What Causes an Epidemic?

Data 698 Masters Research Project

CUNY School of Professional Studies
Michele Bradley

Abstract: Epidemics possess a significant ability to threaten society: both through the loss of lives and through the loss of economic growth. Combatting their emergence should be a societal priority, but our understanding of how outbreaks and epidemics originate are often hyper-localized. Researchers focus on the outbreak of a single epidemic but do not typically attempt trend analysis to better understand how global metrics correlate to their emergence. Through this paper, we will explore if various metrics (forest loss, climate data, population density, cropland, and surface water) will impact the zoonotic type of epidemic emergence. Findings include that Rodent epidemics tend to originate in warmer climates while Bird epidemics tend to occur in regions with more surface water. In addition, hotspots for epidemic emergence were conducted using machine learning techniques and the previously named five metrics. Hotspots identified include regions of Europe, East Asia, South East Asia, Eastern United States, Mexico, West Africa, East Africa, and Brazil.

Keywords: Global Health, Epidemics, Climate Change, Google Earth Engine, Machine Learning

Literature Review

Google Earth Engine (GEE) is a web portal established in 2010 that provides “global time-series satellite imagery and vector data, cloud-based computing, and access to software and algorithms for processing such data” (Kumar, 2018). The data repository is a collection of over 40 years of satellite imagery. As of 2017, 300 papers were published using this data, most of it being for understanding “forest and vegetation category (17%) and land use and land cover studies (10%)” (Kumar, 2018). A small proportion of studies used this data for purposes other than ecological and earth sciences, with only 2% using it to better understand natural hazard and disaster studies” (Kumar, 2018). There is an incredible opportunity to use this data to better understand hazards and disasters that are not solely ecological, such as understanding the mechanisms for epidemic origins.

Projects that use GEE data and machine learning techniques have been encouraging and helpful to learn from. While many data science projects do not utilize GEE, popular data science algorithms and machine learning techniques have been brought into the GEE data environment. This is largely due to the fact that GEE data is mainly used in programming languages such as JavaScript and recently, Python. Being able to merge Python based machine learning techniques along with the repository of GEE data allows for the emergence of new and important global discoveries.

GEE papers largely fell into two categories: those using Landsat satellite imagery to categorize global data and those that use pre-categorized data for data analysis and interpretation. Using

Landsat satellite imagery and random forest algorithms were common use cases that help researchers classify how land is being used. For example, Kelly and team used it to Map Complex Shade-Grown Coffee Landscapes in Northern Nicaragua while Hu and team used it to perform land classification and obtain the annual land cover datasets of Central Asia from 2001 to 2017. Using pre-categorized data was also common and most aligned with this project. Hu and team not only performed land classification but also added interesting socioeconomic data layers such as GDP and population density to determine the mechanisms for land change. The team for example, noted that the capital of Kazakhstan moved from Almaty to Astana in 1998 and with it, regional population shifts caused shifts in occupation of cultivated land that impacted annual precipitation (Hu). Another team used pre-categorized GEE data to explore relations between soil moisture, precipitation, and streamflow for watersheds in order to predict how water flows in watersheds across America. With data on 601 watersheds, the team used (ARIMA) and support vector machine (SVM) regression models (SVR) to determine that there was a significant correlation between precipitation, soil moisture, and streamflow (Sazib). Papers that convey that ARMIA, SVM, and SVR models can be used in order to accurately predict global weather patterns and demonstrate that using global GEE data for predictions were useful research to add to the literature and better understand what GEE is capable of. In addition, viewing how research teams developed underlying data sources and classified land datasets embedded in GEE is useful as it adds to the rigor of the datasets available.

One paper, recently published in late September studied a similar topic in relation to deforestation and the COVID-19 pandemic. Titled *Emerging threats linking tropical deforestation and the COVID-19 pandemic*, this team calculated, using GEE and R, the total deforestation for 2019 and 2020 per country along with the Government Response Stringency Index to COVID-19 (a composite index accounts for nine response indicators, like travel bans and workplace closures) to determine if topical deforestation has increased because of COVID-19 (Brancaliona). Their analysis provides evidence that there was an immediate increase in tropical deforestation following policies aimed at minimizing the impact of COVID-19 (Brancaliona). Their methods for tracking tropical deforestation through GEE and utilizing a 100 km × 100 km grid distributed across the global tropics was useful in understanding how to perform big data analysis in GEE. Looking through the literature, this was commonly used when analysis was performed in more localized regions but were not performed on the global level. Therefore, this approach for data analysis was ultimately not used but could be used in the future.

Another paper, titled *Global hotspots and correlates of emerging zoonotic diseases,* provided and published in 2017 provided an alternative way of analyzing large GEE datasets globally. This paper used EcoHealthAlliance's Global Health Epidemic Emergence location data to generate 5km circular spatial polygons surrounding each event with significant data points and existed after 1970. The team used 20 different spatial indicators and R to determine statistics within these polygons and ultimately, to determine global hotspots for emerging zoonotic diseases. This approach was highly influential and their methods were studied and incorporated into this research paper. The team used boosted regression trees (BRT) to model emerging epidemic occurrences. They also used a bootstrap resampling regime to fit the models and generate more datapoints.

In addition to studying GEE data analysis, the literature surrounding the presence of zoonotic diseases were also studied in order to determine which GEE variables would be useful dependent variables that could predict epidemic emergence. Bloomfield writes that “viruses that jump from animals to people, like the one responsible for COVID-19, will likely become more common as people continue to transform natural habitats into agricultural land”. Rohr found that land changes were associated with 25% of all diseases and nearly 50% of zoonotic diseases that emerged in humans since 1940 and that the more that we transform our natural environment, the more that viruses will spread and continue to jump from animals to people. As we manipulate our landscapes, pandemics that spread might become more commonplace, as the natural habitat of animals become more and more cramped. MacDonald and Mordecai (2019) found that deforestation significantly increases malaria transmission in the Amazon. Through this research, data surrounding land use and land use changes seemed particularly relevant to the question surrounding epidemic emergence.

In addition, climate change’s ability to alter ecosystems can cause significant changes to the animals that live in various ecosystems, and result in “alterations in the natural habitat of pathogens….and human hosts, as well as in the transmission dynamics and geographic distribution of infectious agents” (Christaki, 2020). This interaction between infectious disease and climate change is “not deeply understood”, but with the increasing amount of data available to us, we might be able to understand their interactions more (Christaki, 2020). Therefore, data surrounding weather patterns would also be important to incorporate into this model.

Methodology

Using Google Earth Engine data, Python’s Google Collab, and epidemic data compiled EcoHealthAlliance, it was possible to determine if there were significant correlations between the start of epidemics and various global data markers. Using EcoHealthAlliance’s epidemic emergence dataset, longitude and latitude coordinates were compiled when origin data was complete. Regions that only had country-level data were removed from the dataset. For analysis, sixty mile and one hundred eighty mile buffer polygons around the origin coordinates were generated.

Global data metrics within those polygons were compiled and utilized for generating insights. In addition to generating these buffer zones, the EcoHealthAlliance dataset also incorporated animal transmission data. Zoonotic diseases originate from animal sources, and these sources were important inputs for understanding how different zoonotic diseases originate. Common animals include Bats and Rodents, but they can also be Human-caused, such as through human drug resistance. Epidemic origins based on animals are likely an environment-related variable, as animals live in different climates and regions of the world, so they were an important input variable to understand.

Many of the assumptions in this research surrounded the accuracy of this data. We must first acknowledge that we are assuming the GEE data is accurate, since that data is primarily generated using various statistical metrics. It is especially significant as society has been questioning the origins of COVID-19. Researchers are still unsure of the true origin of this epidemic, especially since it became so global in nature. However, most of the epidemics covered in this dataset were significantly more localized and therefore the region of interest is significantly smaller. By reducing the interest of this project to epidemics as opposed to pandemics, we reduce the risk of reporting on incorrect origination information.

In addition to utilizing polygons surrounding epidemic emergence, it was important to also incorporate sample data from locations that did not have any epidemic emergence present. In order to achieve this, global coordinates were simulated and removed if they were not land-based coordinates or if they were located within the boundaries of the epidemic origin dataset. Sixty mile and one hundred eighty mile buffer polygons were generated surrounding these coordinates as well, and the global data metrics within those polygons were also compiled and utilized for generating insights.

The Google Earth Engine metrics that were used included:

1. Forest Loss data compiled by the University of Maryland. This dataset was generated from time-series analysis of Landsat images in characterizing global forest extent and change from 2000 through 2019.
2. Annual Mean Temperature data compiled by the University of California at Berkeley. This dataset was generated using global climate data compiled between 1950 and 2000.
3. Human Population Data from 2010 compiled by NASA, which uses census data.
4. Global Cropland Data compiled from the 30m Project that includes if the land is largely irrigated by a farmer or rainified. This data was gathered using remote sensing techniques.
5. Location of surface water and the extent and change of the water from 1984 to 2019. Data is compiled by the European Commission and Google and utilized three million LANDSAT satellite images.

Metrics within each of these variables were compiled using GEE's reducer function. This function flattens datasets and summarizes pixel data within a set polygon. The reducer function allows for these large map datasets to be summarized into numeric values that can be used for performing various statistical analysis.

Using the associated metrics found and compiled within polygons of both buffer zones, ANOVA and Tukey tests were used to determine the importance of the animal variable against epidemic emergence. The ANOVA test was able to determine if there were significant differences between epidemics that originate from various animals (Bats, Rodents, Primates, Humans, Unknown, etc.,) along with the other metrics compiled from Google Earth Engine. ANOVA tests are useful for understanding if there are significant differences between all of the groups, while the Tukey Test is a good post-hoc metric that is used to determine which of those pairs of variables have the most statistically significant difference.

After these tests were administered, machine learning techniques such as Binary Linear Regression using Gradient Descent and Random Forest with bootstrapping were used to model and predict the emergence of epidemics. Predicting when and where the next epidemic occurred was not the goal of this paper, but rather to understand hotspots and areas of focus for researchers and scientists.

Finally, coordinates from 15,000 cities in every country of the world were found from simplemaps.com and were used to determine hotspots for epidemic origination. Smaller 60 mile polygons were formed around the 15,000 cities and GEE data was compiled using the reducer function. The chosen machine learning algorithm was used to predict and data from the 15,000 cities served as inputs to determining hotpots of epidemic origination around the world.

Results

Experimentation surrounded adding variables from Google Earth Engine and trying different statistical methods to interpret and understand the data, such as ANOVA, Tukey Test, Binary Linear Regression with Gradient Descent, Random Forest Classifiers, and Random Forest Regression.

It was also important to map the variables in order to better understand the interaction between the variables. Figure 1 maps the location of all the input data, where colors of the icons indicate the origin type of the epidemic: Green = Primates, Blue = Tick, Red = Bat, Dark Purple = Rodent, Black = Human/Domestic Animal, Unknown = Orange, Pink = Birds, Dark Blue = Flea, and Light Grey = None.

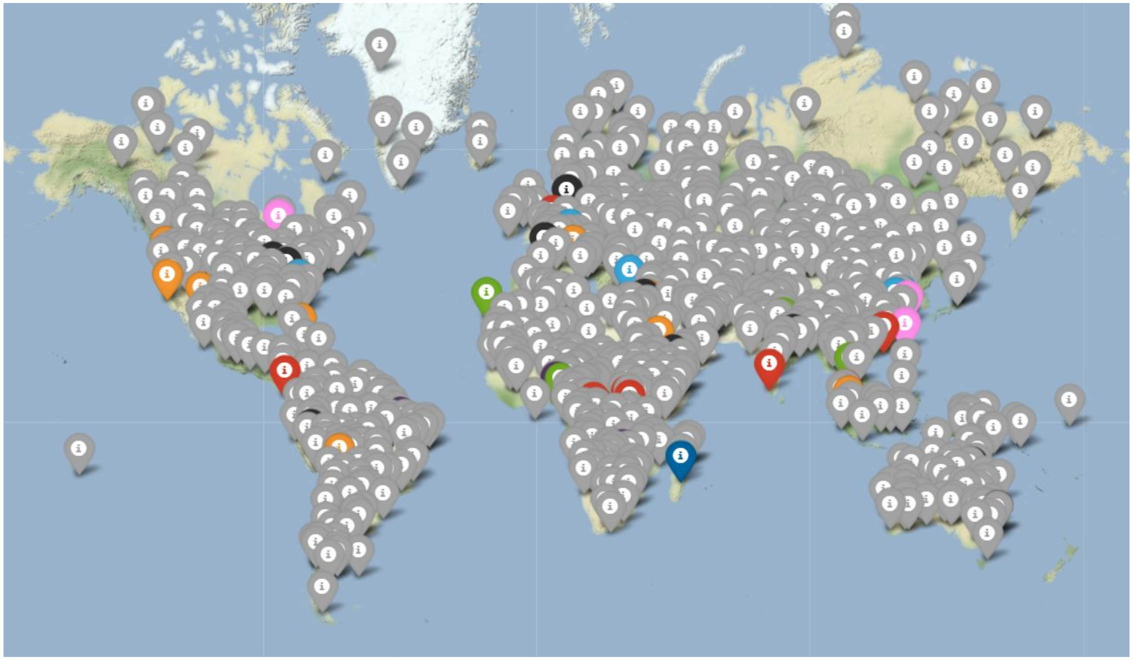

Fig 1

Adding independent variables to the map was the next step. Figure 2 maps the annual mean temperature, where red are hotter climates and turquoise is colder climates. The small black dots indicate areas in which forest has been deforested and lost. The point icons once again indicate the origins of epidemics.

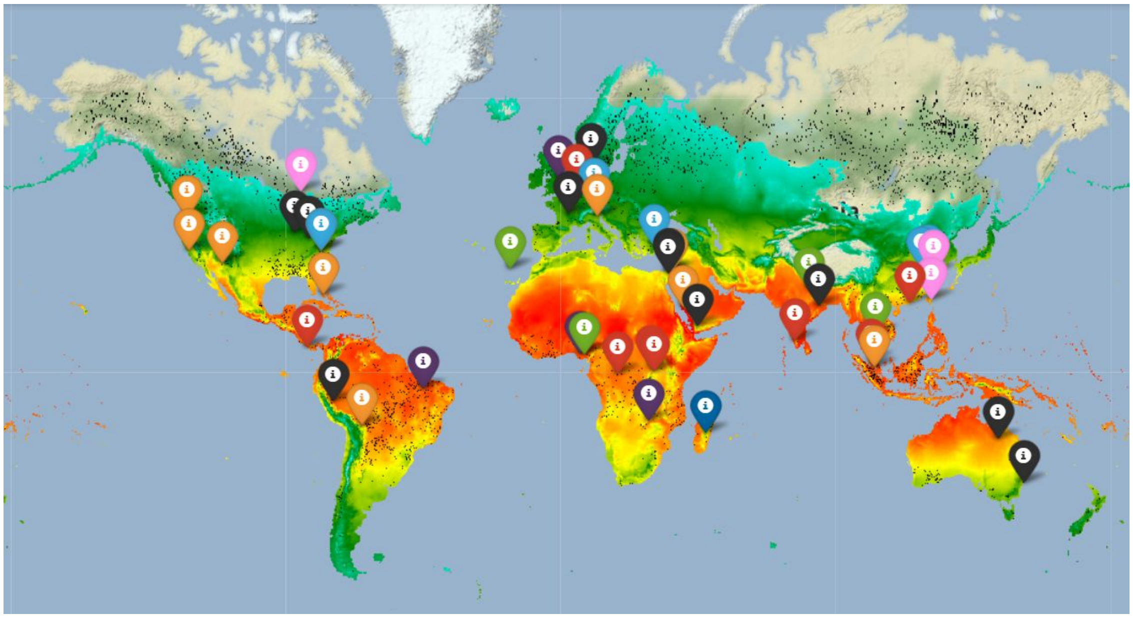

Fig 2

Figure 3 displays human population data alongside the areas in question, where dark green regions indicate higher regions of human population. This map is zoomed in to more clearly the distribution of population.

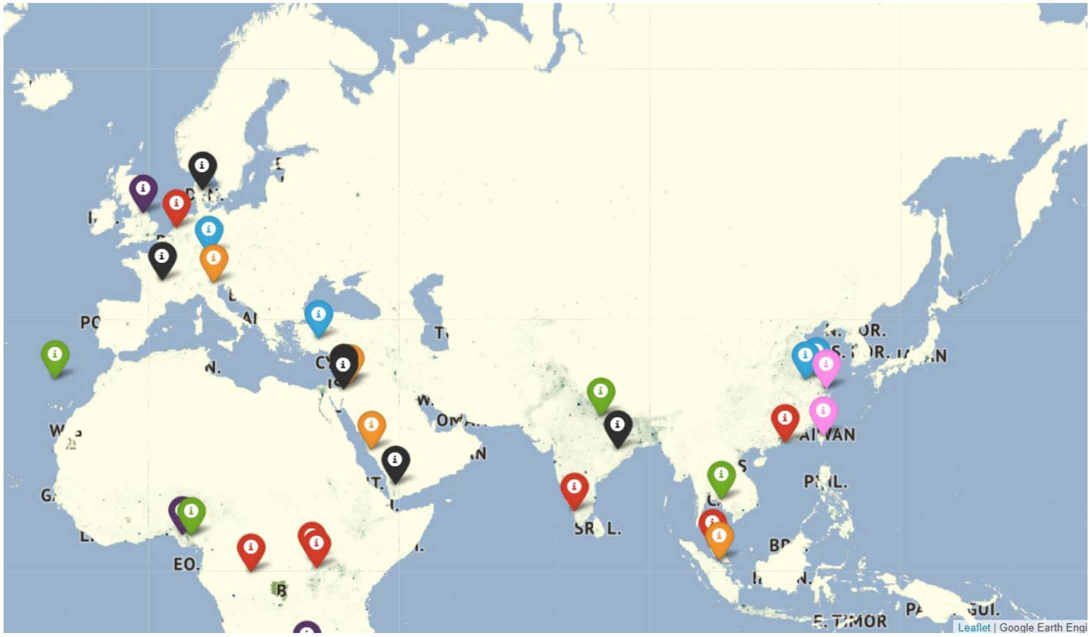

Fig 3

Figure 4 displays cropland distribution in relation to epidemic emergence. Areas that are yellow indicate low concentration of rainfed cropland. Green areas indicate high concentrations of rainfed cropland. Brown areas indicate low concentrations of irrigated cropland. Orange areas indicate high concentration of irrigated cropland.

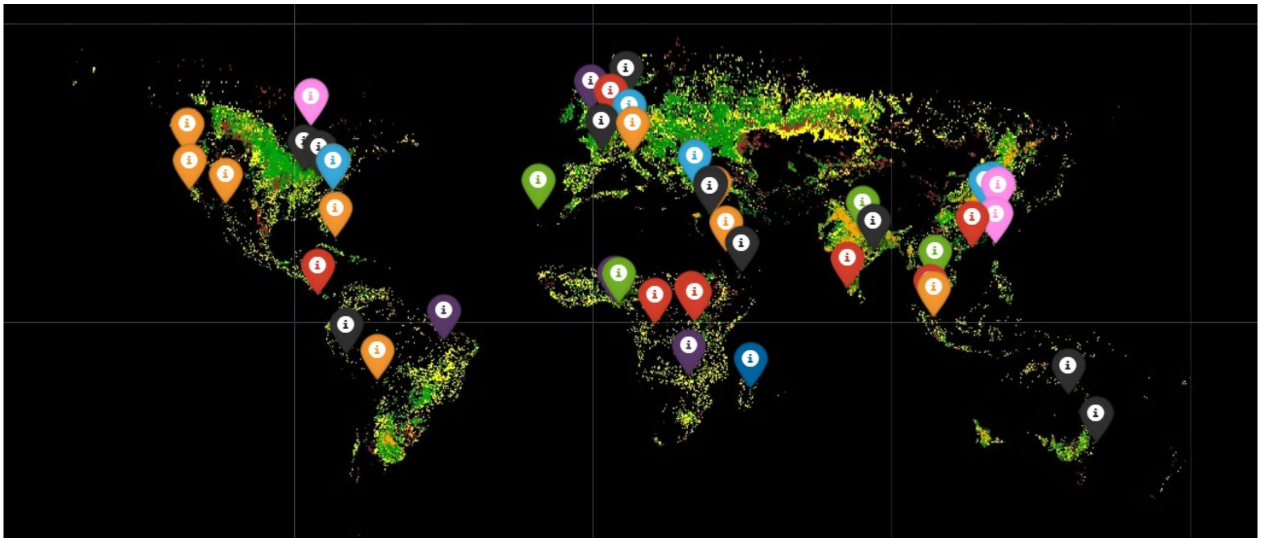

Fig 4

Figure 5 displays surface water to epidemic emergence. Areas that are white indicate presence of surface water, such as lakes, rivers and streams.

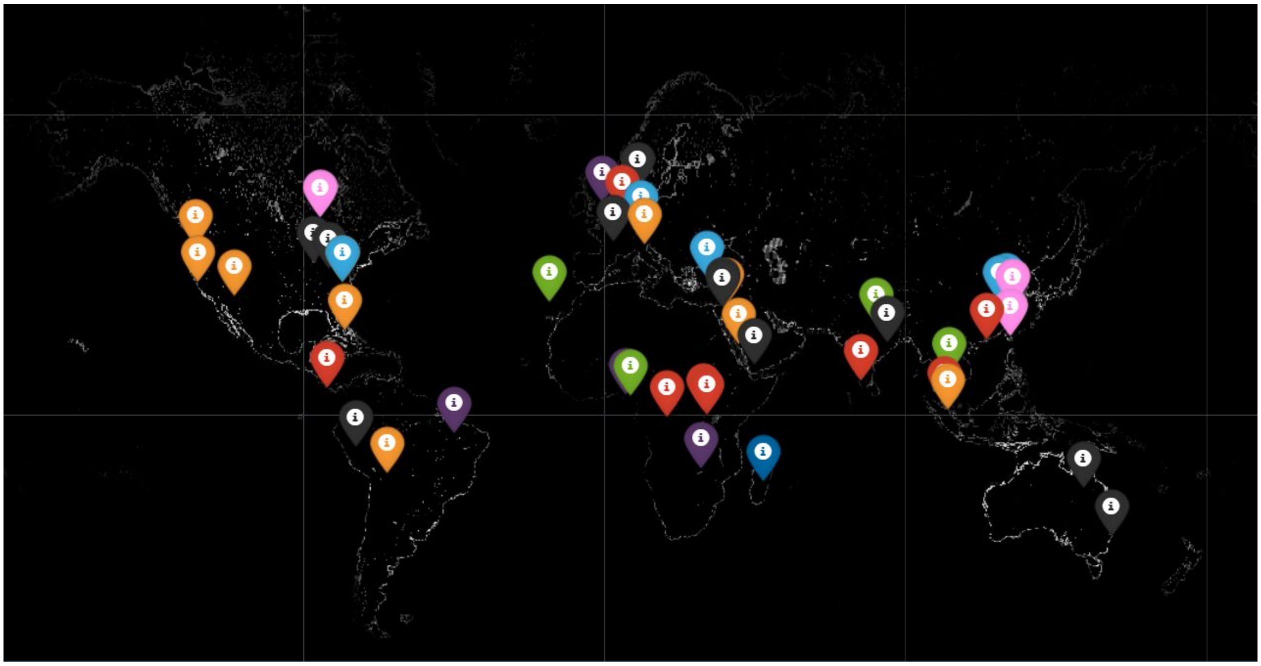

Fig 5

Each of these maps is displaying millions of data points. Each pixel represents a datapoint of information about that particular region. Utilizing global map data is an effective and efficient way to model global data trends.

Each of the animal types were used in the ANOVA tests using the five different variables using a 60-mile and a 180-mile radius. Each ANOVA test was then analyzed using a Tukey Test to determine where the differences lied between the groups. Forest Loss did not differ significantly between the animal origins. Climate differed significantly between Human and Rodent origins with a p-value of .039, where the average temperature for Human-origin epidemics were 33 degrees fahrenheit where the average temperature for Rodent-origin epidemics were 52 degrees fahrenheit. Population density differed significantly between Unknown-origin and Bat-origin epidemics with a p-value of .0333, None and Unknown epidemics with a p-value of .001, and Rodent-origin and Unknown-origin epidemics with a p-value of .03. Unknown-origin epidemics had higher population densities. In addition, there were significant differences between no epidemic locations and Primate-origin epidemics with a p-value of .0048, where primates again had higher population densities. Cropland data did not have significant differences between the animal origins. Surface water differed significantly in 180 mile radius polygons between Birds and various animal origins: Bats, Domestic Animals, None, Primates, Rodent, Tick, and Unknown. In these regions, regions with Birds had more water surface area. Those differences are outlined in figure 6.

Variable	**Radius**	**Animal 1**	**Animal 2**	**P-value**
Surface Water	180	Bird	Bat	.0043
Surface Water	180	Bird	Domestic Animal	.011
Surface Water	180	Bird	None	.001
Surface Water	180	Bird	Primate	.0131
Surface Water	180	Bird	Rodent	.0048
Surface Water	180	Bird	Tick	.0016
Surface Water	180	Bird	Unknown	.0106
Population Density	60	Bat	Unknown	.0333
Population Density	60	None	Unknown	.001
Population	60	Rodent	Unknown	.03

Density				
Population Density	180	None	Primate	.0048
Climate	60	Human	Rodent	.039

Fig 6

In addition, Binary Linear Regression Models were generated using a 60 mile buffer and climate data and population density were incorporated, since those were shown to be significantly important from the ANOVA tests. These models had 92.42% accuracy.

Random Forest Classification was also performed, using a 60 mile buffer and the inclusion of all of the global variables. These models had 93.98% accuracy. The importance of the variables are outlined in Fig 7, with surface water locations being 44% important, climate data being 19% important, population density data being 13% important, forest loss data being 12% important, and crop cover being 11% important.

Variable	**Importance**
Forest Loss	.124454
Climate	.189330
Population Density	.129017
Crop Cover	.112136
Surface Water	.445063

Fig 7

While utilizing this binary classification random forest model, there were not a lot of locations that were being suggested as locations for epidemic origination. This does make sense, since the general probability of epidemics occurring on the day to day at any moment in time is incredibly slim. Using input longitude/latitude data from 15,000 cities in the Random Forest Classification models resulted in only one region of epidemic emergence: Deland, Florida.

However, using a classification model was potentially undercounting the possibility of epidemic occurrence and the probability of emergence would be a more interesting and accurate way of displaying regions that would be hotspots for epidemic origination. The question of epidemic emergence would depend on external variables coming into play and chance playing it's part. Therefore, the question that would be more useful to ask would be: "What areas are ripe for epidemic emergence?" Therefore, Random Forest Regression machine learning models were used instead with data from a 60 mile buffer of data and incorporating the inclusion of all of the global variables. These models had a .25 mean squared error. The importance of the variables are outlined in Fig 8, with surface water locations being 46% important, climate data being 15% important, population density data being 14% important, forest loss data being 12% important, and crop cover being 12% important.

Variable	Importance
Forest Loss	.124512
Climate	.150382
Population Density	.138558
Crop Cover	.122074
Surface Water	.464474

Fig 8

Using the Random Forest Regression models, the possibility that various cities might have to deal with an epidemic emergence was determined. The table below shows sample cities with the highest likelihood and the lowest likelihood of epidemic emergence.

City	Country	Prediction
Juncitan	Mexico	78%
Kabankalan	Philippines	75%
Chai Nat	Thailand	75%
Kpalime	Togo	75%
Bierun Stary	Poland	73%
Pati	Indonesia	0%
Forbe Oroya	Peru	0%
Ta`izz	Yemen	0%

Heydarabad	Azerbaijan	0%
Andorra la Vella	Andorra	0%

Fig 9

Summary and Future Works

Figure 10 graphically displays areas of hotspots for epidemic emergence. Points that are colored orange represent that there was a 40% to 60% change of epidemic emergence, while points that are colored red indicate a 60% to 80% change of epidemic emergence. Clusters of data form around various cities in Europe, East Asia, South East Asia, Eastern United States, Mexico, West Africa, East Africa, East Asia, and Brazil.

In terms of policy recommendation, it would be ideal for these regions to monitor variables that are relevant to their region regarding epidemic origination. Warmer climates should be paying attention to ensuring Rodent-based epidemics do not occur by checking rodent populations and keeping them in check, while cooler climates should pay special attention to drug resistance measures and ensure that they do not cause epidemics. Regions with high levels of population density should check for primate-originated epidemics, while areas with surface area water should pay special attention to Bird populations and presence of the avian flu.

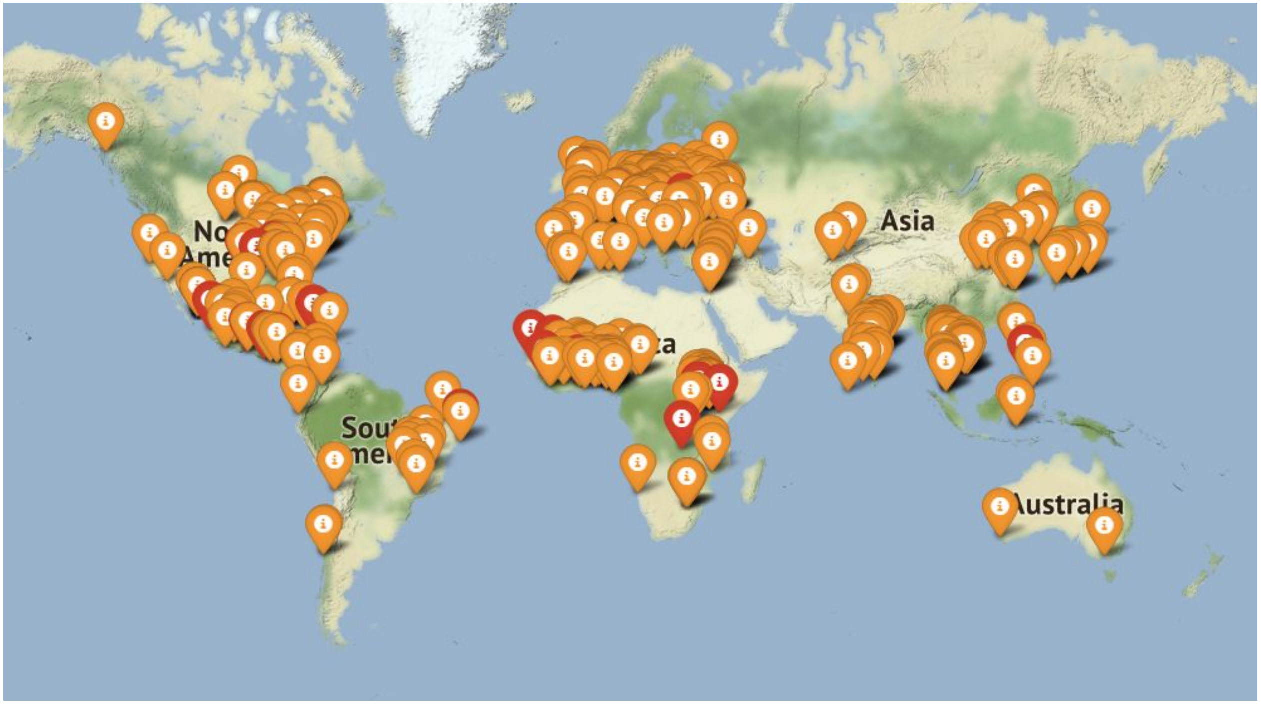

Fig 10

There is ample possibility for future research in this topic. In the future, use of the elements of time within both the independent and dependent variables, such as time of year for temperature data, the year of the outbreak, as well as population change data throughout the years could be useful. Many of these variables are time-dependent, such as temperatures, which vary significantly with time of year.

In addition, incorporating more datasets from Google Earth Engine as independent variables would be interesting, such as Topography or Species Data, in order to see how those variables would improve the model.

Finally, we need to better understand the relationships between our food production, diseases, poverty, and climate change in order for us to better prepare for future global disasters. As Phillps puts it, we’re facing “compound risk”: climate change is causing climate hazards around the world, and these intersections are causing major disruptions that require that we determine how to move forward. The more that we can understand surrounding the interactions between climate and global health, the more that we can prepare and ensure the health and safety of society. Future research would benefit from better understanding how climate change could impact these projections and determine if areas will be under more or less pressure. Through this research it became clear that temperature (and therefore biomes as well as location of animals) were an important component of epidemic emergence. Therefore, as climate change becomes a more and more abundantly clear threat, adding climate projection data that demonstrates temperatures in the future would be a useful component of this research, so we can understand how temperature changes and increasing temperatures will alter the emergence of epidemics in the future.

Appendices

Detailed code found on github.

References

Allen, T., Murray, K.A., Zambrana-Torrelio, C. *et al.* Global hotspots and correlates of emerging zoonotic diseases. *Nat Commun* 8, 1124 (2017). https://doi.org/10.1038/s41467-017-00923-8

Bloomfield, Laura & McIntosh, Tyler & Lambin, Eric. (2020). Habitat fragmentation, livelihood behaviors, and contact between people and nonhuman primates in Africa. Landscape Ecology. 35. 10.1007/s10980-020-00995-w.

Brancalion, P., Broadbent, E., De-Miguel, S., Cardil, A., Rosa, M., Almeida, C., . . . Almeyda-Zambrano, A. (2020, September 30). Emerging threats linking tropical deforestation and the COVID-19 pandemic.

Cappucci, Matthew. Running out of hurricane names, we'll soon switch to the Greek alphabet. That could present a problem., Washington Post, 9/16/2020

Christaki, E., Dimitriou, P., Pantavou, K., & Nikolopoulos, G. K. (2020). The Impact of Climate Change on Cholera: A Review on the Global Status and Future Challenges. *Atmosphere*, *11*(5), 449. MDPI AG. Retrieved from http://dx.doi.org/10.3390/atmos11050449

Hu, Yunfeng & Hu, Yang. (2019). Land Cover Changes and Their Driving Mechanisms in Central Asia from 2001 to 2017 Supported by Google Earth Engine. Remote Sensing. 11. 554. 10.3390/rs11050554.

Kelley, Lisa & Pitcher, Lincoln & Bacon, Christopher. (2018). Using Google Earth Engine to Map Complex Shade-Grown Coffee Landscapes in Northern Nicaragua. Remote Sensing. 10. 952. 10.3390/rs10060952.

Kumar, Lalit & Mutanga, Onisimo. (2018). Google Earth Engine Applications Since Inception: Usage, Trends, and Potential. Remote Sensing. 10. 1509. 10.3390/rs10101509.

MacDonald, Andrew & Mordecai, Erin. (2019). Amazon deforestation drives malaria transmission, and malaria burden reduces forest clearing: a retrospective study. The Lancet Planetary Health. 3. S13. 10.1016/S2542-5196(19)30156-1.

Phillips, C.A., Caldas, A., Cleetus, R. *et al.* Compound climate risks in the COVID-19 pandemic. *Nat. Clim. Chang.* 10, 586–588 (2020). https://doi.org/10.1038/s41558-020-0804-2

Rohr, J.R., Barrett, C.B., Civitello, D.J. *et al.* Emerging human infectious diseases and the links to global food production. *Nat Sustain* 2, 445–456 (2019). https://doi.org/10.1038/s41893-019-0293-3

Sazib, N., Bolten, J., & Mladenova, I. (2020). Exploring Spatiotemporal Relations between Soil Moisture, Precipitation, and Streamflow for a Large Set of Watersheds Using Google Earth Engine. *Water*, *12*(5), 1371. MDPI AG. Retrieved from http://dx.doi.org/10.3390/w12051371

Data Source References

Center for International Earth Science Information Network - CIESIN - Columbia University. 2018. Gridded Population of the World, Version 4 (GPWv4): Basic Demographic Characteristics, Revision 11. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H46M34XX.

Jean-Francois Pekel, Andrew Cottam, Noel Gorelick, Alan S. Belward, High-resolution mapping of global surface water and its long-term changes. Nature 540, 418-422 (2016). (doi:10.1038/nature20584)

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. 2013. "High-Resolution Global Maps of 21st-Century Forest Cover Change." Science 342 (15 November): 850–53. Data available on-line at: http://earthenginepartners.appspot.com/science-2013-global-forest.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones and A. Jarvis, 2005. Very High Resolution Interpolated Climate Surfaces for Global Land Areas. International Journal of Climatology 25: 1965-1978.

Teluguntla, P., Thenkabail, P.S., Xiong, J., Gumma, M.K., Giri, C., Milesi, C., Ozdogan, M., Congalton, R., Tilton, J., Sankey, T.R., Massey, R., Phalke, A., and Yadav, K. 2014. Global Cropland Area Database (GCAD) derived from Remote Sensing in Support of Food Security in the Twenty-first Century: Current Achievements and Future Possibilities. Chapter 7, Vol. II. Land Resources: Monitoring, Modelling, and Mapping, Remote Sensing Handbook edited by Prasad S. Thenkabail. In Press.