Scraping Unstructured Data to Explore the Relationship between Rainfall Anomalies and Vector-Borne Disease Outbreaks

Ethan Joseph
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY
josepe2@rpi.edu

Thilanka Munasinghe
Department of Information Technology and Web Science
Rensselaer Polytechnic Institute
Troy, NY
munast@rpi.edu

Heidi Tubbs
University Space Research Association
Goddard Earth Science and Technology Research
& Biospheric Sciences Laboratory, Code 618
NASA/Goddard Space Flight Center
Greenbelt, MD
heidi.c.tubbs@nasa.gov

Bhaskar Bishnoi
University Space Research Association
Goddard Earth Science and Technology Research
& Biospheric Sciences Laboratory, Code 618
NASA/Goddard Space Flight Center
Greenbelt, MD
bhaskar.bishnoi@nasa.gov

Assaf Anyamba
University Space Research Association
Goddard Earth Science and Technology Research
& Biospheric Sciences Laboratory, Code 618
NASA/Goddard Space Flight Center
Greenbelt, MD
assaf.anyamba@nasa.gov

***Abstract*—According to the World Health Organization (WHO), vector-borne diseases such as malaria and dengue account for 17% of all infectious disease cases and lead to more than 700,000 deaths per year. Tracking and predicting the spread of vector-borne diseases is a vital task that could save hundreds of thousands of lives annually. Oftentimes, the first reports of vector-borne disease outbreaks occur through emails and online reporting systems long before they are officially documented. Tracking and predicting the emergence and spread of vector-borne disease outbreaks requires extracting data from these unstructured sources in combination with historical weather and climate data to understand the underlying background triggers and disease dynamics. In this work, we develop a data extraction pipeline for the online outbreak reporting website ProMED-mail that utilizes a web scraper, transformer neural network summarizer, and named entity recognizer to obtain a dataset of malaria, dengue, zika, and chikungunya outbreaks over the last 30 years. This scraped dataset was further analyzed in association with global rainfall anomalies derived from NASA's Integrated Multi-satellitE Retrievals for GPM [Global Precipitation Mission] (IMERG) dataset. This preliminary analysis was to understand the effect of global rainfall patterns on the spread of vector-borne diseases. Analysis of the ProMED-mail and GPM data shows that vector-borne disease outbreaks are clustered towards the tropics and outbreaks are often amplified during the rainy seasons. Our scraped dataset can be a valuable tool in creating comprehensive georeferenced disease records for modeling and predicting future outbreaks.**

***Index Terms*—Web scraping, data mining, epidemiology, vector-borne disease, ProMED, NLP, transformers**

I. Introduction

Recent epidemics of infectious diseases–such as zika, bird flu, and most prominently Covid-19–have shown the need for public health information systems that can detect the onset of an epidemic, assess regional risk, and update decision makers in real-time as the situation evolves. Historically, outbreak surveillance and forecasting have relied on reports from sentinel healthcare providers and laboratory results [1]. Unfortunately, that data is usually not made publicly available until weeks after the data was first collected. This poses a problem, as physicians, researchers, and computational epidemiologists are sometimes unable to react to an outbreak until after it has passed the critical tipping point of becoming an epidemic. Oftentimes the first public notices of infectious disease outbreaks occur through disparate Internet-based passive surveillance sources long before they are officially documented. However, this online data tends to be either unstructured or semi-structured, which makes collecting and interpreting the data a difficult task. In this work, we focus on extracting valuable information from this unstructured data and creating data records of selected vector-borne disease

outbreaks from internet sources, though our methods could be extended for other infectious diseases.

Vector-borne diseases are infectious diseases transmitted through vector organisms. These organisms include mosquitoes, ticks, etc. which can transfer infectious pathogens from animals to humans. Vector-borne diseases account for 17% of all infectious disease cases and cause over 700,000 deaths each year [2], mainly in tropical and sub-tropical regions. Examples of prominent vector-borne diseases include malaria, dengue, zika, and chikungunya all of which are spread through a mosquito vector.

There are specific conditions needed for the survival of the mosquitoes. For the *Aedes* mosquito species–which transmits Zika, Chikungunya, and Dengue–to breed successfully, it requires stagnant water (breeding site), warm temperatures, low altitudes, and high humidity. As a result, research has shown that heavy rainfall as an indicator of an outbreak can be used to prepare for epidemics as early as one month [3]. Other findings indicate that drought and higher than normal temperatures can be a trigger of outbreaks by concentrating breeding sites in densely populated urban environments [4, 5].

Tracking and predicting the emergence and spread of vector-borne disease outbreaks requires extracting data from these unstructured sources and combining it with historical weather and climate data. Doing so is critical to reduce the impact on healthcare systems and save lives.

To aid in this task, we develop a data extraction pipeline for the online outbreak reporting service ProMED-mail (ProMED) that utilizes a web scraper, abstractive summarizer, and named entity recognizer to obtain a dataset of malaria, dengue, zika, and chikungunya outbreaks over the last 30 years. We further analyze this data with global precipitation data from NASA's Global Precipitation Mission (GPM) by examining the derived rainfall anomalies to determine their effect on geographic patterns and seasonality of these vector-borne disease outbreaks. We analyze the results of this work, provide source code, and discuss further work.

II. Related Work

To address the need for early detection of epidemics, researchers have studied the use of disparate Internet-based sources of epidemiological information, such as Google search queries [6, 7], social media websites like Twitter [8], and online news reports [9].

Ginsberg et al. [6] developed a system for generating comprehensive influenza surveillance models using billions of searches from Google web search logs, while Milinovich et al. [7] showed a significant correlation between official vector-borne disease notifications in Australia and specific search terms. Jain and Kumar [8] developed a social media-based mosquito-borne disease surveillance and outbreak management system using spatial and temporal information extracted from Twitter and RSS feeds. Zhang et al. [9] examined the potential of news articles in predicting and modeling outbreaks of dengue in India and zika in Brazil.

Other works have also looked into using the online outbreak reporting service and mailing-list ProMED-mail as a source for vector-borne disease data. CHIKRisk [10] is an online chikungunya virus monitoring web application that utilizes manually extracted ProMED data to visualize risk for chikungunya outbreaks on a map. Anyamba et al. [11] manually extracted ProMED data to explore the relationship between land temperature, vegetation, and vector-borne disease outbreaks.

Web applications have been developed that combine multiple internet sources of data. These systems are commonly referred to as event-based surveillance and have emerged as a complementary method to the traditional indicator-based surveillance that relies on routine reporting by healthcare facilities. Prominent examples of event-based systems include Project Argus [12], EpiSPIDER [13], HealthMap [14], and BioCaster [15]. These applications often aggregate data from multiple unstructured text sources. For example, Project Argus employed web scraping tools, native speaker analysts, and a formal social disruption taxonomy to identify indicators of bird flu's global spread from open media sources on the Internet. HealthMap integrates outbreak data from online news sites (e.g. Google News), RSS feeds, curated mailing lists and reporting systems (e.g. ProMED), multinational surveillance reports (e.g. Eurosurveillance), and validated official alerts (e.g. from WHO). EpiSPIDER specifically extracts location and topic data from ProMED to display on a map. Biocaster continuously analyzes documents reported from over 1700 RSS feeds and plots them onto a Google map. Project Argus, EpiSPIDER, and Biocaster are non-operational as of the time of writing. Together, these applications show how valuable utilizing disparate online sources can be in obtaining epidemiological data. However, these tools are usually only designed to detect outbreaks as they emerge. The scraped data is not made public, and the code used to extract that data is closed source. While these tools are valuable in detecting outbreaks, it can be difficult to collect data from them for use in predicting and modeling future outbreaks.

Related work on NLP data extraction for epidemiology includes EventEpi [16] and work by DILLER [1]. EventEpi developed an open-source document annotator EpiTator that was trained on WHO Disease Outbreak News and ProMED to annotate an article's disease, location, date, and confirmed-case counts. We build off the named entity recognizer used in EpiTator in this work and supplement it with further training and more annotation fields. DILLER [1] also developed an open-source named entity recognizer, trained to annotate case counts and locations in avian influenza reports extracted from the Avian Influenza News Archive. He further supplements it with a rule-based entity resolver that links data between reports to determine if they are part of the same epidemic.

Rainfall and wetness have been shown to be a useful indicator of a mosquito-borne disease outbreak and can be used to prepare for epidemics as early as one month [3]. Glancey et al. [17] showed strong support that the proportion of days with precipitation and the number of water sources were positively associated with increased severity of Rift

```
        MALARIA - JAMAICA (KINGSTON) (07)
********************
A ProMED-mail post
ProMED-mail is a program of the
International Society for Infectious Diseases

Date: Tue 1 May 2007
Source: Jamaica Observer edited

New cases of malaria reported
---------------------
A total of 11 new cases of Malaria has been
reported since the beginning of April 2007,
the Ministry of Health said on 30 Apr 2007.
According to a release from the ministry, 2
cases were reported between 15-21 Apr 2007,
3 the previous week 15-21 Apr 2007 while a
total of 6 cases was reported between 1-7 Apr
2007. The ages of the affected persons range
from 10 to 59.
...
Meanwhile, the ministry reiterated that it was
in the process of seeking alternative
insecticides to prevent further outbreaks after
recent tests confirmed some resistance of
the _Anopheles albimanus_ mosquito taken from
the Duhaney River to malathion insecticide,
which it was hoping to use to eliminate the
parasites. The tests were conducted by
consultants from the United States-based
Centers for Disease Control and Prevention
(CDC).
...
```

Fig. 1: Example ProMED Alert on Malaria (truncated). Data desirable for extraction underlined.

Valley Fever, another mosquito-borne disease that primarily infects domestic animals. Campbell et al. [18] also support that increased wetness, precipitation, and land surface temperature had a positive effect on the abundance of the *Aedes mcintoshi* species of mosquito, a primary vector for Rift Valley Fever in Kenya.

Predicting disease outbreaks using geographical information systems (GIS) has also been explored in a number of papers, and is often the primary source of temperature and precipitation data. Bergquist [19] called for the need for a broadening of GIS data usage in epidemiological studies in 2001, but GIS usage is now fairly common among epidemiological research as evidenced by its widespread usage across the papers cited previously. The problem is that these systems are not flexible enough to process and analyze unstructured data sources and create models for early warning.

III. Data Sources

To explore the relationship between rainfall anomalies and vector-borne disease outbreaks, we combine data from ProMED-mail and rainfall data from NASA GPM Integrated Multi-satellitE Retrievals for GPM (IMERG).

A. ProMED-Mail

ProMED-mail was launched in 1994 as an email service to identify unusual health events related to emerging and re-emerging infectious diseases. Since then, it has grown into a full infectious disease alert system used daily by public health leaders, government officials at all levels, physicians, veterinarians, and other healthcare workers, researchers, private companies, journalists, and the general public. Reports are produced by a global team of experts in a variety of fields including virology, parasitology, epidemiology, and entomology [20]. ProMED's website offers a map displaying the locations where alerts originated from as well as a search page through archives of previous alerts.

Alerts on the ProMED-Mail website can take a variety of formats. For example, some alerts describe a single outbreak while others will describe multiple outbreaks, possibly over the span of multiple years. Alerts format data in different ways, such as in a raw ASCII table or incorporated into sentences. There are also alerts that don't describe outbreaks at all, and are simply case studies into a single infection and treatments. The majority of alerts take the format of a report, see Fig. 1 for an example of a malaria report. The differences in alert formatting and the variety of topics make automated data extraction very challenging.

B. NASA IMERG

To supplement outbreak data from ProMED-mail, we analyze historical rainfall data from NASA's NASA's Integrated MultisatellitE Retrievals for GPM [Global Precipitation Mission] (IMERG) dataset. IMERG is an algorithm that intercalibrates, merges, and interpolates satellite microwave precipitation estimates together with microwave-calibrated infrared satellite estimates, precipitation gauge analyses, and potentially other precipitation estimators at fine time and space scales for the Tropical Rainfall Measuring Mission (TRMM) [21] and GPM eras over the entire globe [22].

Thus, the IMERG dataset contains the best estimates of precipitation across the Earth's surface, and is particularly valuable over the majority of the Earth's surface that lacks rain-gauge measurements. There are multiple formats for the IMERG data, ranging from half-hourly collection rates to monthly. We utilize the monthly IMERG data in this work which covers a time-range of over 20 years spanning from June 2000 to February 2021. This data was accessed using the new NASA GES DISC API.

IV. Data Collection Methods

We develop an extraction pipeline that utilizes a web scraper, abstractive text summarizer, and named entity recognizer to extract data from ProMED-mail, then combine the data with NASA IMERG global precipitation data before analysis. See Fig. 2 for an overview of the process.

A. Web Scraping

The first tool developed was a web scraper for ProMED-mail alert post. The scraper allows for the input of a search term. The content and various metadata about each of the

ProMED-mail alert posts related to the search term are extracted and added to a Pandas DataFrame, then saved to a comma-separated-value (.csv) format.

The initial version of the web scraper extracted data directly from the HTML on the ProMED website using Selenium. Selenium was necessary since ProMED-mail doesn't expose a public API and posts are loaded via JavaScript, so static web scraping tools (such as BeautifulSoup, etc.) are not able to retrieve all the posts. The Selenium scraper works by simulating clicks on the website to load subsequent alert posts. As a result, performance was poor with the Selenium web scraper (up to 4 hours to scrape a query), so we began looking into alternative ways to download posts.

In the second version of the web scraper, we achieve better performance by sending POST requests directly to the ProMED-mail's PHP backend. After monitoring network requests on ProMED's search page after inputting a search query, it was straightforward to decode the necessary requests to the backend, which returns a JSON response containing the posts.

Performance of the backend querying scraper was improved by up to 500% compared to the original Selenium version. Additionally, the backend provides metadata about the alert article in the response that is not easily accessible in the HTML format. This metadata includes publication date, location of the alert, duplicate posts, among other useful fields. These tags were included as part of the textual description in the HTML article, which would have required the development of a separate tool to extract them, but this metadata is separated in the JSON response from the backend.

B. NLP to Extract ProMED-Mail Data

Once we extract an alert article's content, we begin extracting the data using natural language processing (NLP) techniques. Namely, we utilize an abstractive summarizer to condense a post's content into only the most relevant information, followed by a named entity recognizer to extract data from that summary into a DataFrame. Depending on the format of the data, we may instead use a regular expression to extract data from tables.

If the post contains a table (primarily dengue posts), we simply use a regex to parse each line of the table into a row in our output DataFrame. Otherwise, we use the abstractive summarization/NER combination to extract the data.

As seen in Fig. 1, the second paragraph of the report format post contains no information about an outbreak, but rather talks about the effectiveness of an insecticide. While these paragraphs might provide useful context to human researchers, they often confuse named entity recognition models which may extract information that does not pertain to an outbreak. To prevent this from happening, we condense the entirety of the alert post's text into its most important sentences using an abstractive text summarizer via the BART transformer neural network [23]. Abstractive text summarization is the task of generating a concise summary that captures the salient ideas of the source text. The summaries may contain new phrases

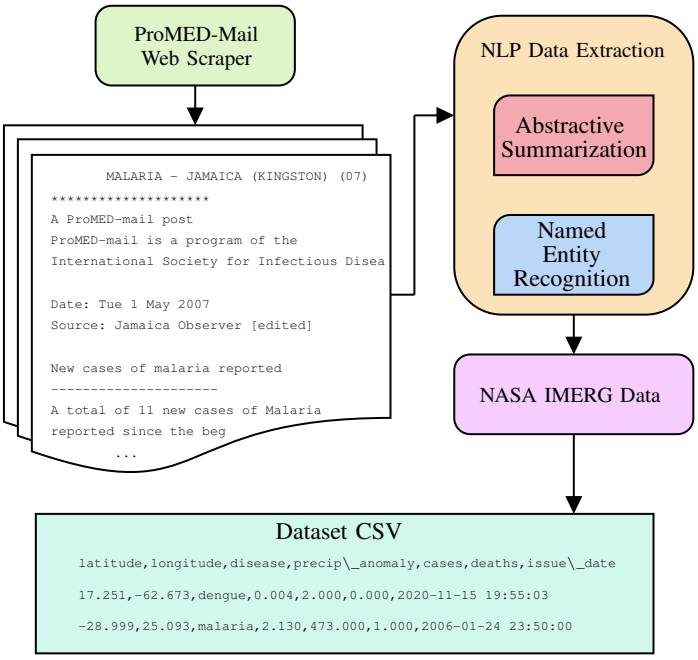

Fig. 2: Data Collection Overview

and sentences not contained in the source text, as long as the phrases capture the main ideas (in our case, number of cases, location, number of deaths, etc.).

For summarization, we utilize a BART transformer neural network [23] pretrained on the CNN/Daily Mail summarization dataset [24]. We choose this model due to its strong performance on summarization baselines, and we pretrain on the CNN/Daily Mail dataset due to slight overlap in subject matter (to our knowledge there aren't any public epidemiology summarization datasets, and the CNN/Daily Mail dataset contains some examples of disease outbreak news articles).

Once summarized, we utilize Named Entity Recognition (NER). NER is a computational method that segments and classifies words/phrases that represent entities in a text according to pre-selected categories. This method is popular for information extraction tasks in the biomedical and public health domains. For our NER model, we build off the work of Abbood et al. [16]. We utilize the NER vocabulary, entities, and initial dataset from the EpiTator project, but further improve extraction results by changing the underlying model to a transformer neural network instead of a conditional random field model (CRF). Transformers have consistently higher F1 scores on the NER task than CRF models [25], which is why we chose to use one. Additionally, we perform light finetuning using annotated examples of vector-borne disease posts from ProMED-mail.

Using this NER model, we extract the following data from the ProMED posts: *number of cases, number of deaths, location, location hierarchy level (i.e. country, state, city),* and the *starting date of the incident.* See the underlined phrases in Fig. 1 for examples of extracted entities.

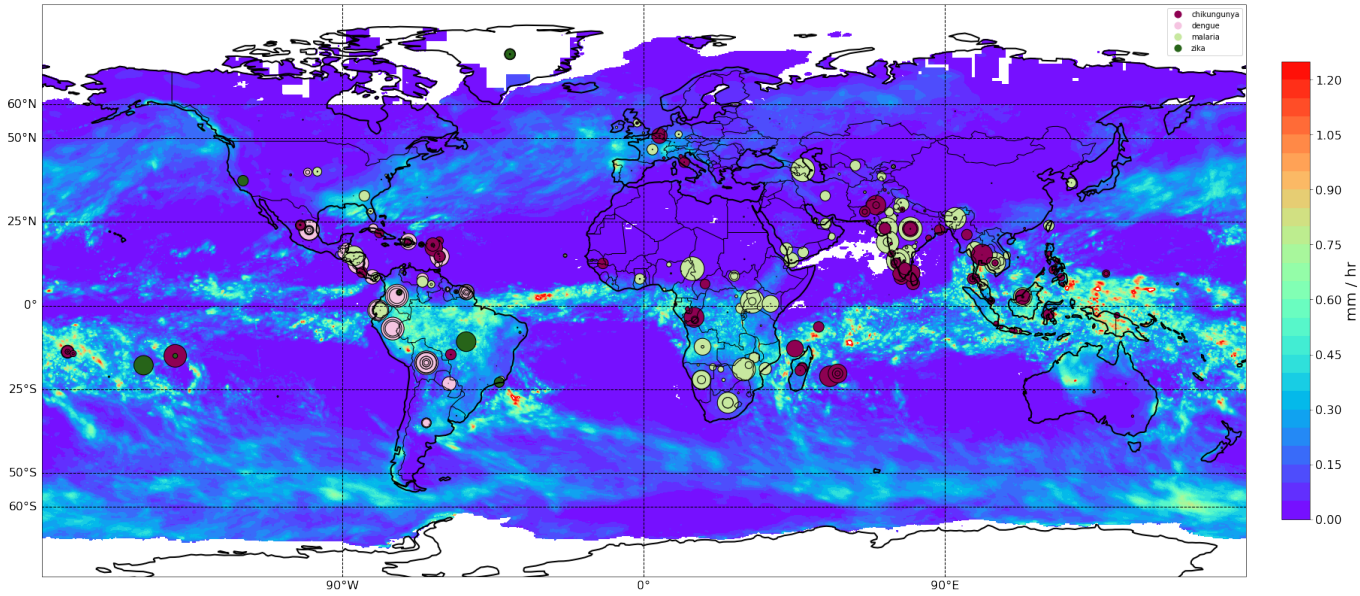

Fig. 3: GPM IMERG Mean Precipitation overlayed with disease outbreak locations (radius represents outbreak magnitude)

C. Rainfall Anomaly Detection

The NASA IMERG dataset provides global gridded precipitation data from 2000 to present, with an accuracy of 2 decimal places for latitude/longitude coordinates. We first round up our extracted latitude/longitude outbreak locations to two decimal places to align with the gridded dataset. The anomaly is then calculated as the percentage difference in precipitation between the month during which the outbreak occurred, and the average rainfall across 20 years for each location in the dataset [11]. If the date of an outbreak occurs outside of the range of available IMERG data, we use the precipitation value from the closest month.

V. Analysis

A. Extracted Dataset

We run our extraction pipeline on the following search queries: 'dengue', 'malaria', 'zika', and 'chikungunya'. This yields a total extracted dataset of 2103 rows with data about vector-borne disease outbreaks over 25 years from March of 1996 to July 2021. The dataset contains 43 columns, the most important of which are *disease name, latitude, longitude, start date, number of cases, number of deaths, precipitation anomaly, precipitation value, and precipitation mean*. Other columns contain extra information that is not directly useful for predictive modeling but may be of interest to researchers, such as post metadata from ProMED, or intermediary data that can be used to recalculate the dataset (such as the IMERG coordinates). See Table I for a sample of the dataset.

There are 788 rows of dengue data, 600 rows of malaria data, 531 rows of chikungunya data, and 184 rows of zika data. There are 523 unique locations in the dataset. The number of cases ranges from 0 to 1.25 million with a heavily skewed mean of 54754.22, while deaths range from 0 to 20000 with a mean of 50.16. Precipitation anomalies range from a difference of -1 to 4.514. -1 in this case signifies that that location received 100% less rain than average, while the latter location received 451% more rain than average. See Fig. 3 for a global plot of outbreaks over average rainfall.

B. Outbreak Distributions and Clustering

To explore the existence of patterns and clusters in the geographical distribution of outbreaks, we run 3 unsupervised clustering algorithms: K-Means, Birch [26], and DBSCAN [27], then compare their similarity using an adjusted rand index metric. For K-Means, we utilize the Elbow Method [28], i.e. plotting an elbow curve for inertia values across different values of k to identify a suitable range for k. In the curve, the elbow exists between k=3 and k=7. We chose k=6 as there are 6 continents excluding Antarctica, and we want to explore if the cases can be clustered roughly into the separate continents.

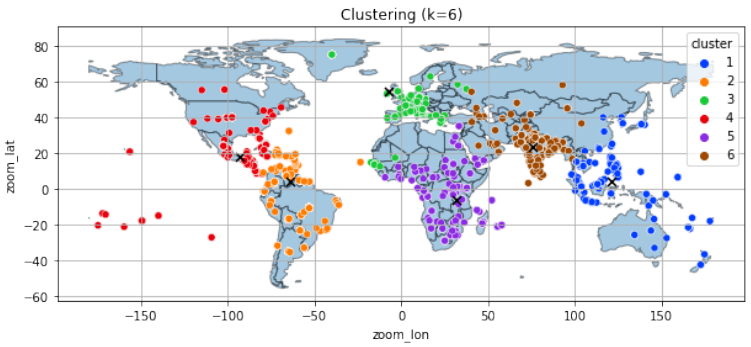

Fig. 4: K-Means clustering results with k=6. X represents cluster centroids.

latitude	longitude	zoom_level	disease	precip_mean	precip_value	precip_anomaly	cases	deaths	issue_date
17.251	-62.673	4	dengue	0.114	0.114	0.004	2.000	0.000	2020-11-15 19:55:03
-28.999	25.093	5	malaria	0.055	0.172	2.130	473.000	1.000	2006-01-24 23:50:00
4.842	-58.642	4	dengue	0.260	0.284	0.091	403.000	0.000	2020-11-15 19:55:03
-15.500	33.000	4	malaria	0.101	0.227	1.258	2015.000	213.000	2016-01-11 14:29:21

TABLE I: Sample of the extracted dataset.

The Birch algorithm does not take a number of clusters as an argument, and subsequently grouped the coordinates into 3 clusters. These clusters roughly correlate with continental landmasses (i.e. there is one cluster for the Americas, one for Afro-Eurasia, and one for Oceania). Lastly for DBSCAN, we utilize the ball tree algorithm and haversine formula to calculate great-circle distances between points as the metric for clustering [29]. We choose epsilon to be such that the maximum distance between points that can be considered part of the same cluster is 1000km [29]. DBSCAN grouped the coordinates into 42 clusters, with no obvious clustering pattern. Additionally, there were 53 "noisy" points that don't fit into a cluster.

To compare the similarity of clustering results, we calculated the rand index adjusted for change (ARI) [30, 31] between the clustering algorithms. K-Means and Birch have an ARI score of 0.523. K-Means and DBSCAN have an ARI of 0.029, and Birch and DBSCAN have an ARI of 0.026. On comparing these results, we observe that K-Means and Birch produced the most similar clusters, while DBSCAN produced dissimilar clusters to both K-Means and Birch.

As seen in Fig. 4, the clusters for K-Means are more or less grouped into the different continents, but slightly skewed towards the tropical band (i.e. between 20 and -20 latitude). There is one cluster centroid in North/Central America, one in South America, one in Europe, one in Africa, one in South Asia, and another in Southeast Asia/Oceania. Another observation is that the cluster centroids mainly fall within the tropics, save for the European centroid. As shown Fig. 3, the tropics receive on average more rainfall than the rest of the world, therefore we would also expect a higher probability of outbreaks. Going back to the dataset and counting the number of outbreaks between ± 20 degrees in latitude, we find that 1315 outbreaks (62.5% of total outbreaks in the dataset) lie within the tropics.

We also explore patterns in the temporal distribution of the outbreaks. Looking into the monthly distribution of outbreaks in Fig. 5, we see that there is a spike in the number of outbreaks that occur between March and April in both the Northern ($> 0°$ latitude) and Southern ($< 0°$ latitude) Hemispheres, and another spike from August to September in only the Northern Hemisphere (there are no points in our dataset directly on the equator, i.e. with latitude=0°). This supports the correlation between rainfall anomalies and outbreaks, since June to September is generally the monsoon season in South/Southeast Asia [32] which lies in the Northern Hemisphere, while the wet season in Central/South America/Tropical Africa can extend from February through April [33]. Central America and the northern part of South America lies within the Northern Hemisphere and the rest of South America and much of Tropical Africa lie within the Southern Hemisphere. Overall, there is a higher frequency of cases in the Northern Hemisphere than in the Southern Hemisphere regardless of month.

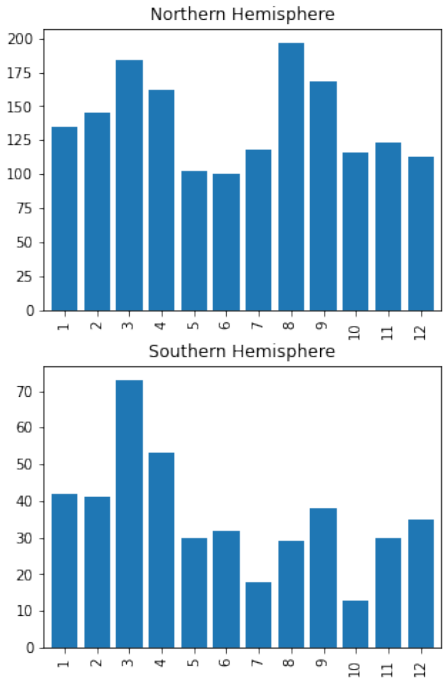

Fig. 5: Month (x-axis) vs Number of outbreaks (y-axis).

C. Rainfall Anomalies and Correlation

Next, we explore the relationship between rainfall, rainfall anomalies, and outbreak magnitude (i.e. case and death counts). Example locations with high rainfall anomaly include Mumbai, India in July 2010 (anomaly=4.11) and Quatre-Bornes, Mauritius in January 2011 (anomaly=4.51). Example locations with low rainfall anomaly include Bié, Angola in June 2007 (anomaly=–1), and Beni Suef, Egypt in July 2011 (anomaly=–1). We calculate a Pearson correlation coefficient [34] of 0.0388 between precipitation anomaly percentage and number of cases, indicating a slight positive correlation; however, this correlation is insignificant at the 0.05 level (p=0.1). Surprisingly, there is a slight negative correlation between deaths and the precipitation anomaly percentage of -0.02, but this is also insignificant at the 0.05 level (p=0.26). However, this requires further detailed investigation.

D. Predictive Models

We train predictive models and evaluate on two tasks: a regression task to predict the number of cases given the coordinates, location hierarchy, month, and precipitation data (anomaly, mean value, current value), and a classification task to predict whether or not there will be a "severe" outbreak (cases > 1000 or deaths > 100) in a given location provided with the same explanatory variables. We use a test-train split of 80/20 and stratify based on disease type to get a representative amount of all disease types in both the training and test data sets. For reproducibility, we set a seed of 69.

Regression modeling is one of the most important statistical techniques used in analytical epidemiology, but our regression task is more challenging than our classification task due to the skewness of our dependent variable (cases). For the regression task, we use a LASSO linear regression model [35]. We normalize values by removing the mean and scaling to unit variance to improve model training. The LASSO model performed very poorly on this data, with an R^2 value of 0.005.

For the easier classification task, we use a random forest classifier [35]. This model performs better than the regression models, with a prediction accuracy of 80.01% and a balanced accuracy of 58.93%. Exploring reasons for the lower balanced accuracy, the model was more accurate in predicting that an outbreak was not severe. The variables with the highest importance were the latitude, longitude, and precipitation data.

Due to the success of the random forest classification model on the classification task, we see if a random forest regression model will work well for the regression task. However, performance of the random forest regressor was even worse than LASSO, with an R^2 of -0.03.

VI. Conclusions and Future Work

A. Conclusion

In this work, we develop an extraction pipeline for ProMED-mail that yielded an initial dataset of 2103 rows with geographical and epidemiological data about malaria, dengue, zika, and chikungunya outbreaks over 25 years from March 1of 1996 to July 2021. We combine this extracted dataset with historical precipitation data from NASA IMERG, and calculate rainfall anomaly percentage for locations and months where an outbreak occurred. Plots of outbreaks over average precipitation and K-Means clustering support that outbreaks are more prevalent and clustered towards the global tropical band (lat ± 20), which also has higher rainfall on average. Outbreaks can also be clustered into separate continents excluding Antarctica, and we see the centroids of these clusters fall mainly within the tropical band. The months with the highest quantities of recorded outbreaks are March, April, August, and September, which align with the rainy seasons of South America, Africa, and South/Southeast Asia respectively. Rainfall anomalies have a slight positive correlation with the number of cases in a recorded outbreak, however, this correlation is insignificant at the 0.05 level. Regression models perform poorly on this data, which may have been due to the skewness of the response variable (cases) and that the relationship between variables is simply not linear. A random forest classification model was fairly successful at predicting if an outbreak is severe or not, with 80% accuracy on our dataset.

B. Future Work

For future work, we would like to expand our data extraction methodology to other websites such as the WHO's weekly epidemiological updates to yield a richer and larger dataset. An improvement to our web scraping methodology could include an entity resolver that could determine if different rows in our dataset correspond to the same actual outbreak to prevent double-counting of outbreaks across administrative hierarchies. We would also perform further time series analysis, e.g. also calculating precipitation anomalies in the months leading up to an outbreak and averaging rainfall over varying lengths of time. Additionally, we would consider combining our scraped data with geographic elevation data and historic land surface temperature data to also explore relationships between elevation/temperature and vector-borne disease outbreaks.

Acknowledgment

This was work conducted as part of Group on Earth Observations (GEO) Health Community of Practice (CoP) activities for Student engagement under Thilanka Munasinghe and Assaf Anyamba. Heidi Tubbs, Bhaskar Bishnoi, Assaf Anyamba were supported under funding from Armed Forces Health Surveillance Branch - Global Emerging Infections Surveillance (GEIS) Project #P0044_20_NS and NASA Applied Sciences Program – Health and Air Quality, Grant #17-HAQ17-0065.

References

[1] M. A. DILLER, "A web scraper and entity resolver for converting public epidemic reports into linked data," Ph.D. dissertation, UNIVERSITY OF FLORIDA, 2018.

[2] WHO, "Vector-borne diseases fact sheet," 2020, accessed: 2021-08-01. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases

[3] U. of California Los Angeles Health Sciences, "Rainfall can indicate that mosquito-borne epidemics will occur weeks later," *ScienceDaily*, 2017. [Online]. Available: https://www.sciencedaily.com/releases/2017/11/171122093117.htm

[4] A. Anyamba, K. J. Linthicum, J. L. Small, K. M. Collins, C. J. Tucker, E. W. Pak, S. C. Britch, J. R. Eastman, J. E. Pinzon, and K. L. Russell, "Climate teleconnections and recent patterns of human and animal disease outbreaks," *PLoS Neglected Tropical Diseases*, vol. 6, no. 1, p. e1465, 2012.

[5] J.-P. Chretien, A. Anyamba, S. A. Bedno, R. F. Breiman, R. Sang, K. Sergon, A. M. Powers, C. O. Onyango,

J. Small, C. J. Tucker *et al.*, "Drought-associated chikungunya emergence along coastal east africa," 2007.

[6] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–4, 12 2008.

[7] G. J. Milinovich, S. M. R. Avril, A. C. A. Clements, J. S. Brownstein, S. Tong, and W. Hu, "Using internet search queries for infectious disease surveillance: screening diseases for suitability," *BMC Infectious Diseases*, vol. 14, no. 1, Dec. 2014. [Online]. Available: https://doi.org/10.1186/s12879-014-0690-1

[8] V. K. Jain and S. Kumar, "Effective surveillance and predictive mapping of mosquito-borne diseases using social media," *Journal of Computational Science*, vol. 25, pp. 406–415, 2018.

[9] Y. Zhang, M. Ibaraki, and F. W. Schwartz, "Disease surveillance using online news: Dengue and zika in tropical countries," *Journal of Biomedical Informatics*, vol. 102, p. 103374, Feb. 2020. [Online]. Available: https://doi.org/10.1016/j.jbi.2020.103374

[10] R. Soebiyanto, A. Anyamba, R. Damoah, W. Thiaw, and K. Linthicum, "Chikrisk app: Global mapping and predicting chikungunya risk," in *100th American Meteorological Society Annual Meeting*. AMS, 2020.

[11] A. Anyamba, J. L. Small, S. C. Britch, C. J. Tucker, E. W. Pak, C. A. Reynolds, J. Crutchfield, and K. J. Linthicum, "Recent weather extremes and impacts on agricultural production and vector-borne disease outbreak patterns," *PLoS ONE*, vol. 9, no. 3, p. e92538, Mar. 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0092538

[12] H. Chen, D. Zeng, and P. Yan, "Argus," in *Infectious Disease Informatics*. Springer, 2010, pp. 177–181.

[13] M. Herman Tolentino, M. Raoul Kamadjeu, M. Michael Matters PhD, M. Marjorie Pollack, and M. Larry Madoff, "Scanning the emerging infectious diseases horizon-visualizing promed emails using epispider," *Adv Dis Surveil*, vol. 2, p. 169, 2007.

[14] C. C. Freifeld, K. D. Mandl, B. Y. Reis, and J. S. Brownstein, "Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports." *J Am Med Inform Assoc*, vol. 15, no. 2, pp. 150–7, 2008 Mar-Apr 2008.

[15] N. Collier, S. Doan, A. Kawazoe, R. M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, M. Shigematsu, and K. Taniguchi, "BioCaster: detecting public health rumors with a web-based text mining system," *Bioinformatics*, vol. 24, no. 24, pp. 2940–2941, Oct. 2008. [Online]. Available: https://doi.org/10.1093/bioinformatics/btn534

[16] A. Abbood, A. Ullrich, R. Busche, and S. Ghozzi, "Eventepi—a natural language processing framework for event-based surveillance," *PLoS computational biology*, vol. 16, no. 11, p. e1008277, 2020.

[17] M. M. Glancey, A. Anyamba, and K. J. Linthicum, "Epidemiologic and environmental risk factors of rift valley fever in southern africa from 2008 to 2011," *Vector-Borne and Zoonotic Diseases*, vol. 15, no. 8, pp. 502–511, 2015.

[18] L. P. Campbell, D. C. Reuman, J. Lutomiah, A. T. Peterson, K. J. Linthicum, S. C. Britch, A. Anyamba, and R. Sang, "Predicting abundances of aedes mcintoshi, a primary rift valley fever virus mosquito vector," *PloS one*, vol. 14, no. 12, p. e0226617, 2019.

[19] N. Bergquist, "Vector-borne parasitic diseases: new trends in data collection and risk assessment," *Acta tropica*, vol. 79, no. 1, pp. 13–20, 2001.

[20] M. Carrion and L. C. Madoff, "Promed-mail: 22 years of digital surveillance of emerging infectious diseases," *International health*, vol. 9, no. 3, pp. 177–183, 2017.

[21] J. Simpson, C. Kummerow, W.-K. Tao, and R. F. Adler, "On the tropical rainfall measuring mission (trmm)," *Meteorology and Atmospheric physics*, vol. 60, no. 1, pp. 19–36, 1996.

[22] G. J. Huffman, D. T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, P. Xie, and S.-H. Yoo, "Nasa global precipitation measurement (gpm) integrated multi-satellite retrievals for gpm (imerg)," *Algorithm Theoretical Basis Document (ATBD) Version*, vol. 4, p. 26, 2015.

[23] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[24] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[25] C. Lothritz, K. Allix, L. Veiber, J. Klein, and T. F. D. A. Bissyande, "Evaluating pretrained transformer-based models on the task of fine-grained named entity recognition," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3750–3760.

[26] T. Zhang, R. Ramakrishnan, and M. Livny, "Birch: An efficient data clustering method for very large databases," in *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '96. New York, NY, USA: Association for Computing Machinery, 1996, p. 103–114. [Online]. Available: https://doi.org/10.1145/233269.233324

[27] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231.

[28] C. Yuan and H. Yang, "Research on k-value selection method of k-means clustering algorithm," *J*, vol. 2, no. 2, pp. 226–235, 2019.

[29] G. Boeing, "Clustering to reduce spatial data set size," Mar 2018. [Online]. Available: osf.io/preprints/socarxiv/nzhdc

[30] T. Munasinghe, A. N. Maheshwarkar, and K. Bhanot, "Socioeconomic and geographic variations that impacts the spread of malaria," 2020.

[31] K. Y. Yeung and W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 17, no. 9, pp. 763–774, 2001.

[32] S. Kumar and G. Bhat, "Vertical structure of orographic precipitating clouds observed over south asia during summer monsoon season," *Journal of Earth System Science*, vol. 126, no. 8, pp. 1–12, 2017.

[33] S. L. Hastenrath, "Rainfall distribution and regime in central america," *Archiv für Meteorologie, Geophysik und Bioklimatologie, Serie B*, vol. 15, no. 3, pp. 201–241, 1967.

[34] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[35] Y. Mansiaux and F. Carrat, "Detection of independent associations in a large epidemiologic dataset: a comparison of random forests, boosted regression trees, conventional and penalized logistic regression for identifying independent factors associated with h1n1pdm influenza infections," *BMC medical research methodology*, vol. 14, no. 1, pp. 1–10, 2014.