We think the reason is that the values of features that count in the distance after standardization are very small compare to those big, developed country. Thus the value difference of each feature between each small country is tiny. Thus, their distances are very close.

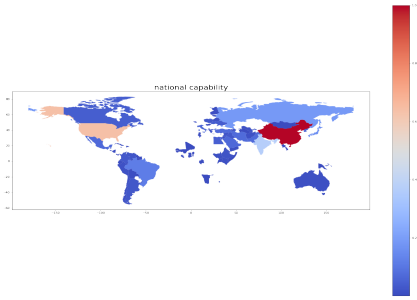


Fig. 4. Worldmap

## IV. Results

The worldmap (Fig. 4) below generates a clear visualization of the "cinc" distribution over the world wide region. The more power imbalance happens around the Middle East regions and some part of Asian, since China has the higher rate of "cinc" than their surrounding countries. The power balance regions are Africa and Europe regions, since the previous is evenly undeveloped and the latter is evenly developed. Based on observations, the Middle West regions may have high probability to engender wars, and this comes true now, since there's a conflict between Ukraine and Russian.