West Nile Virus Prediction - Kaggle

Gabrielle | Wei Ern | Karunya | Dawn

Agenda

- Problem Statement
- The data
  - Any challenges or patterns?
- Built the model to make predictions
- Results
- Key predictors for WNV

Problem Statement

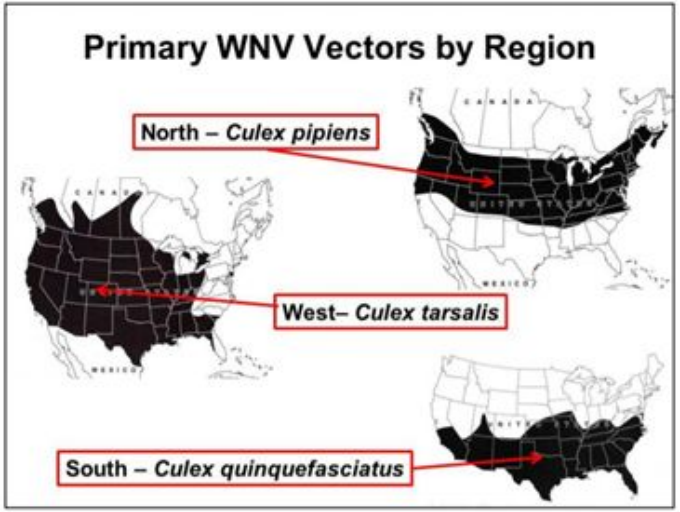
- Predict probability of the presence of the West Nile Virus for a given location, date, and species
- Understand key factors (features) for making accurate predictions...
  ... is it because of the weather ie rainfall (total precipitation) or tempe?
  ... is it the location?
  ... is it the frequency of spraying?

Diving into weather/spray patterns..

Just as importantly... how effective is the spray on curbing the number of Mossies?

West Nile Virus (WNV)

- First emerged in the eastern U.S. in 1999. In 2002, WNV reached Chicago for the first time with 225 human cases reported that summer.



1/5

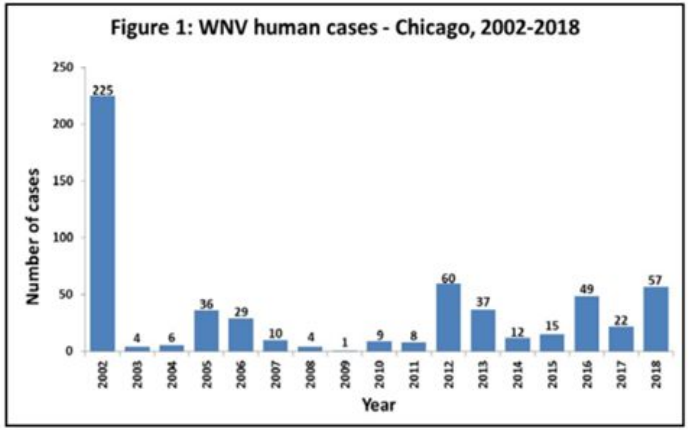
Develop fever or mild symptoms

1/150

Develop a serious illness that affects the central nervous system and can result in death

> Age 60

At greater risk



Chicago Department of Public Health (CDPH)
City-wide Surveillance and Mosquito Control Measures

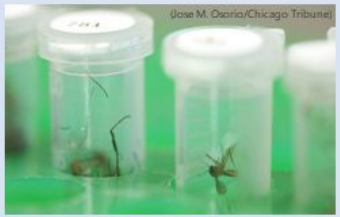
Collection

Sorting & Pooling

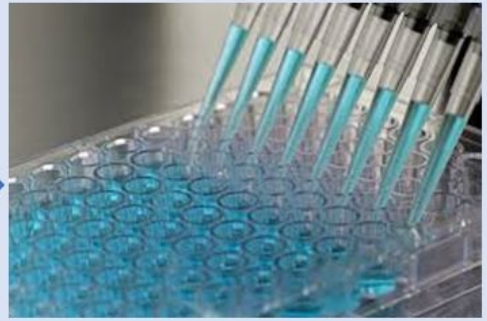
- Species 1
- Species 2
- Species 3
- Species 4
- Species 5
- Species 6

By species
By sex

Homogenisation & Centrifugation



Testing



Laboratory

Overview of the datasets

Train.csv

- 10,506 rows with 12 columns
- No null rows
- 1,062 duplicate rows (~10.11%)

Test.csv

- 116,293 rows, 11 columns

Spray.csv

- 10 unique spray dates
  - 2 spray dates in 2011
  - 8 spray dates in 2013
- 14,835 rows with 4 columns
- 584 null rows (~3.94%)
- 543 duplicate rows (~3.66%)

Weather.csv

- Daily records May-Oct, 2007 -2014
- 2 weather stations
- 2944 rows, 22 columns
- No duplicate rows

Let’s Get in the EDA EDA EDA

Train; aka trap records

- 10,506 rows with 12 columns
- No null rows

1) Multiple rows for the same
   - date
   - trap
   - location (latitude and longitude)
   - species

WHY?

- mosquitoes collected from traps
- sorted by 7 species
- bundled into “pools” of <= 50 with records for each “pool”

—> combined the split records

Train; aka trap records

- 10,506 rows with 12 columns
- No null rows

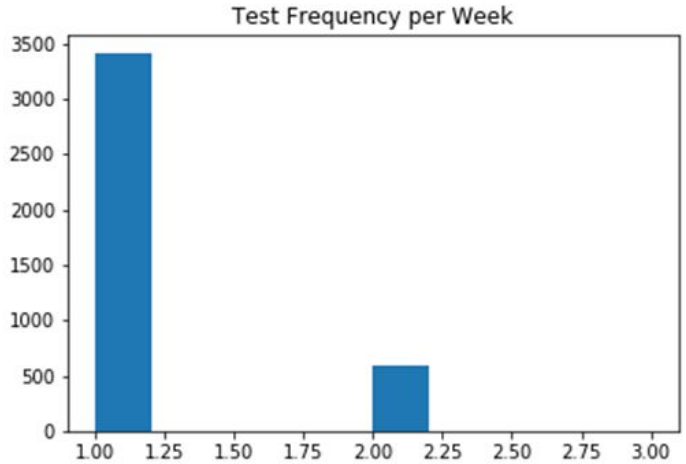
2) WnvPresent

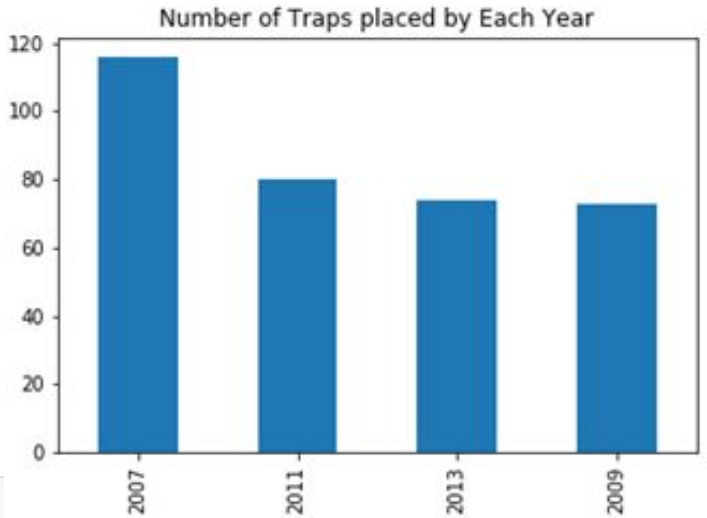
- does not indicate proportion of WNV-carrying mosquitoes in each test tube
- binary indicator; threshold unknown

WHY?

- 1 test tube: 1 “pool” of <= 50 mosquitoes
- Mosquito diluent to “homogenise” all specimens in each test pool

Train; aka trap records






```
wt_freq_trap = num_pools.groupby(['year','month','week','Trap']).count()
wt_freq_trap.Species.value_counts()
```

```
1    3411
2     598
3       3
Name: Species, dtype: int64
```

Train; aka trap records

Average number of Mosquitos peaks

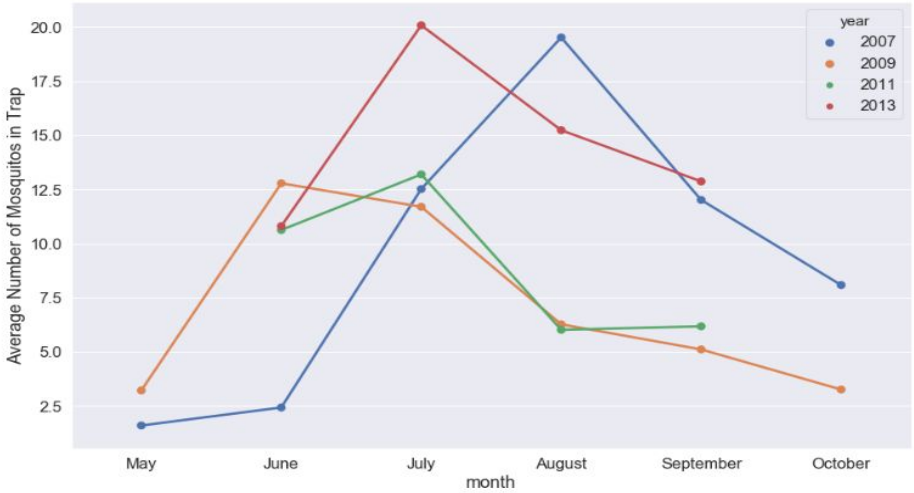
- June - 2009
- July 2011, 2013
- August - 2007

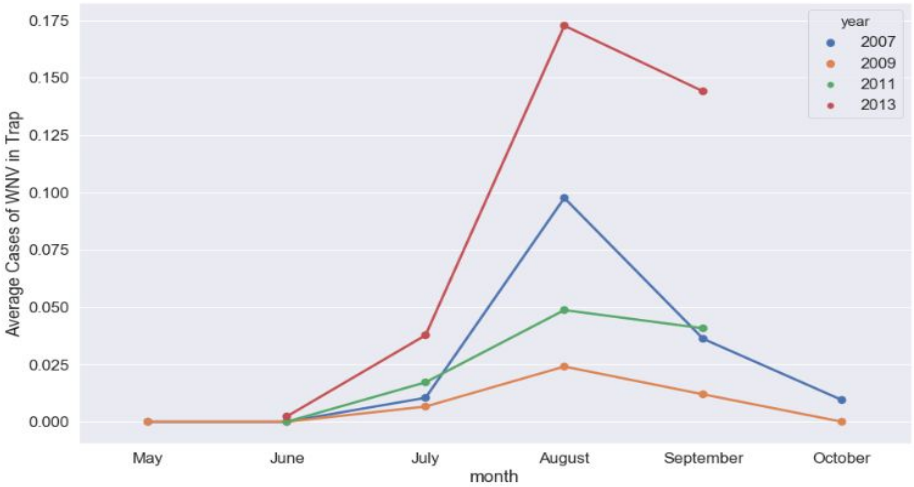
Average number of WNV cases

- August

```
1  merged.groupby('year')['Trap'].count()
```

```
year
2007    3811
2009    2249
2011    2054
2013    2392
Name: Trap, dtype: int64
```





Spray

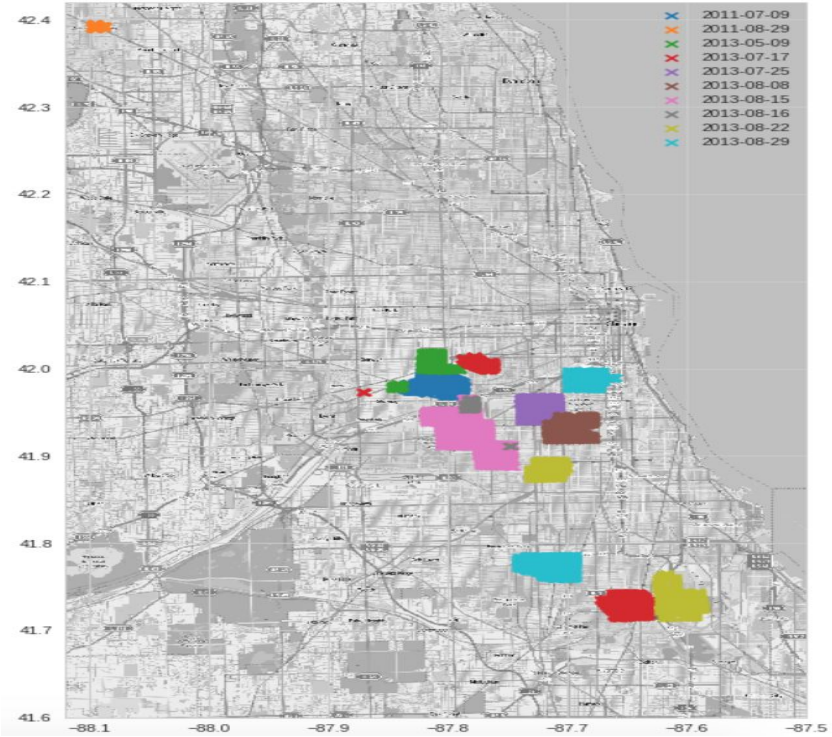
- 14,835 rows with 4 columns
- 584 null rows (~3.94%)
- 543 duplicate rows (~3.66%)

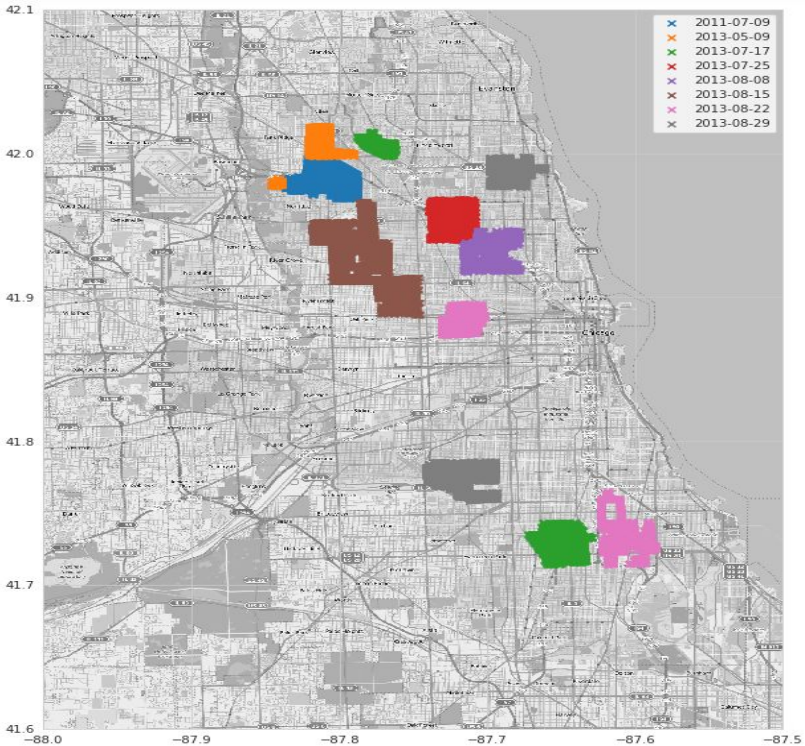
1) null cells
   - Caused by missing data in ‘Time’ column —> drop column
2) duplicate cells
   - Caused by data entry error —> drop duplicates

Spray

Coloured-coded clusters according to the spray dates in Years 2011 and 2013

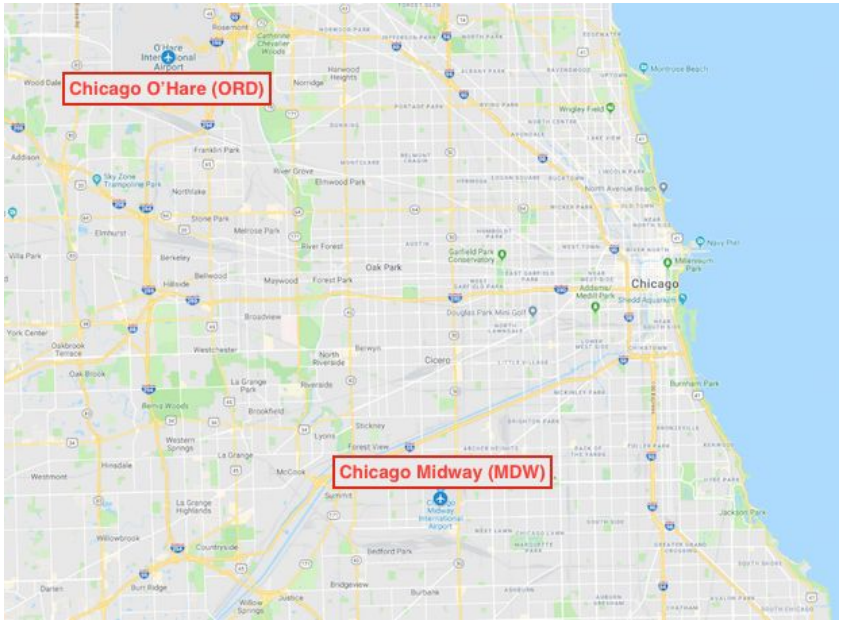
- Dropped 2011-08-29 (outlier to trap locations) + 5 entries recorded in the 10:49 timeframe (insignificant)





Weather

- Daily records May-Oct, 2007 -2014
- 2 weather stations
- 2944 rows, 22 columns
- No duplicate rows



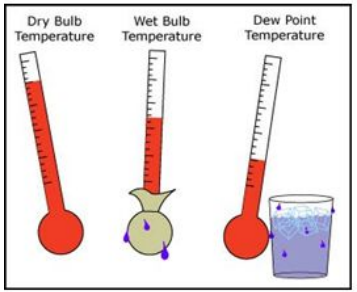
Trace values represented by T:

- Replace with 0

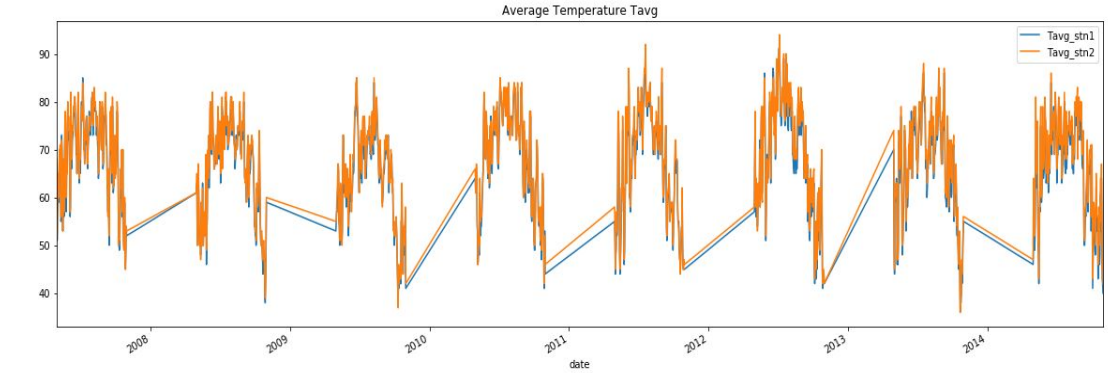
Missing values represented by M:

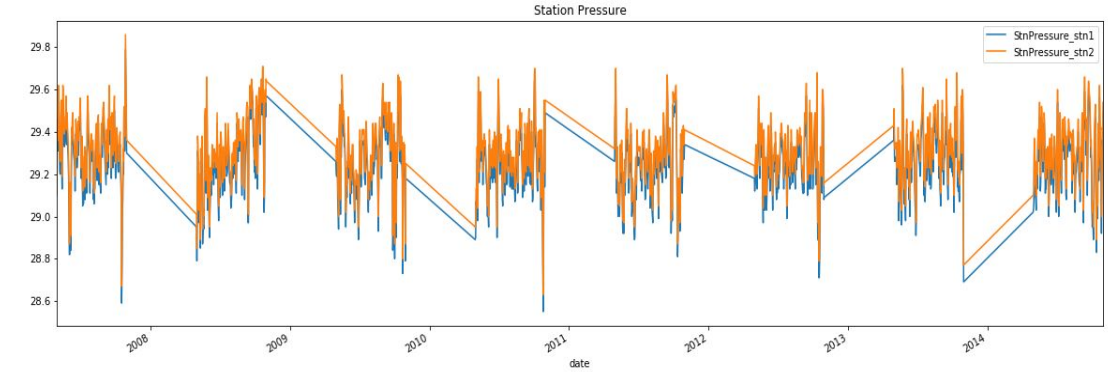
- Calculate from other features (e.g. Tavg = (Tmax + Tmin)/2)
- Impute from records of Station 1
- Forward fill (only 2 records for StnPressure)

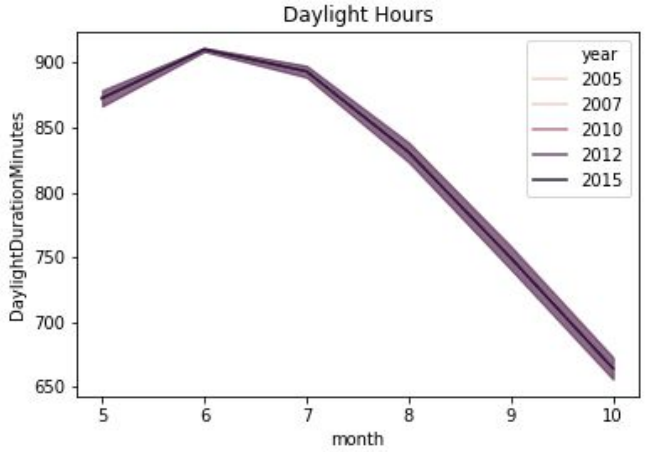
Weather



Station	Station 1: Chicago O'Hare INTL Airport at (41.995, -87.933) Station 2: CHICAGO Midway INTL Airport at (41.786, -87.752)
Date	Data collection date
Tmax, Tmin, Tavg	Daily extremes and averages of temperature
Depart	Temperature Departures from normal The difference between the average temperature and the 30-year normal temperature for this date (climate change measurement)
DewPoint	Average dew point temperature
WetBulb	Average wet bulb temperature
Heat	Heating Degree Days = Tavg – baseline
Cool	Cooling Degree Days = baseline – Tavg A gauge of the energy demand for heating or cooling a building (baseline = 65 ^{o}F)
Sunrise, Sunset	SUNSET (Calculated, not observed)
CodeSum	Significant weather condition (coded remarks)
Depth	Depth of rainfall and melted snow
Water1	Water equivalent of the rainfall and melted snow
SnowFall	The depth of snowfall
PrecipTotal	Total precipitation
StnPressure, SeaLevel	station and sea level pressure in unit of INCHES OF HG
ResultSpeed	Resultant Wind Speed (Calculated speed based on other wind speed measurements)
ResultDir	Resultant Wind Direction (calculated based on other wind direction measurements)
AvgSpeed	Average Speed


Average Temperature Tavg
Tavg_stn1
Tavg_stn2
90
80
70
60
50
40
2008
2009
2010
2011
2012
2013
2014
date


Station Pressure
StnPressure_stn1
StnPressure_stn2
29.8
29.6
29.4
29.2
29.0
28.8
28.6
2008
2009
2010
2011
2012
2013
2014
date

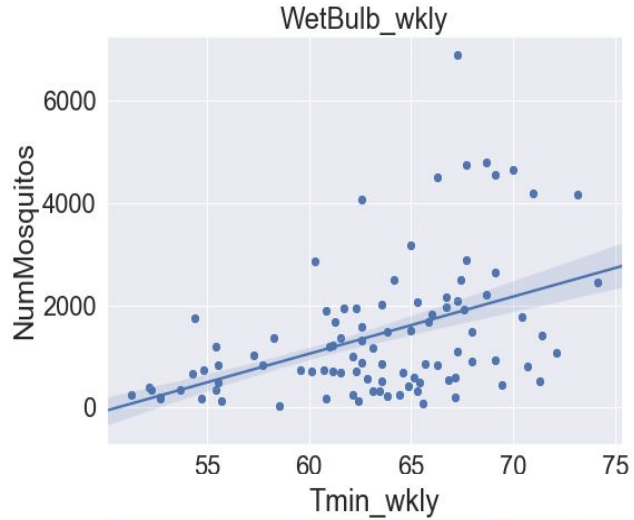

Daylight Hours
year
2005
2007
2010
2012
2015
DaylightDurationMinutes
900
850
800
750
700
650
5
6
7
8
9
10
month

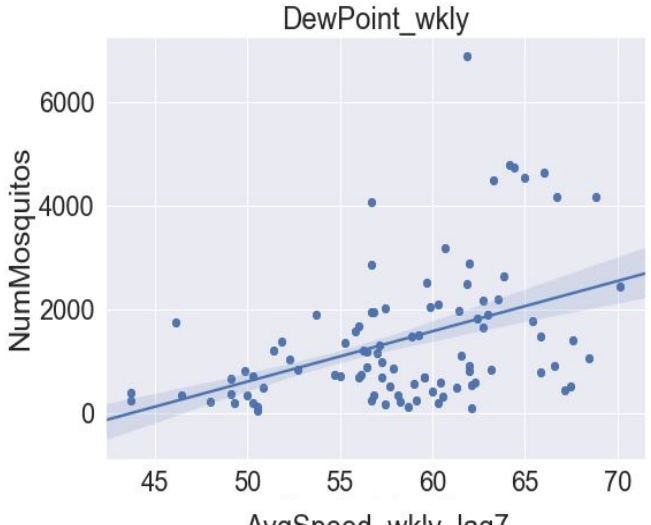
Weather

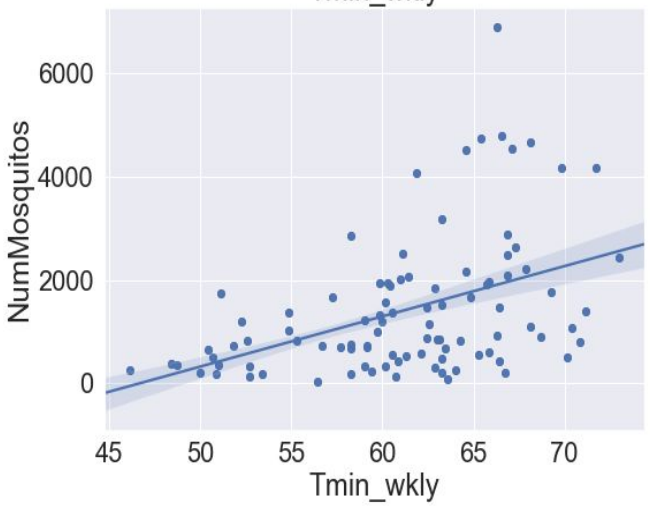
```
print(mosquitos_weather.corr()['NumMosquitos'].sort_values(ascen
print()
print(mosquitos_weather.corr()['NumMosquitos'].sort_values(ascen
```

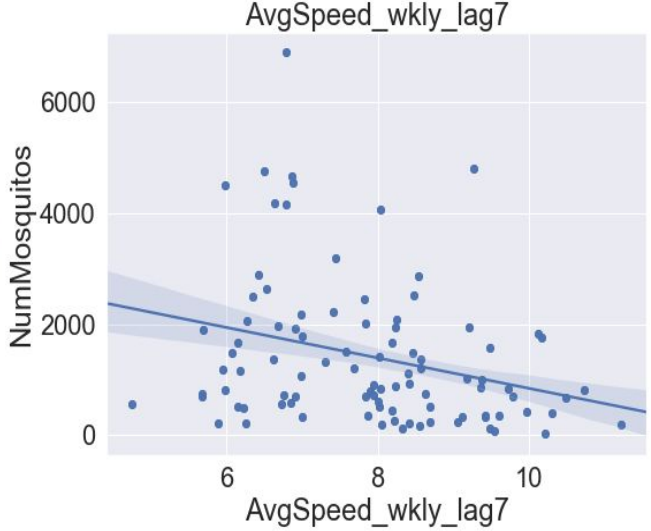
```
NumMosquitos       1.000000
WetBulb_wkly       0.446993
Tmin_wkly          0.439562
DewPoint_wkly      0.428584
Tavg_wkly          0.424924
Tmax_wkly          0.387870
Tavg               0.385492
Tmin               0.383440
WetBulb            0.371709
Tavg_wkly_lag21    0.354208
Name: NumMosquitos, dtype: float64

SeaLevel_wkly_lag14      -0.141479
AvgSpeed_wkly_lag21      -0.165419
ResultSpeed_wkly_lag21   -0.173634
ResultDir_wkly_lag7      -0.176042
AvgSpeed_wkly_lag14      -0.184429
ResultSpeed_wkly_lag14   -0.203734
ResultSpeed_wkly         -0.213070
AvgSpeed_wkly            -0.223811
AvgSpeed_wkly_lag7       -0.285431
ResultSpeed_wkly_lag7    -0.308009
Name: NumMosquitos, dtype: float64
```



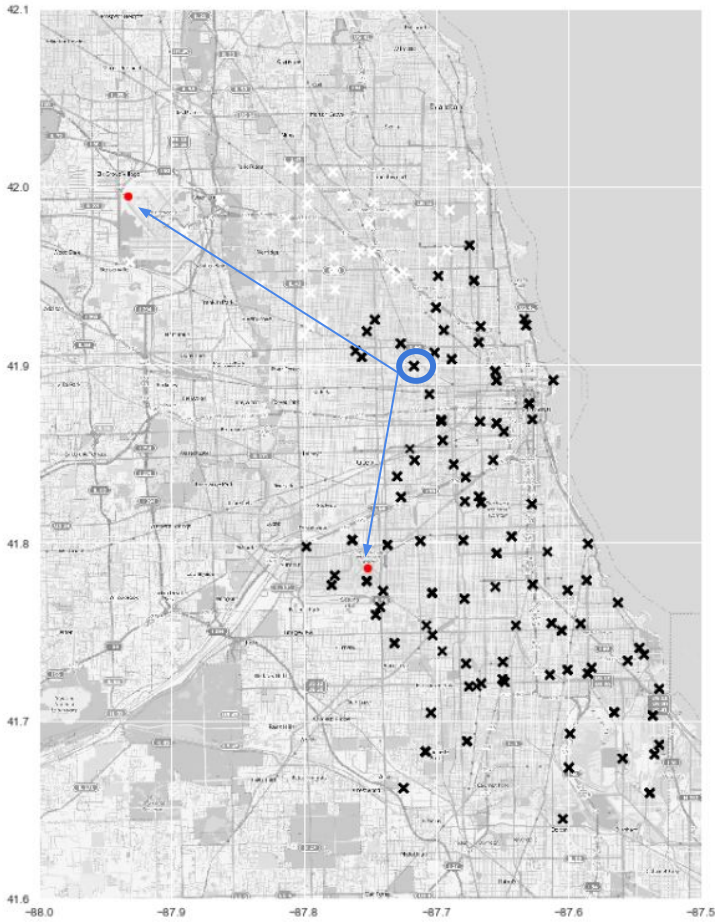






Pre-processing summary

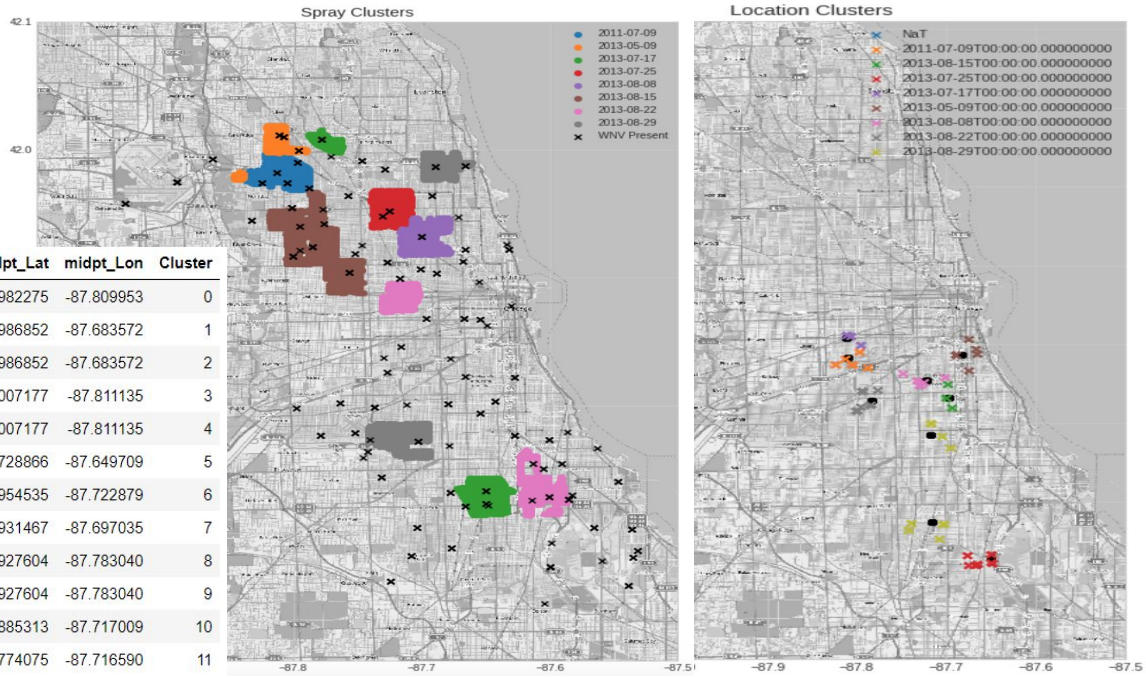
- Parsed dates into **year**, **month**, **week of year**, **day of year**
- Dropped all address-related columns, except **latitude** and **longitude**
- Mapped nearest **weather station** for each **trap record** to retrieve weather conditions



Pre-processing summary

- Mapped each **trap record** to a **spray cluster** to determine if the trap location has been sprayed
  - Number of days between Spraying and Testing
  - Binary imputation if an observation has been sprayed

	Date	midpt_Lat	midpt_Lon	Cluster
95	2011-07-09	41.982275	-87.809953	0
13911	2013-05-09	41.986852	-87.683572	1
14606	2013-05-09	41.986852	-87.683572	2
2537	2013-07-17	42.007177	-87.811135	3
2209	2013-07-17	42.007177	-87.811135	4
4411	2013-07-25	41.728866	-87.649709	5
6018	2013-08-08	41.954535	-87.722879	6
7213	2013-08-15	41.931467	-87.697035	7
10022	2013-08-22	41.927604	-87.783040	8
10867	2013-08-22	41.927604	-87.783040	9
11609	2013-08-29	41.885313	-87.717009	10
12980	2013-08-29	41.774075	-87.716590	11



Pre-processing summary

- Engineered the following features:
  - Dummy columns for mosquito Species

CULEX ERRATICUS	CULEX PIPIENS	CULEX PIPIENS/RESTUANS	CULEX RESTUANS	CULEX SALINARIUS	CULEX TARSALIS	CULEX TERRITANS

  - Average the weather elements by using 7-day rolling average

Tmax_wkly	Tmin_wkly	Tavg_wkly	Depart_wkly	DewPoint_wkly	WetBulb_wkly

PrecipTotal_wkly	StnPressure_wkly	SeaLevel_wkly	ResultSpeed_wkly	ResultDir_wkly	AvgSpeed_wkly

Imbalanced dataset

WnvPresent (0 for Absent; 1 for Present)	Proportion
0	0.947554
1	0.052446

Oversampled the underrepresented minority class - Virus Not Present -> balance the class distribution in the dataset for modeling

Balance Classes by: Oversampling (SMOTE)

In [38]:
```
# Oversampling on training data only
sm = SMOTE(random_state=42)
X_res, y_res = sm.fit_sample(Xs_train, y_train)
```

Modelling

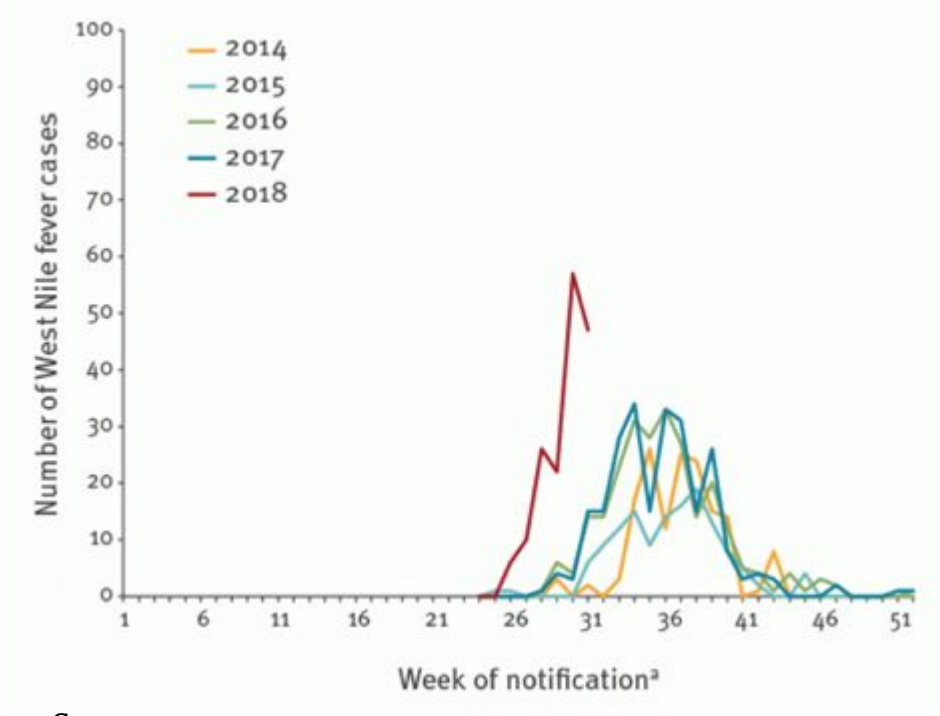
	model	parameters	scores	recall	precision	ROC - AUC
0	Decision Tree	{'clf__max_depth': None, 'clf__min_samples_spl...	0.052348	1.000000	0.052348	0.5
1	RandomForest	{'clf__max_depth': None, 'clf__min_samples_spl...	0.931155	0.012121	0.035714	0.512
2	LogReg	{'clf__C': 5, 'clf__max_iter': 50, 'clf__solve...	0.052348	1.000000	0.052348	0.5
3	KNN	{'clf__n_neighbors': 3, 'clf__p': 1, 'clf__wei...	0.267766	0.684848	0.047699	0.5
4	SVC	{'clf__kernel': 'rbf'}	0.947652	0.000000	NaN	0.5
5	AdaBoost	{'clf__n_estimators': 500}	0.947652	0.000000	NaN	0.5
6	BaggingClass	{'clf__max_features': 10, 'clf__max_samples': ...	0.052348	1.000000	0.052348	
7	GradientBoost	{'clf__loss': 'deviance', 'clf__max_features':...	0.052348	1.000000	0.052348	0.518
8	XGBoost		0.748	0.770	0.144	0.837

Final model: XGBoost!

Roc-Auc Score: 0.837

Cost & Benefit Analysis

At least 104 confirmed cases in EU countries as well.



2014 to 2017

- 5 to 25 cases occurred from week 25 to week 31

2018

- first case reported in week 26
- total cases amount to 168
- 231 cases in EU member states

Source: https://www.ecdc.europa.eu/en/news-events/unusual-early-start-west-nile-fever-season-and-rise-cases-ecdc-assessment

Cost & Benefit

Benefits

- Lowers the number of mosquitos which are carriers of not only West Nile virus but also other vector bourne diseases such as Zika

- Lower the high costs of possible medical and hospitalisation fees

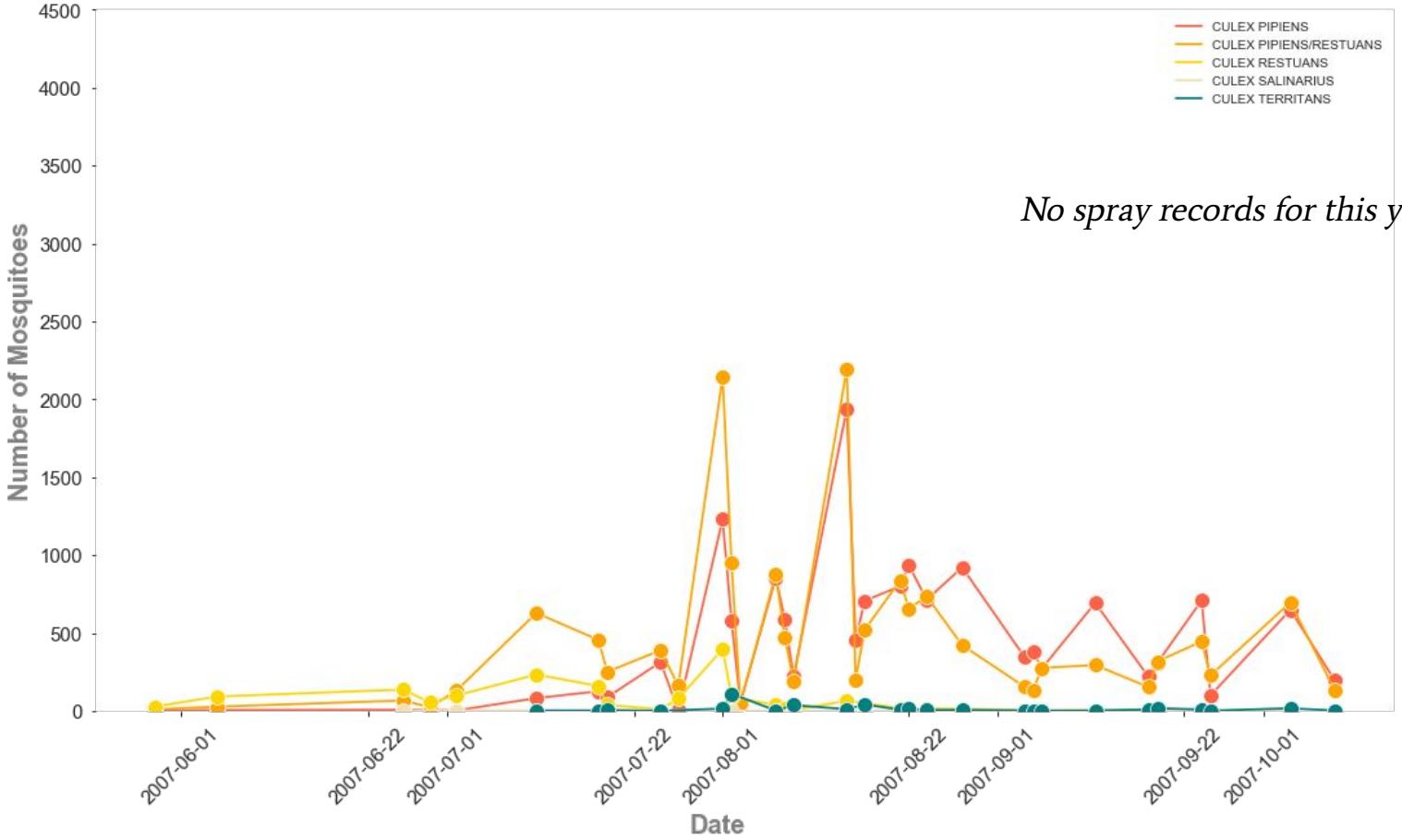
- Reduce costs of the state to handle such outbreaks and emergencies

Costs

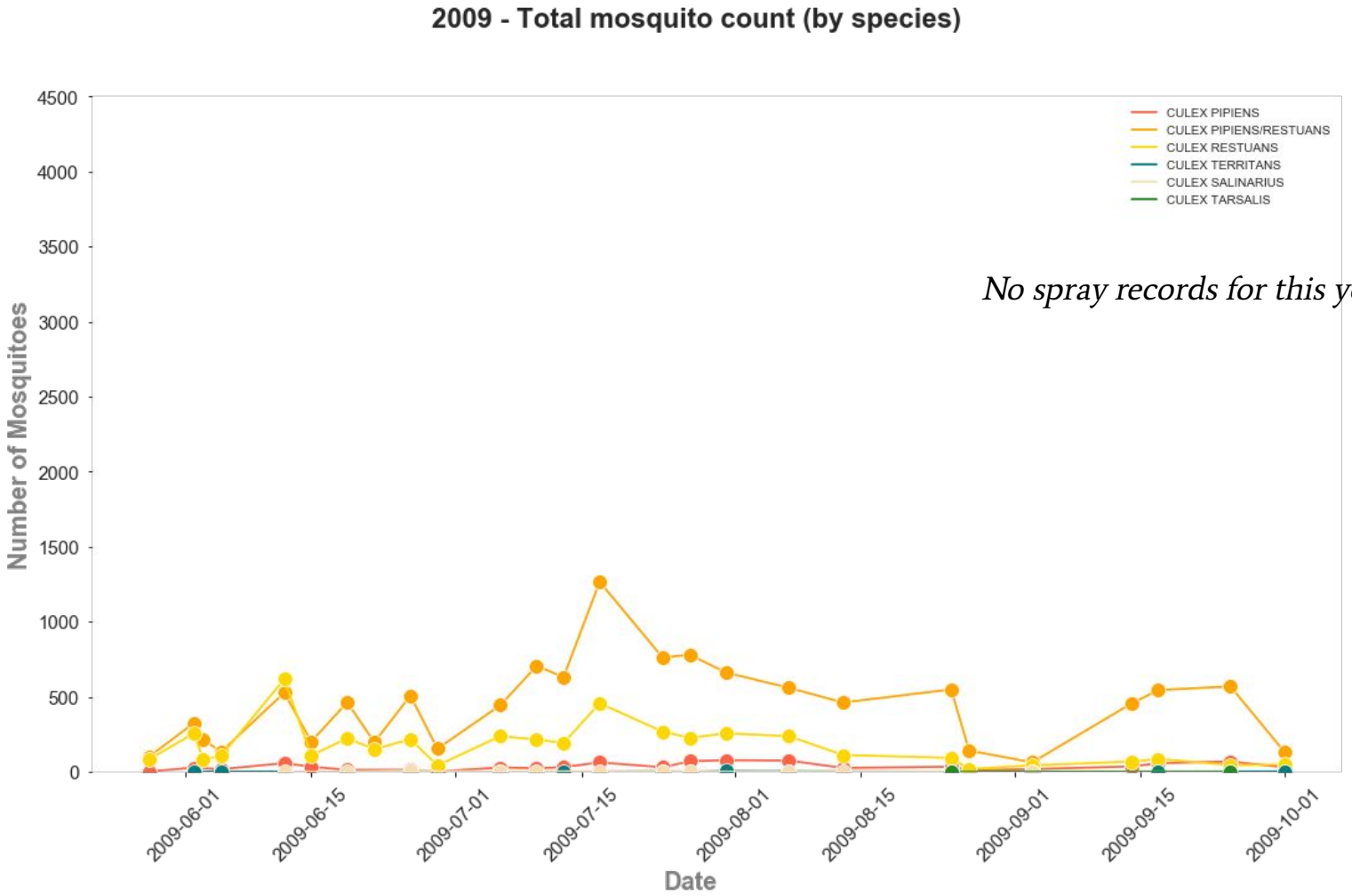
- Medical and hospitalisation fees could amount to up to $250 million per year

- Vector control measures - Spraying insecticide (spraying procedure and overtime hours)

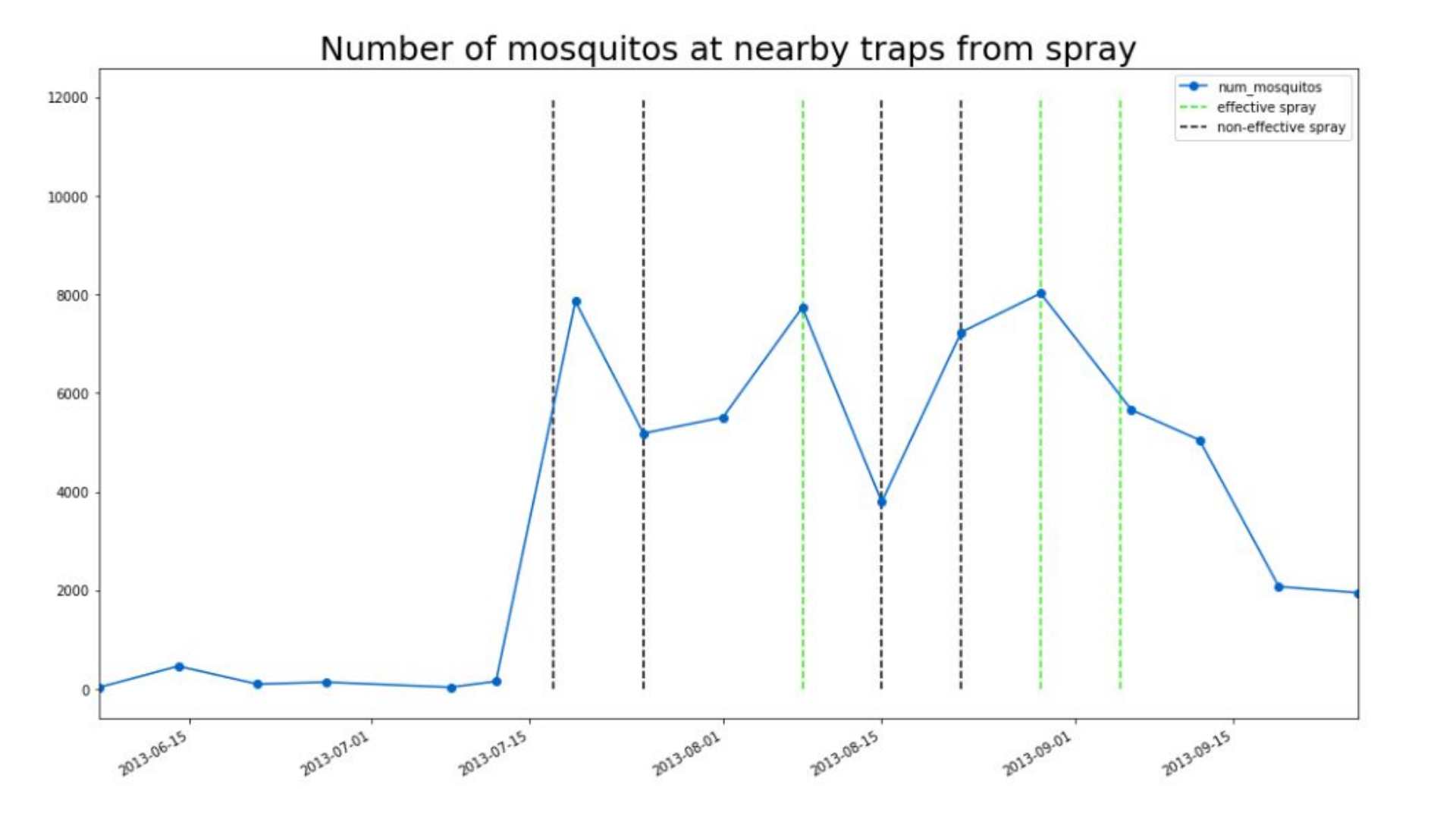
2007 - Total mosquito count (by species)

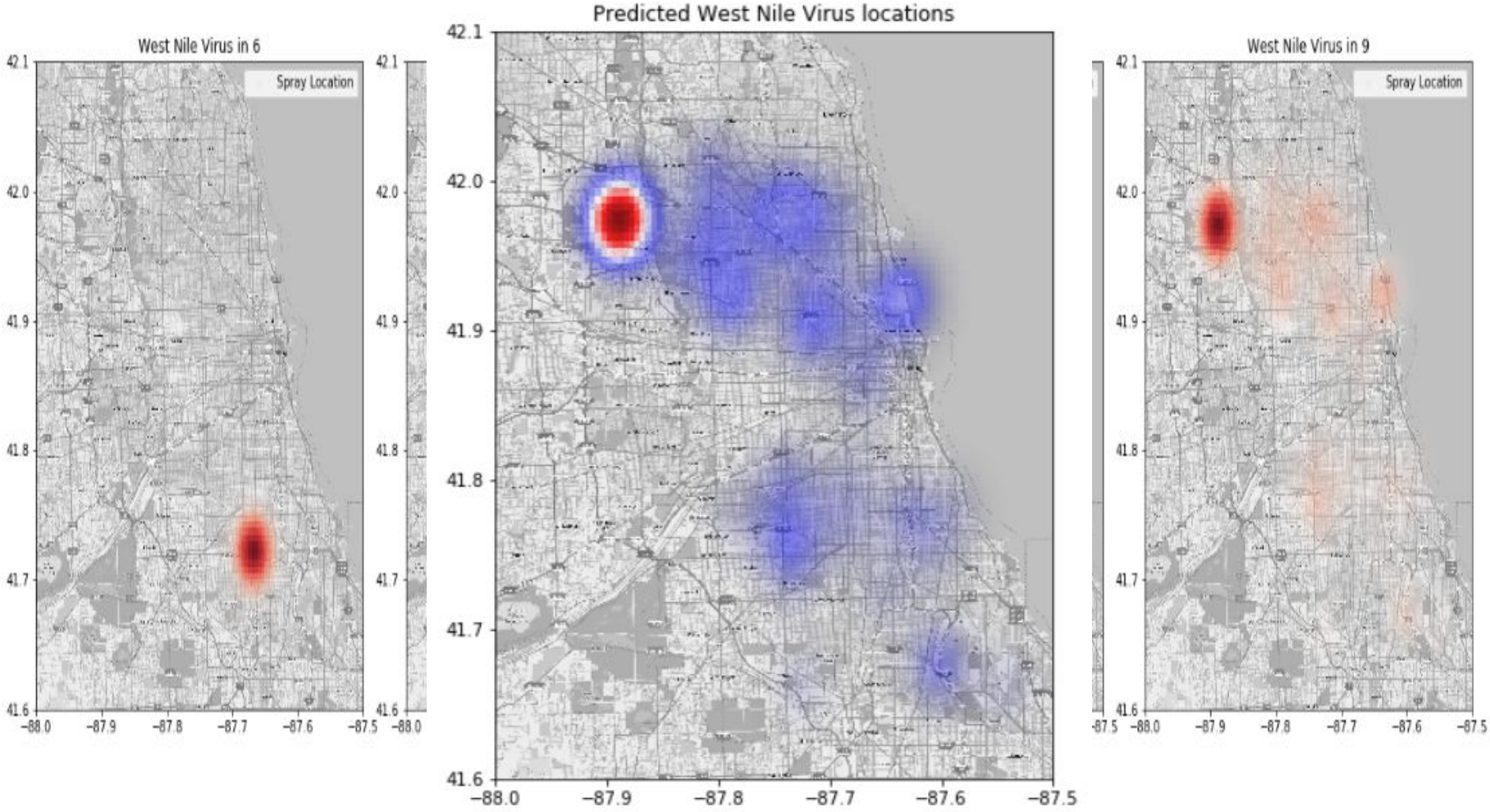


No spray records for this year



No spray records for this year


Number of mosquitos at nearby traps from spray
num_mosquitos
effective spray
non-effective spray
12000
10000
8000
6000
4000
2000
0
2013-06-15
2013-07-01
2013-07-15
2013-08-01
2013-08-15
2013-09-01
2013-09-15


West Nile Virus in 6
Spray Location
Predicted West Nile Virus locations
West Nile Virus in 9
Spray Location
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

Limitations | Recommendations

Overlapping Clusters

- Might not be very definitive if that particular point has been sprayed

Practice and Exercise Societal Responsibility

Consider including bird migratory data for affected birds