Predicting West Nile Virus

Andy, Hong Yee, Ivan, Jeremy

Agenda

1. Problem Statement
2. Exploratory Data Analysis
    a. Correlation graphs (number of mosquitos with weather, Wnv present etc)
    b. Probabilities of each species getting the virus
    c. Trends by year/week, like sunrise sunset
    d. Other interesting findings
        i. Urgent clusters to fog identified
3. Modelling and Results
4. Cost and Benefit
5. Conclusion and Recommendation

Problem Statement

- Recent West Nile Virus outbreak from mosquitos in Chicago City.
- Seeking a cost effective plan to make predictions on when and where to spray pesticides in the City of Chicago.
- This is a binary classification problem which predicts whether different species of mosquitoes will be tested positive for West Nile virus, based on where and where they were captured

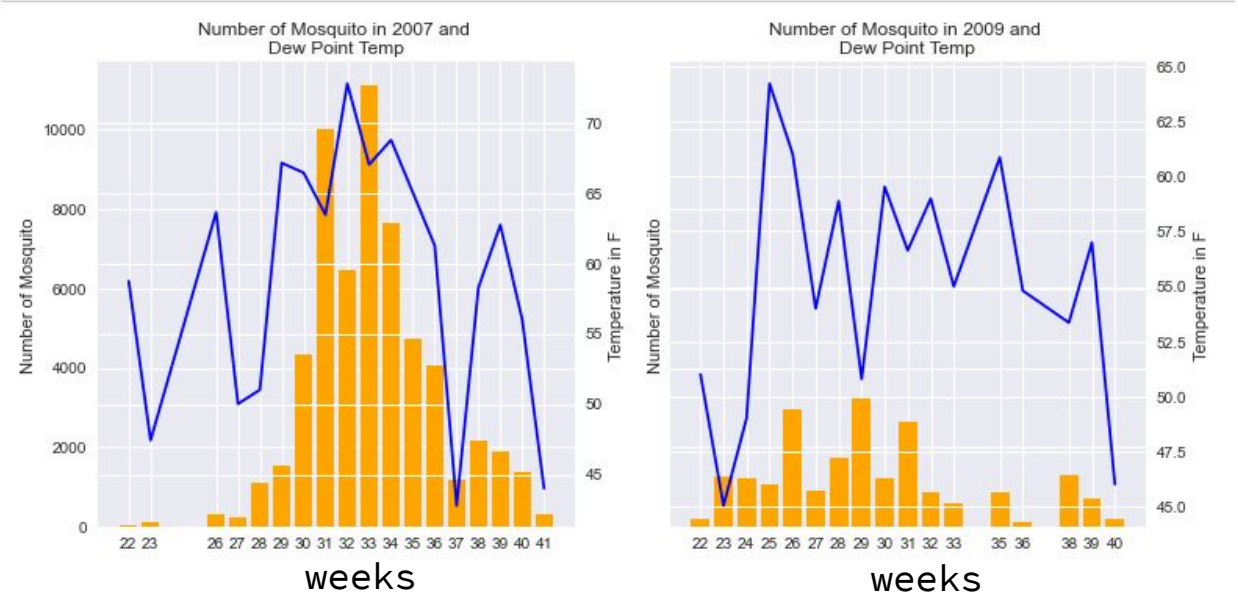
- Chicago Department of Public Health (CDPH), CDC, General Public of Chicago City

Exploratory Data Analysis

Findings from Exploring Data

Temperature

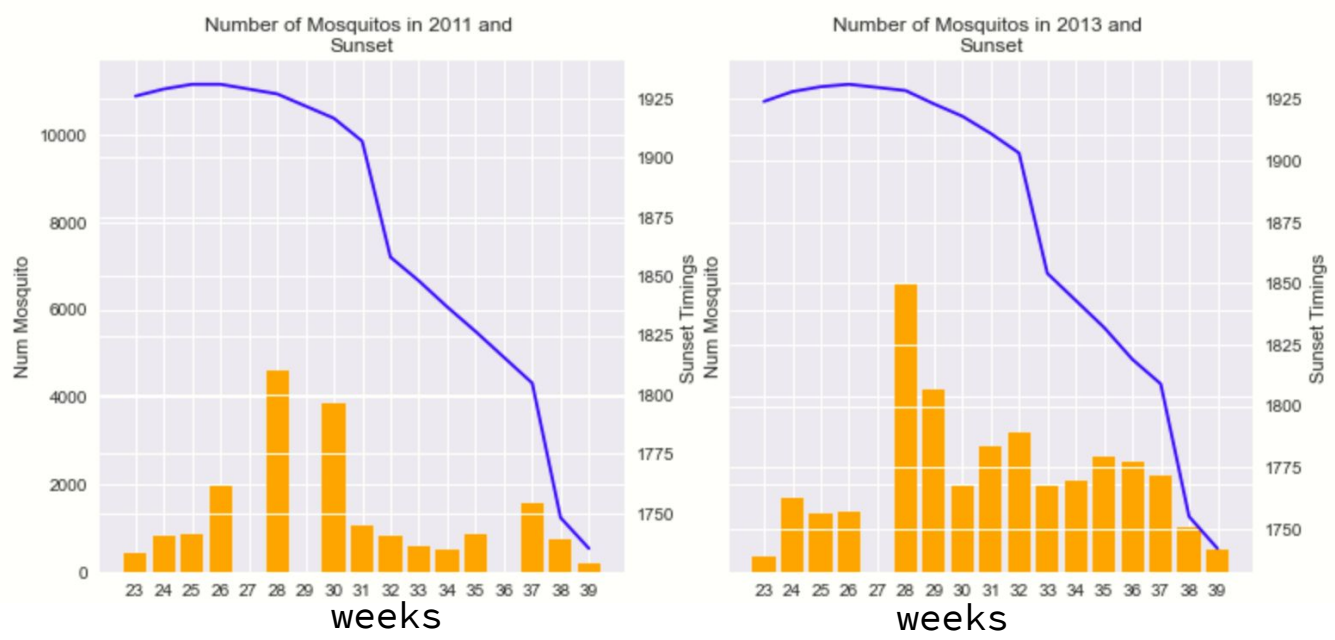
Number of Mosquitos



Mosquitos Likes Longer Days

Sunset timings

Less Mosquitos towards the end of the year, where days are shorter, night time is longer.

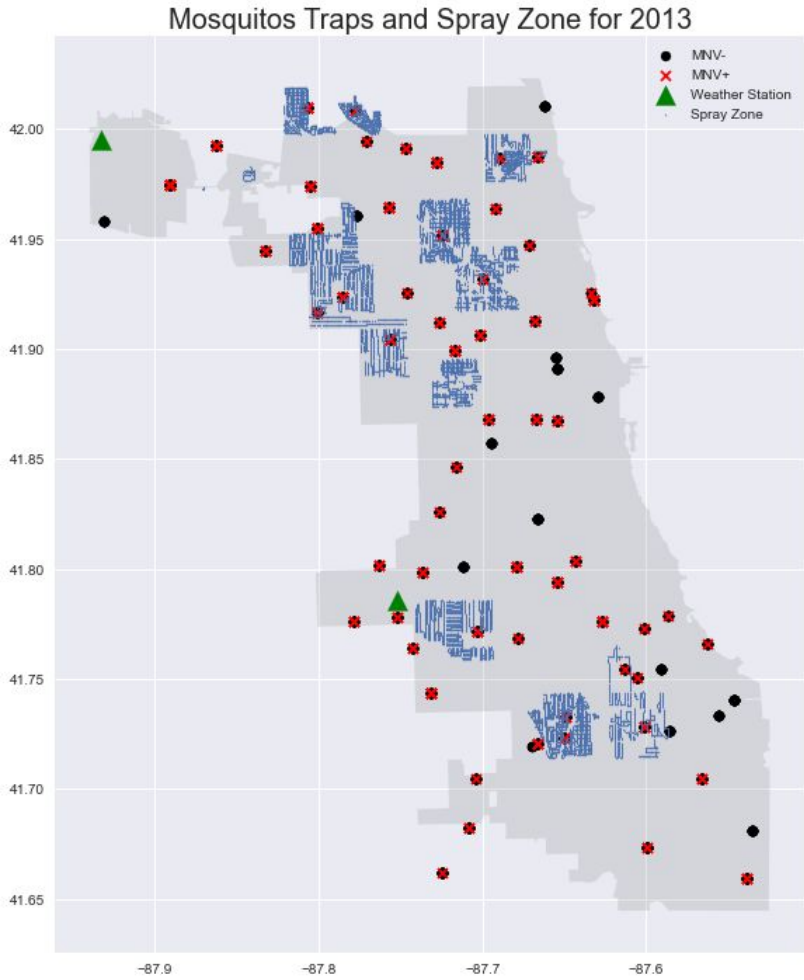


2013 Traps and Spray Zone

In 2013, WNV was found in most traps across the city.

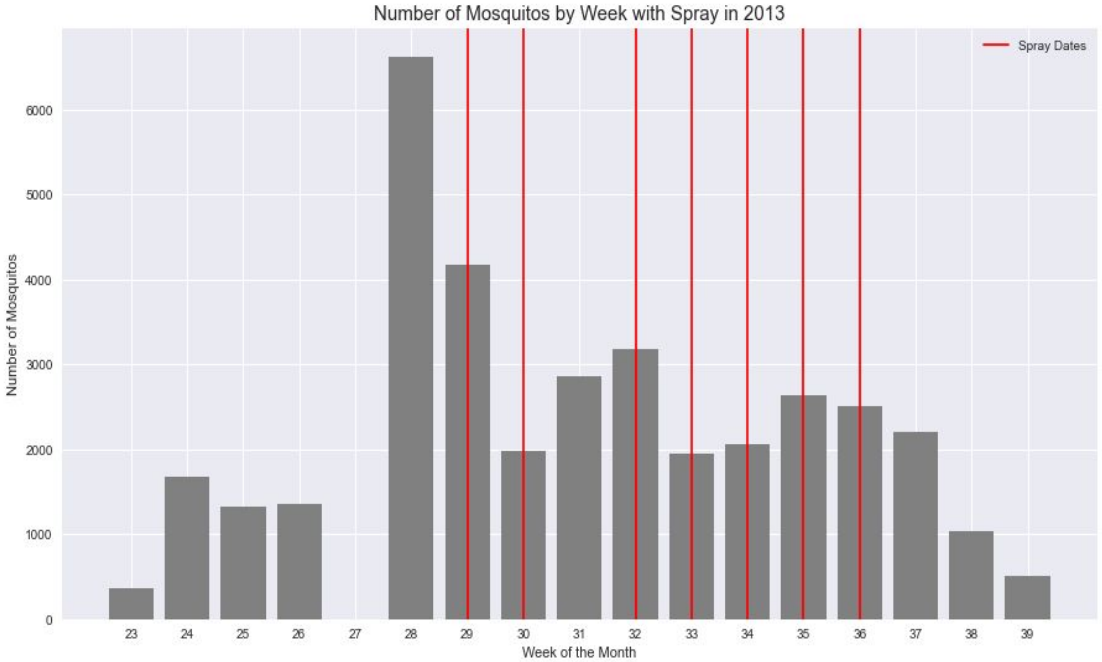
The northern area seems to be the hotspot for WNV.

Most of spray is deployed in this region.



To Spray or not to Spray?

Spraying prove to be quite **effective** generally

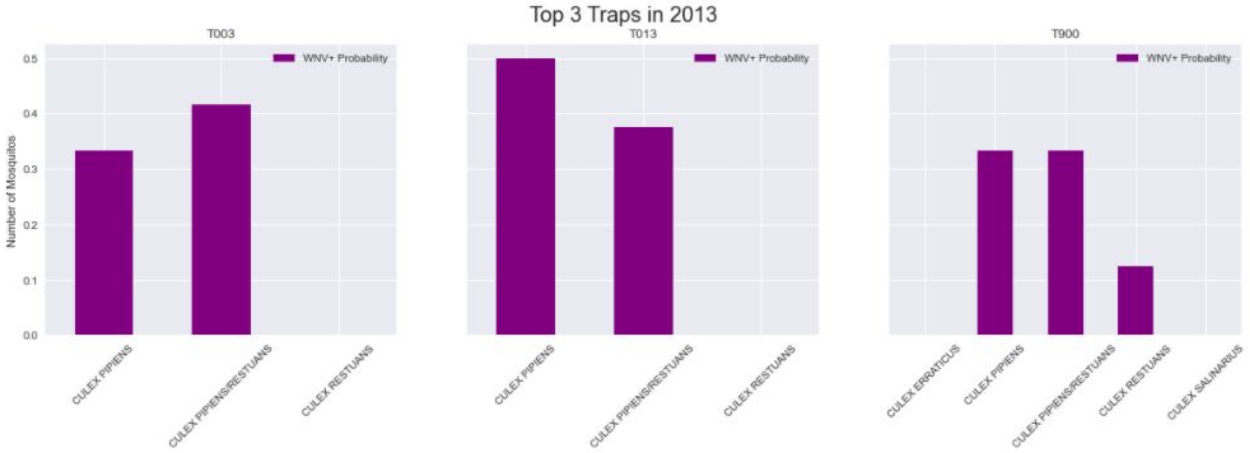


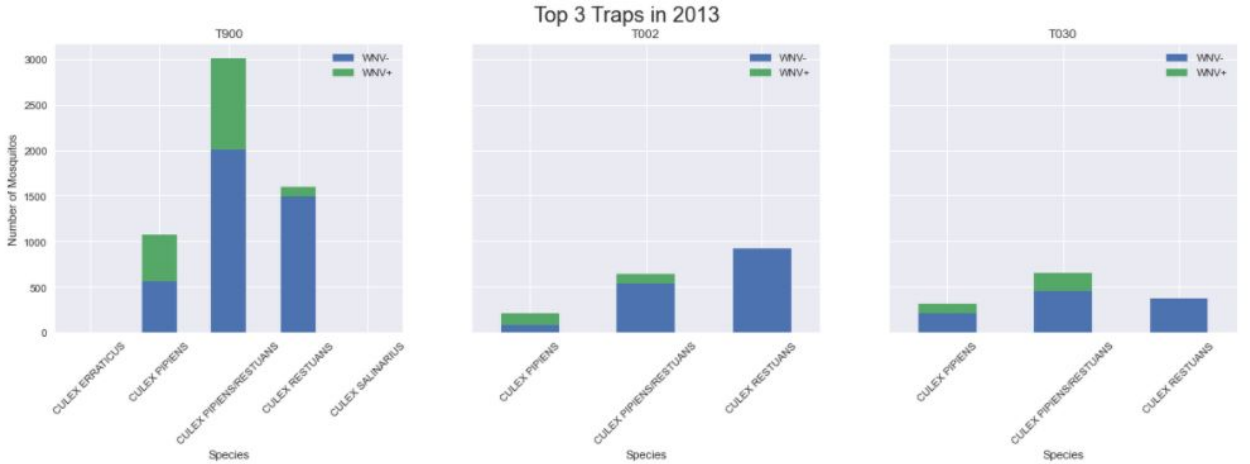
Identify Urgent Clusters

Highest Probability

+

Highest Population





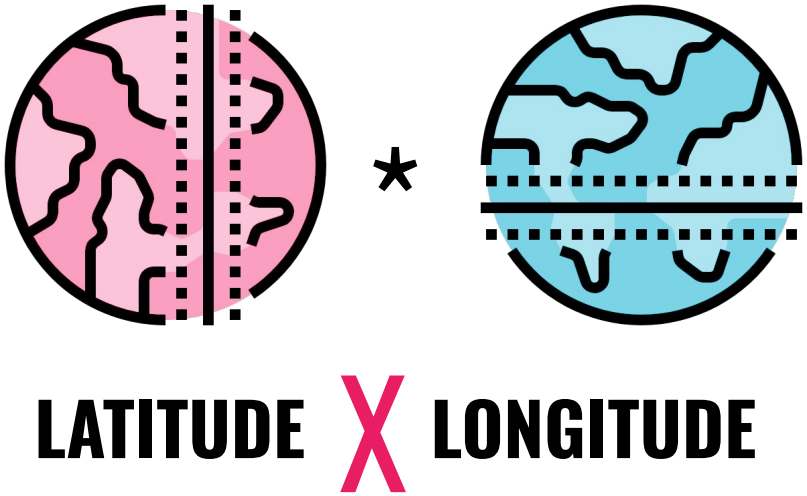
Urgent Clusters Based on 2013 Data EDA

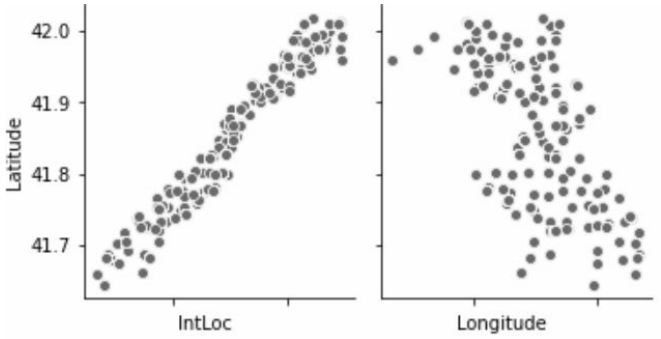
T900, T002, T030, T008, T225,
T066, T233, T013, T028, T228

What will our Model say?

Modeling

Preprocessing/Feature Engineering

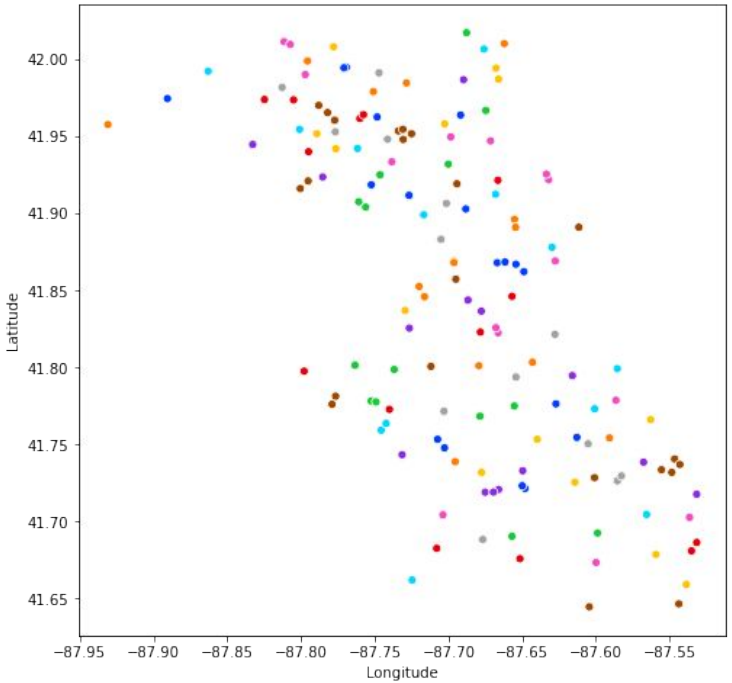


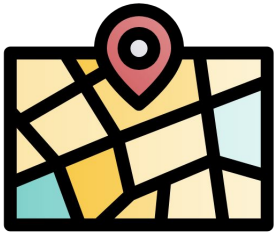


How to describe Location as a feature?

https://www.flaticon.com/free-icon/equator_1899102?term=latitude&page=1&position=6

Preprocessing/Feature Engineering





CLUSTERING

LOCATIONS USING DBSCAN

https://www.flaticon.com/free-icon/map-location_3203013?term=location&page=1&position=35

Preprocessing/Feature Engineering

Mosquitoes take ~ 2 weeks to breed

Weather fluctuations from season changes

Habitability for mosquitoes breeding season

MEAN WEATHER

DATA OF 14, 30, 90 DAYS

https://www.flaticon.com/free-icon/weather_2383684?term=weather&page=1&position=84

Feature Selection

LESS IS MORE
WHEN GENERALIZING

- **Manually Select Features (human brain)**
- **Recursive Feature Elimination (DTC, SVC)**
- **Cross Validated Recursive Feature Elimination (DTC, SVC)**
- **Principal Component Analysis**
- **SelectKBest (f classif)**

6 Combinations of features programmatically selected

MANY Combinations of manual selections

Manually selecting 38/188 features was the best

https://www.flaticon.com/free-icon/scissor_1803244?term=scissors&page=1&position=61

Model Selection

	Train	Validation
1. Logistic Regression	75	66
2. K Neighbors Classifier	94	67
3. Decision Tree Classifier	99	64
4. Random Forest Classifier	99	65
5. Extra Trees Classifier	99	65
6. XGBoost	79	76
7. Scalar Vector Classifier	89	72
8. MultinomialNB	75	66
9. Gradient Boosting Classifier	89	77

GENERALIZING

CLOSEST NUMBERS WINS

https://www.flaticon.com/free-icon/general_286642?term=general&page=1&position=20

Final model: XGBoost

BEST PARAMS : XGBOOST

- eval_metric='auc'
- subsample = 1
- colsample_bytree = .15
- learning_rate = .05
- max_depth=3
- scale_pos_weight=19
- n_estimators=500
- reg_alpha=.9
- reg_lambda= 5
- gamma=0.01

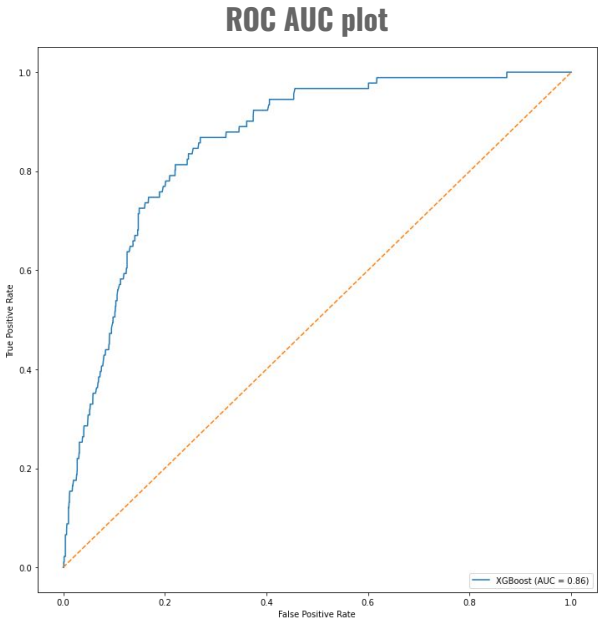
SMOTE CERTIFIED

ROC AUC

0.699

TEST SCORE

WE ARE ABLE TO **DISTINGUISH WNV+** AREAS WITH 70% CONFIDENCE



Feature Importance

rank	feature	cumulative	importance
1	species_CULEX RESTUANS	0.134	0.134
2	Sunrise	0.199	0.0649
3	Week	0.2588	0.0599
4	Month	0.3161	0.0572
5	species_CULEX PIPIENS/RESTUANS	0.372	0.0559
6	Sunset	0.427	0.055
7	species_CULEX PIPIENS	0.4774	0.0504
8	species_CULEX TERRITANS	0.5187	0.0412
9	code_ra	0.5566	0.0379
10	Year	0.5875	0.0309
11	ResultSpeed	0.6177	0.0302
12	code_vcts	0.6469	0.0292
13	AvgSpeed	0.6753	0.0283
14	Tmax	0.702	0.0267
15	SeaLevel	0.7267	0.0247
16	Cool	0.748	0.0213
17	Tavg	0.7687	0.0207
18	Tmin	0.7887	0.02
19	code_ts	0.8081	0.0194
20	code_br	0.826	0.0179

SPECIES OF 

WEEK OF 

TYPE OF 

https://www.flaticon.com/free-icon/calendar_306804?term=calendar&page=1&position=64 https://www.flaticon.com/free-icon/seasons_1405877?term=season&page=1&position=8 https://www.flaticon.com/free-icon/mosquito_1779544?term=mosquito&page=1&position=9

Cost and Benefit Analysis

Cost and Benefit Analysis

Costly :

Additional Spraying Cost:

- Spraying when risk of WNV is low or present

Economic Cost:

- Not Spraying when risk of WNV is high
- Hospitalisation Cost and Productivity suffer

Beneficial:

Save Spraying Cost:

- Targeted Spraying at Urgent Clusters

Lower Economic Cost:

- Less people with WNV-related illnesses

Case study from California, Sacramento

Cost of spray

Man hours = $41,790

Pesticide cost = $660,000

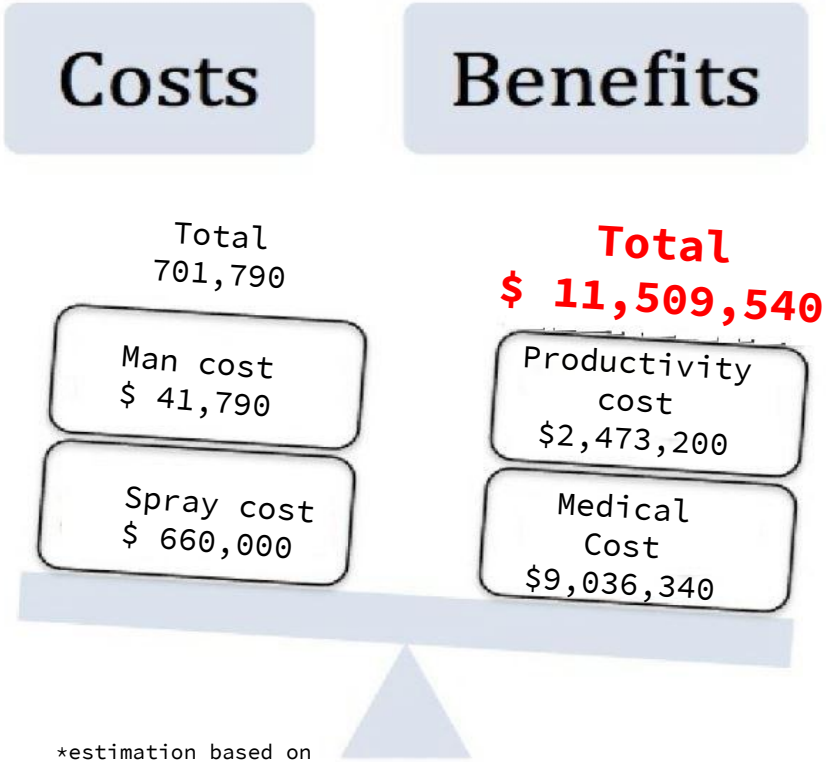
Monetary loss due WNV

Productivity cost/pax = $10,800

Medical cost/pax = $39,460

Total cost/pax = $50,260

Estimated for 229(ref yr 2012) cases = $ 11,509,540



*estimation based on
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7241786/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322011/

Cost Benefit Analysis of Prevention in Chicago

Cost

Prevention cost estimated $5.3 million over 56 districts, $9,000 on average per square mile.

- Spraying cost
- Raising Awareness (education, public campaigns etc)

Environmental impact/cost

- Affect crops and other animals

Benefit

Reduce Burden on Healthcare System and on People - ~$30,000 - 40,000 per person.

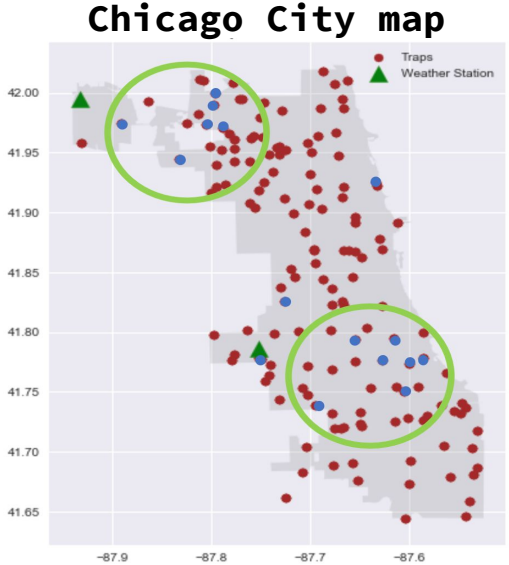
People feel confident and assured to be outdoors.

Preventing an outbreak

- Cost $2.98 million in California

Conclusion and Recommendations

Answers to our problem statement

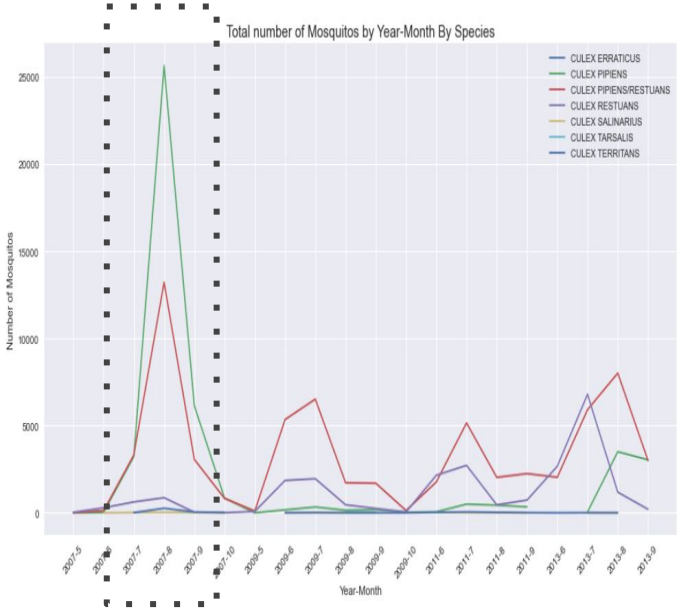


Where?

Culex Pipiens

Culex Restuans

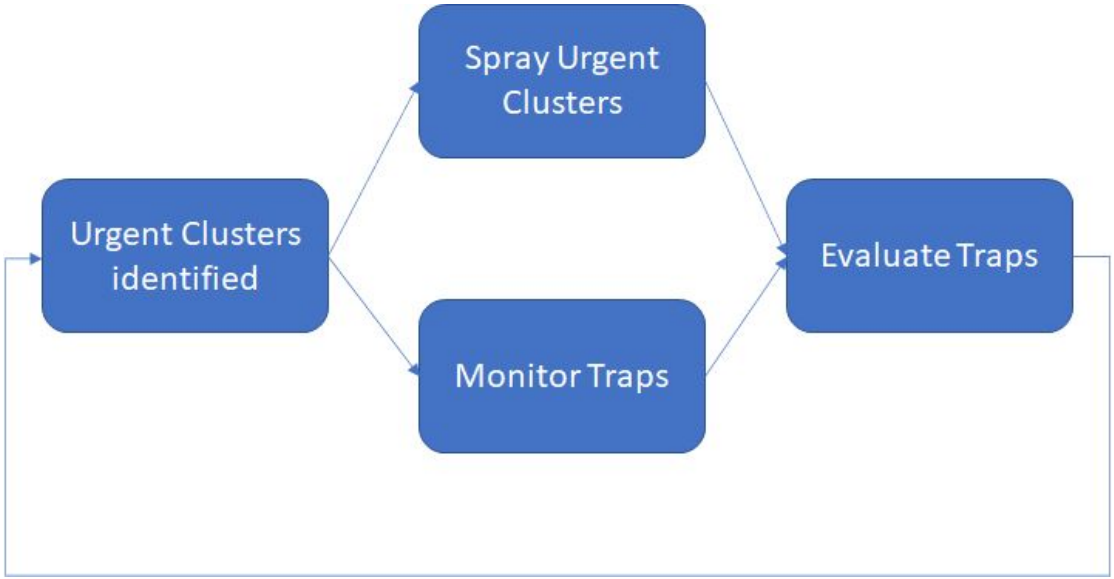
Species



When? Summer July and August

Recommendation for Spraying method

- A systemic-approach to Spraying, rather than a catch-all approach



Recommendations to general public

Use insect Repellent

Wear long sleeve shirt

Remove standing water

Fit windows and doors tightly

Future plans

- Moving forward, the model can be continuously improved by introducing more new data.
- Incorporate inhabitants population density of areas into the dataset in order to assess the importance of eradicating mosquitoes.
- Data from other mosquito-borne illness such as more threatening Zika virus or Aedes aegypti can also be added to expand the model to not limited just to West Nile virus use
- Expand the model to other mosquito infested cities

Limitations

- Our model has a ROC AUC test score of 0.699 hence there will be risk and cost incurred for falsely spraying areas (ones that are not infested).

- This modeling process only works for binary classification. For Multi classification problems, the process will have to be modified and reevaluated.