Correlates of War Lab 2

Xubin Lou and Jiayue Meng

***Abstract*— Through summarizing data about the history of wars, conducting descriptive analysis with ridge plot and heatmap of correlation, and analyzing causal inference by calculating the Manhattan and Euclidean distances and creating colored world map with different power of countries, the report of the Correlation of Wars project tries to figure out the conflict among states and further study the cause of warfare.**

I. INTRODUCTION

We expected to find that the power of countries and the national power gap might lead to conflicts. We found the distributions of the top 10 "most capable" countries. We also found the correlations between each pair of variable and used the correlation to find effects that would have on a potential outcome variable. We pointed out the most similar countries are TUV Tuvalu) and NAU (Nauru), two island countries in Oceania, by using calculating the Euclidean and Manhattan distance. Besides, by mapping the variable cinc (Composite Index of National Capability), we discovered that regions with larger difference of cinc are more power-imbalanced, so that they are more likely to have conflicts. We used history to support this idea. The history showed that some regions that we expected to have conflicts did have conflicts in the past decades.

II. DATASET

The Correlates of War project is an academic study of the history of warfare. It was started in 1963 at the University of Michigan by political scientist J. David Singer. It's an Academic resource.

Many individuals contributed to the collection of national capabilities data and this documentation/coding manual over many years. Below are the people and institution related for the COW research dataset:

version 3:Reşat Bayer, Diane Dutka, Faten Ghosn, and Christopher Housenick

version 1990:Paul Williamson, C. Bradley, Dan Jones, and M. Coyne.

Update in June 2010:J. Michael Greig and Andrew J. Enterline directed the update with the assistance of graduate students, Christina Case and Joseph Magagnoli.

Update carried out at UNT that concluded in October 2017: J. Michael Greig and Andrew J. Enterline directed the update with the assistance of graduate student Chis Macaulay.

Institution is the UNT Department of Political Science.

Data sets developed by this project are :

COW Country Codes: The list of states with COW abbreviations and ID numbers

State System Membership (v2016): This data set records the fluctuating composition of the state system since 1816. It also identifies countries corresponding to the standard Correlates of War country codes.

COW War Data, 1816 - 2007 (v4.0): The new list of wars that will be included in the COW war databases is available. Non-State War data set (v4.0), Intra-State War data set (v5.1), Inter-State War data set (v4.0), and Extra-State War data set (v4.0) are now available.

Militarized Interstate Disputes (v5.0): This data set records all instances of when one state threatened, displayed, or used force against another. Version 5.0 covers the 1816-2014 period.

National Material Capabilities (v6.0): Power is considered by many to be a central concept in explaining conflict, and six indicators - military expenditure, military personnel, energy consumption, iron and steel production, urban population, and total population - are included in this data set. It serves as the basis for the most widely used indicator of national capability, CINC (Composite Indicator of National Capability) and covers the period 1816-2016.

Militarized Interstate Dispute Locations (v2.1): This data set records the geographic locations of MIDs in latitude/longitude coordinates, per dispute and per incident. Dispute version 2.1 (MIDLOC-A) covers the 1816-2010 period. Incident version 2.1 (MIDLOC-I) covers the 1993-2010 period. The population of disputes and incidents in both datasets match the population of disputes and incidents in MID v4.3.

World Religion Data (v1.1): This data set aims to provide detailed information about religious adherence worldwide since 1945. This data set is hosted by Zeev Maoz, University of California-Davis, and Errol A. Henderson, Pennsylvania State University.

Formal Alliances (v4.1): This data set records all formal alliances among states between 1816 and 2012, including mutual defense pacts, non-aggression treaties, and ententes. This data set is hosted by Douglas Gibler, University of Alabama.

Direct Contiguity (v3.2): The Direct Contiguity data set registers the land and sea borders of all states since the Congress of Vienna, and covers 1816-2016. This data set is hosted by Paul Hensel, University of North Texas.

Territorial Change (v6): This data set records all peaceful and violent changes of territory from 1816-2018. This data set is hosted by Steven V. Miller, Clemson University.

Colonial/Dependency Contiguity (v3.1): The Colonial/Dependency Contiguity data set registers contiguity relationships between the colonies/dependencies of states

(by land and by sea up to 400 miles) from 1816-2016.

Intergovernmental Organizations (v3): Although the number of intergovernmental organizations (IGOs) grew dramatically during the late 20th century, they have been part of the world scene for much longer. This data set tracks the status and membership of such organizations from 1815-2014. Access information about this data here. This data set is hosted by Timothy Nordstrom, University of Mississippi, Jon Pevehouse, University of Wisconsin and Megan Shannon, Colorado-Boulder.

Defense Cooperation Agreement Dataset: This dataset covers bilateral defense treaties from 1980-2010. This data set is hosted by Brandon Kinne at University of California, Davis.

Diplomatic Exchange (v2006.1): The Diplomatic Exchange data set tracks diplomatic representation at the level of chargé d'affaires, minister, and ambassador between states from 1817-2005. This data set is hosted by Reşat Bayer, Koç University.

Trade (v4.0): This data set tracks total national trade and bilateral trade flows between states from 1870-2014. This data set is hosted by Katherine Barbieri, University of South Carolina, and Omar Keshk, Ohio State University.

v.5.0 is not the latest version of the dataset, the latest version v.6.0, and version v.4.0 of National Material Capabilities are available. The updates that v.5.0 includes are adding more rows of data from 14199 to 15171 rows, and the updated version column is shown from 4 to 2011.

The variables in the dataset NMC50-wsupplementary represent:

statenme is State name

stateabb means the 3 letter country Abbreviation

ccode is the COW Country code

year means Year of observation

milex means Military Expenditures (For 1816-1913: thousands of current year British Pounds. For 1914+: thousands of current year US Dollars.)

milexsource is Source of military expenditures

milexnote is Notes of military expenditures

milper means Military Personnel (thousands)

milpersource is id Source of military personnel

milpernote is Notes of military personnel

irst means Iron and steel production (thousands of tons)

irstsource is Source of Iron/Steel

irstnote is Notes of Iron/Steel

irstqualitycode is Quality Code of Iron/Steel

irstanomalycode is Anomaly Code of Iron/Steel

pec means Primary energy consumption (thousands of coal-ton equivalents)

pecsource Source is energy consumption

pecnote Notes is energy consumption

pecqualitycode is Quality Code of energy consumption

pecanomalycode is Anomaly Code of energy consumption

tpop means Total Population (thousands)

tpopsource is Source of total population

tpopnote is Notes of total population

tpopqualitycode is Quality Code of total population

tpopanomalycode is Anomaly Code of total population

upop means Urban population (In thousands. For 1816-2001: population in cities w/>=100k; For 2002-2012: population in cities w/>=300k)

upopsource is Source of urban population

upopnote is Notes of urban population

upopqualitycode is Quality Code of urban population

upopanomalycode is Anomaly Code of urban population

upopgrowth is Growth rate of urban population

upopgrowthsource is Growth rate source of urban population

cinc means Composite Index of National Capability (CINC) score

version means Version number of the data set

The variables in the dataset v.5.0 represent:

"stateabb" means the 3 letter country Abbreviation

"ccode" is the COW Country code

"year" means Year of observation

"irst" means Iron and steel production (thousands of tons)

"milex" means Military Expenditures (For 1816-1913: thousands of current year British Pounds. For 1914+: thousands of current year US Dollars.)

"milper" means Military Personnel (thousands)

"pec" means Primary energy consumption (thousands of coal-ton equivalents)

"tpop" means Total Population (thousands)

"upop" means Urban population (In thousands. For 1816-2001: population in cities w/>=100k; For 2002-2012: population in cities w/>=300k)

"cinc" means Composite Index of National Capability (CINC) score

"version" means Version number of the data set

According to showing the distinct values of the source columns, we can see there are so many sources for milex, milper, upop, and upopgrowth. Thus, the data was collected from various sources such as governments, official departments, institutions, and recording materials from different countries. This poses a problem since there are too many sources for the data in supplementary dataset. Such many data sources will cause problems like measurement bias and exogenous factors in the causal inference studies. Since there are many various sources, it's hard for people to check the authorization, some measurements of the data might be poor due to poor-trained skills and unexact tools. Besides, the method of measure some kind of data will be also different by different sources, thus hard to unite the measurement scale and completely correct some measurement errors on time.

III. RESULT

The top 10 "most capable" countries are CHN, USA, IND, RUS, JPN, BRA, ROK, GMY, IRN, and UKG.

In the plot, the general range of the cinc for the top "most capable" countries overtime was from 0 to 0.4. The cinc of UKG had a general decrease trend, the cinc data of IRN and BRA were the smallest, which are around 0.02; the cinc of USA fluctuated a lot from around 1900 to around 1975 and reach the highest cinc which is around 0.4 in about 1925

and 1950. JPN had the third-lowest cinc, which is around 0.05; other countries' cinc fluctuated between 0.05 and 0.2. The cinc of CHN only had data after around 1925, and had a fluctuate around 0.1 to 0.15 during 1925-1975, and a obviously increasing trend after 1975. The cinc of IND only had data after about 1950 and had a slightly increasing trend.

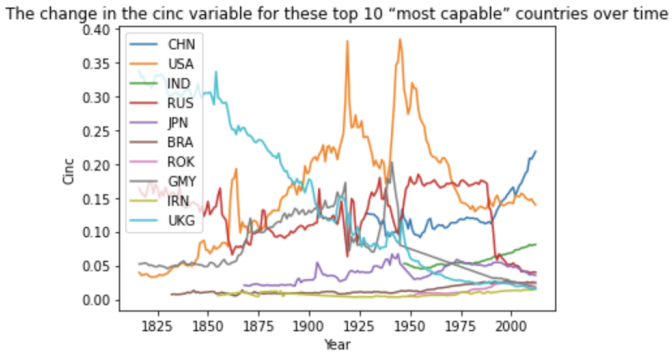

Fig. 1. Plot

In the ridge plot above, we can see the cinc of UKG, USA, GMY, RUS, and CHN are in nearly normal distribution since they are all nearly bell-shaped, while the cinc of other countries are obviously not normally distributed.

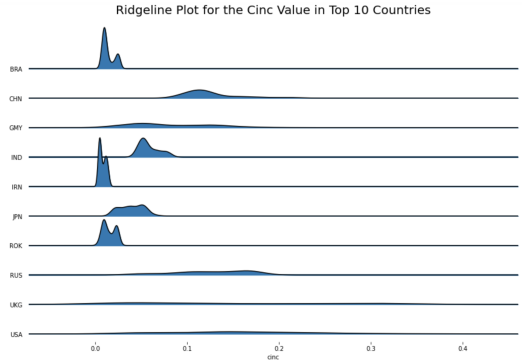

Fig. 2. Ridge plot

The variables cinc and pec might correlate by nature since the Composite Index of National Capability (CINC) score mainly shows the power of a country, which is highly close to the Primary energy consumption (thousands of coal-ton equivalents). The higher pec means higher energy consumption in various fields like military, industry, production, and technology, which are the main source of the power of a country. Thus, it makes sense that the higher energy consumption comes with a higher Composite Index of National Capability. The cinc and pec have a strong positive correlation by nature.

According to the correlation heatmap Figure 3, we found cinc and pec have a strong positive correlation near to 1, which also proves my analysis of their correlation by nature. Besides, in the heatmap, we can also find the correlations among milper, irst, pec, tpop, upop, and cinc are generally high, which means there are relatively positive correlations among these variables.

If a variable is positively correlated by nature with another explanatory variable X in the same predictive model: If X is not close to zero, then X may get closer to zero when the positively correlated explanatory variable is added. It will mask the strength of the effect of explanatory variable X to the potential outcome variable Y. If X is close to zero, then X may change its sign when the positively correlated explanatory variable is added. This will compensate for the strength of the effect of X on the potential outcome variable Y.

If a variable is negatively correlated by nature with another explanatory variable X in the same predictive model: If X is not close to zero, then X may get further away from zero when the negatively correlated explanatory variable is added. This will overestimate the strength of the effect of X on the potential outcome variable Y. If X is close to zero, then X may change its sign/get further away when the negative correlated explanatory variable is added. This will mask the effect of X on the outcome variable Y.

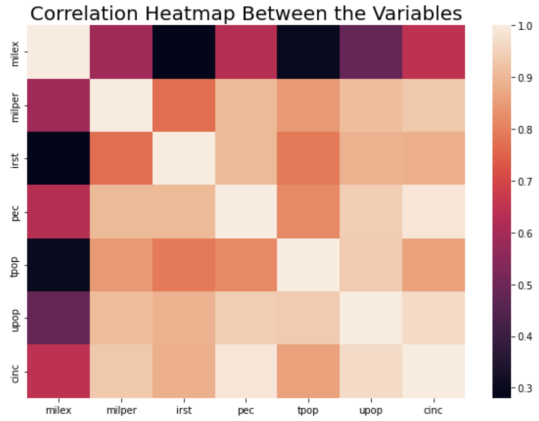

Fig. 3. Heatmap

Figure 4 is the standardized variables between 0 and 1 of the dataset.

stateabb	milex	milper	irst	pec	tpop	upop	cinc
USA	1.000000	0.686652	0.121327	0.592435	0.230561	0.417286	0.638890
CAN	0.028144	0.028884	0.018476	0.064312	0.025292	0.046898	0.041973
BHM	0.000084	0.000438	0.000000	0.000049	0.000263	0.000000	0.000113
DOM	0.000554	0.010941	0.000000	0.001378	0.007456	0.007480	0.003712
JAM	0.000212	0.001313	0.000000	0.000471	0.002004	0.001331	0.000787

Fig. 4. Standard

The top 10 smallest Manhattan and Euclidean distance pairs are shown in Figure 5.

From the data shown below, TUV (Tuvalu) and NAU (Nauru) seem to be most similar to each other. I expect these results. TUV is an island country in the Polynesian subregion of Oceania in the Pacific Ocean. NAU is an island country and microstate in Oceania, in the Central Pacific.

It further lies northwest of Tuvalu. Both countries have a similar geographic location. They are both island countries in Oceania, and they are very small. They both have very small populations. They are both one of the least developed countries in the world. It is reasonable that they are very similar.

```
Euclidean
TUV_NAU    3.820816e-07
SKN_MSI    2.243880e-06
KIR_TON    3.001378e-06
MNC_LIE    3.127367e-06
GRN_SVG    3.792130e-06
SVG_TON    3.901145e-06
LIE_SNM    4.770614e-06
GRN_TON    4.968829e-06
MNC_SNM    5.269978e-06
SLU_WSM    6.243583e-06
dtype: float64
Manhhatan
TUV_NAU    4.483291e-07
SKN_MSI    3.292861e-06
KIR_TON    3.971528e-06
MNC_LIE    4.230296e-06
GRN_SVG    5.387926e-06
GRN_TON    5.837448e-06
SVG_TON    6.259021e-06
MNC_SNM    6.854191e-06
LIE_SNM    7.007295e-06
SLU_WSM    8.471966e-06
dtype: float64
```

Fig. 5. Top 10 Euc and Manh

From other data of these two countries, we can see that TUV and NAU have the same milex, milper, irst, tpop, and upop. They also have similar pec and cinc. That also explains why they are the countries that are similar to each other the most.

stateabb	milex	milper	irst	pec	tpop	upop	cinc
TUV	0.0	0.0	0.0	0.000000e+00	0.0	0.0	0.000000e+00
NAU	0.0	0.0	0.0	3.749738e-07	0.0	0.0	7.335534e-08

Fig. 6. TVU-NAU

Asian and North America seem to have a great amount of power imbalance. We can see from the map that China is in dark red while countries around it are in light yellow or light orange. This means that there is a power imbalance in this region. Also, the USA is in dark orange while all other countries in North and South America are in light yellow. This shows a very clear pattern of power imbalance. Regions like the Middle East, South America, West Europe, East Africa, and Australia are with more power balance because there are no obvious color differences in those regions. I think conflicts may happen in regions where imbalance power appears. I expect conflicts between Russia and West Europe, China and other Asian countries, and India and countries in the Middle East. I also expect conflicts between the USA and Canada, and the USA and countries in South America. Comparing my guesses with reality, I found some of my expectations came true. Since CINC is not the only factor that will influence world conflicts, and conflicts could be affected by countries' policies, attitudes, trades, and other things, it is reasonable that some of my expectations didn't come true. It came true that there were conflicts between the USA and North American countries like Mexico about Mexico–United States border. There are also conflicts between Russia and Ukrainian like Russo-Ukrainian War, an ongoing and protracted conflict between Russia and Ukraine that began in February 2014, and the ongoing military conflict in Donbas. China and Japan also have an unstable relationship.

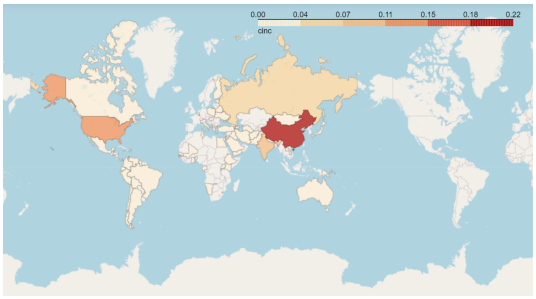

Fig. 7. Colored World Map