Contents

SURF2017 Daily Record

Zhuoqun Liu, Jiaqi Yu, Zhong Chu

August 23, 2017

1 This is a diary of SURF 2017

Let's write something about our daily progress.

1.1 Today is July 12, 2017

We had a meeting today and talked about "data washing"
The location dataset (lat,lng) is from Geonames.org
City names are provided by Geonames and geodataset
conflict data is from COW (correlations of war) inter state war
our tasks now, is to wash the data merge the data, do any possible criticism research
The first step I will take is to design such a table with reasonable fields describing wars, cities so that the data can be utilized easily.

1.2 Today is July 13, 2017

We had a discussion about what should be included in our criticism research on wars since World War One (1914) based on COW Inter-state war dataset.
There are about 112 wars in our scope, while 226 countries are included in the wars.
COW dataset provides Start Date, End Date, Death Count, Result
For our research, we need The reason of the war, The reason of the end of the war, The cities that are influenced by the war, locations (lat,lng) of the cities.
The reasons of wars varies, thus we three plan to take World War One as a trail so that we can perform the calibration. We want to decide how detailed the reason should be for the sake of feasibility and then we will encode each of the reasons(begin or end). We may also write an abstract about each war including the reasons and other attributes that are logged in the spreadsheet.
Define "cities influenced by the war"
The cities should be where the war's (first) battle begins. These cities are generally special ones in their countries in terms of traffic, economy, culture, geography, that the combatant

countries chose to start there. For a similar reason, the cities that a commander will choose to start a campaign or some important operation are also worthy of our notice. We should note down the names of the cities and their countries, location of the cities (location in this diary means latitude and longitude coordinate). By searching about certain war, reading any reliable history record or research, the influenced cities should be found. Then we may use Geonames cities dataset or google map API to find the exact location of the cities. Tableau, excel, gephi can perform the data blending which is a process matching city names with their coordinates.

1.3 Today is July 15, 2017

A reduced version of this classification, I think, should be:
For nature resource, for faith of religion, for the genuineness of a regime.
Reasons of this classification should be given. (i.e.the references) Hope I can find them tomorrow.
Cities that are influenced by the war can be divided in to groups in terms of the size of combat at that place. The location where the war begins must be in our list. Then comes the cities that can be found in records due to some important military operation or campaign. Maybe three cities for each war is enough but still, we need to find supportive material to make this decision.

1.4 Today is July 17, 2017

Discussion about classification
In our dataset, the reason will focus on the apparent reasons not the aim underneath the declaration and behavior.
Territory: Without knowing and looting nature source
Colony: Invade to build colony
Invade for plunder (nature resource and residents' wealth)
Diplomatic reason (international Alliances or military organization)
Strong Faith of certain religion (fight for taking ownership of some holy places)

1.5 Today is July 18, 2017

For each interstate war in COW dataset, we will find the influenced cities, forming a spreadsheet including city names and the corresponding events. The fields defined in "cities" spreadsheet is decided as:

- Fields in "cities" spreadsheet
    1. Warname
    2. continent
    3. country or region

4. city or region
5. latitude
6. longitude
7. event
8. year
9. month

- Fields in "Reason_and_result" spreadsheet

1. warname
2. reason
3. result(winner)
4. sidea
5. sideb
6. result(event)

- the field "reason" has a standardized categorization
  The reasons are described from the starter's perspective.

1. Economic Gain
   One country wishes to take control of another's wealth, such as precious materials, oil and minerals.
2. Territory Gain
   A country might decide that it needs more land, either for living space, or for agricultural use, or for other purposes.
3. Religion
   Religious wars can be related to the ownership of holy places or tied in with other reasons for conflict.
4. Nationalism
   Nationalism essentially means attempting to prove that country is superior to another by violently subjugationC this often takes the form of an invasion.
5. Revenge
   Seeking to punish, redress a grievance, or simply just strike back for a perceived previous slight can often be a factor in wars.
6. Diplomatic Reason
   the country is in some International military Alliances, thus is involved in the war without initiating it.

7. Independence
   Liberation war, war of freedom, war for establishing a national government. Including unitize a part of the country which has some power at back.
8. other reasons
   There are wars that do not apparently belong to the categories mentioned above, for instance, some wars of "country level" like China(PRC) against Taiwan(RoC), however, both sides are claiming China is one country which implies the war is a civil war. Further more, the nationalist party of China has lost its domination of Taiwan, thus Taiwan is no longer the RoC which appears to compete with PRC for the name of "China".

The classification is based on "Classification of Interstate Wars"
I suddenly think of checking whether the city names have changed through this hundred years.
The common process of finding out which cities were affected is firstly search on both wikipedia and Britannica Encyclopedia so that the process and details of battles can be confirmed. Afterwards details such as date and locations will be compared through different academic files.
Up to now three battles are confirmed as temples, which are Estonian Liberation, Latvian Liberation and Russo-Polish war.

1.6 Today is July 23, 2017

According to my working process this week, I'd like to conclude a typical procedure that is used to find required information we need.

- Use of Britannica Encyclopedia
  Our work follows COW interstate war list. Each war has a name given by COW dataset, thus the name is the keyword to search. At most cases, terms of countries in Britannica can provide information about conflict history of the country or a region or there are even terms about famous warfare. Some of the cities or regions that are involved in such wars can be found by reading the article, the spreadsheet named "cities" comes from this method. For now, therefore, the reference material is only Britannica and Google map.

- Google Map for Geo Coordinate
  When a city is confirmed by Britannica, I will search its name with Google Map. A name of a city or region is not exactly a point, therefore, the coordinate I chose doesn't represent strictly a point, but a point within the region described by Britannica, especially for some "places" that are actually borders of countries.

- Questions to ask

1. Reason categorization. is our standard acceptable
2. MATLAB structure field design
3. Geo coordinate reasoning. e.g. border dispute
4. what kind of algorithm to use
5. How to write reference about cities and warfare events
6. WWI and WWII are special
7. Overview of this project, is the most important attribute the locations that wars and conflicts happen?
8. Ways to present data and result analysis, like network analysis(diagram)?

Examine these cities first. algorithm: betweenness

1.7 Today is July 27, 2017

The term Vietnam War from 1954 to 1975 in Britannica Encyclopedia may be divided by COW as Vietnam War, Phase 2 and Vietnam War, Phase 1. In COW interstate war dataset, Vietnam war, Phase 2 is from 1965 to 1975 (for South Vietnam) which is not the complete range according to Britannica. The war is referred to as Indochina Wars (the first and the second) by Britannica in the term of Vietnam
This war also led me to think about our field design and definitions of each field.

- Country or region
  means this region has a definite government which is recognized as a country level regime at least by another generally recognized country, for instance, Vietnam(a country), Palestine (a region that is recognized as a country by some other countries).

- City or region
  means that the region is at city level by convention (according to history or some generally accepted classifications), for instance, Jerusalem(a city under dispute), Washington D.C.(a city that is definitely a city).

For the case of Vietnam war, I decided to amend a row to COW table, called Vietnam War, Phase 1 from 1946 to 1954 describing the end of French colonial rule in Vietnam, Indochina. We may also develop attributes for Result(winner) field. Because sometimes both sides claims they are the winner and there may be some cases that no agreement is achieved whereas the battle stops.

1.8 Today is July 28, 2017

Proposal about "result (winner)" of wars:
At the beginning of writing down results of wars, we considered that each war must have a winner officially or a winner factually. After reading some record of wars, we realized that wars may have different results rather than just a winner side. There may be ceasefire agreement, treaty to stop war and finally there may be a definite winner side. Under the field "result(winner)", I'd like to add the following classifications:

1. winner side's name
   simply write down the name of the very winner,i.e. there should be definite winner and loser countries or sides.
2. cease-fire agreement
3. treaty to end war
4. draw
   nothing is signed, while both sides claim that they have won the war, or neither of the sides claim victory.

1.9 Today is July 30, 2017

Constructing a MATLAB structure array of our war data is the aim. We now have these many fields:

1. Warname
2. continent
3. country or region
4. city or region
5. latitude
6. longitude
7. event
8. year
9. month
10. reason
11. result(winner)
12. sidea

13. sideb
14. result(event)
15. source
16. URL&citation

1.10 Today is July 31, 2017

In MATLAB, csv(txt) file can be read by read().
The function will read csv spreadsheet as a table object. Then MATLAB can perform table2struct() to make the table object a struct array. The problem for now is we don't know the difference between 1 by 58 and 58 by 1 struct object with the same fields.

1.11 Today is August 6, 2017

Now we have produced a concatenated spreadsheet from the three spread sheets cities, reason and results, references. War names and city names can form unique pairs. reasons and results will go with war names while the reference informations goes with city names. In other words, the names of cities is absolute. There could be multiple reference sources for one city thus reference sources field leads to empty rows in cities column. Today's discussion will provide more about concatenation and field arrangement. After the final version of data cleansing, I wish we three have a read through to check whether there are apparent mistakes.We will then try to calculate the distance between places and try to find positions of world top 200 cities according to population and GDP. The distance between these cities should also be calculated.(in matlab) The task of calculation is for today and tomorrow.
The following is result(winner) classification proposed by Jiaqi on Aug 3rd result:

1. Battle continues (even with cease-fire agreement)
2. One government collapses (applied for civil war or one country is occupied)
3. Invader is driven out.

reason1 Border conflict could be categorized into territory gain. 2 Civil war could be categorized into independence.
calculate distance is our main task.
For our SURF project, we plan to give some geographical analysis base on our spreadsheet and the GTD data. We may be able to give some intuitive prediction based on our map of wars and terrorism. Betweenness is a concept that can be used to support the prediction according to the article we have read. The task for us is to understand betweenness centrality and try to use the concept. According to our last meeting with our supervisor, the very first step to obtain betweenness is to calculate the distance between each of our sampled cities. There are 115 cities(including some regions that are classified as city level in this research)

in our spreadsheet, thus a 115*115 matrix(table) can be used to represent distance between each two cities. We can also have geographical data of top 200 cities in the world in terms of city area or population, and the distance can also be calculated. The results will probably help with our prediction on where the wars are not likely to star or to affect, that is, which cities are good for people to live peacefully.

1.12 Today is Aug 8, 2017

MATLAB script to calculate distance between locations with give latitude and longitude is implemented. Requirement on input data is also clearer that those coordinates of cities should be unique though the names of cities may be the same in different countries. A better practice it to concatenate city name and its country tag so that all names are unique in the dataset. The next question is to understand "betweenness (centrality)".
We are planning to build a time line base presentation of wars and cities involved. Tableau can do this with filter. Gephi seems to have time line function built-in.
MATLAB can be used to calculate betweenness centrality, we shall then understand the concept and try to give some analytic conclusion and compare our results with the article we have read at the beginning.

1.13 Today is Aug 9, 2017

we may want some traffic data and trade data. IMF and World Bank may have the data, but the earliest data in the set is 1948. All the data are logged in names of countries not cities. There are videos on YouTube showing abbreviated air traffic in 24 hours, I wonder how can we do a similar thing to get data of transportation on land(train) and in the sea(by ship). Here are two tutorials on tableau time line implementation,time line by dragging a cursor and time line with details on a plane.
Wikipedia has explained "Betweenness Centrality" clearly. The problem is, what is the relation ship between this attribute and our calculated distance. Betweenness is a concept used in network, thus we need to find the network based on distance first, that is, to determine whether two places are connected or not using a distance criteria.

- we can try to verify the conclusion of betweenness
- if the result is not satisfying, we will find some other attributes.

1.14 Today is Aug 12, 2017

We had a discussion about SURF direction. We will firstly try to verify the results about betweenness mentioned in Weisi Guo's article that those cities with relatively high betweenness tend to experience violence. We currently have 127 samples(cities) extracted from recorded wars since 1914(the end of WWI), the next step is therefore finding the betweenness of these cities and compare our results with Guo's prediction.
However, there are other attributes for wars, like nature resource and trade record and even

grain storage which can reflect war preparations. New data sets are needed in the following step, but firstly we'd deal with betweenness.

1.15 Today is Aug 13, 2017

csvread() function in MATLAB is for reading numbers. I previously design scripts based on string and structure array but it really seems not convenient in MATLAB. data cleansing functions like merge and join really give me a better choice that to use numbers as id corresponding to the different rows. Basically, the structure of our location data is:

- id(number, unique i.e. enumerate)
- name(unique)
- latitude
- longitude

each row should be unique. id, lat, lng are friendly with MATLAB.

1.16 Today is Aug 14, 2017

To get the betweenness information for all the 127 cities, we need more generalized location data because connectivity is determined by distance threshold and the calculation of betweenness and betweenness centrality should use objective data rather than just 127 cities with betweenness with each other because all of the 127 cities are tagged as "battle field" which means they are special cases, there betweenness won't contribute to our expected conclusion.
Thus far, a MATLAB script is implemented to calculate distance between locations and sift out connected pairs using numerical id representing the locations, however, larger data set like cities15000.csv from geonames has over 25,000 rows of data(positions in lat,lng pairs) challenges my pc capability. Handling 127 cities betweenness looks fine, that the results can be seen in 3 seconds with 127*127 distance matrix behind, whereas 20,000*20,000 leads to about one billion terms to deal with, it's horrible for a laptop.

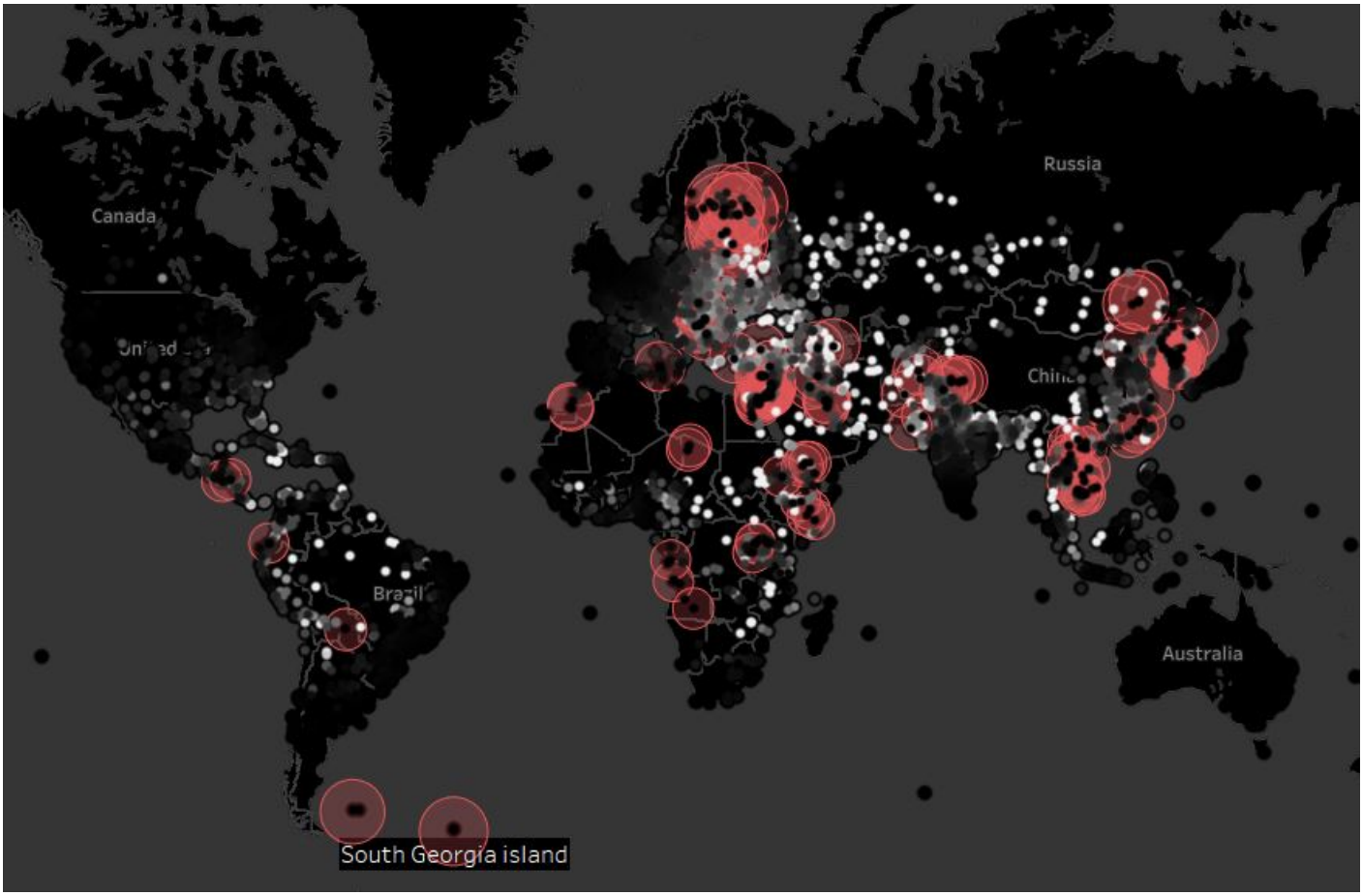
1.17 Today is Aug 20, 2017

An affordable location dataset of size 4000 cities is used. The dataset is selected from over 20,000 locations in geonames dataset cities15000.csv by population criteria that the top 2000 places and the bottom 2000 location are selected in order to balance location concentration due to high population density areas.
4000 cities information was thus formated for MATLAB input so that MATLAB script id_proto.m can calculate the distance (dependency: invoking id_dist.m inside the function) and output a csv file indicating connectivity of the locations according to the criteria – a radius, inside which will be considered connected.

The next step is to calculate (betweenness) centrality of each "node", any of the 4000 cities. MATLAB can handle this, however, I chose to use Gephi because it looks simpler(It Has GUI!!!...). By running Avg. Path length in statistic window ins Gephi, centrality should be assigned to nodes (4000), then the nodes information was exported.
In Tableau, 4000 locations are shown on world map. Betweenness centrality can be shown in color gradients so that the high or low centrality places are depicted clearly on that map.(the gradient restriction is adjustable)
Another layer of conflict affected area was add to the graph yesterday. Take each of the 127 recorded conflict location as the centers, choose 500km or so as area radius, 127 circular area covering some objective locations can be drawn in tableau using its polygon mark.
This demonstration needs vertexes information when data preparation. I used circlem script in MATLAB. The script can can convert a point with given latitude, longitude pair and radius to 100 constituting a contour of certain circle. What I've done is to construct an input and output shell for this core algorithm. There is one exception worth mentioning that the circle estimated by circlem.m script may not be shown as the same size on a planar world map. The reason for this observation is not clear yet, whereas here I have some hypothesis:

1. The algorithm have some defections inside, thus we may need to look into the code.

2. There are some mathematical transformation between coordinates unknown to us, then we may have to learn about some geographical basics to validate the visualization shown in tableau.



1.18 Today is Aug 21, 2017

Since our tableau demonstration can show certain level of correctness of that conclusion "there is correlation between war and geographic feature determined betweenness centrality of a location" intuitively, we may want to go further to predicting conflict or violence.
The first step is some "self-amending" that there are only 127 place found to have experienced war and they are covering a significant amount of white points (high centrality), however, there are some noticeable white points not covered by any of the red circles. Those "war-proof" areas may have some conflict or even wars that we didn't considered. Those place may have been in wars or under war threat now, if we can find some more evidence about whether those locations have experienced violence, we probably can push the map demonstration more accurate. Further more, this process is actually a trail of predicting regions of high vulnerability, be cause we suppose we only know the 4000 cities' centrality and their geographical distribution and we trace back to history for conflict records due to high centrality of such location or high concentration of high centrality places in some region (like northern-east India).
Here is the plan for the next step:

1. Take the 4000 cities data set as the finalized basic data set for our SURF
2. Filter out all points of high ($\geqslant 20000$) centrality (245 locations are extracted).
3. Filter out locations that are not covered by any wars according to our reddish war circular buffer.
4. Let's call them "Pseudo exceptions". And we give some "predictions" on the highly vulnerable areas.
5. We can check those predicted area manually, simultaneously when designing script and predicting.

This process may take 1 day, and currently I'm at the beginning of term number 3.

1.19 Today is Aug 22, 2017

For now, there are several variables appeared in our research and demonstration:

1. Number of Geographical position (number of base locations on map)
   We have tried 340, 100, 2000, 4000. The latest demonstration is created based on 4000 locations dataset. The volume of location dataset will change betweenness centrality and degree in terms of order of magnitude, thus the distribution shown by the dataset may be different (tend to be more detailed and easier to classify).
2. Criteria of colorization according to betweenness centrality
   By observing the distribution of centrality or degree, we can adjust restrictions of colorization in Tableau and Gephi when demonstrating our result (to make it more acceptable and persuasive).
3. Number of recorded conflict points
   We currently have found 127 locations recorded to have experienced war according to raw data provided by COW dataset of interstate war. This dataset is used to verify our hypothesis that centrality and degree of a location have some correlation with its risk of being directly affected by war or other kinds of violence.
4. Radius of estimated war-affected area
   We use circular buffer around each conflict point to depict each area affected by wars. Tableau is able to generate illustration based on this idea while a MATLAB script if_affected_confl.m will sift out all the locations circulated by conflict area estimation.

We want to find some threshold like properties for betweenness centrality and degree, thereby to predict some regions under high war or violence risk.
My idea thus far is to find concentration of high centrality points and the similar to degree.