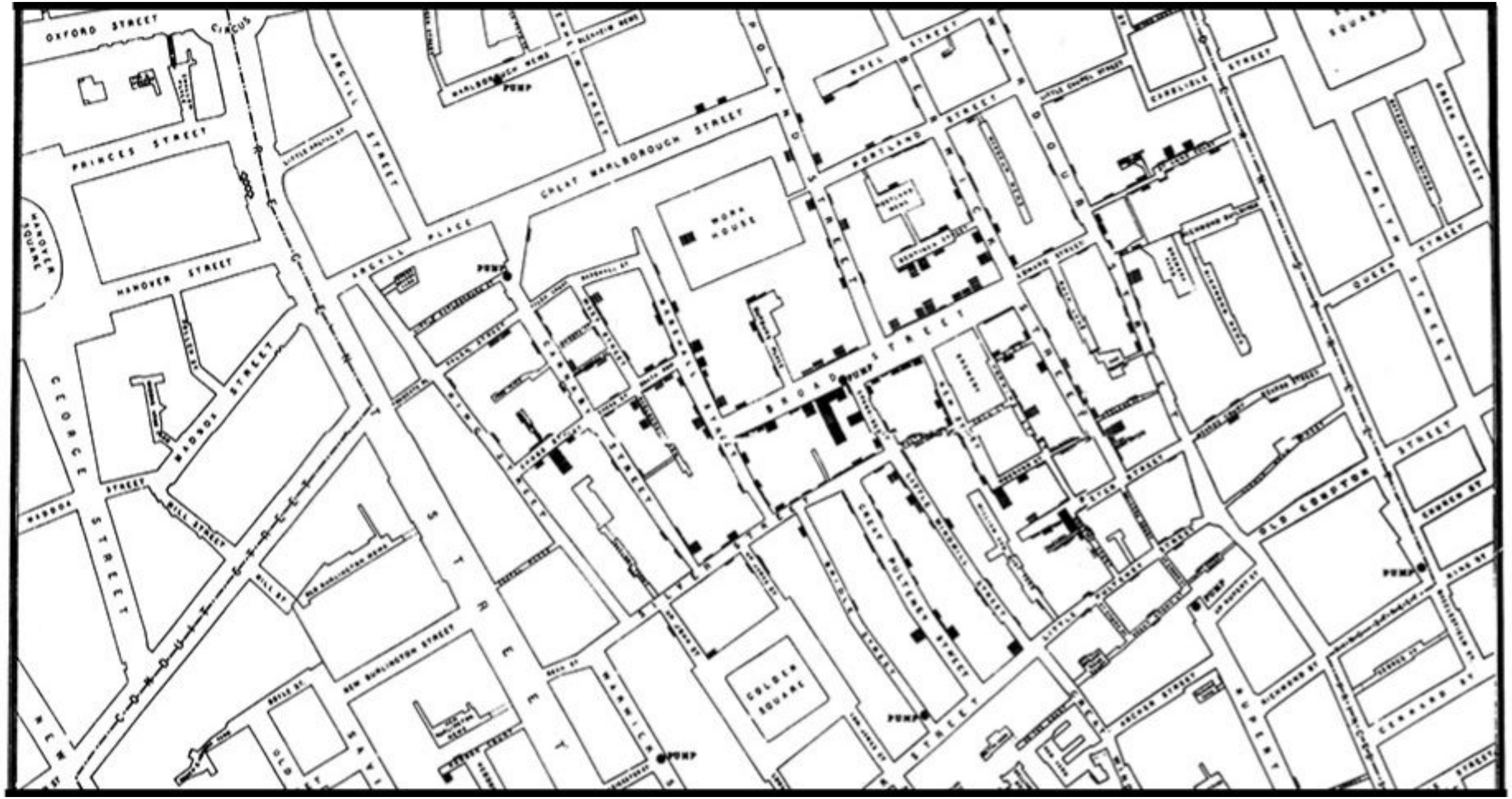
What Causes an Epidemic?

DATA 698 Analytics Masters Research Project

Michele Bradley

CONTAGION

John Snow’s 1854 London Cholera Map



Objective

Understand outbreaks of epidemics utilizing global data

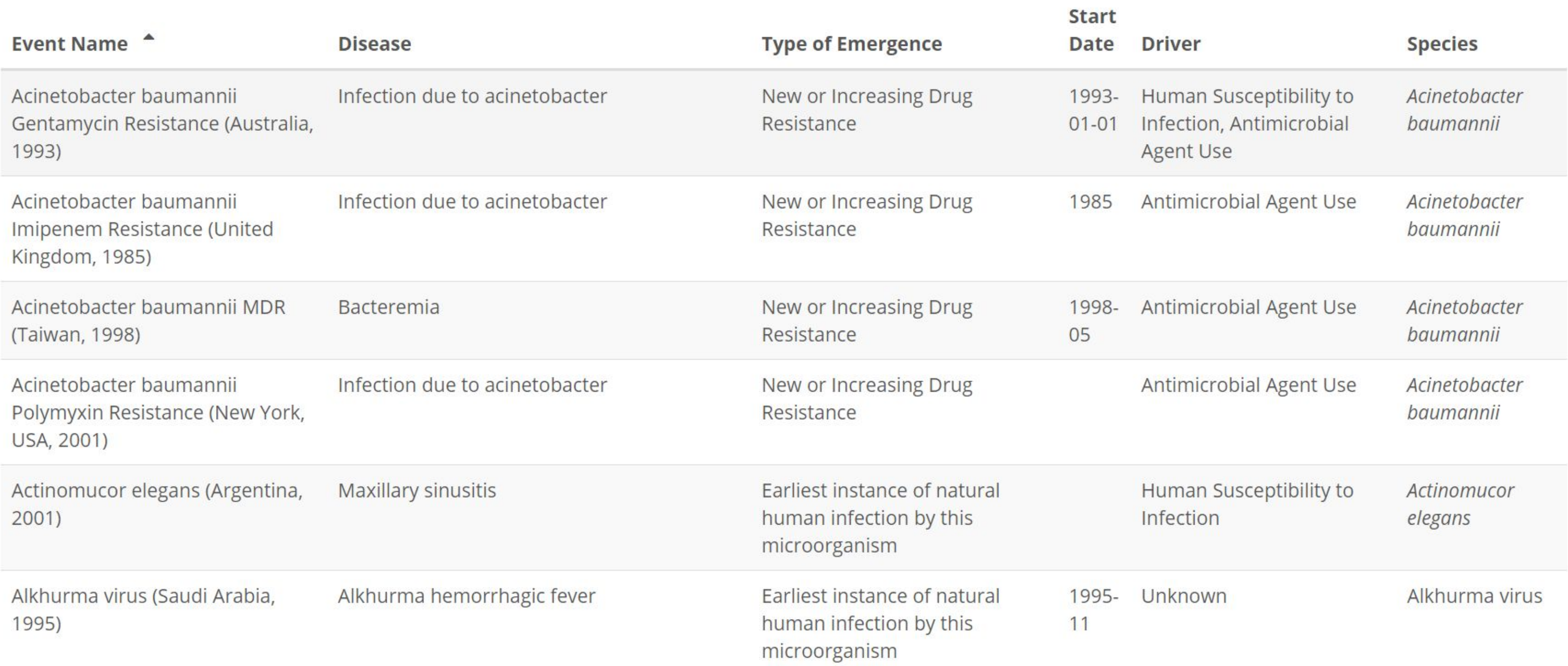
1. Are certain epidemics more likely to occur in certain regions?
2. Do epidemics correlate to global metrics?
3. Can we find hotspots for epidemic emergence?

Epidemic vs Pandemic

Epidemic: A widespread occurrence of an infectious disease in a community at a particular time

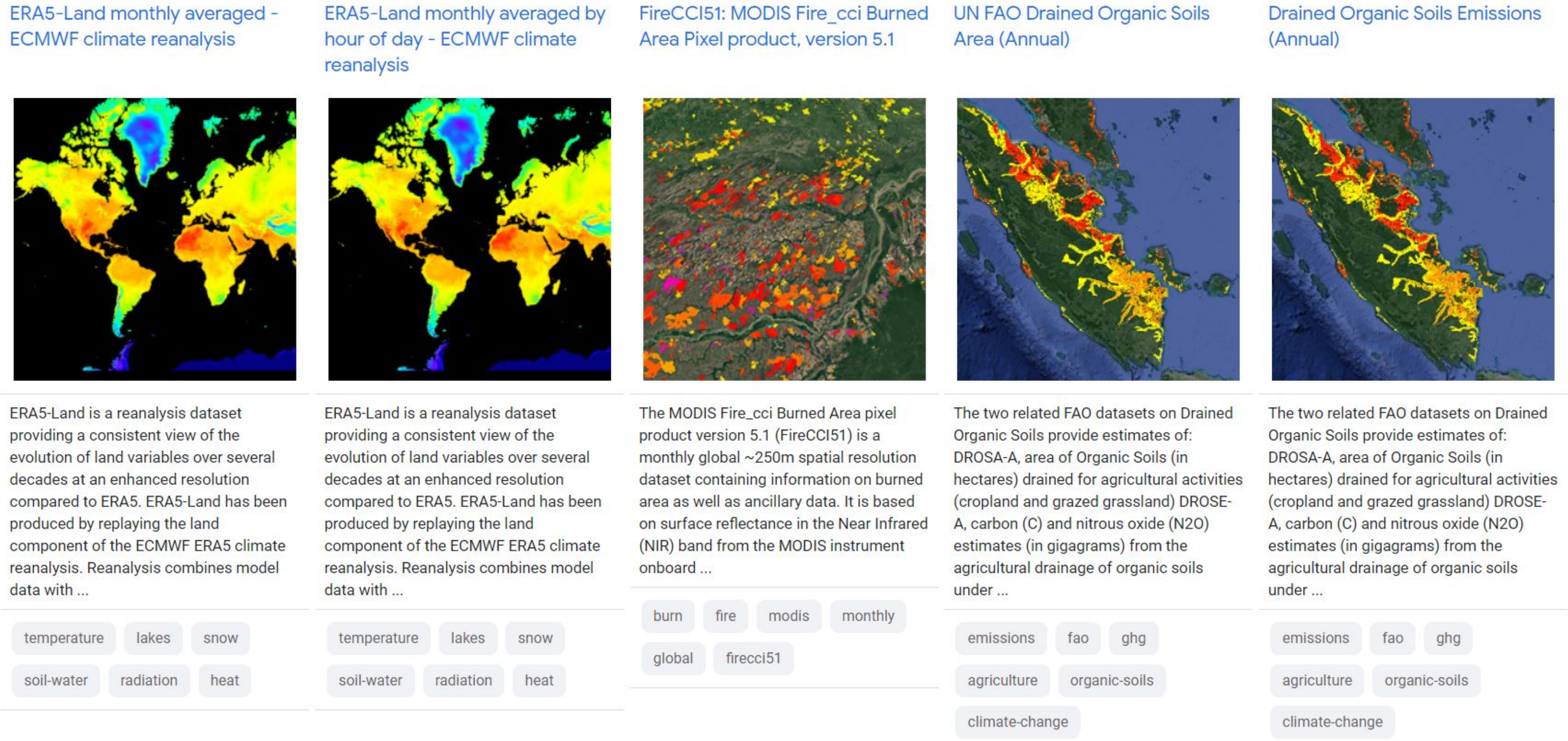
Pandemic: An epidemic that’s spread over multiple countries or continents

Infectious Diseases Repo (EcoHealthAlliance)



Event Name ▲	Disease	Type of Emergence	Start Date	Driver	Species
Acinetobacter baumannii Gentamycin Resistance (Australia, 1993)	Infection due to acinetobacter	New or Increasing Drug Resistance	1993-01-01	Human Susceptibility to Infection, Antimicrobial Agent Use	*Acinetobacter baumannii*
Acinetobacter baumannii Imipenem Resistance (United Kingdom, 1985)	Infection due to acinetobacter	New or Increasing Drug Resistance	1985	Antimicrobial Agent Use	*Acinetobacter baumannii*
Acinetobacter baumannii MDR (Taiwan, 1998)	Bacteremia	New or Increasing Drug Resistance	1998-05	Antimicrobial Agent Use	*Acinetobacter baumannii*
Acinetobacter baumannii Polymyxin Resistance (New York, USA, 2001)	Infection due to acinetobacter	New or Increasing Drug Resistance		Antimicrobial Agent Use	*Acinetobacter baumannii*
Actinomucor elegans (Argentina, 2001)	Maxillary sinusitis	Earliest instance of natural human infection by this microorganism		Human Susceptibility to Infection	*Actinomucor elegans*
Alkhurma virus (Saudi Arabia, 1995)	Alkhurma hemorrhagic fever	Earliest instance of natural human infection by this microorganism	1995-11	Unknown	Alkhurma virus

Google Earth Engine Data



ERA5-Land monthly averaged - ECMWF climate reanalysis

ERA5-Land is a reanalysis dataset providing a consistent view of the evolution of land variables over several decades at an enhanced resolution compared to ERA5. ERA5-Land has been produced by replaying the land component of the ECMWF ERA5 climate reanalysis. Reanalysis combines model data with ...

temperature · lakes · snow · soil-water · radiation · heat

ERA5-Land monthly averaged by hour of day - ECMWF climate reanalysis

ERA5-Land is a reanalysis dataset providing a consistent view of the evolution of land variables over several decades at an enhanced resolution compared to ERA5. ERA5-Land has been produced by replaying the land component of the ECMWF ERA5 climate reanalysis. Reanalysis combines model data with ...

temperature · lakes · snow · soil-water · radiation · heat

FireCCI51: MODIS Fire_cci Burned Area Pixel product, version 5.1

The MODIS Fire_cci Burned Area pixel product version 5.1 (FireCCI51) is a monthly global ~250m spatial resolution dataset containing information on burned area as well as ancillary data. It is based on surface reflectance in the Near Infrared (NIR) band from the MODIS instrument onboard ...

burn · fire · modis · monthly · global · firecci51

UN FAO Drained Organic Soils Area (Annual)

The two related FAO datasets on Drained Organic Soils provide estimates of: DROSA-A, area of Organic Soils (in hectares) drained for agricultural activities (cropland and grazed grassland) DROSE-A, carbon (C) and nitrous oxide (N2O) estimates (in gigagrams) from the agricultural drainage of organic soils under ...

emissions · fao · ghg · agriculture · organic-soils · climate-change

Drained Organic Soils Emissions (Annual)

The two related FAO datasets on Drained Organic Soils provide estimates of: DROSA-A, area of Organic Soils (in hectares) drained for agricultural activities (cropland and grazed grassland) DROSE-A, carbon (C) and nitrous oxide (N2O) estimates (in gigagrams) from the agricultural drainage of organic soils under ...

emissions · fao · ghg · agriculture · organic-soils · climate-change

Independent Variables from Google Earth Engine

- Forest Loss Data (University of Maryland):
  - Results from time-series analysis of Landsat images in characterizing global forest extent and change from 2000 through 2019
  - * Research published from University of Notre Dame found that forest changes were associated with 25% of all diseases and nearly 50% of zoonotic diseases that emerged in humans since 1940
- Annual Mean Temperature (University of California at Berkeley)
  - Interpolated temperatures using data from 1950 to 2000
- 2010 Human Population Density Data (NASA)
  - Census Data
- Global Cropland Data (30m Project)
  - Includes if the land is cropland and if it is largely irrigated by a farmer or if it is rainified
  - Gathered through remote sensing techniques
- Location of surface water (European Commission and Google)
  - Used three million Landsat satellite images to quantify extent of and changes in global surface water over the past 32 years

Challenges

1. Recent integration (in 2019) for Google Earth Engine Data and Python (main users and documentation are Javascript so it required translating resources between two languages)
2. A lot of data (pixel data)
3. Data where time and space are important variables
    a. Epidemics occur in different regions at different times
    b. Variables such as forest loss, temperature and population have time periods where it might be more or less relevant (ex: Temperature changes throughout the year)

Assumptions

Assuming accurate data

a. Google Earth Engine Data
   i. Some of this data utilizes machine learning data processing techniques
b. Knowledge surrounding Epidemic Origin Data
   i. COVID-19 origin story is currently being analyzed by researchers and is not definitive
   ii. Do not have exact locations of where epidemics started

Collecting and Analyzing Data

- Filtered epidemic data for past 20 years
- Using general location of origins, created 60 mile and 180 mile circular polygons surrounding origin
- Utilized Google Earth Engine’s Reducer function to obtain statistics from maps within the specified regions of interest (epidemic and non-epidemic origins)
  - Pixel Data

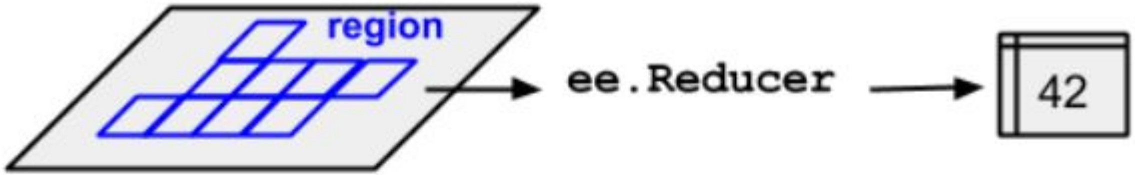
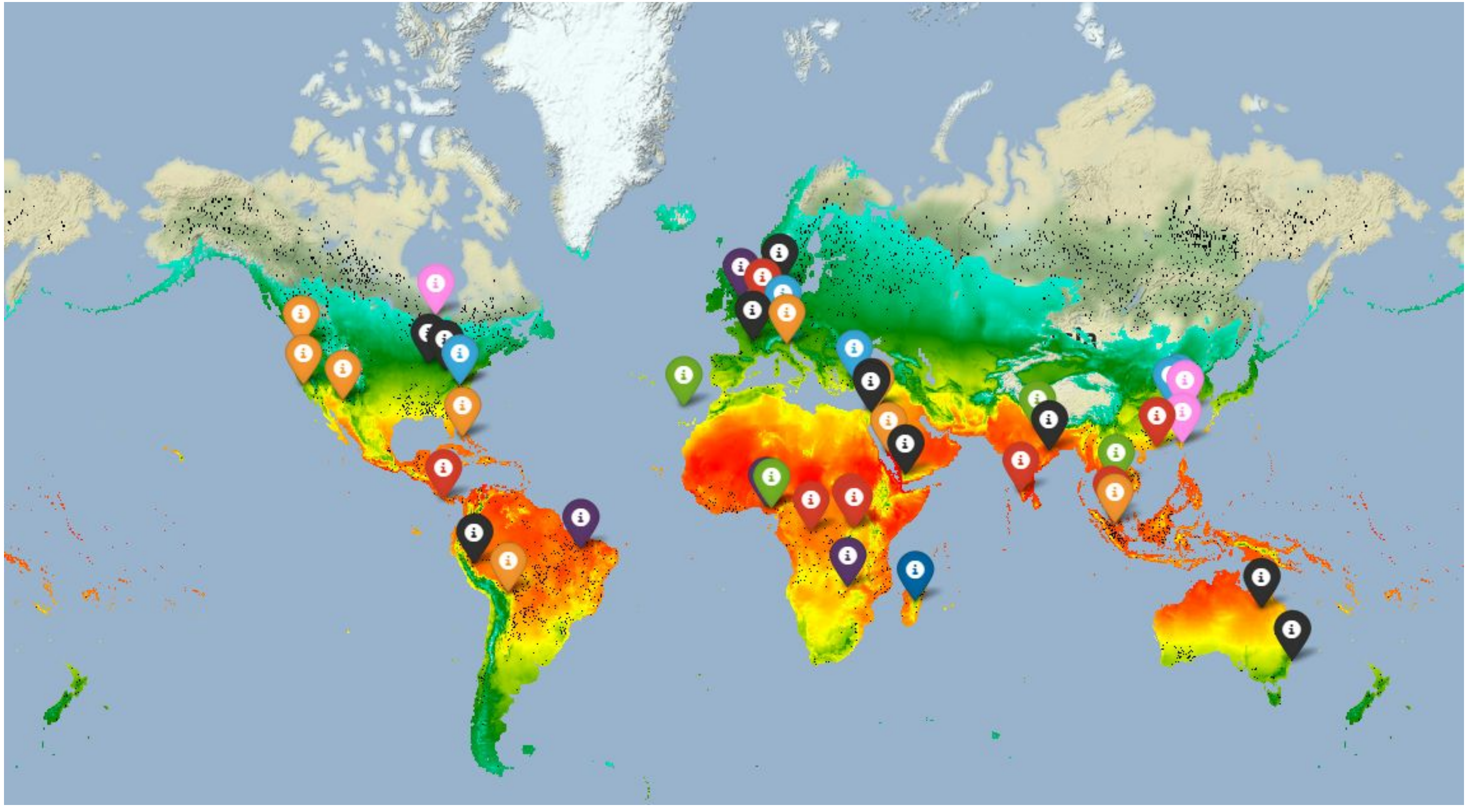
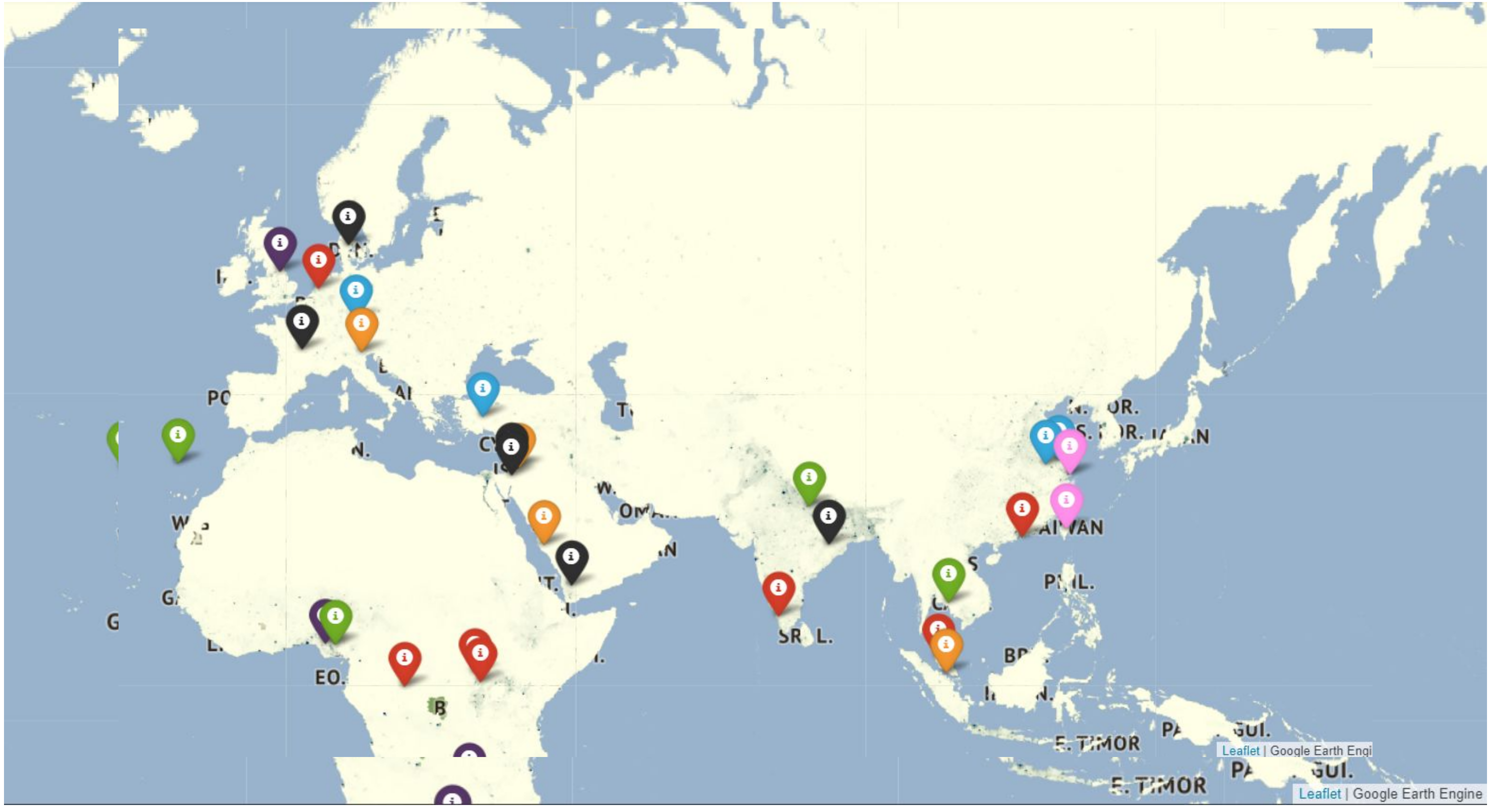


Figure 1. An illustration of an `ee.Reducer` applied to an image and a region.

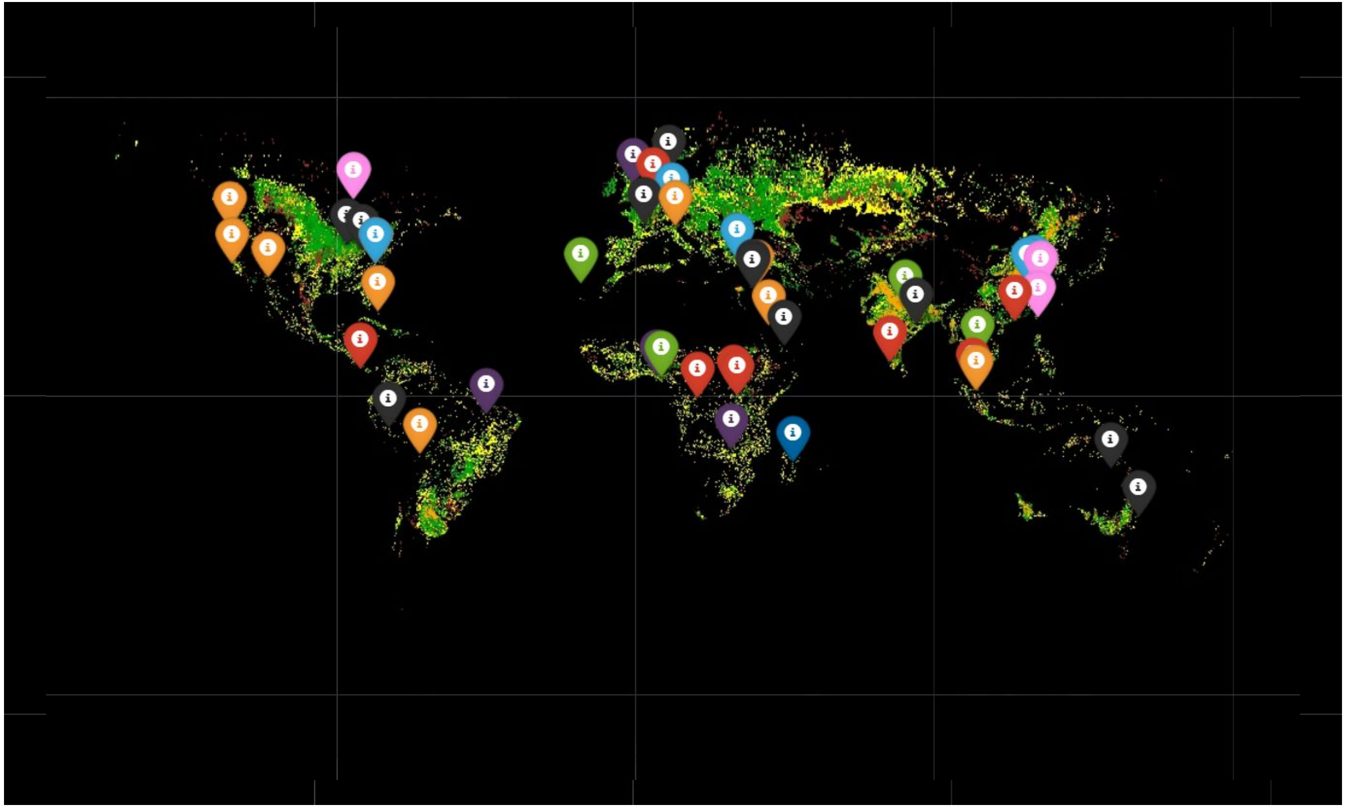
Forest Loss, Temperature, and Epidemic Origins



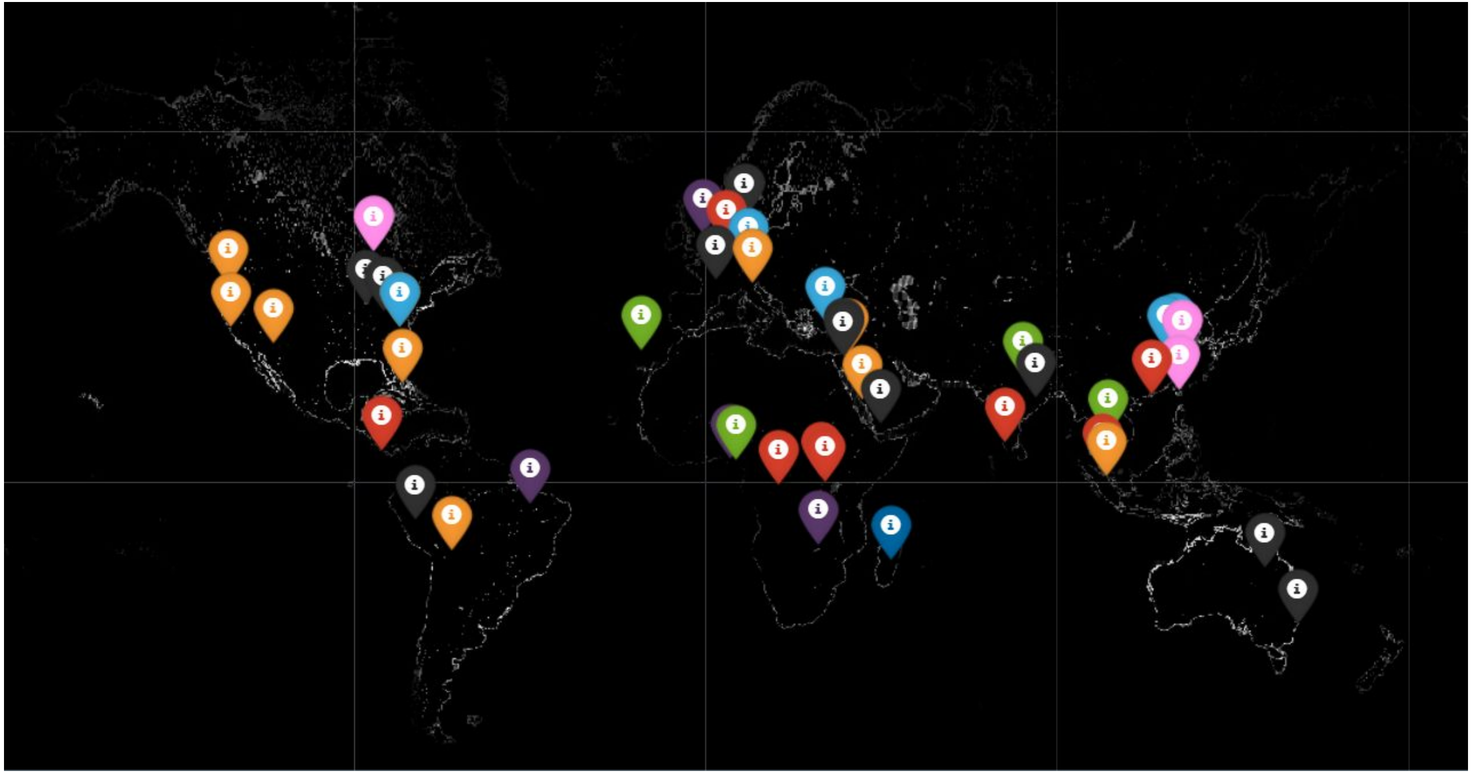
Population and Epidemic Origins



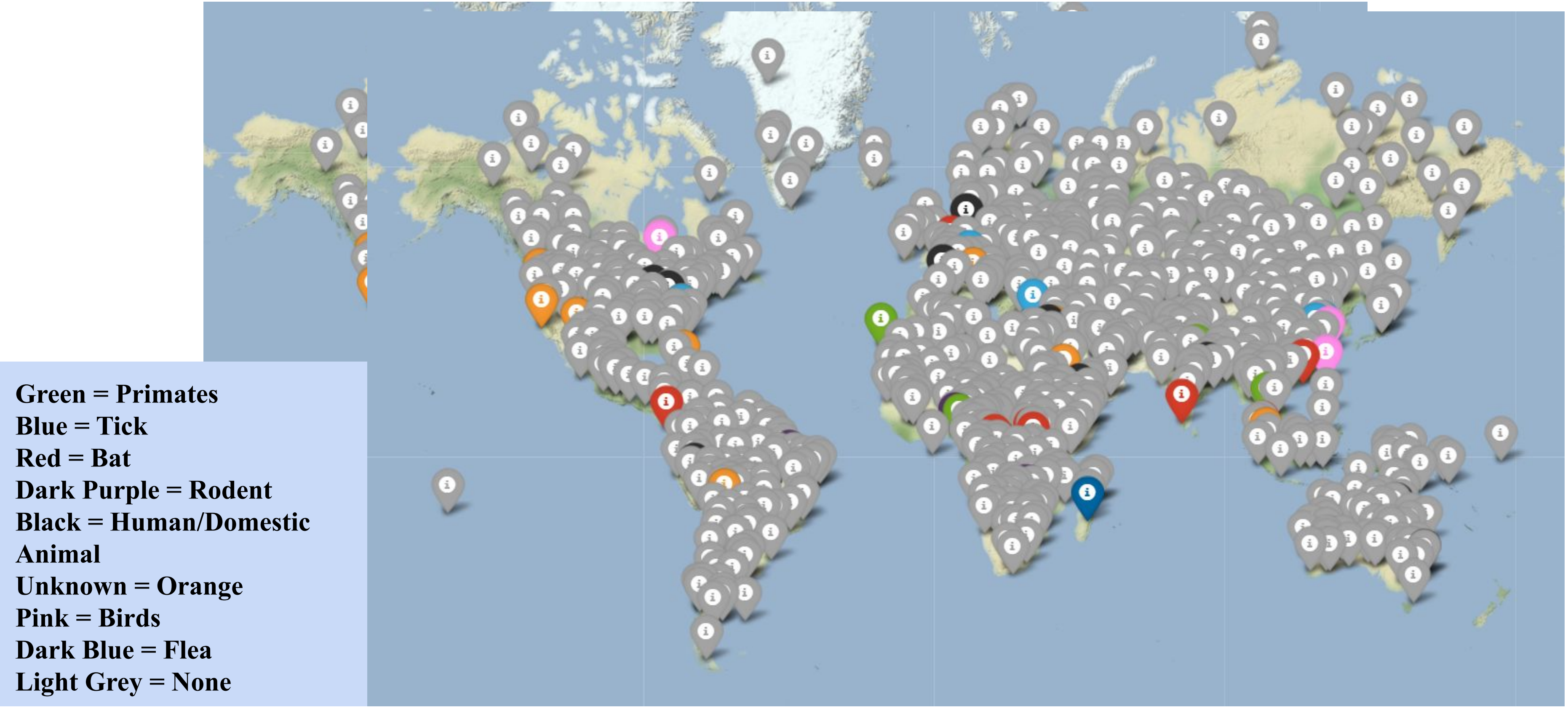
Cropcover and Epidemic Origins



Surface Water and Epidemic Origins



Simulated land-based latitude/longitude coordinates



Green = Primates
Blue = Tick
Red = Bat
Dark Purple = Rodent
Black = Human/Domestic Animal
Unknown = Orange
Pink = Birds
Dark Blue = Flea
Light Grey = None

ANOVA and Tukey Test

Performed ANOVA tests on each variable and used a post-hoc Tukey test to explore differences between the groups in relation to animal origin

1. Forest Loss
    a. 60 mile radius overall ANOVA p-value = .961
    b. 180 mile radius overall ANOVA p-value = .381
2. Climate
    a. 60 mile radius overall ANOVA p-value = .004
        i. Significant Difference Between Origins:
            1. Human and Rodent, p-value = .039 (Average temp for Humans = 33F, Rodent 52F)
    b. 180 mile radius overall ANOVA p-value = .157

ANOVA and Tukey Test (cont.)

3. Population Density
    a. 60 mile radius overall ANOVA p-value = 2.3464695177999518e-08
        i. Significant Difference Between Origins:
            1. Bat and Unknown, p-value = .0333
            2. None and Unknown, p-value = .001
            3. Rodent and Unknown, p-value = .03
                a. Unknown had higher population density
            4. None and Primates, p-value = .0048
                a. Primates had higher population density
    b. 180 mile radius overall ANOVA p-value = .87
4. Cropland
    a. 60 mile radius overall ANOVA p-value = .94
    b. 180 mile radius overall ANOVA p-value = .298

ANOVA and Tukey Test (cont.)

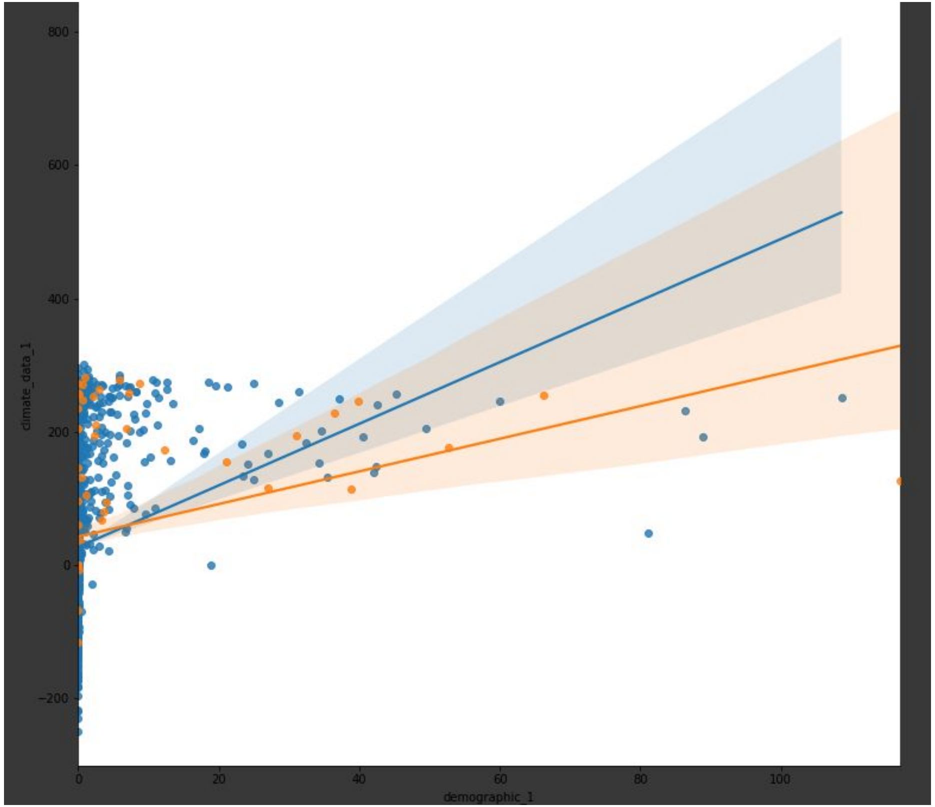
5. Surface Water
    a. 60 mile radius overall p-value = .009
        i. Significant Difference Between Origins:
            1. None and Unknown, p-value = .008
                a. More surface water in regions with Unknown causes
    b. 180 mile radius overall p-value = .002
        i. Significant Difference Between Origins:
            1. Birds and Bat, p-value = 0.0043
            2. Birds and Domestic Animal, p-value = 0.011
            3. Birds and None, p-value = 0.001
            4. Birds and Primates, p-value = 0.0131
            5. Birds and Rodent, p-value = 0.0048
            6. Birds and Tick, p-value = 0.0016
            7. Birds and Unknown, p-value = 0.0106
                a. More water surface area in regions with Bat epidemic emergence

Binary Linear Regression Models

Predicting epidemic emergence using

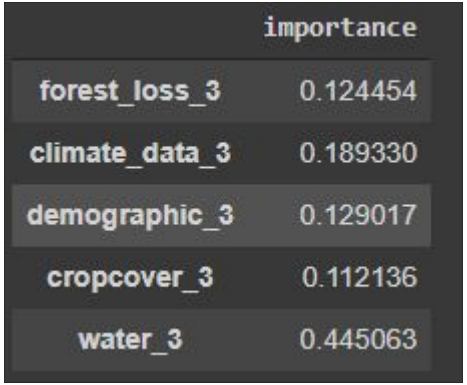
- 60 mile buffer data
- Climate data
- Population data

`92.42% Accuracy`

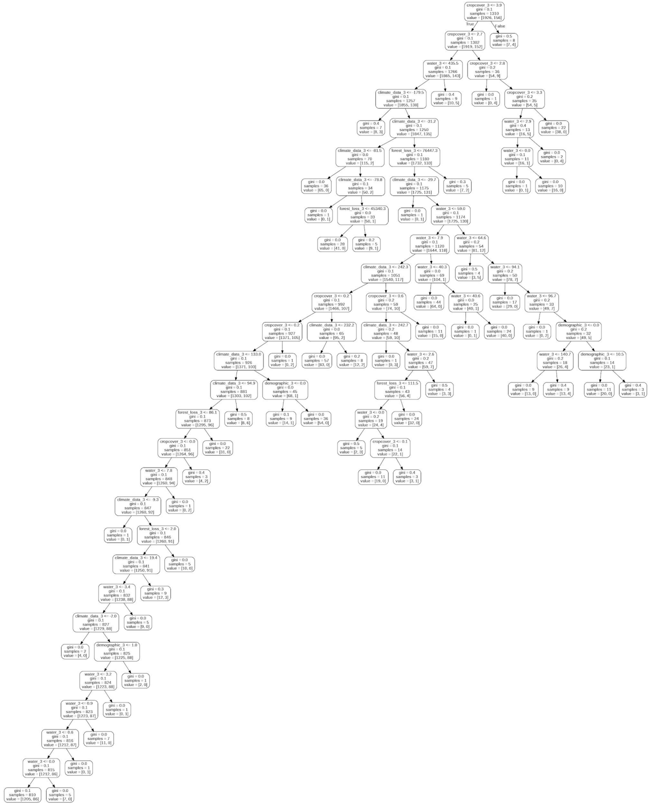


Random Forest Classifier

Incorporated all Global Variables within 60 mile buffer which resulted in 93.98% Accuracy



	importance
forest_loss_3	0.124454
climate_data_3	0.189330
demographic_3	0.129017
cropcover_3	0.112136
water_3	0.445063


cropcover_3 <= 3.9 gini = 0.1 samples = 1310 value = [1926, 156]
True
False
cropcover_3 <= 2.7 gini = 0.1 samples = 1302 value = [1919, 152]
gini = 0.5 samples = 8 value = [7, 4]
water_3 <= 435.5 gini = 0.1 samples = 1266 value = [1865, 143]
cropcover_3 <= 2.8 gini = 0.2 samples = 36 value = [54, 9]
climate_data_3 <= -179.5 gini = 0.1 samples = 1257 value = [1855, 138]
gini = 0.4 samples = 9 value = [10, 5]
gini = 0.0 samples = 1 value = [0, 4]
cropcover_3 <= 3.3 gini = 0.2 samples = 35 value = [54, 5]
gini = 0.4 samples = 7 value = [8, 3]
climate_data_3 <= -31.2 gini = 0.1 samples = 1250 value = [1847, 135]
water_3 <= 2.9 gini = 0.4 samples = 13 value = [16, 5]
gini = 0.0 samples = 22 value = [38, 0]
climate_data_3 <= -81.5 gini = 0.0 samples = 70 value = [115, 2]
forest_loss_3 <= 76447.3 gini = 0.1 samples = 1180 value = [1732, 133]
water_3 <= 0.0 gini = 0.1 samples = 11 value = [16, 1]
gini = 0.0 samples = 2 value = [0, 4]
gini = 0.0 samples = 36 value = [65, 0]
climate_data_3 <= -78.8 gini = 0.1 samples = 34 value = [50, 2]
climate_data_3 <= -29.7 gini = 0.1 samples = 1175 value = [1725, 131]
gini = 0.3 samples = 5 value = [7, 2]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 10 value = [16, 0]
gini = 0.0 samples = 1 value = [0, 1]
forest_loss_3 <= 45340.3 gini = 0.0 samples = 33 value = [50, 1]
gini = 0.0 samples = 1 value = [0, 1]
water_3 <= 59.0 gini = 0.1 samples = 1174 value = [1725, 130]
gini = 0.0 samples = 28 value = [41, 0]
gini = 0.2 samples = 5 value = [9, 1]
water_3 <= 7.9 gini = 0.1 samples = 1120 value = [1644, 118]
water_3 <= 64.6 gini = 0.2 samples = 54 value = [81, 12]
climate_data_3 <= 242.3 gini = 0.1 samples = 1051 value = [1540, 117]
water_3 <= 40.3 gini = 0.0 samples = 69 value = [104, 1]
gini = 0.5 samples = 4 value = [3, 5]
water_3 <= 94.1 gini = 0.2 samples = 50 value = [78, 7]
cropcover_3 <= 0.2 gini = 0.1 samples = 992 value = [1466, 107]
cropcover_3 <= 0.6 gini = 0.2 samples = 59 value = [74, 10]
gini = 0.0 samples = 44 value = [64, 0]
water_3 <= 40.6 gini = 0.0 samples = 25 value = [40, 1]
gini = 0.0 samples = 17 value = [29, 0]
water_3 <= 96.7 gini = 0.2 samples = 33 value = [49, 7]
cropcover_3 <= 0.2 gini = 0.1 samples = 927 value = [1371, 105]
climate_data_3 <= 232.2 gini = 0.0 samples = 65 value = [95, 2]
climate_data_3 <= 242.7 gini = 0.2 samples = 48 value = [59, 10]
gini = 0.0 samples = 11 value = [15, 0]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.0 samples = 24 value = [40, 0]
gini = 0.0 samples = 1 value = [0, 2]
demographic_3 <= 0.0 gini = 0.2 samples = 32 value = [49, 5]
climate_data_3 <= 133.0 gini = 0.1 samples = 926 value = [1371, 103]
gini = 0.0 samples = 1 value = [0, 2]
gini = 0.0 samples = 57 value = [83, 0]
gini = 0.2 samples = 8 value = [12, 2]
gini = 0.0 samples = 1 value = [0, 3]
water_3 <= 2.6 gini = 0.2 samples = 47 value = [59, 7]
water_3 <= 140.7 gini = 0.2 samples = 18 value = [26, 4]
demographic_3 <= 10.5 gini = 0.1 samples = 14 value = [23, 1]
climate_data_3 <= 94.9 gini = 0.1 samples = 881 value = [1303, 102]
demographic_3 <= 0.0 gini = 0.0 samples = 45 value = [68, 1]
forest_loss_3 <= 111.5 gini = 0.1 samples = 43 value = [56, 4]
gini = 0.5 samples = 4 value = [3, 3]
gini = 0.0 samples = 9 value = [13, 0]
gini = 0.4 samples = 9 value = [13, 4]
gini = 0.0 samples = 11 value = [20, 0]
gini = 0.4 samples = 3 value = [3, 1]
forest_loss_3 <= 86.1 gini = 0.1 samples = 873 value = [1295, 96]
gini = 0.5 samples = 8 value = [8, 6]
gini = 0.1 samples = 9 value = [14, 1]
gini = 0.0 samples = 36 value = [54, 0]
water_3 <= 0.0 gini = 0.2 samples = 19 value = [24, 4]
gini = 0.0 samples = 24 value = [32, 0]
cropcover_3 <= 0.0 gini = 0.1 samples = 851 value = [1264, 96]
gini = 0.0 samples = 22 value = [31, 0]
gini = 0.5 samples = 5 value = [2, 3]
cropcover_3 <= 0.1 gini = 0.1 samples = 14 value = [22, 1]
water_3 <= 7.8 gini = 0.1 samples = 848 value = [1260, 94]
gini = 0.4 samples = 3 value = [4, 2]
gini = 0.0 samples = 11 value = [19, 0]
gini = 0.4 samples = 3 value = [3, 1]
climate_data_3 <= -9.3 gini = 0.1 samples = 847 value = [1260, 92]
gini = 0.0 samples = 1 value = [0, 2]
gini = 0.0 samples = 1 value = [0, 1]
forest_loss_3 <= 2.0 gini = 0.1 samples = 846 value = [1260, 91]
climate_data_3 <= 19.4 gini = 0.1 samples = 841 value = [1250, 91]
gini = 0.0 samples = 5 value = [10, 0]
water_3 <= 3.4 gini = 0.1 samples = 832 value = [1238, 88]
gini = 0.3 samples = 9 value = [12, 3]
climate_data_3 <= -2.0 gini = 0.1 samples = 827 value = [1229, 88]
gini = 0.0 samples = 5 value = [9, 0]
gini = 0.0 samples = 2 value = [4, 0]
demographic_3 <= 1.0 gini = 0.1 samples = 825 value = [1225, 88]
water_3 <= 3.2 gini = 0.1 samples = 824 value = [1223, 88]
gini = 0.0 samples = 1 value = [2, 0]
water_3 <= 0.9 gini = 0.1 samples = 823 value = [1223, 87]
gini = 0.0 samples = 1 value = [0, 1]
water_3 <= 0.6 gini = 0.1 samples = 816 value = [1212, 87]
gini = 0.0 samples = 7 value = [11, 0]
water_3 <= 0.0 gini = 0.1 samples = 815 value = [1212, 86]
gini = 0.0 samples = 1 value = [0, 1]
gini = 0.1 samples = 810 value = [1205, 86]
gini = 0.0 samples = 5 value = [7, 0]

Challenge

The general probability of epidemics occurring on the day to day in any moment in time is incredibly slim

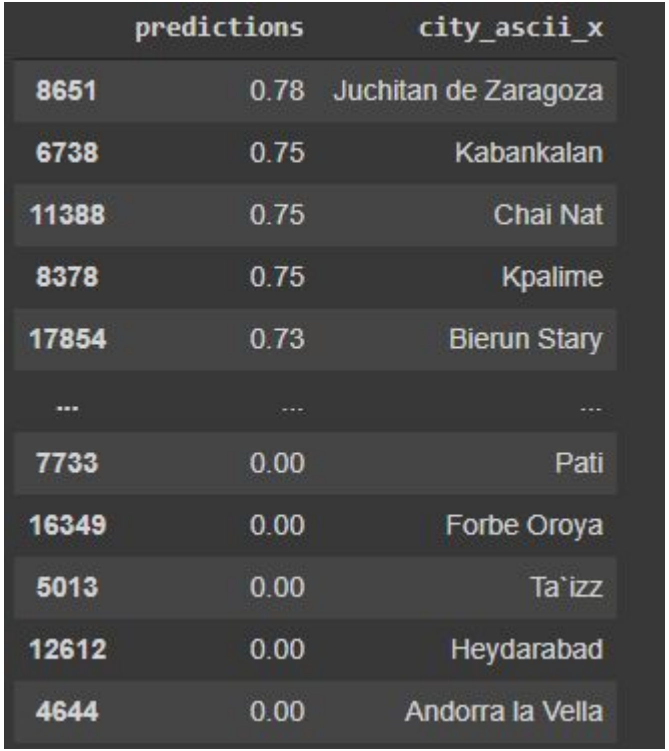
- Input longitude/latitude data from 15,000 cities into the previous models and one of them resulted in being epidemic emergence locations (Deland, Florida)

Random Forest Regressor

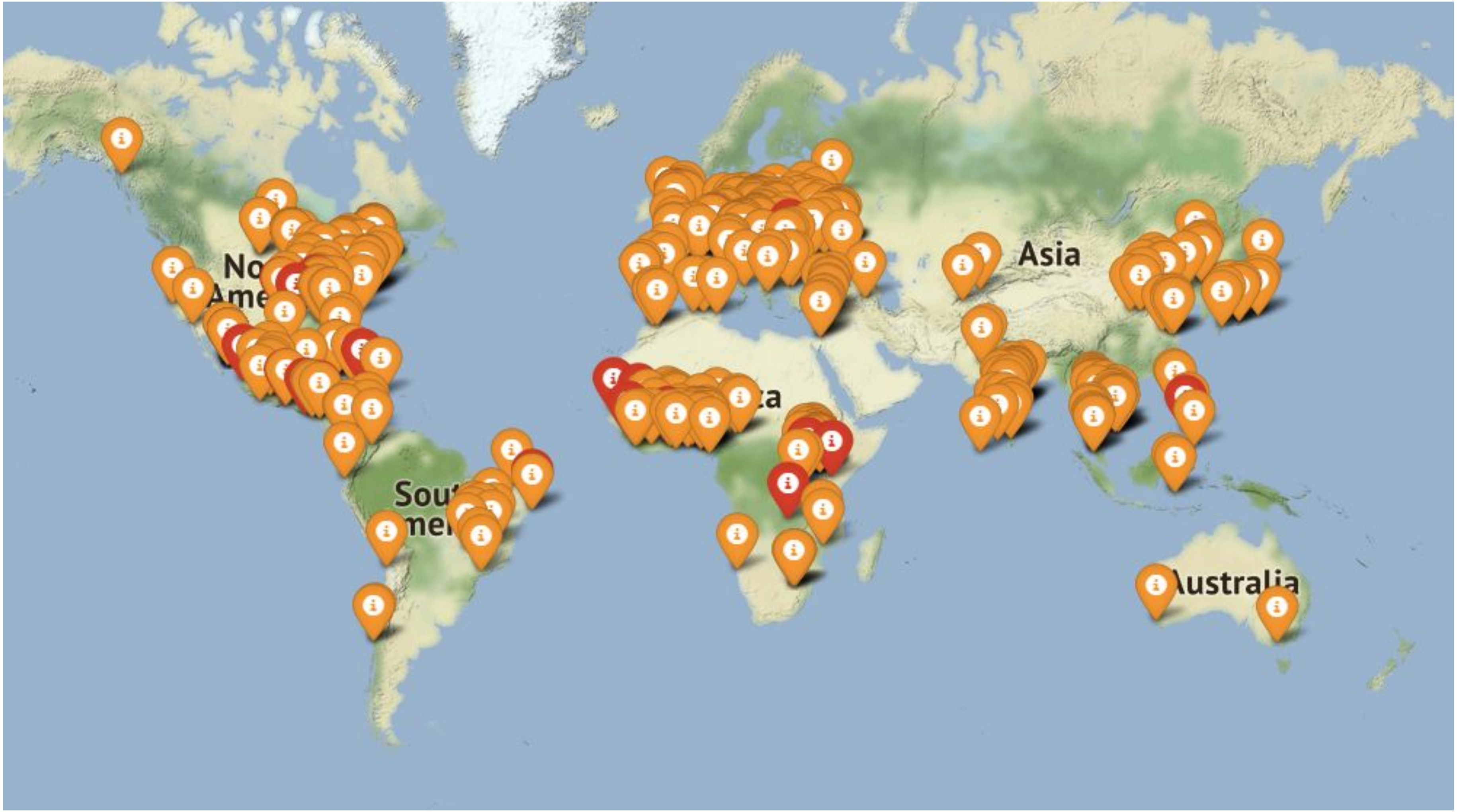
Incorporated all Global Variables within 60 mile buffer with mse = 0.25

- Juncitan = Mexico
- Kabankalan = Philippines
- Chai Nat = Thailand
- Kpalime = Togo
- Bierun Stary = Poland

- Forbe Oroya = Peru
- Ta’izz = Yemen
- Heydarabad = Azerbaijan
- Andorra la Vella = Andorra



	predictions	city_ascii_x
8651	0.78	Juchitan de Zaragoza
6738	0.75	Kabankalan
11388	0.75	Chai Nat
8378	0.75	Kpalime
17854	0.73	Bierun Stary
...	...	...
7733	0.00	Pati
16349	0.00	Forbe Oroya
5013	0.00	Ta`izz
12612	0.00	Heydarabad
4644	0.00	Andorra la Vella


North America
South America
Africa
Asia
Australia

Future Research

- Incorporate element of time within independent and dependent variables
    - Time of year
    - Year of outbreak
    - Population Data
- Incorporating more datasets from Google Earth Engine as independent variables
    - Topography
    - Species Data
- Incorporate climate change projection data to show if rising temperatures will increase/decrease