The Task: Cross Reference Ground Truths

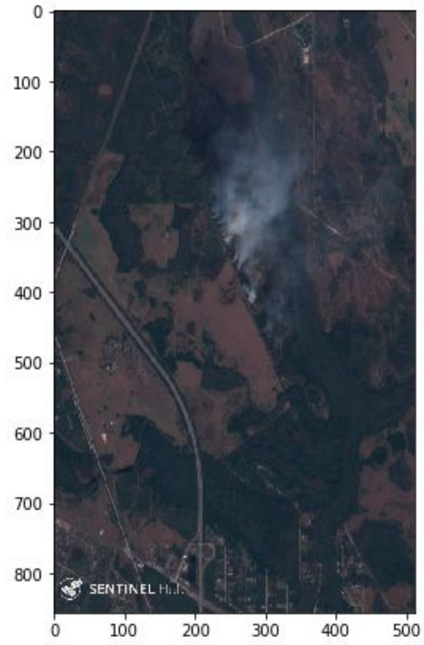
MODIS

Many institutions and papers rely on data gathered by an instrument on NASA’s Terra and Aqua Satellites called MODIS (MOderate Resolution Imaging Spectroradiometer) .

MODIS detects thermal anomalies using combinations of spectral bands and reports them as possible wildfires.

Kaggle

Referred to internally as “Kaggle” (after the site from which we acquired it) this dataset contains 1.88 Million Wildfires reported in the United States by government fire organisations.



MODIS

+ Updated daily
+ 1-2 day resolution
+ Commonly used by academics and institutions
- Only a heuristic

Kaggle

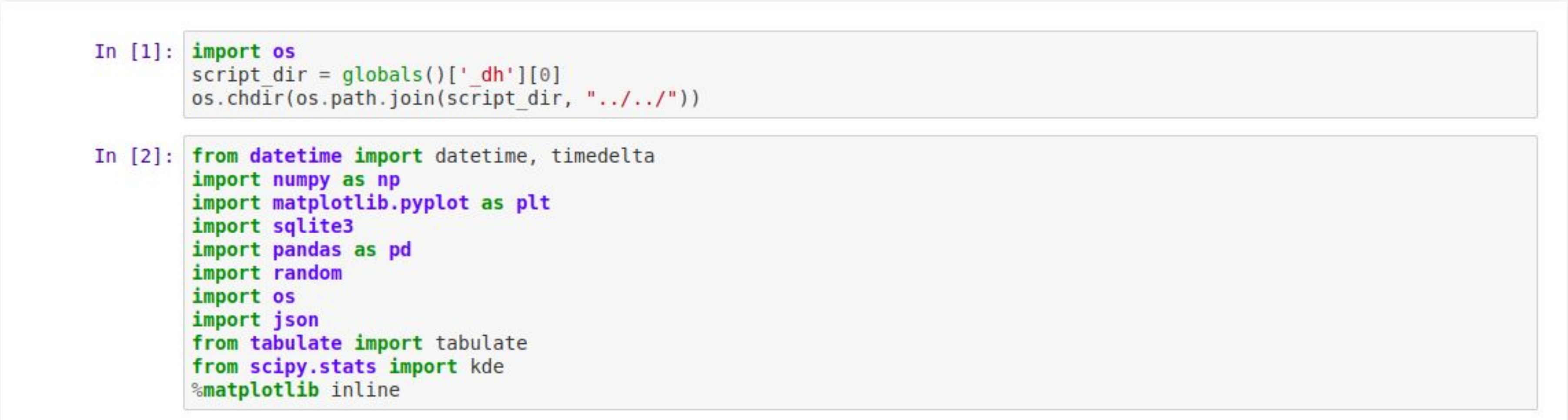
+ Reported by on-the-ground humans
+ More data about each fire
- No data after 2015 (around the time the image satellites were launched)

The solution:
Cross referencing

Cross reference MODIS against Kaggle and, if they agree where they overlap, we might trust MODIS as a ground truth in more recent year and use it to can label our models

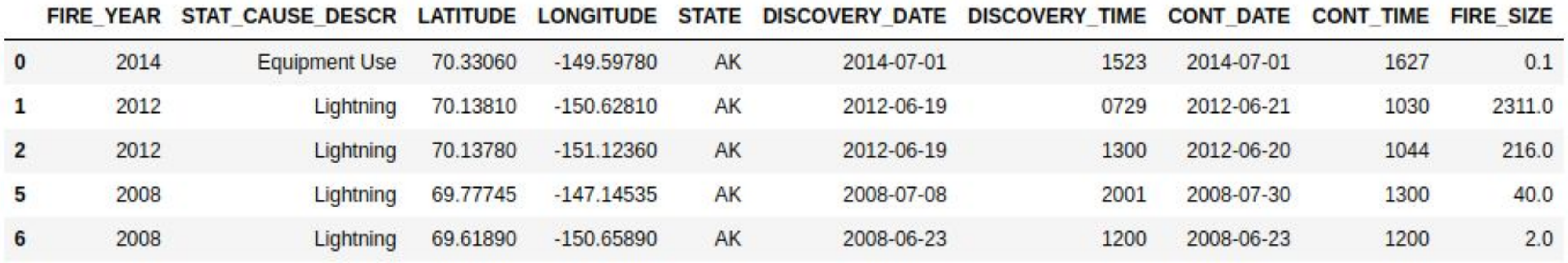
Method

Analyse and visualise the datasets using a jupyter notebook, pandas and matplotlib



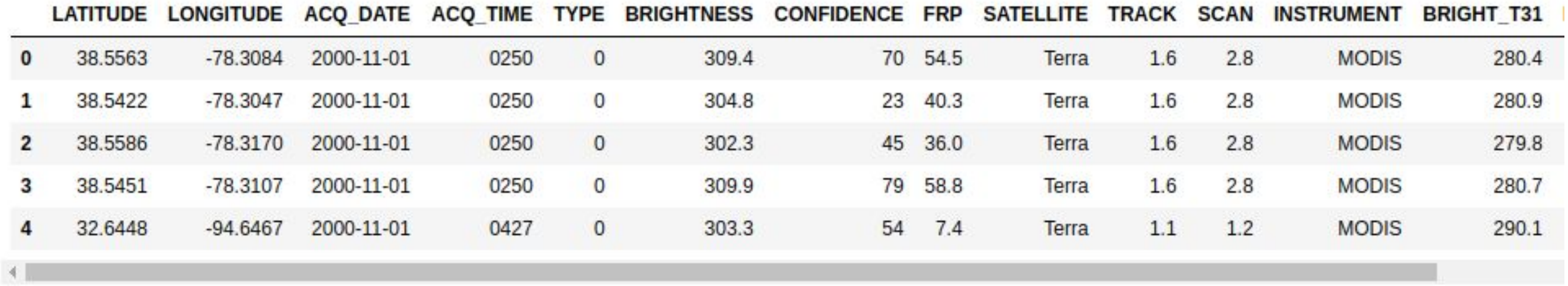
```
In [1]: import os
        script_dir = globals()['_dh'][0]
        os.chdir(os.path.join(script_dir, "../../"))

In [2]: from datetime import datetime, timedelta
        import numpy as np
        import matplotlib.pyplot as plt
        import sqlite3
        import pandas as pd
        import random
        import os
        import json
        from tabulate import tabulate
        from scipy.stats import kde
        %matplotlib inline
```

Kaggle (SQL Database → Pandas Dataframe)



	FIRE_YEAR	STAT_CAUSE_DESCR	LATITUDE	LONGITUDE	STATE	DISCOVERY_DATE	DISCOVERY_TIME	CONT_DATE	CONT_TIME	FIRE_SIZE
0	2014	Equipment Use	70.33060	-149.59780	AK	2014-07-01	1523	2014-07-01	1627	0.1
1	2012	Lightning	70.13810	-150.62810	AK	2012-06-19	0729	2012-06-21	1030	2311.0
2	2012	Lightning	70.13780	-151.12360	AK	2012-06-19	1300	2012-06-20	1044	216.0
5	2008	Lightning	69.77745	-147.14535	AK	2008-07-08	2001	2008-07-30	1300	40.0
6	2008	Lightning	69.61890	-150.65890	AK	2008-06-23	1200	2008-06-23	1200	2.0

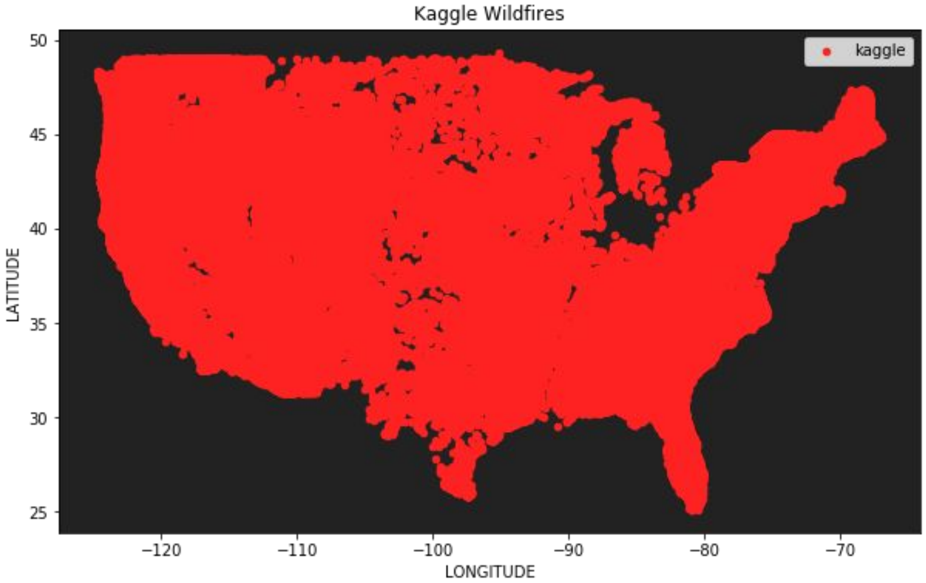
MODIS (JSON → Pandas Dataframe)

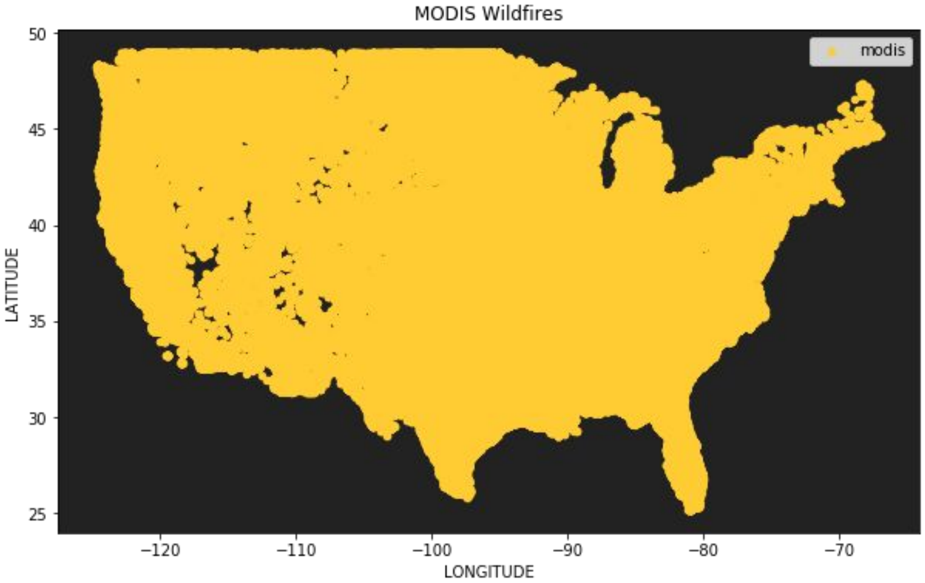


	LATITUDE	LONGITUDE	ACQ_DATE	ACQ_TIME	TYPE	BRIGHTNESS	CONFIDENCE	FRP	SATELLITE	TRACK	SCAN	INSTRUMENT	BRIGHT_T31
0	38.5563	-78.3084	2000-11-01	0250	0	309.4	70	54.5	Terra	1.6	2.8	MODIS	280.4
1	38.5422	-78.3047	2000-11-01	0250	0	304.8	23	40.3	Terra	1.6	2.8	MODIS	280.9
2	38.5586	-78.3170	2000-11-01	0250	0	302.3	45	36.0	Terra	1.6	2.8	MODIS	279.8
3	38.5451	-78.3107	2000-11-01	0250	0	309.9	79	58.8	Terra	1.6	2.8	MODIS	280.7
4	32.6448	-94.6467	2000-11-01	0427	0	303.3	54	7.4	Terra	1.1	1.2	MODIS	290.1

1) Area of Coverage

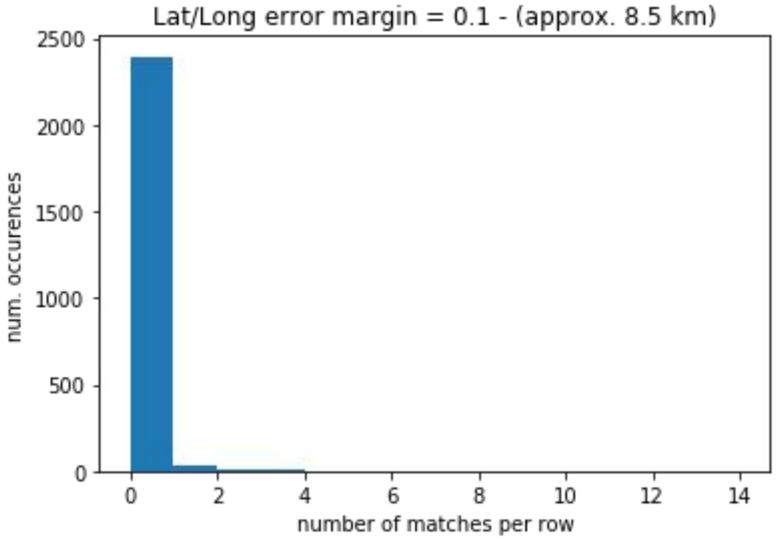
Visualise the datasets by plotting the lat/long coordinates of the data points of each wildfire in the contiguous states.

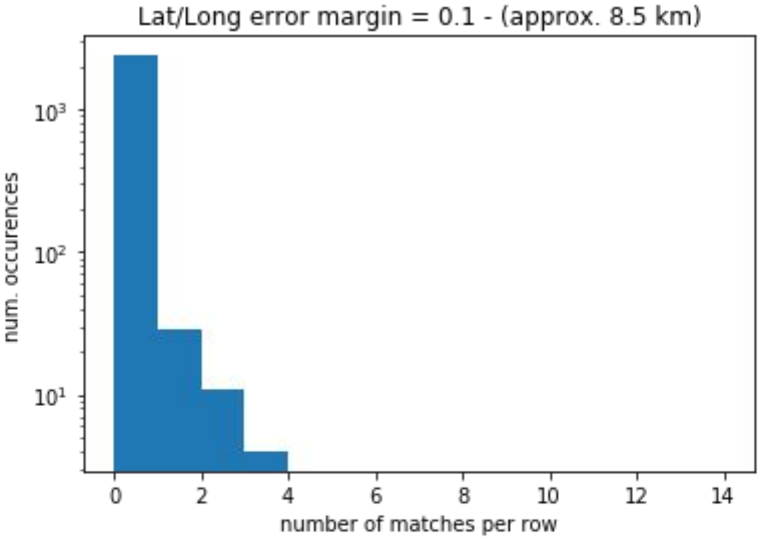




2) Match Data Points

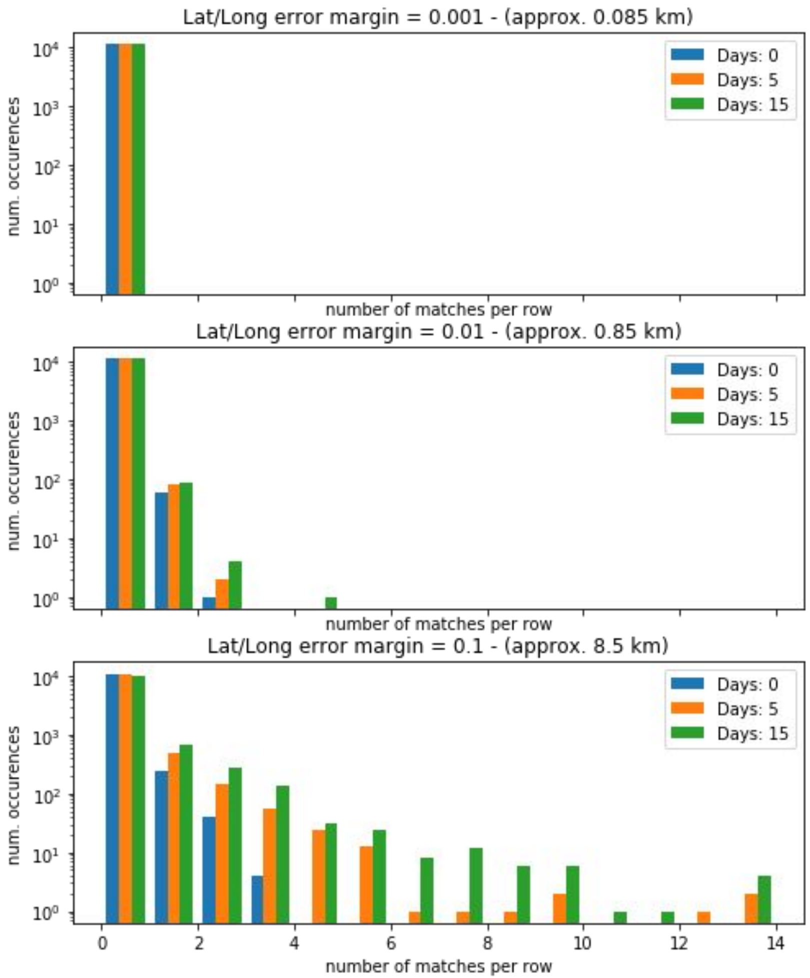
- Investigate the tp/fp/tn/fn rate for each dataset relative to the other
- Match up wildfires identified by kaggle with thermal anomalies detected by MODIS
    - Using a 2 month time-slice and a spatial error margin of 0.01 degrees (approx. 1 km).





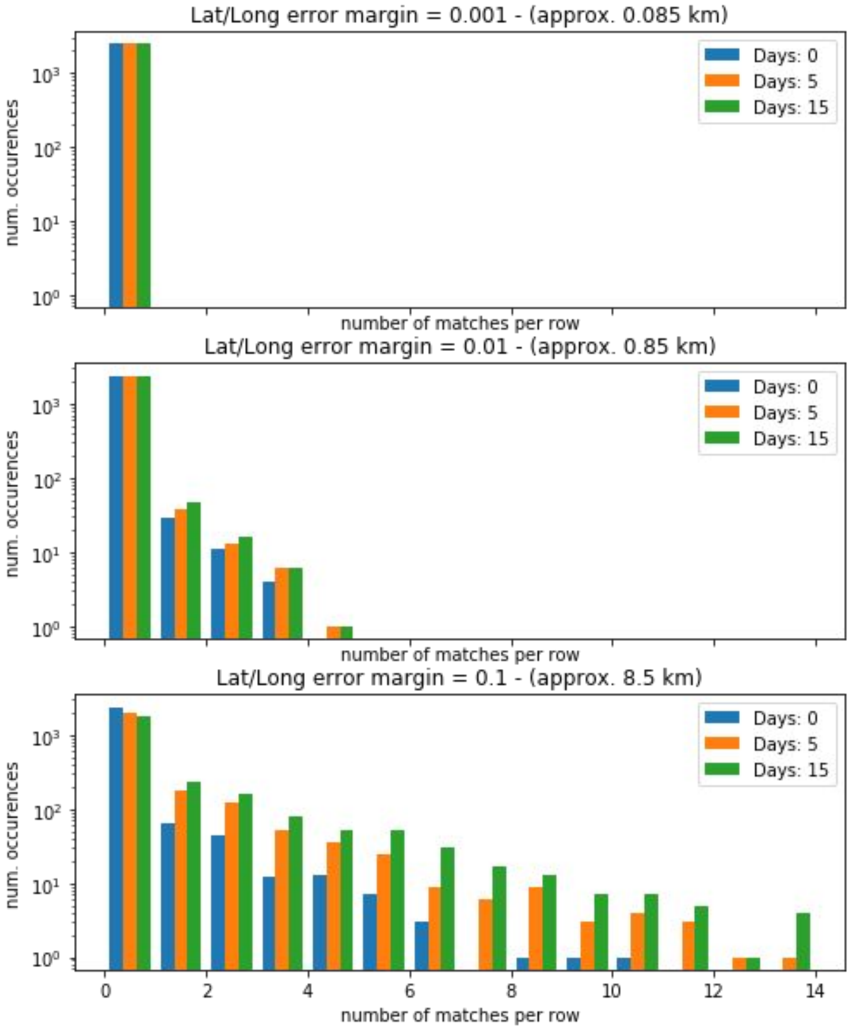
3) Explore Hyperparameter Space

- The Kaggle database only gives us the **discovery** and **containment** dates of fires.
- Fires burn both before they are discovered and after they are contained.
- We should therefore experiment with hyperparameters corresponding to both temporal error (days) and spatial error (degree long/lat).

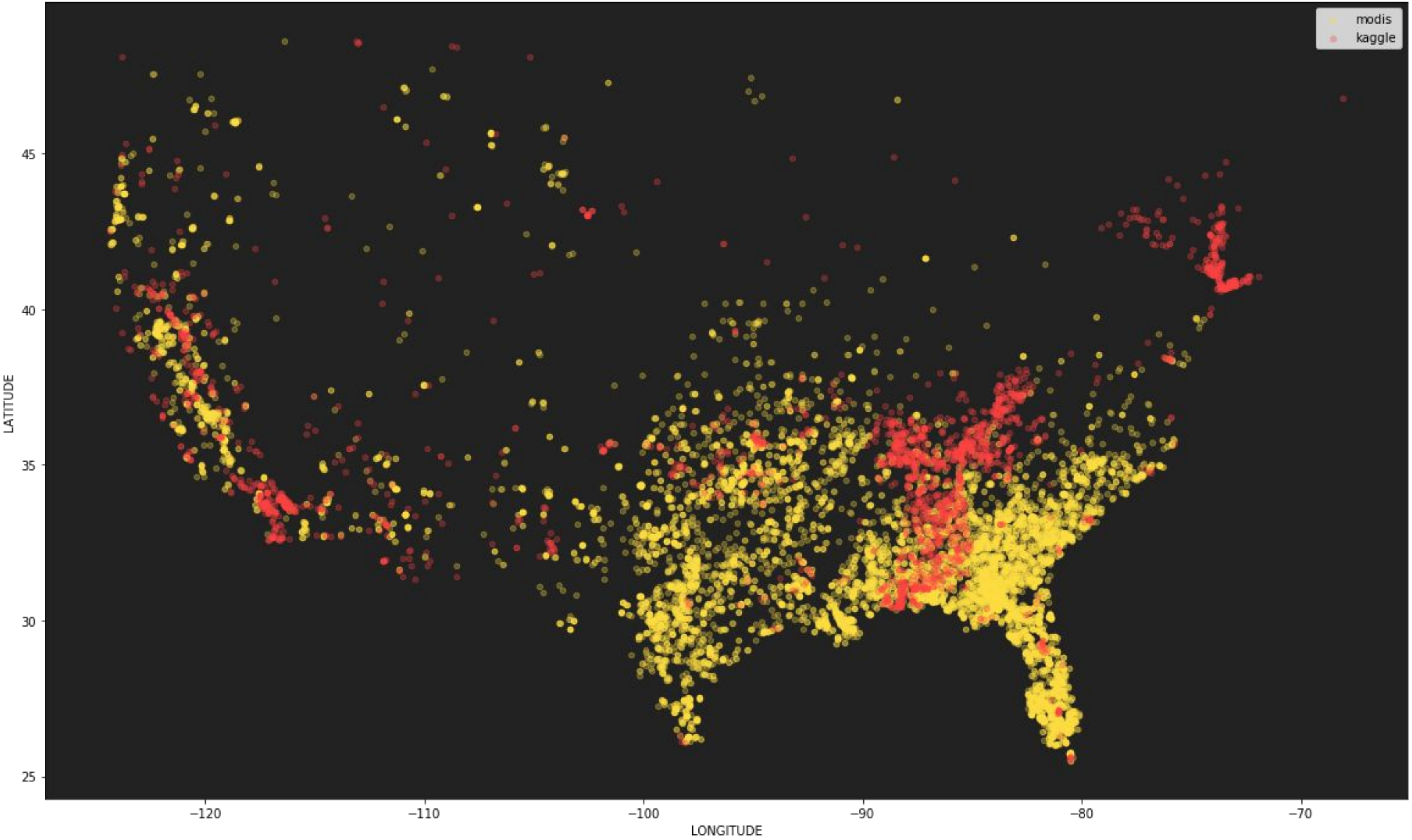


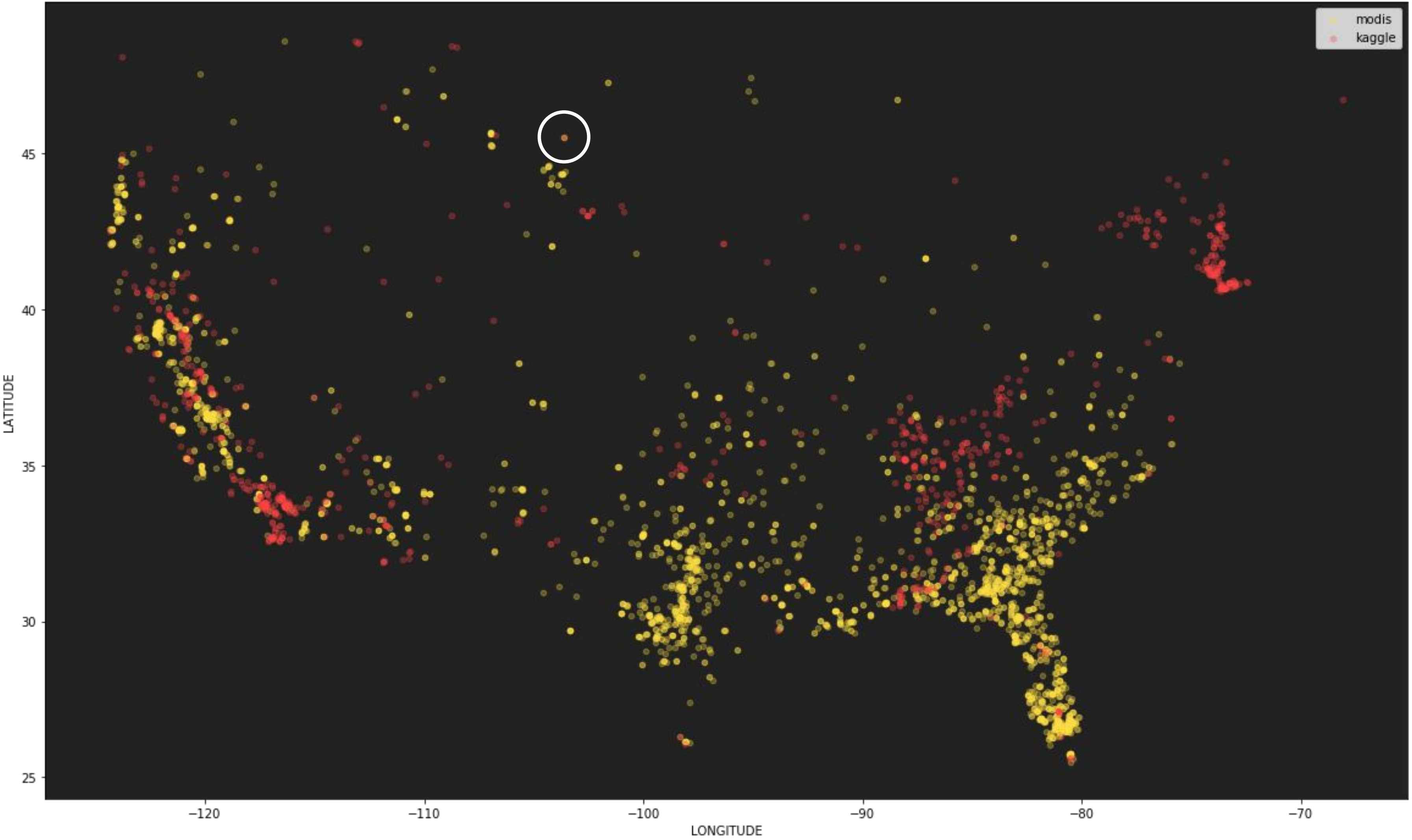
3.5) Matching MODIS data points against kaggle

For each thermal anomaly which MODIS detects, how many kaggle-reported wildfires could it be referring to?

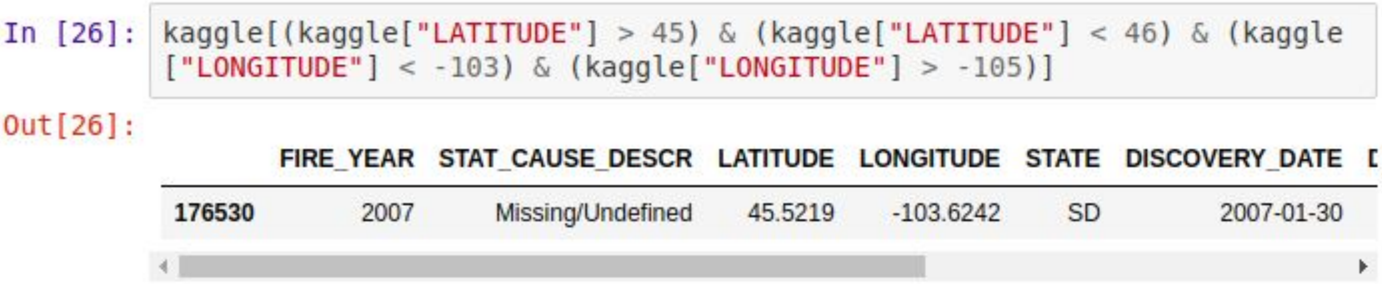


4) Map Sample


modis
kaggle
LATITUDE
45
40
35
30
25
LONGITUDE
−120
−110
−100
−90
−80
−70


modis
kaggle
LATITUDE
45
40
35
30
25
LONGITUDE
−120
−110
−100
−90
−80
−70

Kaggle

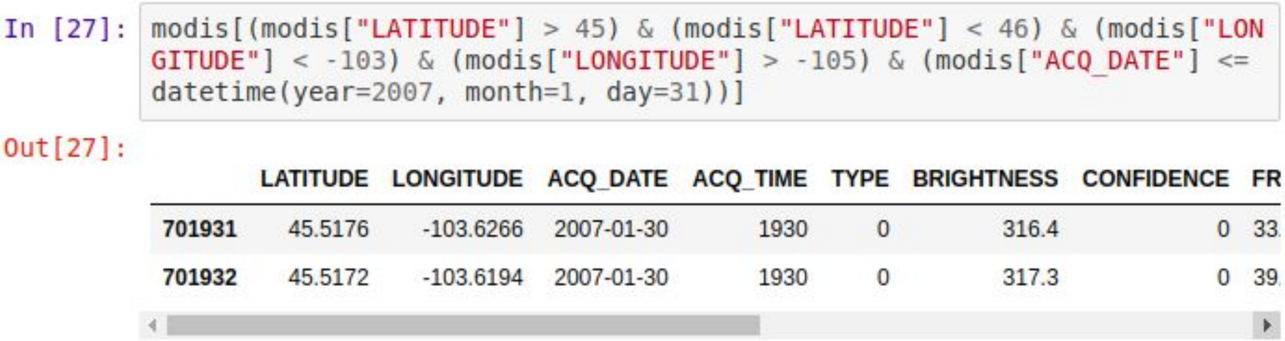


In [26]:
```
kaggle[(kaggle["LATITUDE"] > 45) & (kaggle["LATITUDE"] < 46) & (kaggle
["LONGITUDE"] < -103) & (kaggle["LONGITUDE"] > -105)]
```

Out[26]:

	FIRE_YEAR	STAT_CAUSE_DESCR	LATITUDE	LONGITUDE	STATE	DISCOVERY_DATE	D
176530	2007	Missing/Undefined	45.5219	-103.6242	SD	2007-01-30	

MODIS



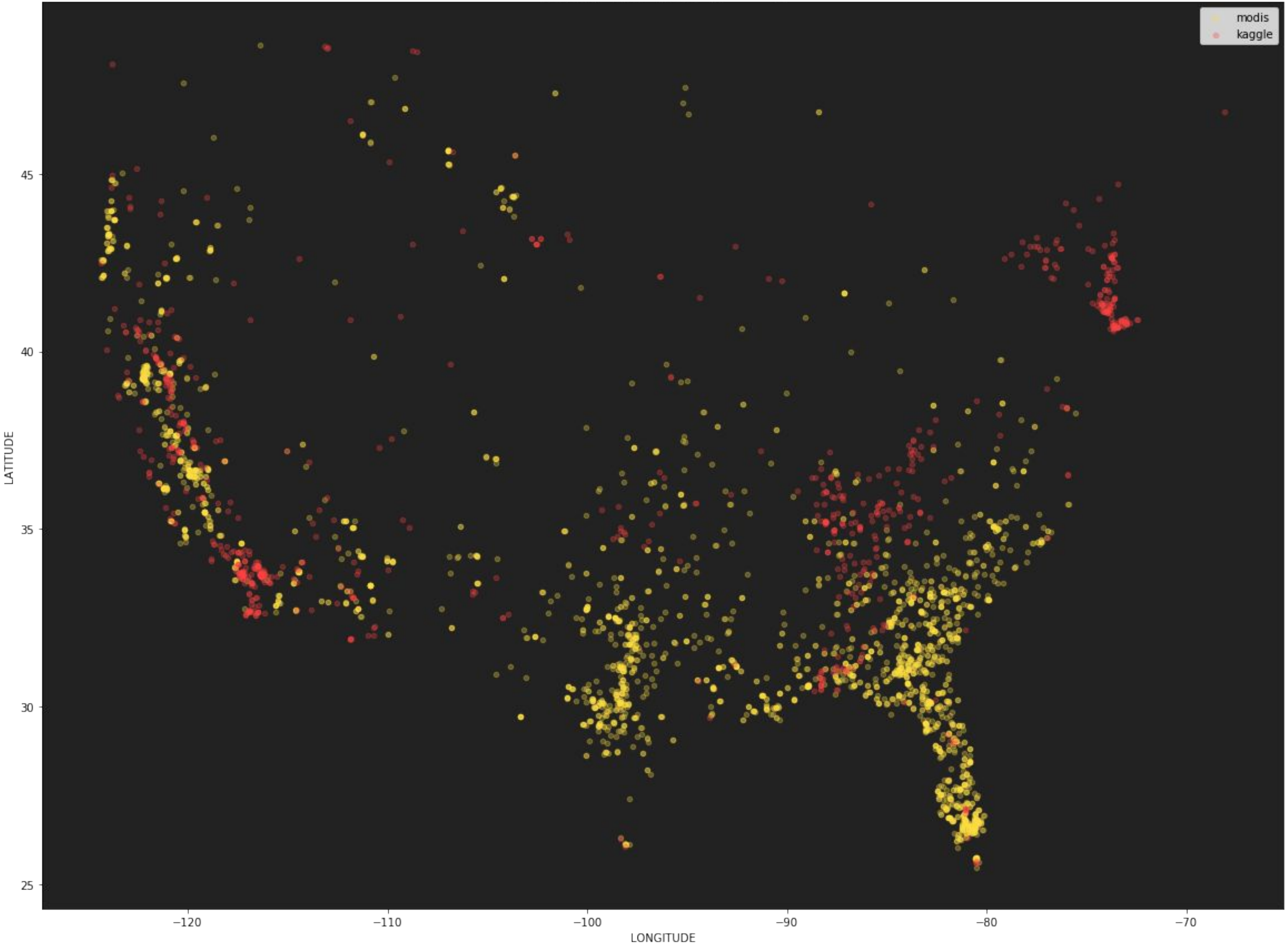
In [27]:
```
modis[(modis["LATITUDE"] > 45) & (modis["LATITUDE"] < 46) & (modis["LON
GITUDE"] < -103) & (modis["LONGITUDE"] > -105) & (modis["ACQ_DATE"] <=
datetime(year=2007, month=1, day=31))]
```

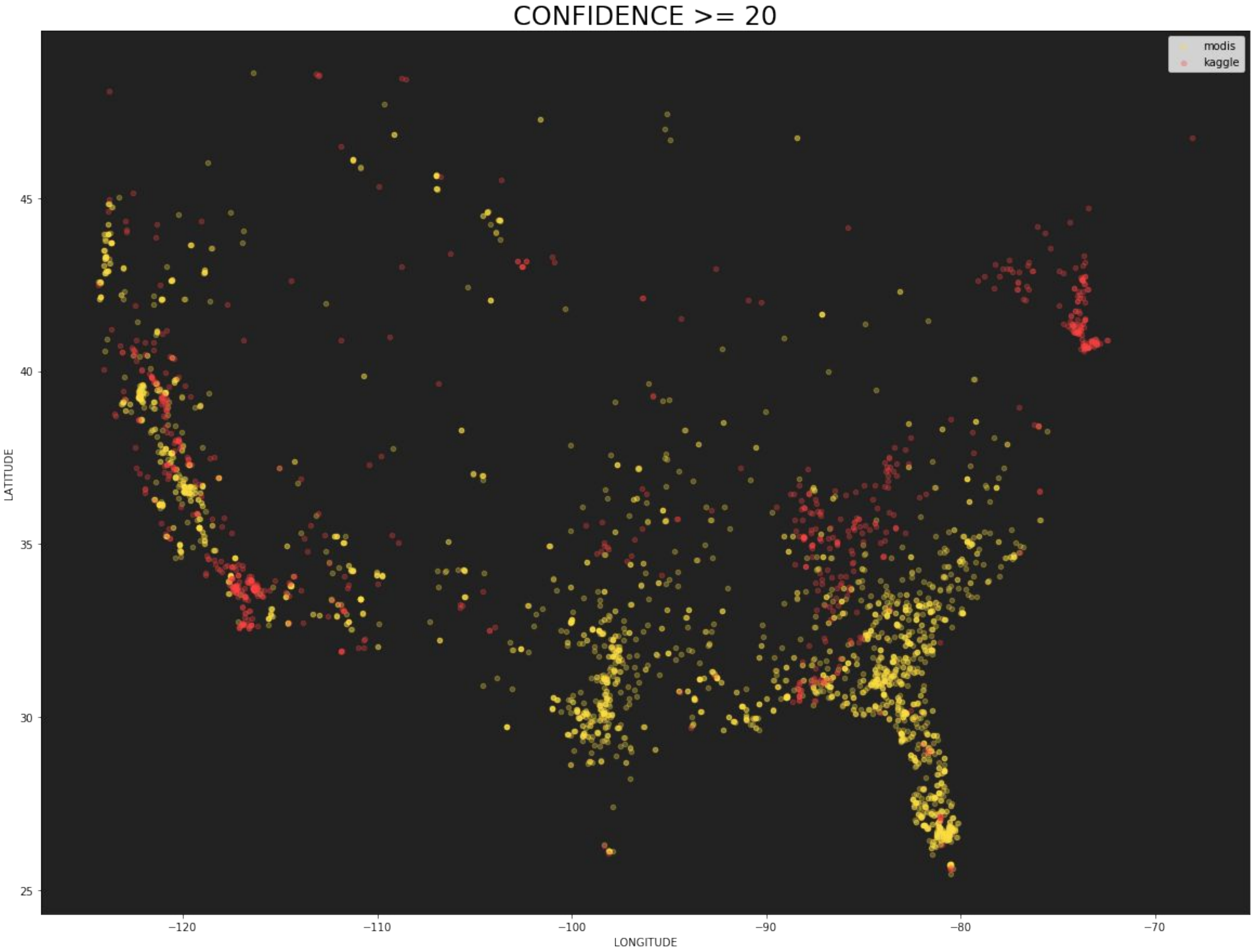
Out[27]:

	LATITUDE	LONGITUDE	ACQ_DATE	ACQ_TIME	TYPE	BRIGHTNESS	CONFIDENCE	FR
701931	45.5176	-103.6266	2007-01-30	1930	0	316.4	0	33.
701932	45.5172	-103.6194	2007-01-30	1930	0	317.3	0	39.

January 2007

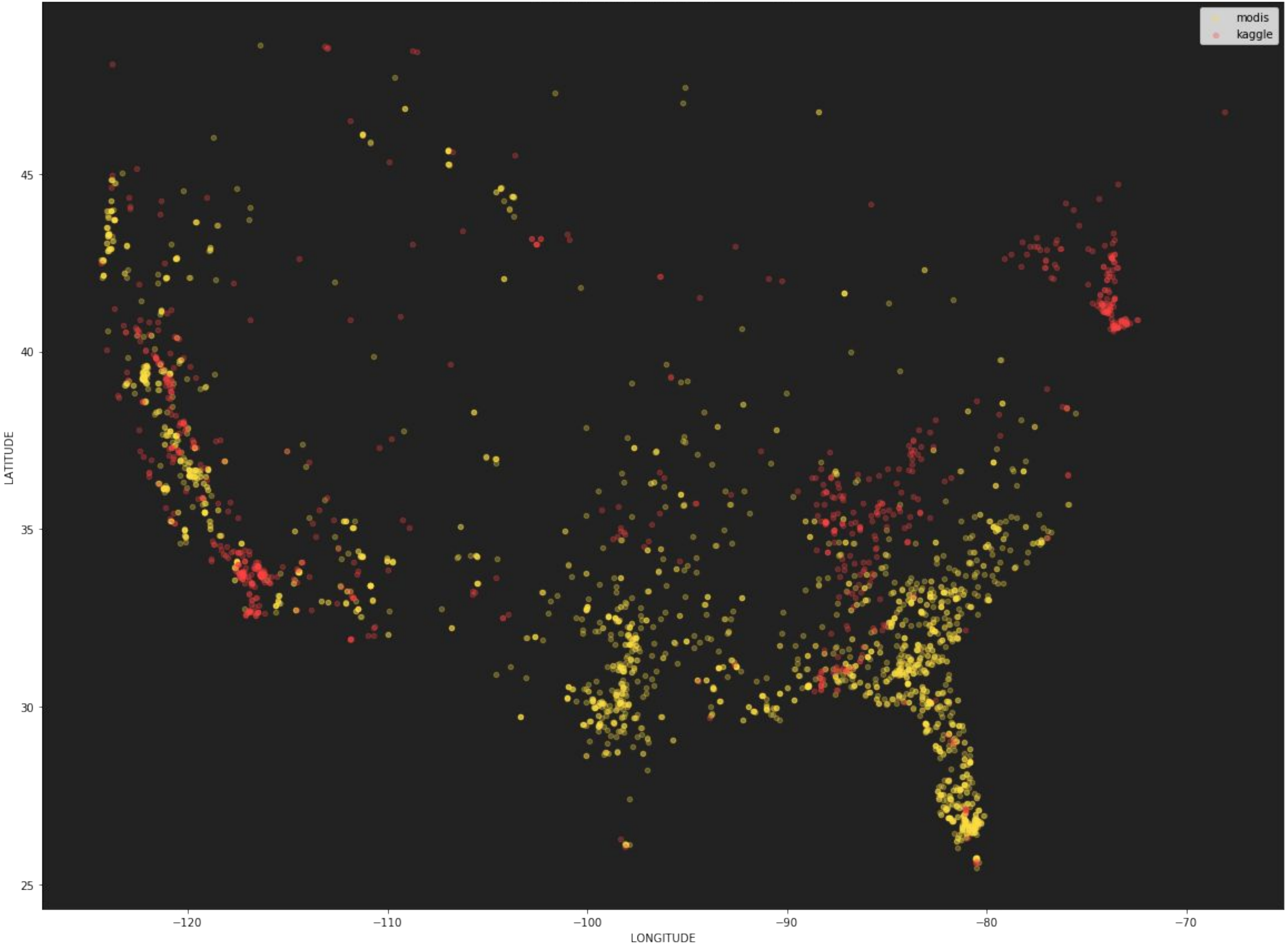
CONFIDENCE >= 0

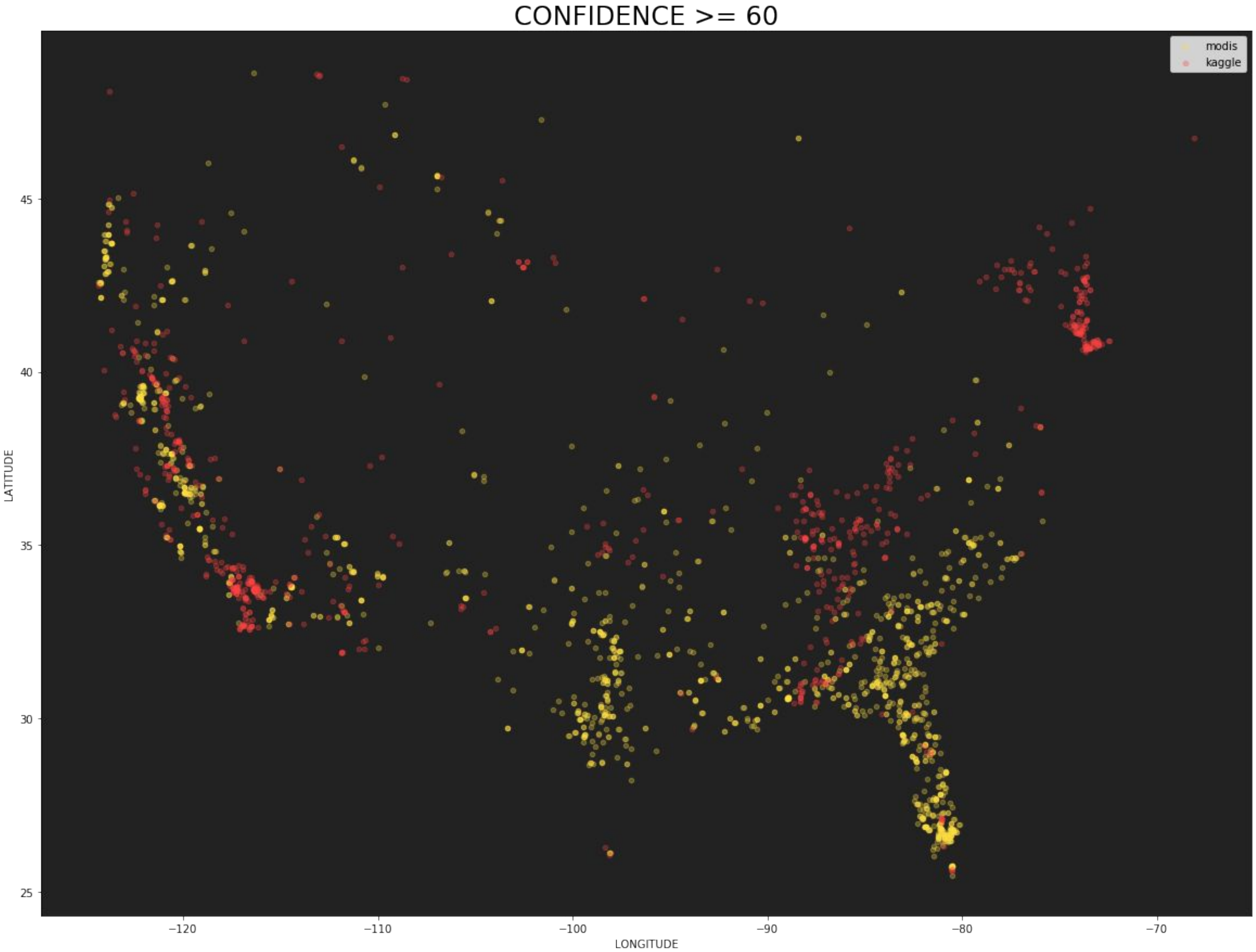




CONFIDENCE >= 20

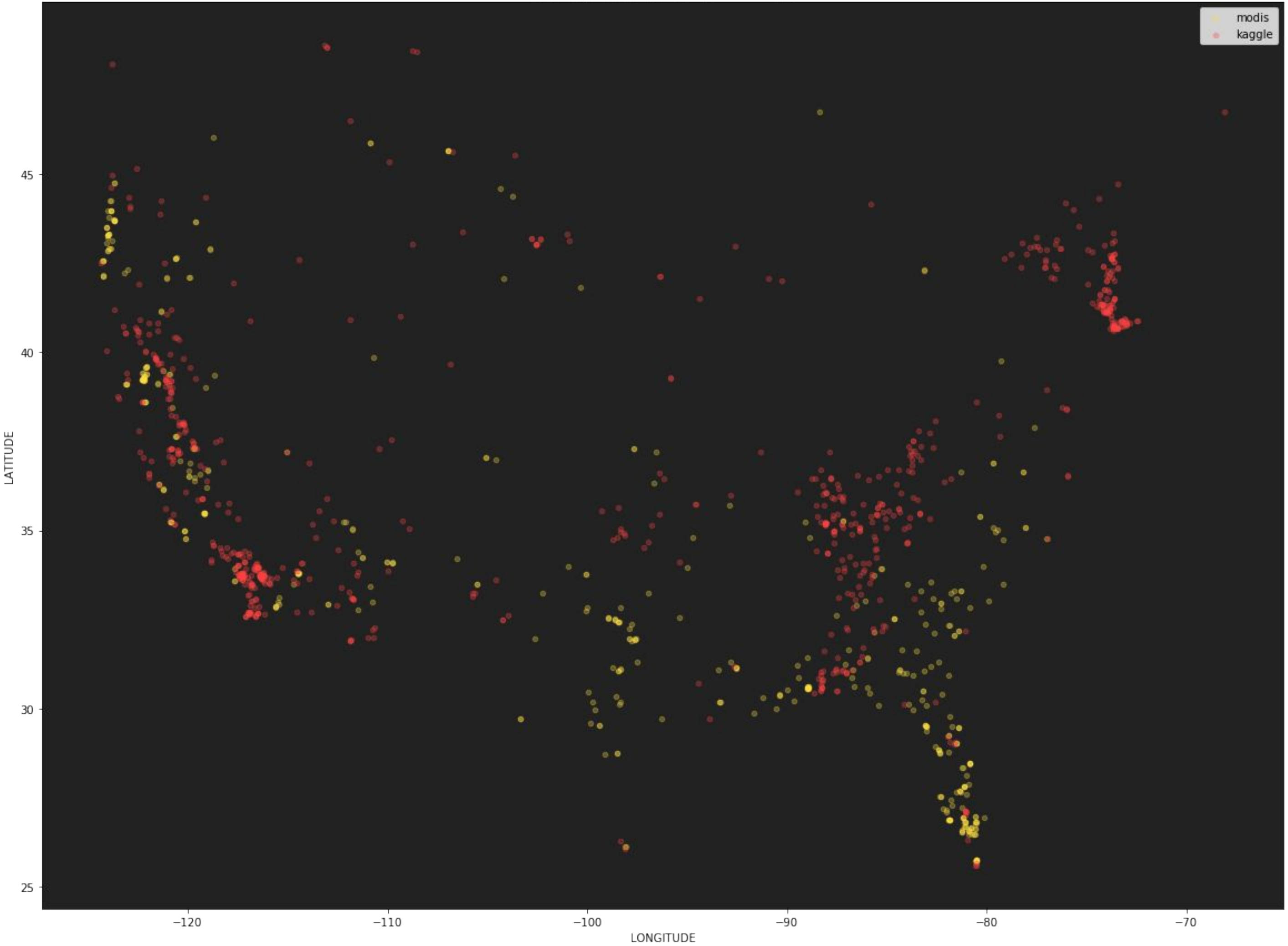
CONFIDENCE >= 40



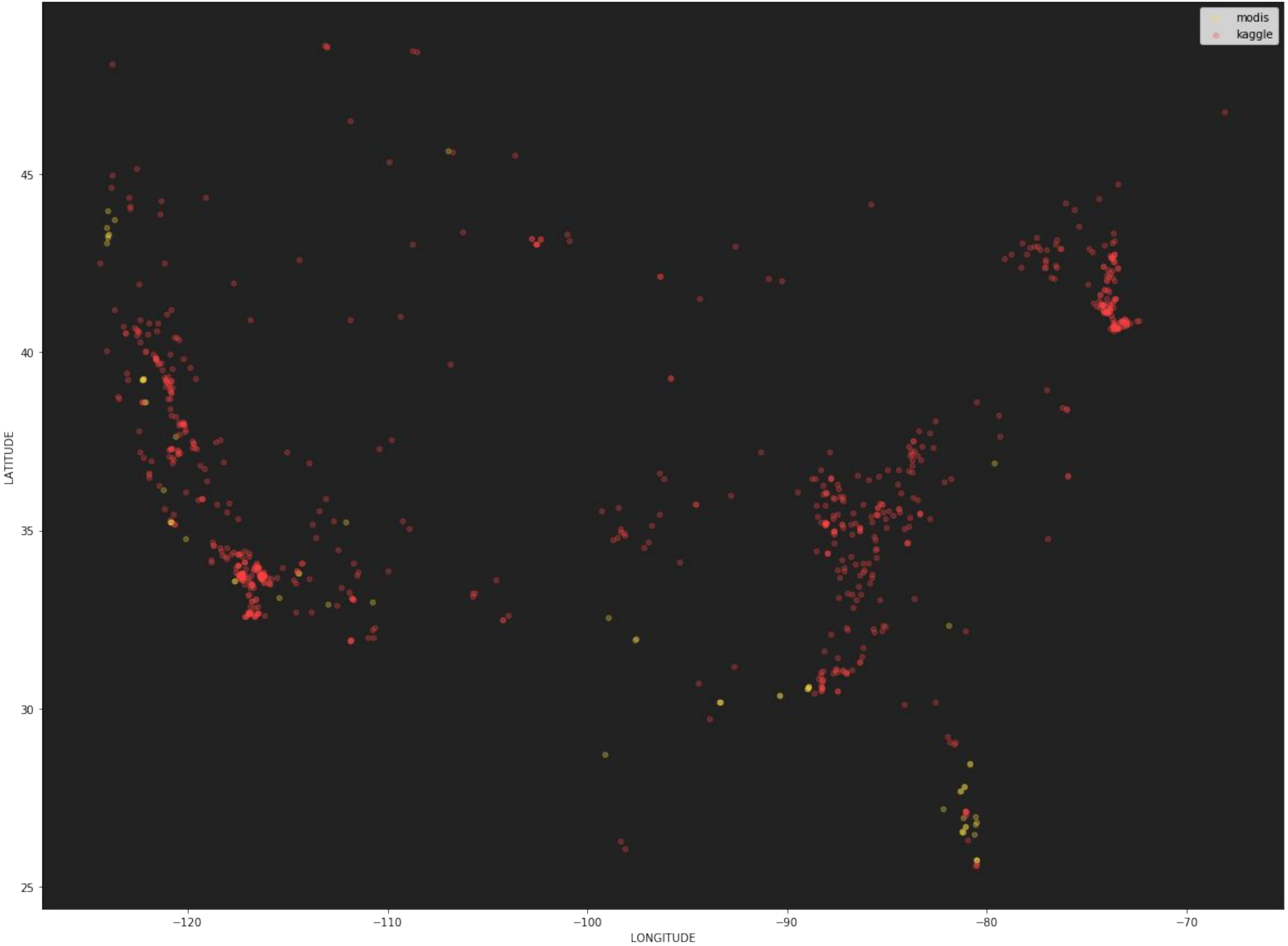


CONFIDENCE >= 60

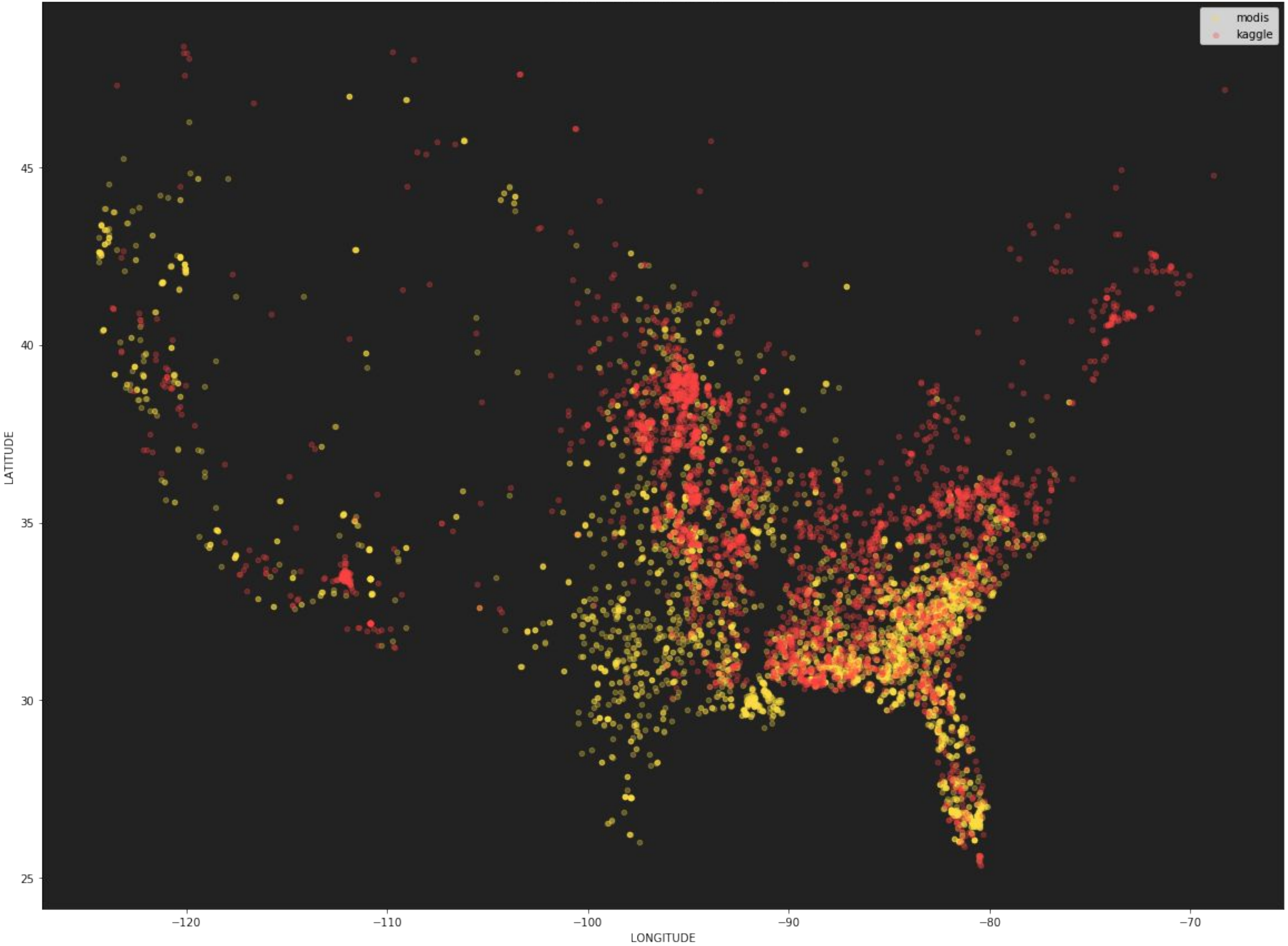
CONFIDENCE >= 80



CONFIDENCE >= 100

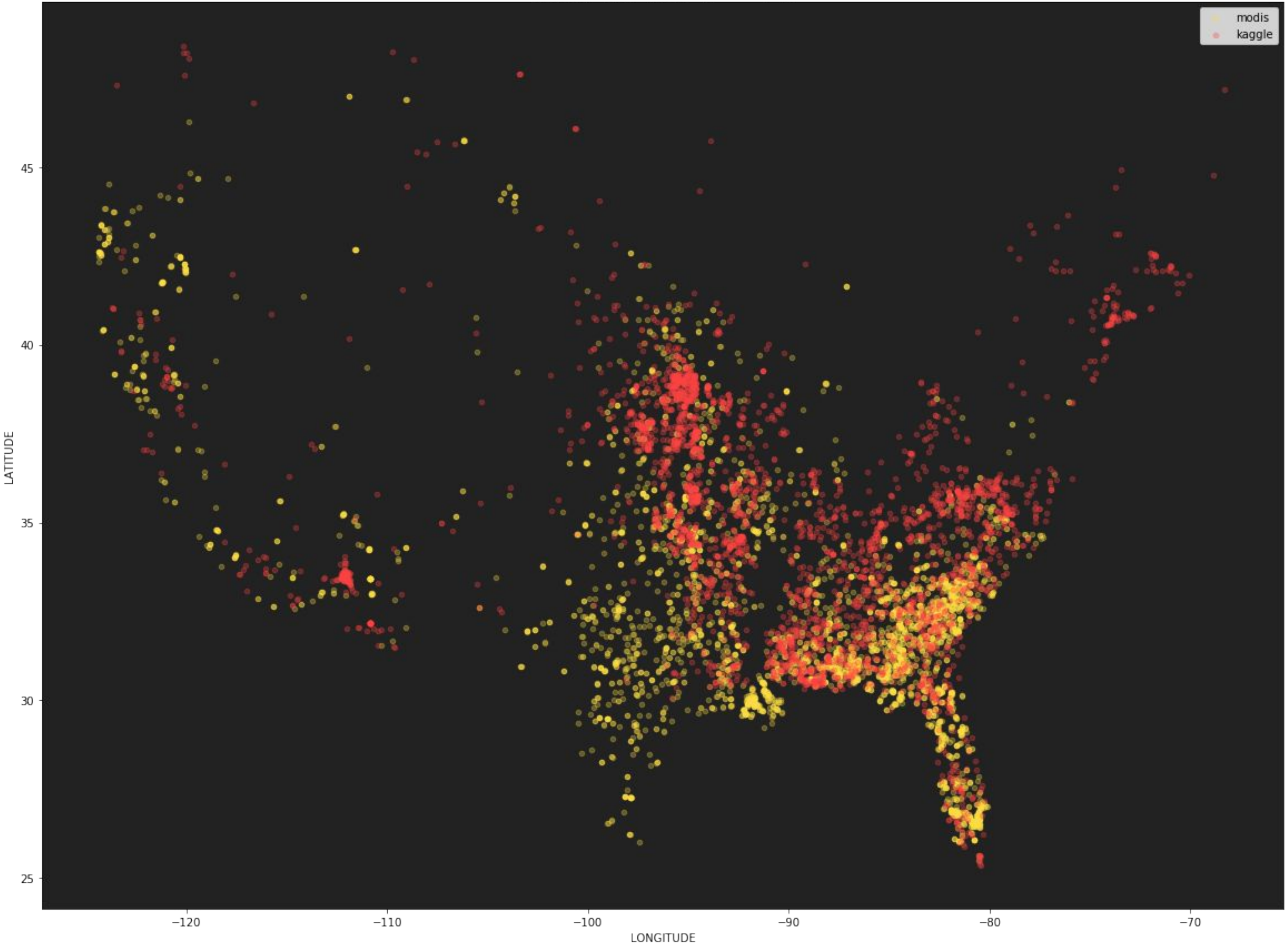


January 2015

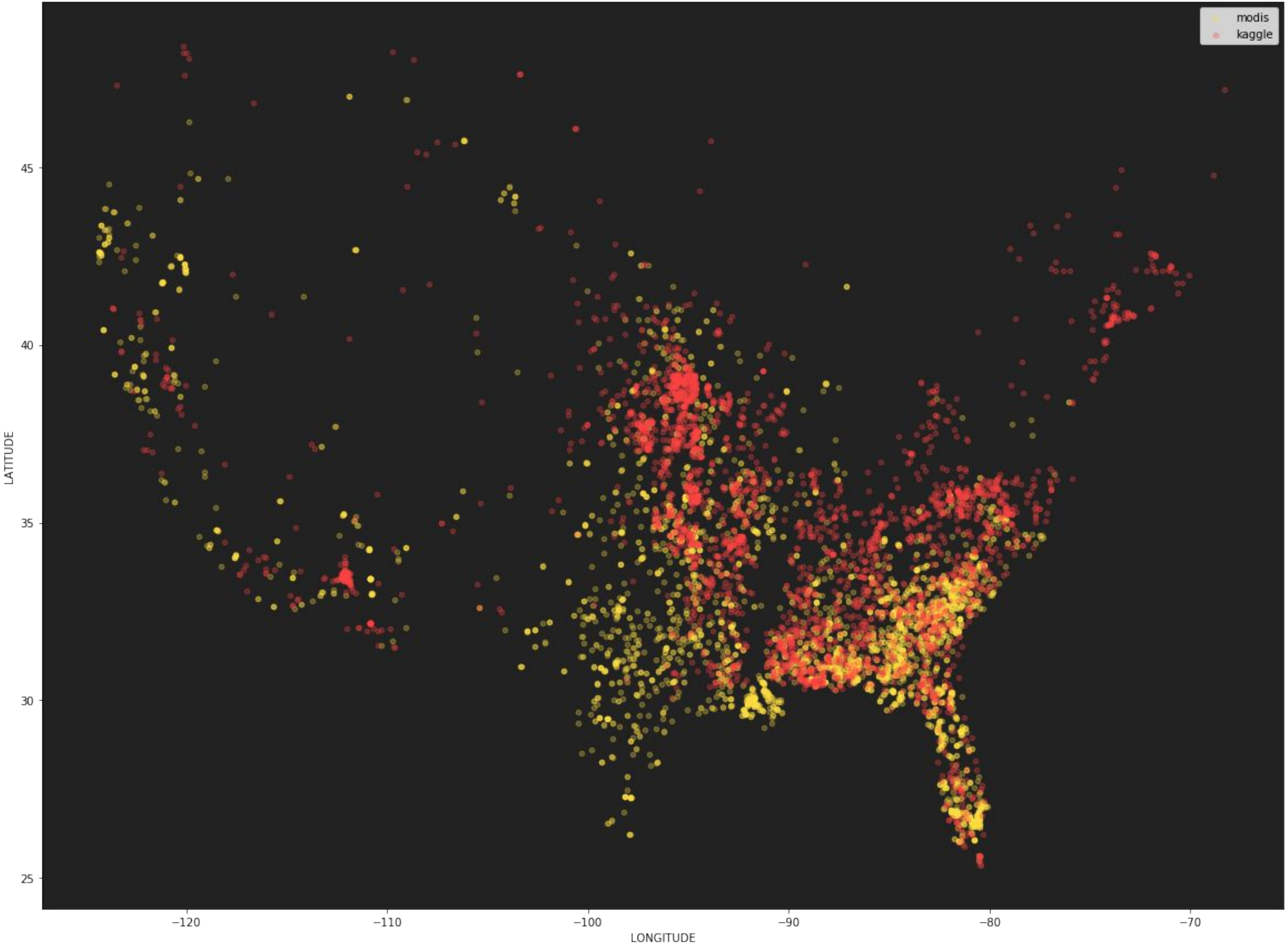


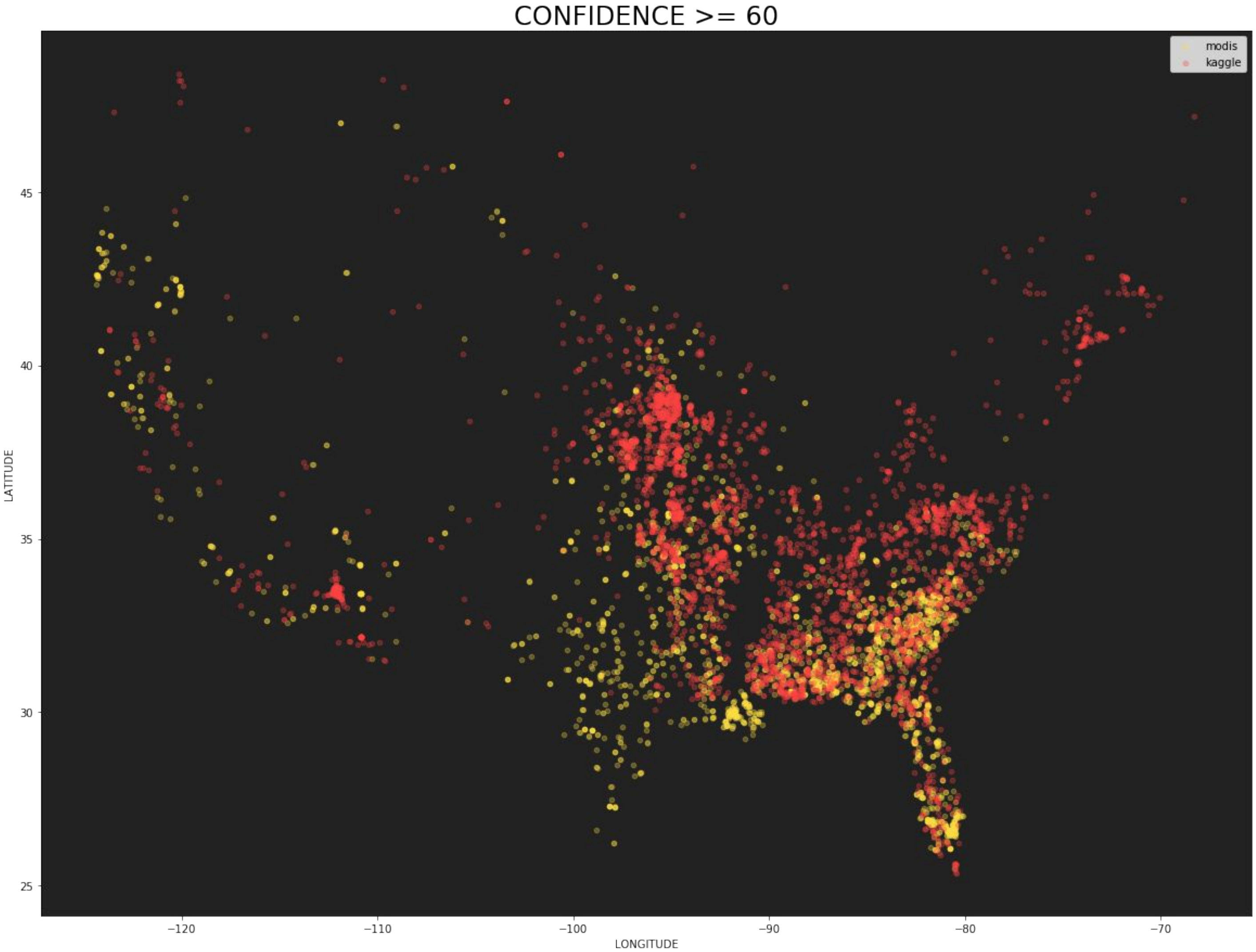
CONFIDENCE >= 0

CONFIDENCE >= 20



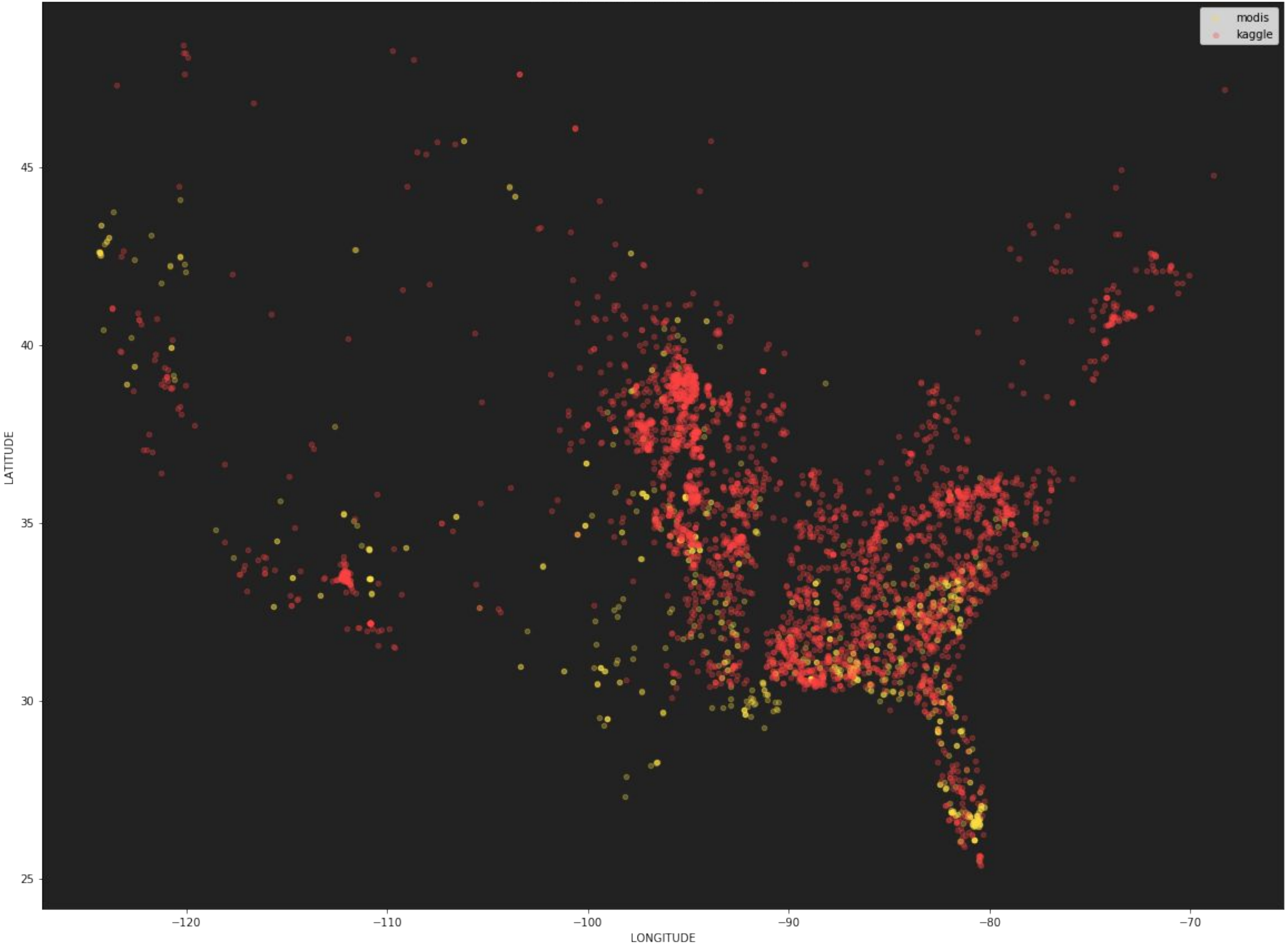
CONFIDENCE >= 40



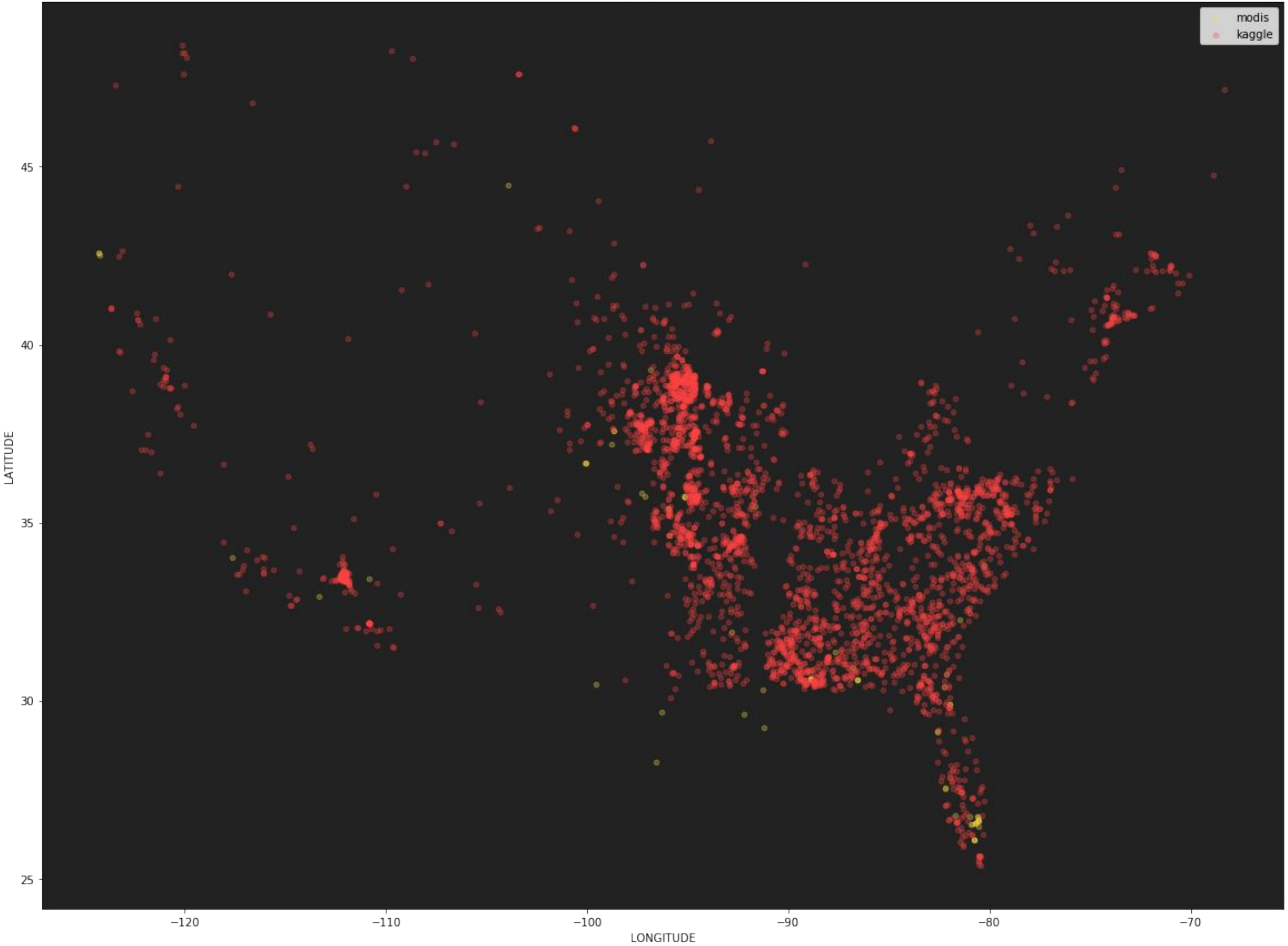


CONFIDENCE >= 60
modis
kaggle
LATITUDE
45
40
35
30
25
LONGITUDE
−120
−110
−100
−90
−80
−70

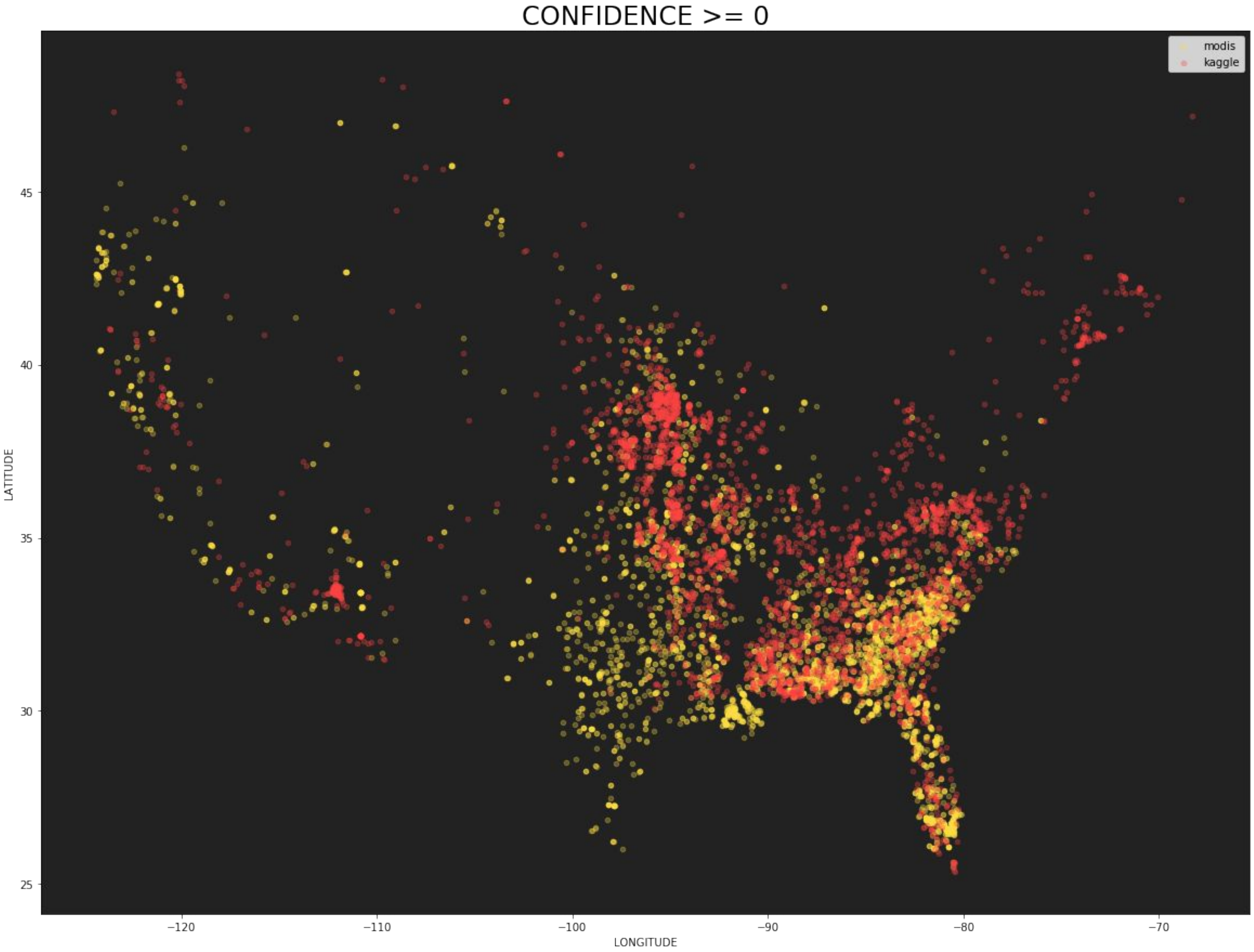
CONFIDENCE >= 80



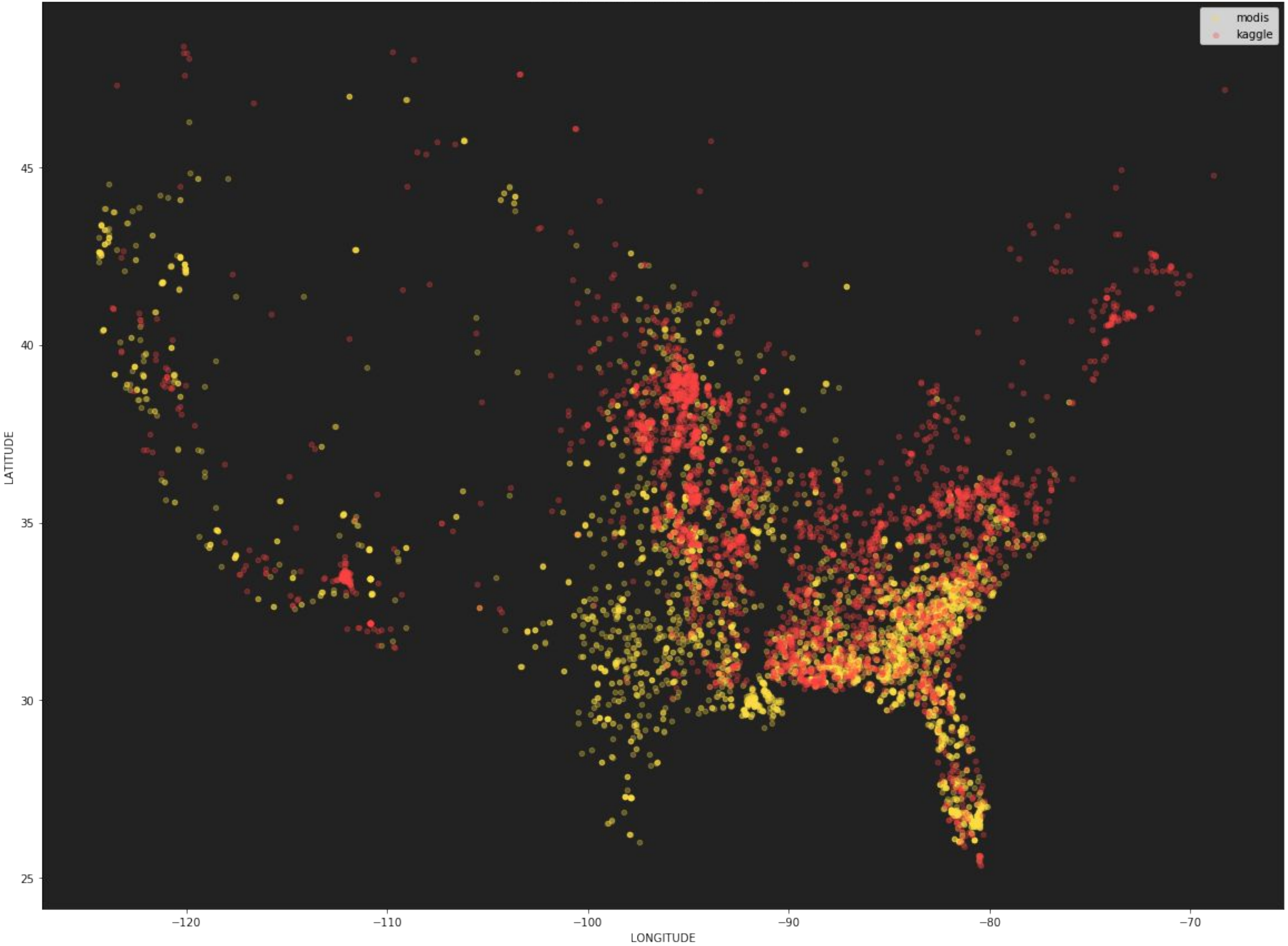
CONFIDENCE >= 100



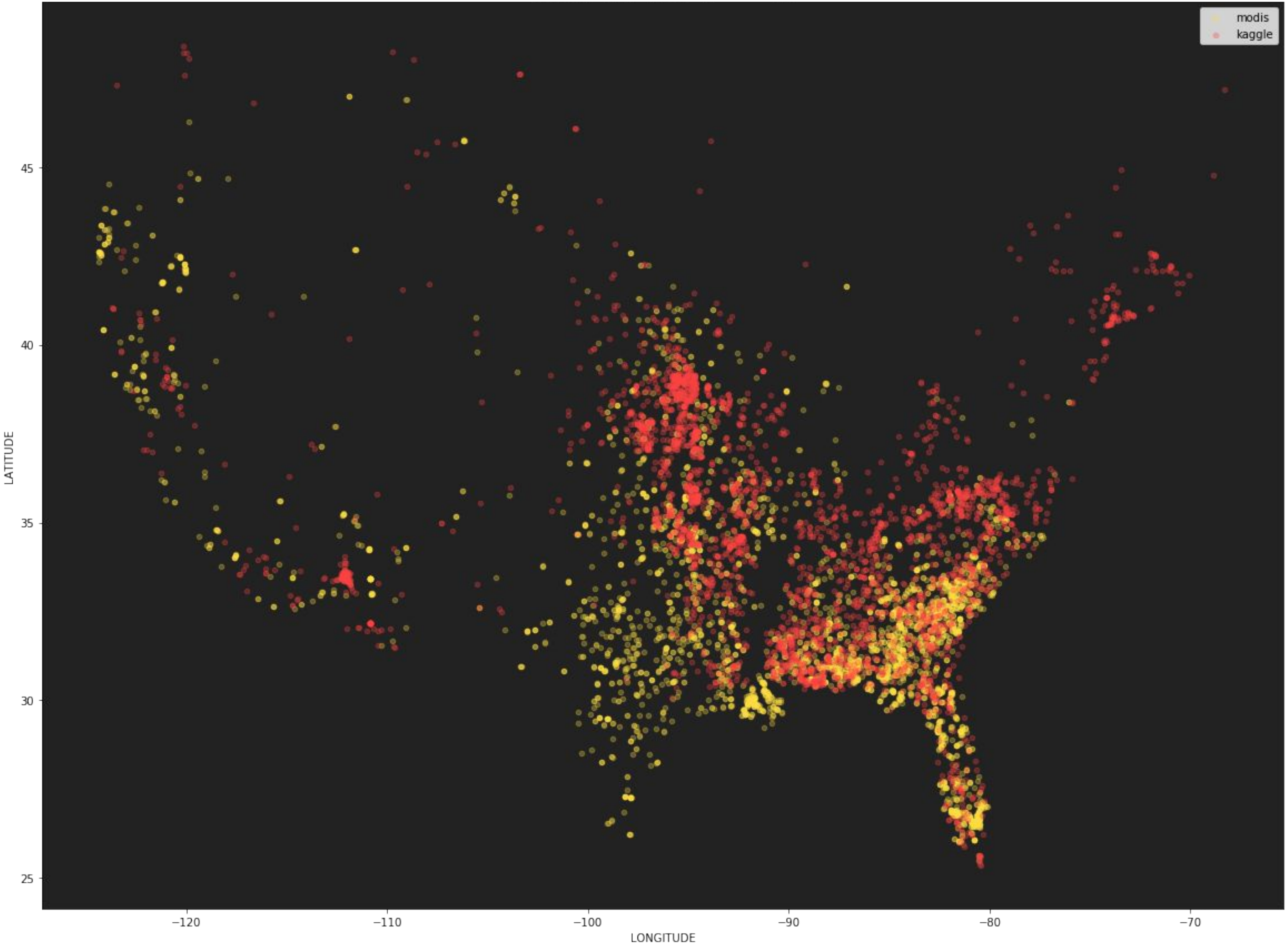
June 2015


CONFIDENCE >= 0
● modis
● kaggle
LATITUDE
45
40
35
30
25
LONGITUDE
−120
−110
−100
−90
−80
−70

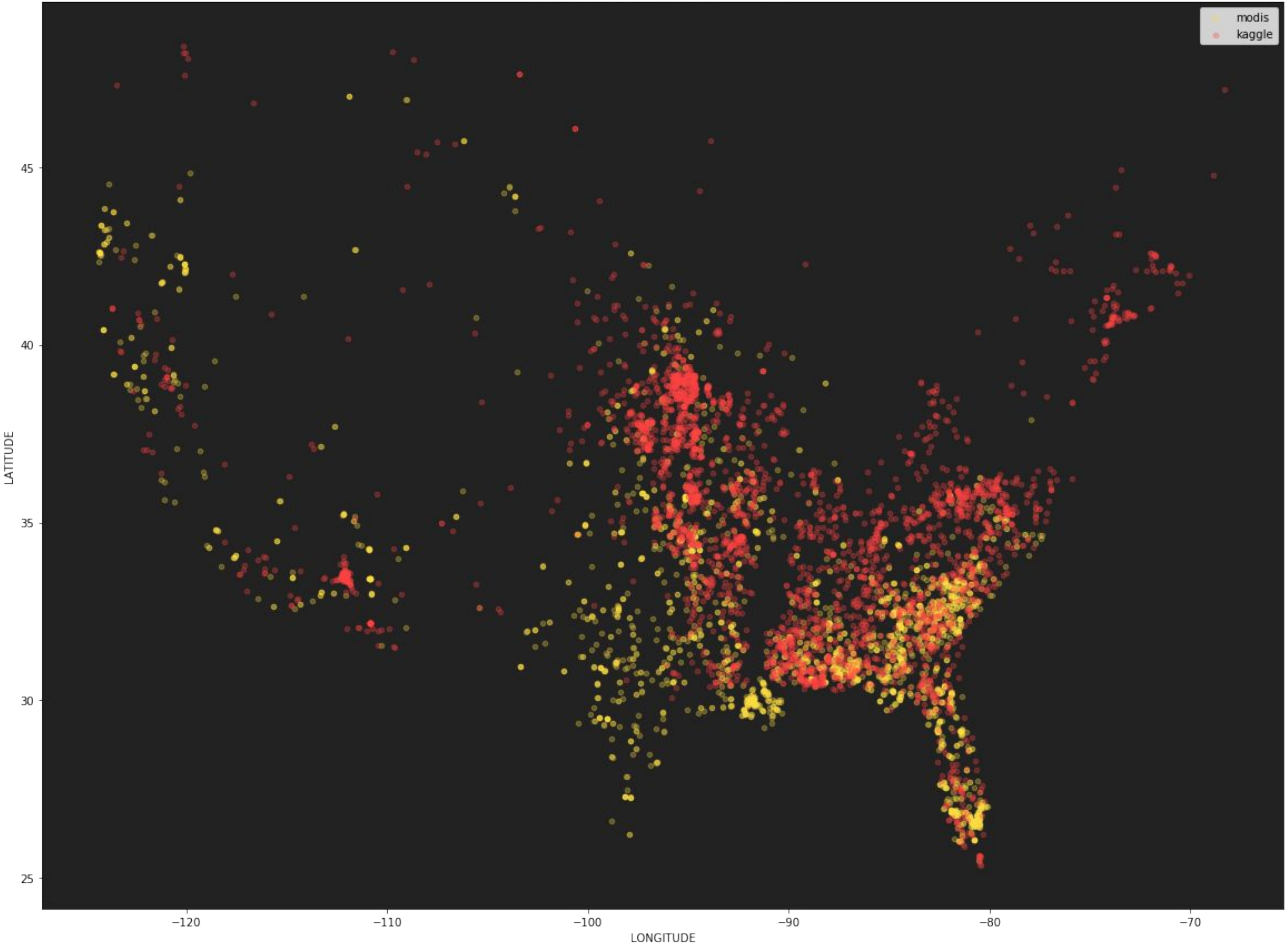
CONFIDENCE >= 20



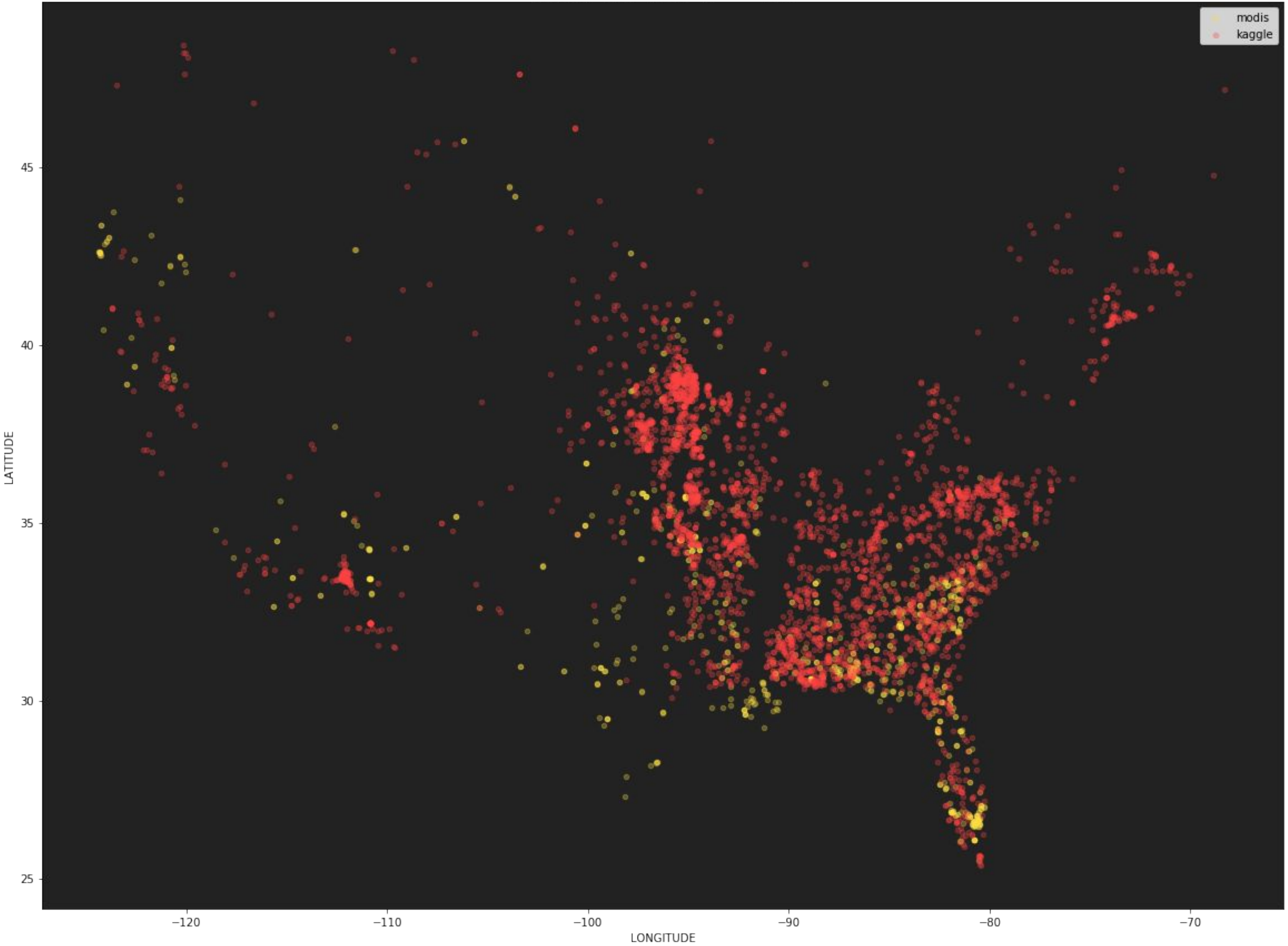
CONFIDENCE >= 40



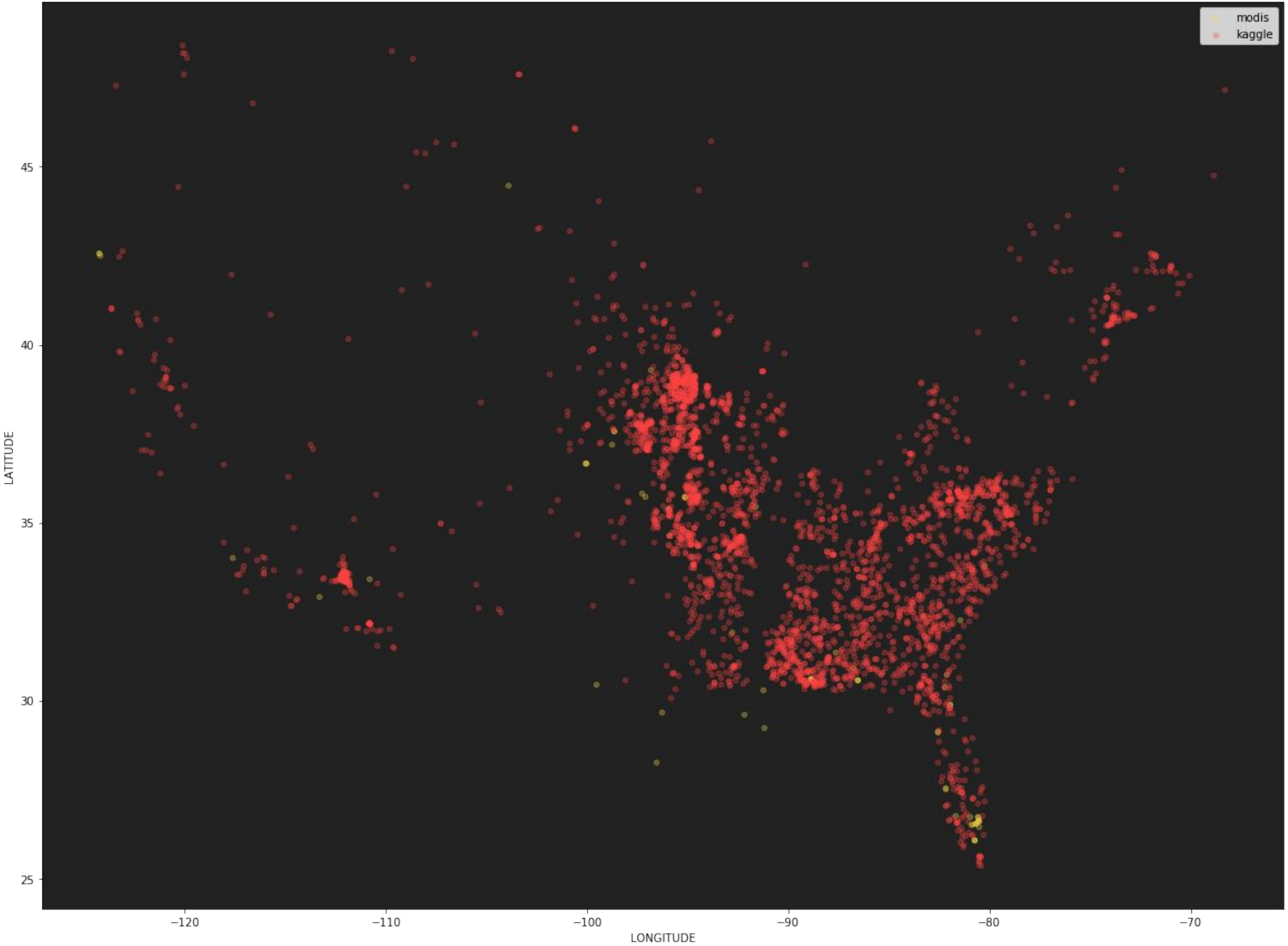
CONFIDENCE >= 60



CONFIDENCE >= 80

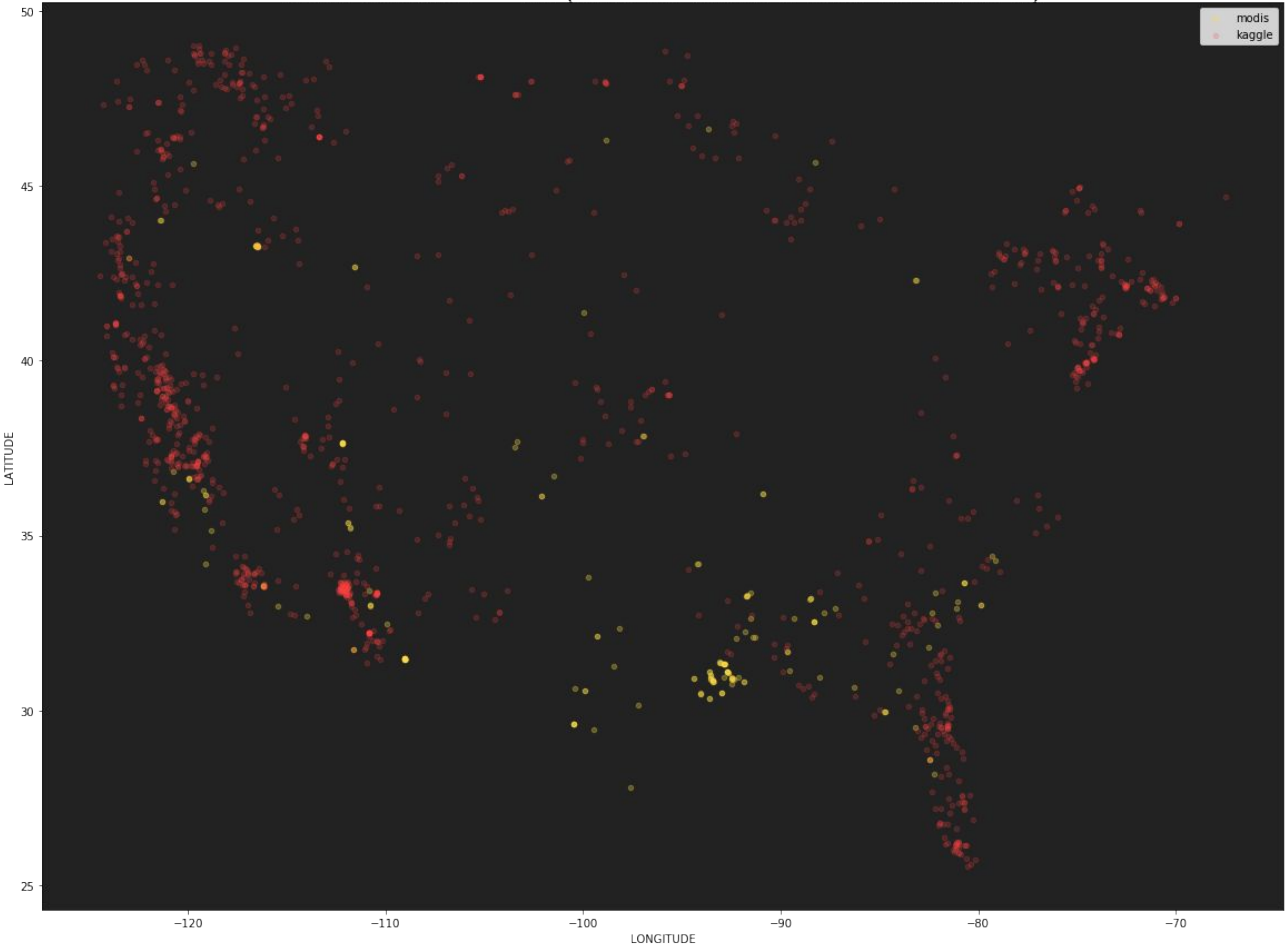


CONFIDENCE >= 100



Wk1 June 2015

CONFIDENCE >= 80 (from 01-06-2015 until 08-06-2015)



Conclusion

- The kaggle and MODIS datasets do not line up.
- Each dataset identifies fires which the other does not.
- MODIS could over-report and flag other thermal anomalies
- Kaggle could under-report and miss fires that are undetected by humans

Implications

The FIRMS dataset that is based on MODIS data is used by governments and academics all over the world. We will continue to examine the datasets, and, if we cannot rectify the discrepancies, we will write-up our findings.