Foreign Media Influence on the South Pacific Environment

Derek Lilienthal*[†]

Advisor: Dr. Elizabeth Gooch[‡][§]

Febuary 2021

Abstract

In this research, we quantify foreign actors media activities in the South Pacific involving an environmental theme. We use the GDELT (Global Database of Events, Language, and Tone) data set to compare the tones of articles produced by Western, Chinese, and South Pacific (Local) media sources that involve an environmental theme and when a great power (United States, China, Australia, New Zealand, Japan, and Russia) is involved as an actor. We found that when comparing Western, Chinese, and Local news sources, the average sentimental analysis of Western tones is negative, the average of Local tones is slightly positive, and the average of Chinese tones are very positive. When comparing the difference in means by each set of sources, we used the Welch's Two Sample t-test because the distribution of Western, Chinese, and Local tones followed a normal distribution but had unequal variances among groups. After conducting our statistical analysis, we found there is strong evidence to conclude the difference in means of tones between the three media sources are statistically significant between each pairwise comparison.

Introduction

The GDELT (Global Database of Events, Language and Tone) Project's goal is to create a platform that monitors the world's news media from nearly every corner of every country in print, broadcast, and web formats [1]. GDELT provides a big data resource to analyze the world's news that allows researchers to explore trends in media that was not easily accessible before. GDELT stores all its data in a relational database that is freely accessible through

*California State University, Monterey Bay, Seaside CA 93955

[†] dlilienthal@csumb.edu

[‡] Naval Postgraduate School, Monterey CA 93943

[§] elizabeth.gooch@nps.edu

Googles BigQuery platform. BigQuery allows researchers to use standard SQL to access the data. GDELT also updates it's database every 15 minutes [2]. This makes it possible to track events as they are happening in almost real time.

The amount of raw data that GDELT gives researchers is monumental. For example, in one of the relational database tables we focus on in this research, there are over five billion unique entries. The GDELT project database itself has over 50 tables available through BigQuery, but for the purpose of this research, we will solely be focusing on three of GDELT's tables. The Events, Eventmentions, and GKG (Global Knowledge Graph).

The Events table is where GDELT creates a new entry for each brand-new event, the Eventmentions table is where GDELT tracks the life of each event as it spans across more media outlets and as the event continues to develop, and the GKG table is a detailed analysis of every news article itself [3][4].

An example on how these three tables relate to each other, if Fox News is the first to report on about a political scandal, the event will first have an entry in the Events table. After 15 minutes, any follow-up news articles, or if a different media outlet like CNN, that report on the same event will all be seperate entries in the Eventmentions table. Each event is uniquely identified by a unique id (*GLOBALEVENTID*) in the Events table that is also present in each entry in the Eventmentions table. This way, we can join both the Event and Eventmentions table on each events' unique id. We can also track individual events by their *GLOBALEVENTID*. The GKG table however uniquely identifies each event by the article's URL. This way, we can join the Eventmentions and GKG tables through each events URL because the Eventmentions table also has the source URL of where the article originated from. Thus, allowing us to join the Events, Eventmentions, and GKG tables collectively to form our entire data set that we will be using throughout this research.

Each event in the Events table is stored in a CAMEO (Conflict and Mediation Event Observation) format where the two actors and the action performed by Actor1 upon Actor2 is recorded [4]. The Geographical location where the event took place is also recorded along with the geographical location where *Actor1* and *Actor2* reside from. If the event only involves one or no actors, then these attributes can be left blank [4].

The Eventmentions table is an extension of the Events table. It records all the mentions of each event as it spans across multiple news sources [4]. It will track an event as it spans past the first initial recording in the Events table. The Eventmentions table also includes more details about each entry compared to the Events table. Some of the note able additional attributes the Eventmentions table gives is the articles tone, actors character offset in the article, confidence level of the reported article is related to the *GLOBALEVENTID*, and many more attributes. While many of those attributes can be used to filter articles out articles based on length or the importance of an actor based on where it was located within the article, in this research the only attribute we are utilizing in the Eventmentions table will be its tone (*MentionDocTone*) [4].

The GKG table "connects every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its connect is and who's involved, and how the world is feeling about it, every single day".[3] In short terms, the GKG table gives an additional level of analysis on each event recorded in GDELT that can be used to filter articles by organizations, persons, themes, tones, locations, and more. For this research, we will be focusing on using the themes (*V2Themes*) and tones (*V2Tone*) of the GKG table.

For the remainder of the report, we will be focusing on analyzing Western, Chinese, and Local media tones across the South Pacific when a great power is either *Actor1* or *Actor2*, the location of the event is a South Pacific country or territory, and the event involves an environmental theme. The great powers of interest in this research are: The United States, China, Australia, New Zealand, Japan, and Russia.

The locations of interest in the South Pacific are: Micronesia, Fiji, Kiribati, Marshall Islands, Nauru, Palau, Papua New Guinea, Samoa, Solomon Islands, Tonga, Tuvalu, Vanuatu, Cook Islands, Niue, American Samoa, Ashmore Reef, Baker Island, Coral Sea, Easter Island, Galapagos Islands, French Polynesia, Guam, Howland Island, Jarvis Island, Johnston Atoll, Kingman Reef, Midway Island, New Caledonia, Norfolk Island, Norther Mariana Islands, Ogasawaramura Japan, Palmyra Atoll, Papua Indonesia, Pitcairn Islands, Tokelau, Wake Island, Wallis and Futuna, West Papua, and Bonin Islands. [i]

Data

In order to gain a perspective of how many articles are contained in GDELT and what portion of those are located in the South Pacific, we tabulated the total number of articles in GDELT and the total number of articles located in the South Pacific using two separate SQL scripts.

Listing 1: Tallying The Total Number Of Articles In GDELT

```sql
SELECT
    count(1)
FROM
    `gdelt-bq.gdeltv2.eventmentions` AS em JOIN `gdelt-bq.gdeltv2.event` AS e
   ON em.GLOBALEVENTID = e.GLOBALEVENTID JOIN `gdelt-bq.gdeltv2.gkg`AS GKG
   ON em.MentionIdentifier = GKG.DocumentIdentifier
```

Listing 2: Tallying The Total Number Of Articles In The South Pacific

```sql
SELECT
    count(1)
FROM
    `gdelt-bq.gdeltv2.eventmentions` AS em JOIN `gdelt-bq.gdeltv2.event` AS e
   ON em.GLOBALEVENTID = e.GLOBALEVENTID JOIN `gdelt-bq.gdeltv2.gkg`AS GKG
```

[i] We did not include every location in the South Pacific (Hawaii, New Zealand and Australia) because those three locations accounted for the majority of all the entries in our South Pacific subset of data from GDELT.

```
  ON em.MentionIdentifier = GKG.DocumentIdentifier
WHERE
  (ActionGeo_ADM1Code like `FM%` -- Micronesia
  OR ActionGeo_ADM1Code like `FJ%` -- Fiji
  OR ActionGeo_ADM1Code like `KR%` -- Kiribati
  OR ActionGeo_ADM1Code like `RM%` -- Marshall Islands
  OR ActionGeo_ADM1Code like `NR%` -- Nauru
  OR ActionGeo_ADM1Code like `PS%` -- Palau
  OR ActionGeo_ADM1Code like `PP%` -- Papua New Guinea
  OR ActionGeo_ADM1Code like `WS%` -- Samoa
  OR ActionGeo_ADM1Code like `BP%` -- Solomon Islands
  OR ActionGeo_ADM1Code like `TN%` -- Tonga
  OR ActionGeo_ADM1Code like `TV%` -- Tuvalu
  OR ActionGeo_ADM1Code like `NH%` -- Vanuatu
  OR ActionGeo_ADM1Code like `CW%` -- Cook Islands
  OR ActionGeo_ADM1Code like `NE%` -- Niue
  OR ActionGeo_ADM1Code like `AQ%` -- American Samoa
  OR ActionGeo_FullName = `Ashmore Reef, Queensland, Australia`
  OR ActionGeo_ADM1Code like 'FQ%' -- Baker Island
  OR ActionGeo_FullName = `Coral Sea, Oceans (general), Oceans`
  OR ActionGeo_FullName like `Easter Island, V%`
  OR ActionGeo_FullName = `Galapagos, Imbabura, Ecuador`
  OR ActionGeo_ADM1Code like `FP%` -- French Polynesia
  OR ActionGeo_ADM1Code like `GQ%` -- Guam
  OR ActionGeo_ADM1Code like `HQ%` -- Howland Island
  OR ActionGeo_ADM1Code like `DQ%` -- Jarvis Island
  OR ActionGeo_ADM1Code like `JQ%` -- Johnston Atoll
  OR ActionGeo_ADM1Code like `KQ%` -- Kingman Reef
  OR ActionGeo_FullName = `Midway Island, Western Australia, Australia`
  OR ActionGeo_ADM1Code like `NC%` -- New Caledonia
  OR ActionGeo_ADM1Code like `NF%` -- Norfold Island
  OR ActionGeo_ADM1Code like `CQ%` -- Norther Mariana Islands
  OR ActionGeo_FullName = `Ogasawaramura, Tokyo, Japan`
  OR ActionGeo_ADM1Code like `LQ%` -- Palmyra Atoll
  OR ActionGeo_ADM1Code = `ID36` -- Papua, Indonesia
  OR ActionGeo_ADM1Code like `PC%` -- Pitcairn Islands
  OR ActionGeo_ADM1Code like `TL%` -- Tokelau
  OR ActionGeo_ADM1Code like `WQ%` -- Wake Island
  OR ActionGeo_ADM1Code like `WF%` -- Wallis and Futuna
  OR ActionGeo_ADM1Code = `ID39` -- West Papua, Indonesia
  OR ActionGeo_FullName = `Bonin Islands, Tokyo, Japan`)
```

With both SQL statements tallying the number of articles captured in all of GDELT and only in the South Pacific, we found that only 0.08% of articles contained in GDELT were located in the South Pacific [*Figure 1*].

Next, I used the Python library (*Pandas*), to further filter and tally the data set in order for us to gain a perspective on how many of those articles in the South Pacific pertain to an environmental theme. But before using Python for data aggregation, I ran another query that returned all the important attributes from each of the three tables (Event, Eventmentions, GKG). The columns of interest are: *Actor1CountryCode, Actor2CountryCode, AvgTone, MentionDocTone, SourceCommonName, V2Themes, and V2Tone*. Running this query uses the same join statements and predicates as the second SQL example, the only difference is replacing the *count(1)* with our columns of interest. [*Listing2*]

With the a data set of every article in the South Pacific, I was able to use Pandas and another Python Library (*Numpy*), to further perform more data wrangling.[ii]

Of all the articles in the South Pacific, only 14.43% of them contain an environmental theme. When further filtering the data set by only containing the great powers as an actor, we found that of the 624,075 articles that contained an environmental theme, 27.02% (168,594 articles) involved a great power [*Figure 1*].

A further breakdown of the six great powers in the South Pacific, shows that the United States (*USA*) is the top actor with being in over 47% (48,561) of the articles where a great power is present and the article involves an environmental theme. Next is Australia (*AUS*) with over 17% (17,919), China (*CHN*) with over 14% (15,057), New Zealand (*NZL*) with over 11% (11,228), Japan (*JPN*) with over 7% (7,919), and Russia *(RUS)* with over 1% (1,253). [*Figure 2*]

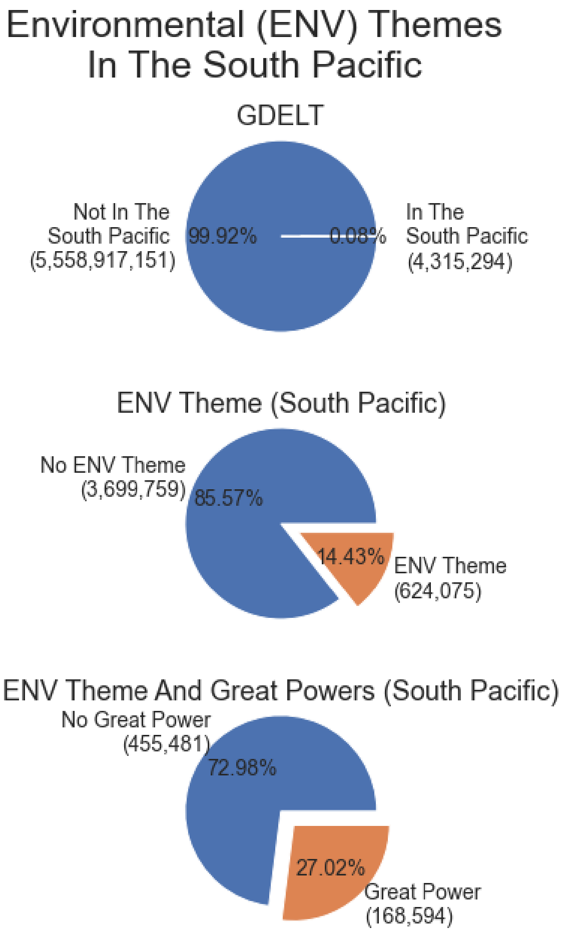

FIGURE 1: Breakdown of GDELT

Methods

Identifying Locations in the South Pacific

To create a subset of data that only contains events in the South Pacific, we had to identify all the location within the South Pacific and there FIPS10-4 country codes. GDELT uses the Federal Information Processing Standards (FIPS) codes to identify each events' unique location. The first place we looked to find all the countries and territories in the South Pacific was Wikipedia because Wikipedia has a list of names of most of the sovereign states and dependent territories in

[ii] Note, there is an increase in 8,540 articles in the middle pie chart in figure 1 because there was a time difference of a few weeks between running queries.

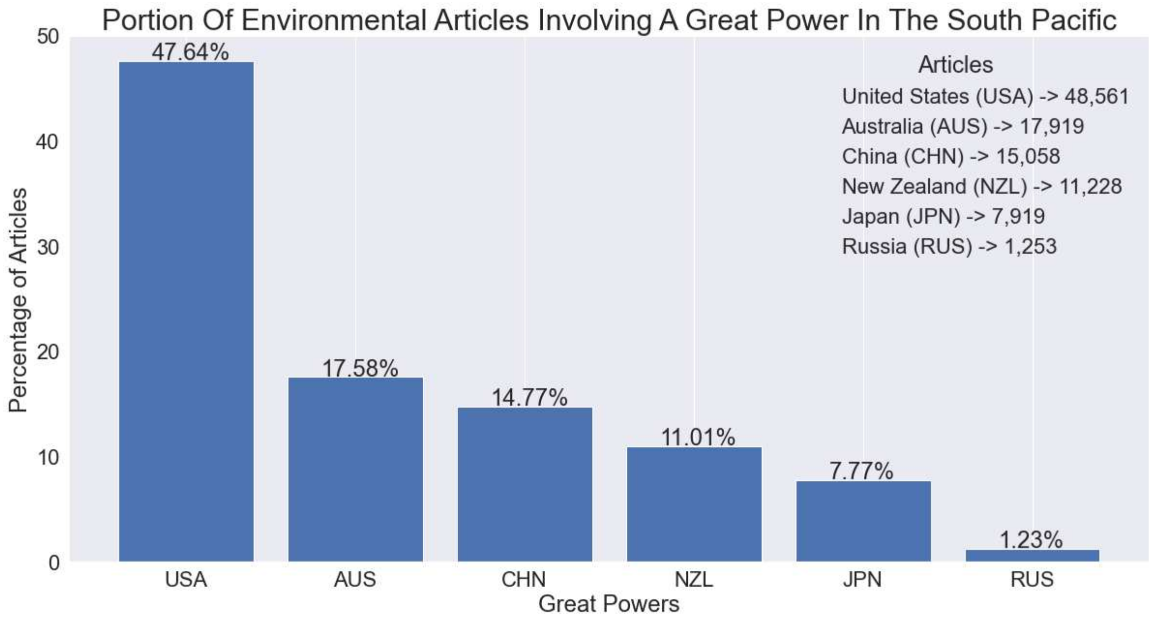

FIGURE 2: Breakdown of Articles by Great Power

the South Pacific [5]. We also decided to excluded Australia and New Zealand from our initial list of nations to include in our data set because both of those nations are already well developed and are major influences on the many of the nations in the South Pacific. After we had a list of countries and territories we wanted to include in our research, we then needed to find each locations corresponding FIPS code. To find each FIPS code, I ran the following query through BigQuery:

LISTING 3: Getting The List Of Countries And Territories In GDELT

```
SELECT
    Actor1Name, Actor1CountryCode, Actor2Name, Actor2CountyCode,
    Actor1Geo_FullName Actor1Geo_CountryCode, Actor2Geo_FullName,
    Actor2Geo_CountryCode, ActionGeo_FullName, ActionGeo_ADM1Code
FROM
    `gdelt-bq.gdeltv2.events`
```

And saved the results into a CSV file. Next, I manually searched the file in Excel for each corresponding locations' two-digit country code (*ActionGeo_CountryCode*). I found this method to be the fastest and most precise way of finding each location in the South Pacific compared to other methods like querying for each location in BigQuery or searching other online resources. We could not trust the FIPS codes found on the internet because there were inconsistencies between what GDELT labeled certain territories compared to what was found online.

Using this method, I was successful in identifying almost every region except for the locations of Territory of Ashmore and Cartier Islands, Galápagos Islands, Midway Islands, Ogasawara village, Papua Province, and West Papua. Searching for these locations returned some results that did not seem to match the location within the south pacific or did not return a result at all. I found the names and their country codes within GDELT after querying GDELT itself. I did however initially miss-identify West Papua's *ActionGeo_ADM1Code* for a different part of Indonesia. I eventually found the correct *ActionGeo_ADM1Code* for West Papua after a series of queries made against the GDELT data set on BigQuery and using Google maps to confirm the regions.

ActionGeo_ADM1Code instead of ActionGeo_CountryCode

I decided to use *ActionGeo_ADM1Code* as the main predicate for filtering by location in the queries ran through BigQuery. There are two reasons that made me decide to use *ActionGeo_ADM1Code* instead of *ActionGeo_CountryCode* or *ActionGeo_FullName* in my query. The first is that I found that when trying to filter a location by explicitly stating the name of the region (*ActionGeo_FullName*) produced a data set with less articles compared to using the *ActionGeo_ADM1Code*. For example, I found that when using the filtering by *ActionGeo_FullName = 'Hawaii, United States'* vs using *ActionGeo_ADM1Code = 'USHI'*, (At the time of the initial query) there was **346,532** more articles that were found using *ActionGeo_ADM1Code = 'USHI'*. The second reason was for consistency in which predicates I was using within the query itself. Because every *ActionGeo_ADM1Code* is anywhere between 2 and 4 characters and the first two characters of the *ActionGeo_ADM1Code* are always the same as the *ActionGeo_CountryCode*. I still capture the same number of articles using *ActionGeo_ADM1Code* as the predicate as we would using *ActionGeo_CountryCode*.

However, some locations within the South Pacific are part of countries that are not entirely considered to be in the South Pacific. For example, Indonesia has two regions that are a part of the South Pacific, but Indonesia itself is not entirely in the South Pacific. To only get those specific regions of Indonesia, I had to explicitly state the whole *ActionGeo_ADM1Code* for West Papua and Papua. Also, the locations of Ashmore and Cartier Islands falls underneath an *ActionGeo_ADM1Code* that captures more than just that region. For this reason, I had to explicitly state the regions, using the *ActionGeo_FullName*, of Ashmore Reef, Queensland, Australia, Coral Sea Islands, Easter Island, the Galapagos Islands, Midway Islands, Ogasawaramura Islands, and the Bonin Islands.

Joining the Events, Eventmentions, and GKG

To access all the attributes about each article (Themes and Tone), I had to join the Events, Eventmentions, and GKG table together. The Events and Eventmentions table joined on the *GLOBALEVENTID*'s and the GKG joined to the Eventmentions table through the Eventmentions *MentionIdentifier* is the same as the GKG *DocumentIdentifier*.

When joining all three tables together and using ActionGeo_ADM1Code as the predicates in the query, the resulting data set produced a table with 4,323,833 entries. When initially filtering which columns to also include in the data set, I decided to pull every column that could be of some relevance for this research. This meant excluding many of the attributes that are in the Events, Eventmentions, and GKG table. Even with excluding most attributes, this still led to a data set that was almost 20 GB (gigabytes) in size. With such a massive data set produced from BigQuery, it presented its own challenges on how to perform certain aggregations and data wrangling.

Because the Pandas library, from Python, embeds the data frame in RAM (Random Access Memory), this meant I could not access more than one instance of this data set at a time. This led to issues when I used the SQLite library using Python to perform queries on the data set. Because SQLite also embeds its data base in memory (instead of the disk drive like a traditional SQL frameworks), I had to load the data set directly into SQLite from Pandas and delete the instance of that data set immediately after it was done to perform queries. To conserve memory within my computer, I eventually made mini data sets of only the attributes used in this research so I could hold both an instance of the data set in Pandas and in SQLite. Doing this allowed me to do data aggregation with SQL and then further analysis using Python without having to drop one instance to use the other.

However, I was eventually able to only use Python for all my data analysis, including the data wrangling, just from using the Pandas library. This allowed me to pull only the columns I needed from the larger 20GB data set. Which saved how much RAM was being occupied at any given time and improved performance.

Pandas to Filter Themes

From the GKG table, the two attributes of interest for this research are the *Themes* and the *Tones*. The *Themes* attribute allowed us to filter our data set of the South Pacific to only contain events that involves the environment. The environmental themes of interest were filtered from the original data set by using Pandas and boolean masks combined with Pythons *str.contains()* method. This allowed us to create a data set that only contained articles where there is an environmental theme associated with it.

There are 21 different environmental themes that GDELT has created: *ENV_CLIMATECHANGE*, *ENV_OIL*, *ENV_FISHERY*, *ENV_MINING*, *ENV_COAL*, *ENV_GREEN*, *ENV_SOLAR*, *ENV_METALS*, *ENV_POACHING*, *ENV_NATURALGAS*, *ENV_DEFORESTATION*, *ENV_OVERFISH*, *ENV_FORESTRY*, *ENV_NUCLEARPOWER*, *ENV_WATERWAYS*, *ENV_SPECIESENDANGERED*, *ENV_HYDRO*, *ENV_BIOFUEL*, *ENV_GEOTHERMAL*, *ENV_WINDPOWER*, *ENV_CARBONCAPTURE*, and *ENV_SPECIESEXTINCT*.

GDELT also has the World Bank themes included in the *Themes* column. The World Bank themes offer even greater details in their themes than what GDELT has provided with its own themes. But for this research, we only are filtering by the themes that begin with *'ENV_'* because the World Bank themes do not follow a uniform structure. Therefore, making it exceedingly difficult to only filter by a certain kind of World Bank theme.

V2Themes vs Themes

Within the GKG table, there are two themes columns. One is labeled *Themes* and the other *V2Themes*. After doing a detailed analysis of both columns, I concluded that both columns produced nearly identical results and choosing one over the other was not important for this research. The major differences between the *Themes* and *V2Themes* is the *V2Themes* states the character offset within the article where the theme was identified. *V2Themes* also allows for the same theme to appear more than once in the column if it appears more than once in the event. The *Themes* is only a list of each unique theme as it appears in the event. For this research, I decided to use *V2Themes* instead of *Themes* when filtering by articles involving the environment. However, we could have used *Themes* instead and gotten the same results for this research.

Tones

Even though the Events and Eventmentions tables also has a tone attribute, I wanted to present all three of them in this report. There is a slight difference on how each of the tones are calculated in each table. The tone in the Events (*AvgTone*) table only represents the average tone of all the article that reported the event within the first 15 minutes it was first seen, the Eventmentions tone (*MentionDocTone*) is the tone of each individual article, and the GKG tone (*V2Tone*) is calculated from taking the average of the positive and negative score of the whole document [3][4]. The tones, and positive and negative scores are automatically generated by GDELT using sentiment analysis. The exact algorithms are not mentioned but GDELT promises that it is an advance Natural Language Processing (NLP) algorithm that generates these scores [4][6].

Great Powers

There is two ways that we can specify when looking for articles only involving the great powers of interest. The first way is by looking for the *Actor1/2CountryCodes* and the other is the *Actor1/2Geo_CountryCode*. While the *ActorGeo_CountryCode* method uses the FIPS country codes to specify who the actors are, the attribute *Actor1CountryCode* (and *Actor2CountryCode*) uses a 3-character CAMEO code for the country affiliations. There are resources online to find the *Actor1/2CountryCode* of each of the countries we are looking for. Instead of turning to the internet, I searched the data sets I had already created using Pandas functions. This way, I was able to get the exact *Actor1/2CountryCode*'s for each of the great powers.

I decided to use the *Actor1/2CountryCode* instead of *ActorGeo_CountryCode* for filtering the data set to only containing the great powers because when analyzing the results of using both, using *Actor1/2Geo_CountryCode* for specifying an actor produced a data set with less than half of the articles compared to using the *Actor1/2CountryCode*. The Events table in GDELT only specifies two actors and a location. So, if multiple actors are present, GDELT will either leave the field blank or will choose which actors take precedence in the article and assign those actors to those attributes [4]. But the *Actor1/2CountryCodes* do not always match. For instance, "if the text refers to 'French Assistant Minister Smith was in Moscow', the CountryCode field will list France in the CountryCode field, while the geographic fields may list Moscow as the location"[4]. While this leads to potentially missing some articles involving certain actors, if we were to look for actors by either there location (*ActorGeo_CountryCode*) or there precedence in the article (*ActorCountryCode*), we may be including articles where we have an actors location who maybe was not an actor in the article itself. On the other hand, when specifying having the *Actor1/2CountryCode* and ActorGeo_CountryCode matching, only produced a data set that only contained a fraction of the number of events (ten's of thousands compared to hundred's of thousands).

Media Sources

When finding the Local media sources of the South Pacific, I created a function in Python to output every unique media source that has involves a great power as *Actor1* or *Actor2*. From there, my research mentor (*Dr. Elizabeth Gooch*) went through the list of sources for each individual location and cross checked each source to see if it was a local source. The Western news sources were selected because they give a range of liberal, conservative, and neutral media outlets. The Chinese news sources were select because each one of those news sources are either state-owned, state-ran, or authorized by the Chinese government.

Calculating Tones

Once I had the geographical locations, the CAMEO codes for each great power, and the proper filtering methods for creating subsets of data, I then created multiple Python functions to compute the average weighted tone for each news source in the South Pacific [*Figure 3*].

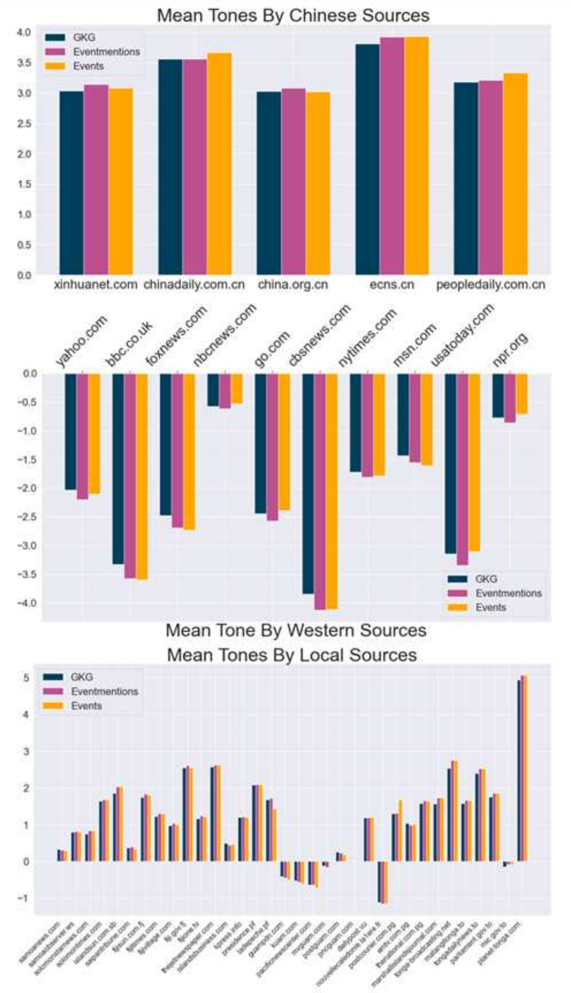

FIGURE 3: Tones by Sources

Results

When comparing each of the three media sources (Chinese, Local, and Western) side by side, Chinese sources have an average tone of around 3.2, Local sources have an average tone of around 0.7, and Western sources have an average tone of 1.8-1.9. The tones in GDELT can have a score of +100 to -100 but most of the tone scores are between +10 and -10 (where -10 is an article written with a very negative sentiment, +10 is written with a very positive sentiment, and 0 indicating neutral) [4][*Figure 4*].

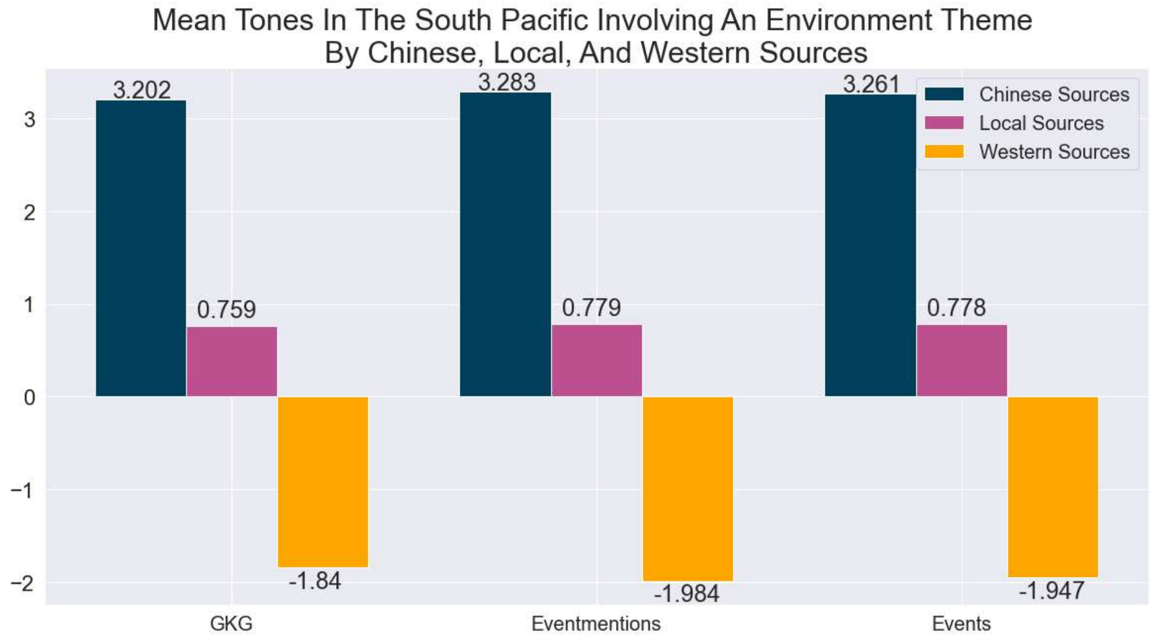

FIGURE 4: Mean Tones By Source

Difference In Means

In order to prove each of the population means are statistically significance, I conducted a difference in means tests for each of our three sources. For the rest of the report, I conducted my statistical analysis using R. I did this for a few reasons. One, R has better interpretability with the results. In Python, similar functions used to conducted statistical analysis using the SciPy library usually produce only single number results. To show the results in a presentable manor, I would need to create additional functions with print statements explaining the results of each test. Instead of doing this, I decided R would be the better choice. Two, the data visualizations produced in R using *ggplot* and *base R* produce very clean graphics with minimal amount of coding compared to Python. I still however used Python for all the data aggregation and pre-processing of the data set and then imported it into R for the data visualizations and statistical analysis.

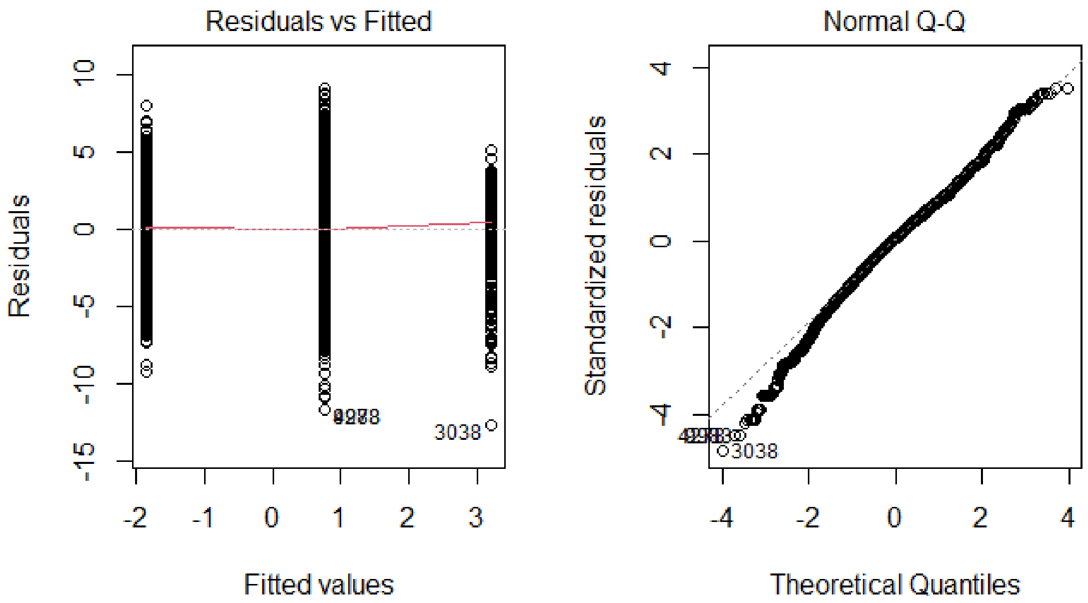

FIGURE 5: Normality and Variance of Tones

In order to prepare the data for statistical analysis using R, I filtered the data set one more time to only contain the *GKG_Tone* and added a new column to indicated if the *Source* was a Chinese source, Local source, or Western Source.

To conduct a difference in means test, we must acknowledge the assumptions to conduct this statistical analysis. First, we will assume that each observation from each of the groups are independent of each other group. Next, I checked the normality by creating three distribution plots and a Q-Q plot [*Figure 5*, 6]. This showed that the data overall follows a normal distribution.[iii] Finally, when checking for equal variances in the Residuals vs Fitted plot, the variances appeared to not be equal among the three groups [*Figure 5*]. Because of this inconsistency, I conducted a Levene Test in R to test the homogeneity of variances of the residuals.

LISTING 4: Levene's Test for Homogeneity of Variance

```
> tones <- read_csv('source_tones.csv')
> leveneTest(GKG_Tone ~ Source, data=tones)
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     2  7.5008 0.0005548 ***
      14650
---
Signif. codes: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1 ` ` 1
```

[iii] Because the number of observations is large (Chinese = 1,013, Local = 10,292, Western = 3,348), violating the normality assumption is okay.

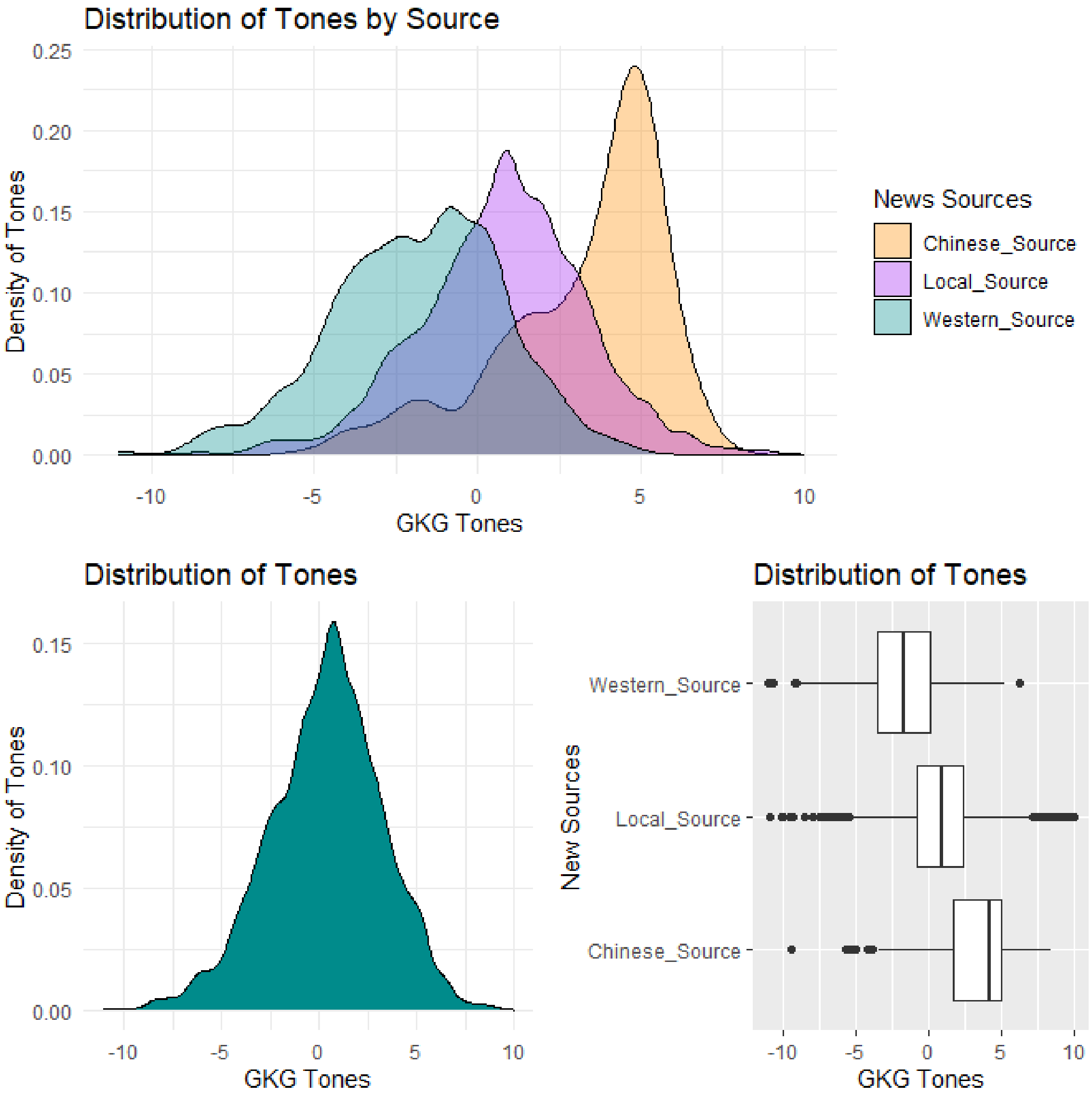

FIGURE 6: Normality of Tones

To interpret the results of the Levene's Test for Homogeneity of Variance, at a confidence level of $\alpha = 0.05$ and p-value of 0.0005, we conclude that the population variances are not equal.

Because the population variances are not equal, we will need to conducted three Welch's t-test's for each of the pairwise groups .[iv]

LISTING 5: Pairwise Comparisons for Tones

```
> data.local <- subset(tones, Source == "Local_Source")
> data.western <- subset(tones, Source == "Western_Source")
> data.chinese <- subset(tones, Source == "Chinese_Source")
```

[iv] Group 1: Local Sources - Chinese Sources, Group 2: Western Sources - Chinese Sources, Group 3: Western Sources - Local Sources

```
> ### Local_Source-Chinese_Source
> t.test(data.local$GKG_Tone, data.chinese$GKG_Tone,
+         mu=0, alternative="two.sided",
+         var.equal = FALSE)

        Welch Two Sample t-test

data:  data.local$GKG_Tone and data.chinese$GKG_Tone
t = -28.536, df = 1218.7, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.610919 -2.275004
sample estimates:
mean of x mean of y
0.7589979 3.2019597

> ### Western_Source-Chinese_Source
> t.test(data.western$GKG_Tone, data.chinese$GKG_Tone,
+         mu=0, alternative="two.sided",
+         var.equal = FALSE)

        Welch Two Sample t-test

data:  data.western$GKG_Tone and data.chinese$GKG_Tone
t = -53.947, df = 1684.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.225113 -4.858498
sample estimates:
mean of x mean of y
-1.839846  3.201960

> ### Western_Source-Local_Source
> t.test(data.western$GKG_Tone, data.local$GKG_Tone,
+         mu=0, alternative="two.sided",
+         var.equal = FALSE)

        Welch Two Sample t-test

data:  data.western$GKG_Tone and data.local$GKG_Tone
t = -49.899, df = 5624.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.700946 -2.496742
sample estimates:
 mean of x  mean of y
-1.8398457  0.7589979
```

Interpreting Welch two-samples t-test's

After conducting these tests in R, we show that the difference in means between each pairwise group is statistically significant.

Local Sources and Chinese Sources

The Welch two-samples t-test between Local Sources and Chinese Sources showed that the difference is statistically significant, t = -28.536, df = 1218.7, p-value < 2.2e-16.

Western Sources and Chinese Sources

The Welch two-samples t-test between Western Sources and Chinese Sources showed that the difference is statistically significant, t = -53.947, df = 1684.2, p-value < 2.2e-16.

Western Sources and Local Sources

The Welch two-samples t-test between Western Sources and Chinese Sources showed that the difference is statistically significant, t = -49.899, df = 5624.3, p-value < 2.2e-16.

Conclusion

In this report, I introduce GDELT and how GDELT can be used to track events in regions of the world. I talked about the three tables (Events, Eventmentions, GKG) we used in this research, how they all relate, and which attributes were important for this research. I showed how I found the exact locations and actors within GDELT using GDELT itself and Wikipedia. I mentioned which countries and territories we included in this research and some that we did not. I showed the queries used to pull our data set from BigQuery. I showed the number of articles within GDELT that have an event located in the South Pacific only accounts for 0.08% of the articles within GDELT itself, of the 4.32 million articles in our data set, only 14.43% involve an environmental theme, and when a great power was an actor in an event in the South Pacific, 27.02% of those events had an environmental theme.

Next, I showed all the methods I used to develop our final data set. I gave explanations on which attributes I decided to choose for our filtering and how I used Python for all of my data wrangling, exploration, and most of the visualizations. I explained why we chose our media sources and how I calculated our average tones for each source. I then showed that Chinese sources have an overall very positive tone in the South Pacific when the event itself involves the environment, Local news sources have a positive to neutral tone, and Western sources have a negative tone. And finally, I proved (using R) the mean tones of Chinese, Local, and Western new sources are statistically different.

Acknowledgment

I would like to thank Dr. Elizabeth Gooch from the Naval Postgraduate School for mentoring me and the Undergraduate Research Opportunity Center (UROC) Program at California State University, Monterey Bay for providing me a scholarship for this research.

References

[1] https://www.gdeltproject.org/about.html

[2] https://www.gdeltproject.org/

[3] http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf

[4] http://data.gdeltproject.org/documentation/GDELT-Event_Codebook-V2.0.pdf

[5] https://en.wikipedia.org/wiki/Oceania

[6] https://blog.gdeltproject.org/introducing-gkg-2-0-the-next-generation-of-the-gdelt-global-knowledge-graph/