Project 4:
West Nile Virus Prediction

Develop a data-driven classification model to pinpoint locations susceptible to West Nile Virus infections in the City of Chicago

By: Yong Khiang, Kayle, Ben, Yun Jie

Contents

- Problem Statement
- EDA
- Modelling
- Cost Benefit Analysis
- Conclusion

West Nile Virus

- First reported case in the United States in 1999
- Leading mosquito-borne disease
- Virus outbreak during seasonal months (May to October)
- Mostly mild symptoms (fever, headache, and body aches, skin rash and swollen lymph glands)
- Can cause life-threatening conditions that include inflammation of the brain and spinal cord
- Integrated Vector Management to curb mosquito numbers

Problem Statement

What is our goal?

- Seasonal outbreaks of the infectious West Nile Virus (WNV) calls for a need to pinpoint the **location** and **time** of spread to identify susceptible areas in the City of Chicago.

Why the need to?

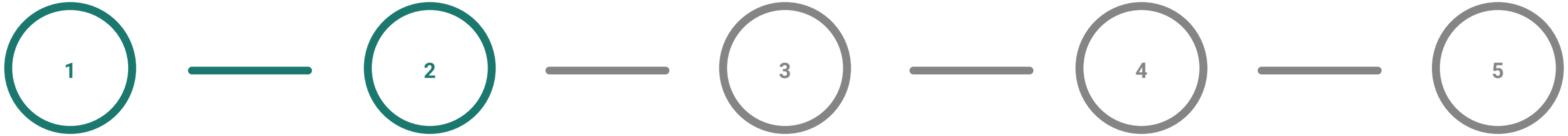
- We aim to assist the **Chicago Department of Public Health(CDPH)** in making **well-informed** and **cost-effective decisions** in the allocation of resources to such vulnerable locations

Problem Statement

How do we achieve that?

- The implementation of an optimized **ExtraTrees Classifier** model, evaluated through **ROC-AUC**, **recall** and **precision** scores

Methodology



Business Understanding

Develop and optimize a classification model to predict WNV in areas of Chicago

Data Analysis

- Imbalanced data
- Remove duplicate
- Missing data from features (Depth, Snowfall)

Feature Selection and Model Optimization

- Feature engineering using external research (Night hours, Humidity)
- Train models
- Tune hyperparameters

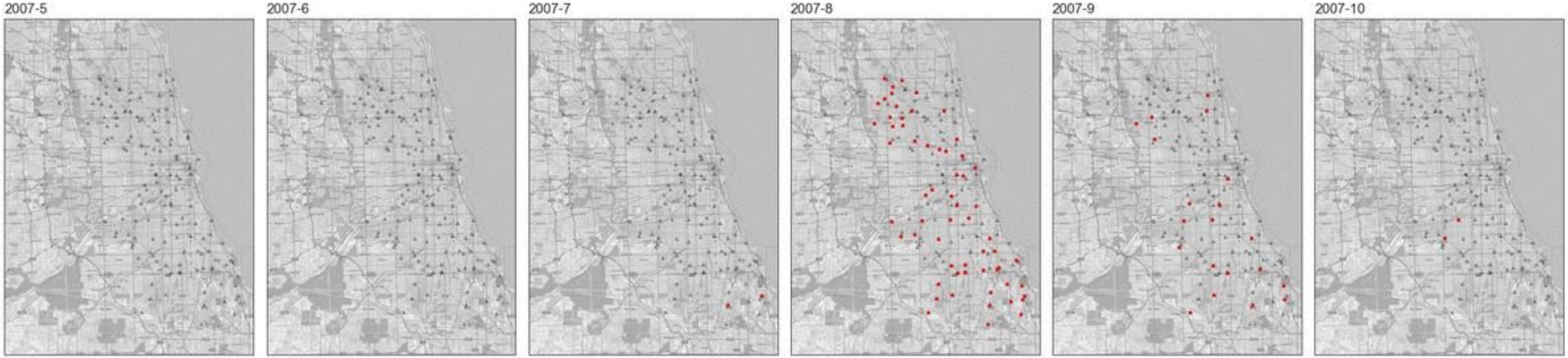
Cost Benefit Analysis

- Find out if sprays are effective in controlling number of mosquitoes
- Weigh costs and benefits associated with using sprays to combat WNV

Conclusion

Provide recommendations

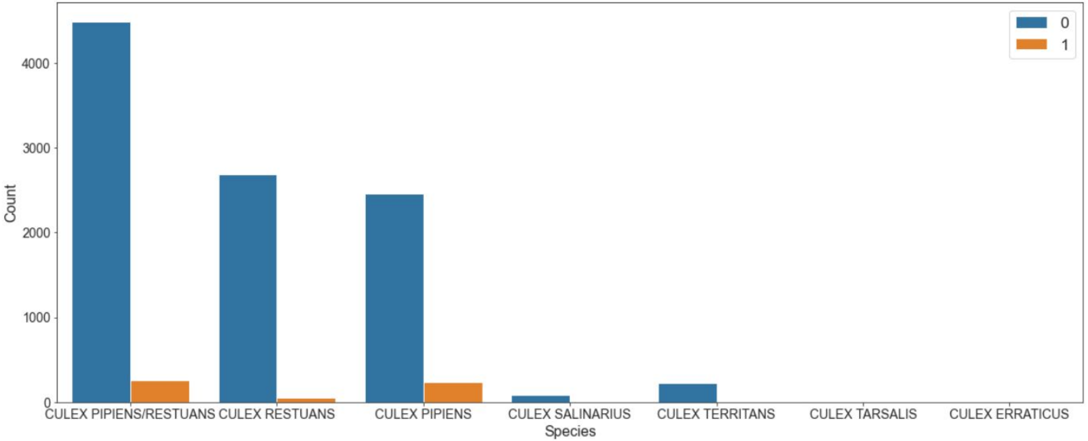
Data Visualization



Data Visualization

Mosquitoes Species:

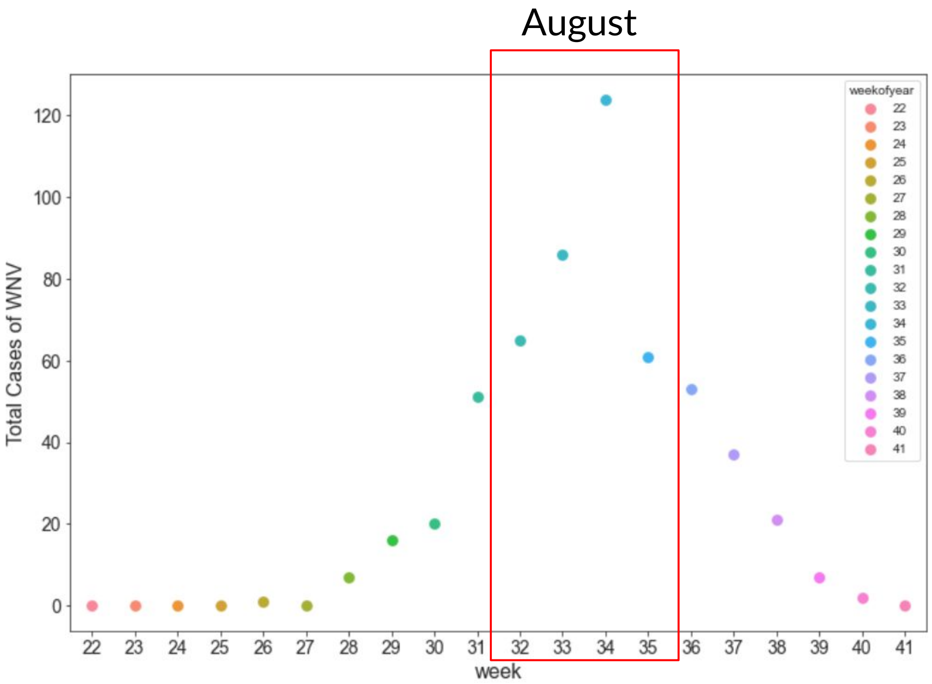
- Commonly found mosquito species: Culex Pipien/ Restuans, Culex Restuans and Culex Pipiens
- Very small proportion of WNV cases



Findings and Insights

1) Monthly seasonality in outbreaks:

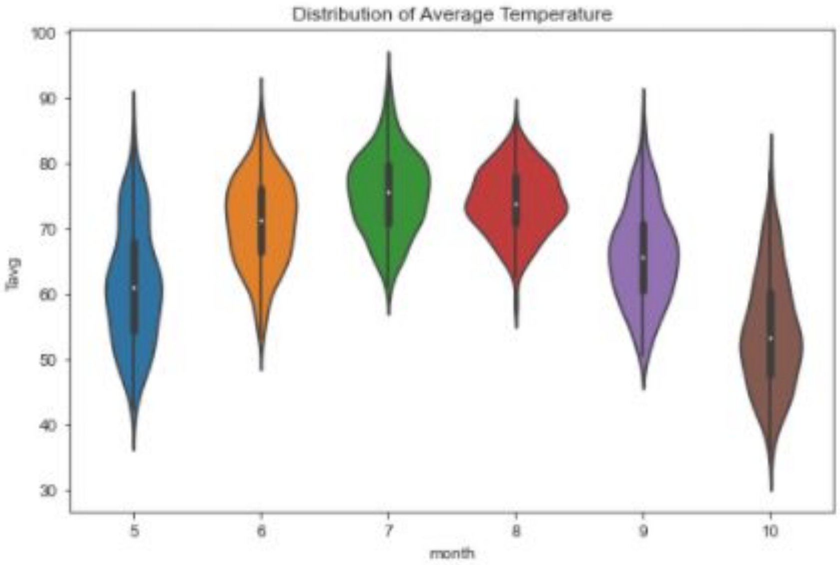
- Highest WNV cases in **Aug(3rd Week)** across all years
- Cyclic nature of outbreaks



Findings and Insights

2) Distribution of average temperature

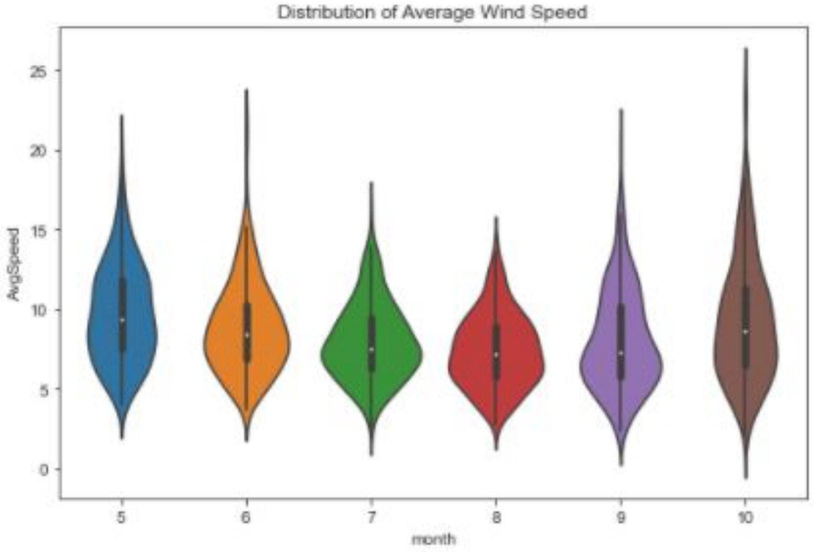
- Peak temperature in July
- Lagging effect of mosquito outbreak in August due to mosquito embryonation process



Findings and Insights

3) Distribution of wind speed

- Low wind speed in July/ August
- Corresponds to mosquito peak season in August

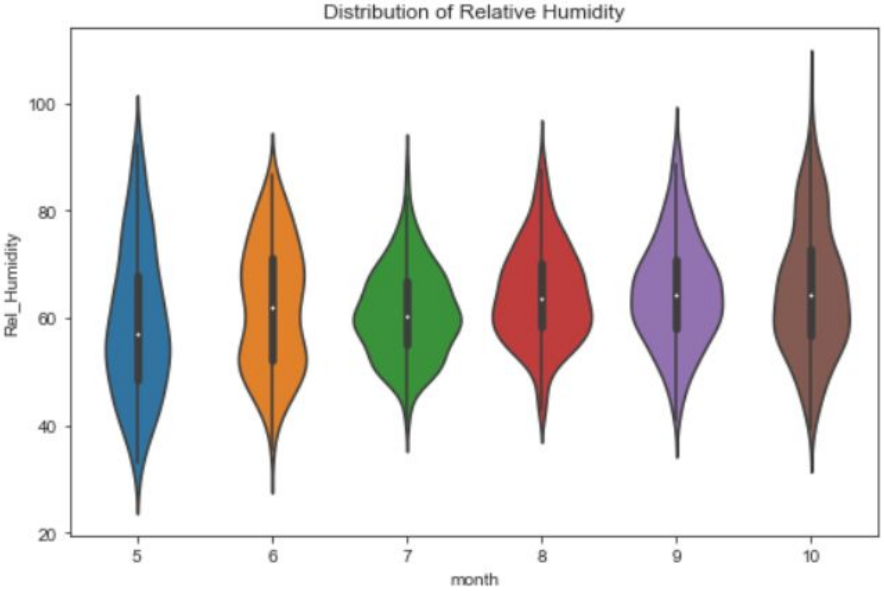


Findings and Insights

4) Distribution of relative humidity

- Slight increase in mean relative humidity going from July to August

- This increase is correlated to the increase in WNV cases



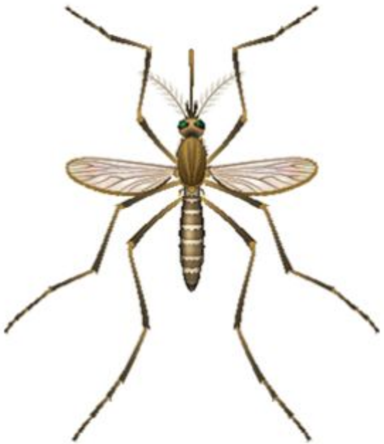
Findings and Insights

5) Length of Night hours

- Night Hours = (Sunset time) - (Sunrise time)
- Studies on *Aedes* species have shown that they prefer to lay eggs **at night**
- Possible effects of night hours on *Culex* species

Factors conducive to mosquito growth

1. Hot and Dry Temperature
2. High Humidity
3. Long Night Hours
4. Low wind conditions



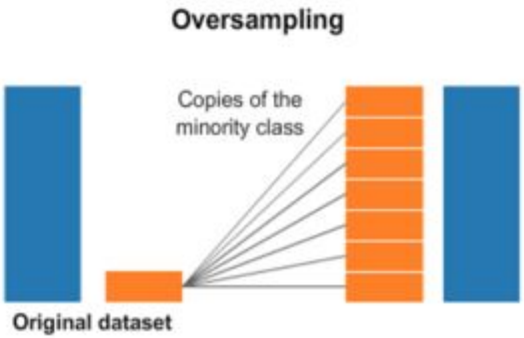
Modelling Approach

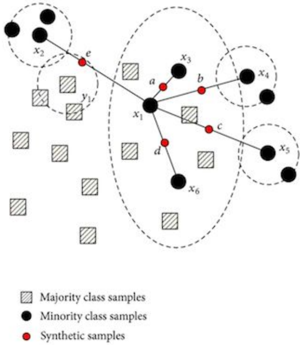
Addressing Imbalanced Data
(5% positive class):

- Oversampling using SMOTE
- Stratify target variable

Optimising Model:

- Grid search on best hyperparameters to optimise AUC metric



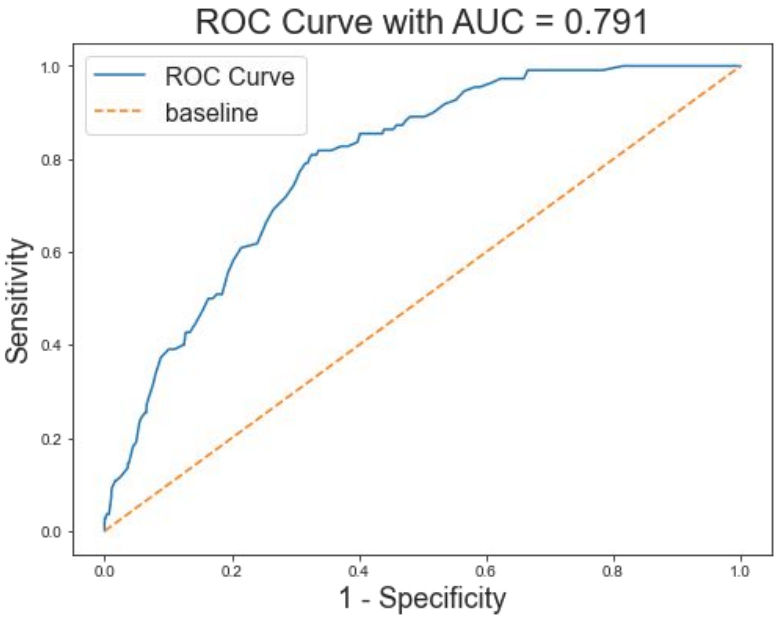


Model Evaluation

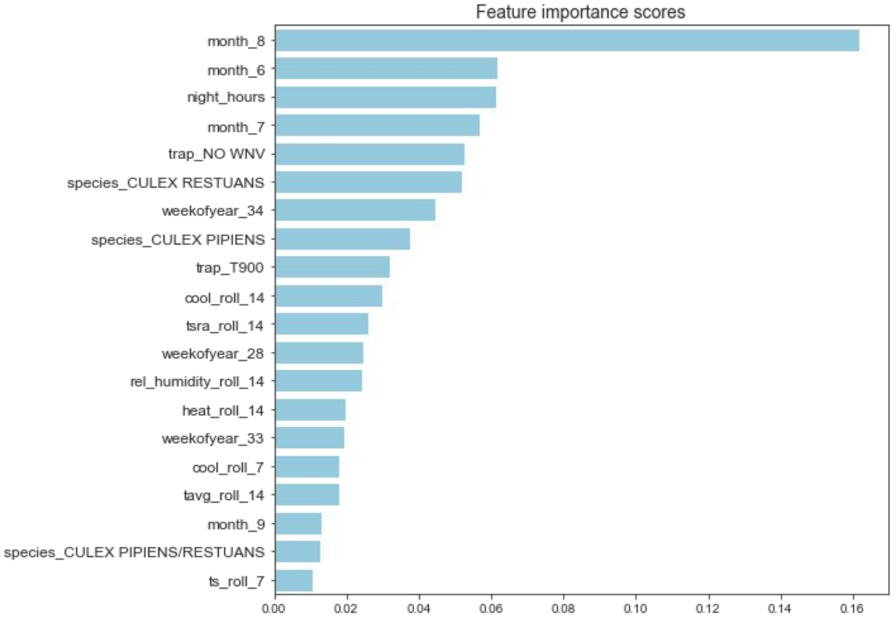
	Logistic Regression	AdaBoost Classifier	Random Forests	Extra Trees Classifier	Gradient Boosted Trees	XG Boost
CrossVal: ROC-AUC	0.813	0.841	0.826	0.827	0.838	0.847
Train Set: ROC-AUC	0.875	0.855	0.839	0.841	0.945	0.923
Test Set: ROC-AUC	0.809	0.827	0.797	0.794	0.830	0.835
Recall	0.664	0.345	0.664	0.745	0.136	0.445
Precision	0.142	0.181	0.142	0.123	0.224	0.189

Extra Trees Classifier Performance

	Predict WNV Negative	Predict WNV Positive
Actual WNV Negative	70.6%	29.4%
Actual WNV Positive	25.5%	74.5%



Feature Importance



- Top feature importance:
  - Seasonal Months
  - Night hours duration
  - Mosquito species (Restuans and Pipiens)
  - Trap locations without WNV presence and Ohare Airport hotspot

Limitations

1) Inclusion of additional data:

   > **Number of mosquitoes** caught has a strong correlation with presence of wnv but was absent in the test set

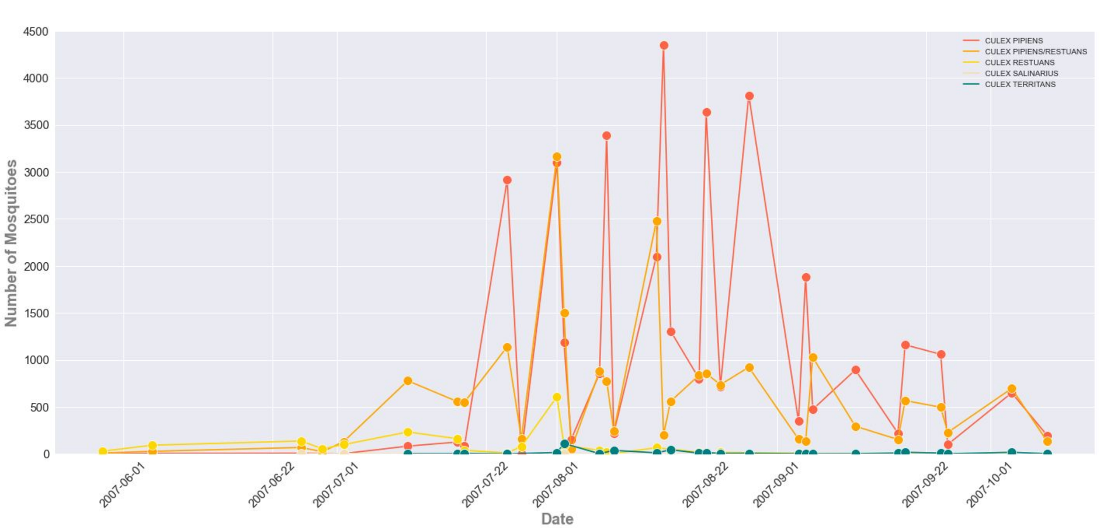
2) Inconsistent findings in external studies:

   > Contradictions in impact of **precipitation** on mosquito survival

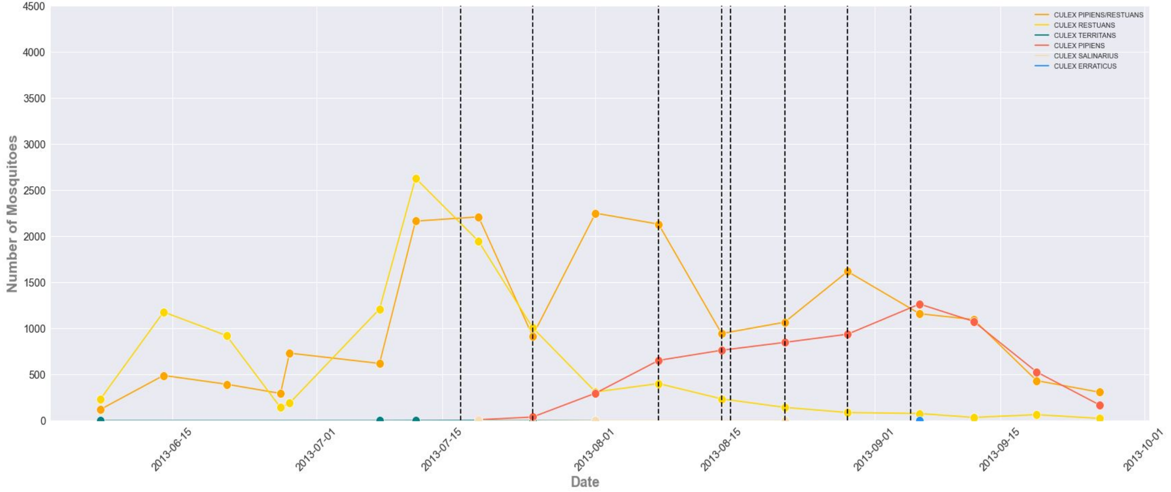
3) Effect of Global Warming

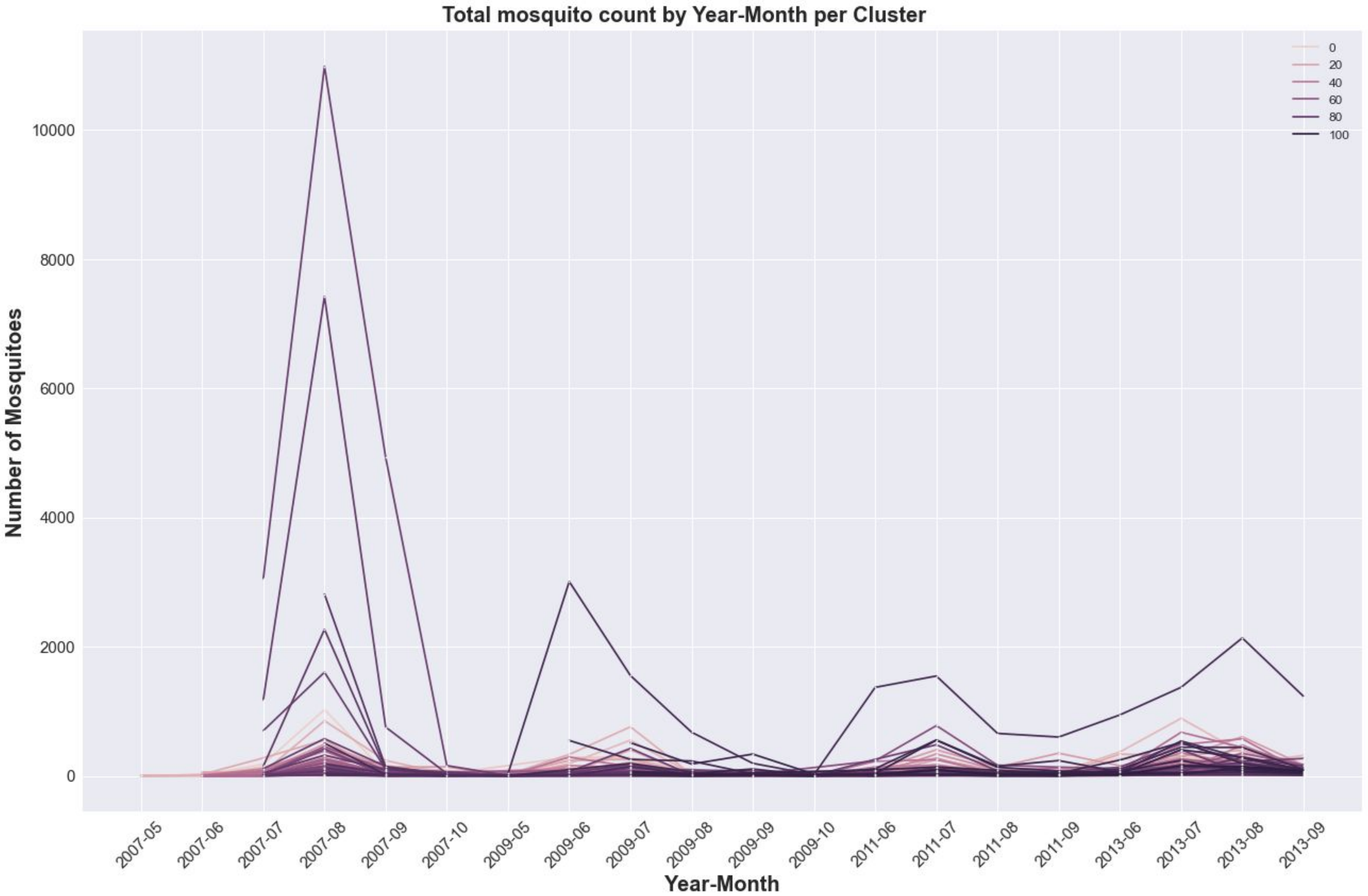
Cost Benefit Analysis

2007 - Total mosquito count (by species)



2013 - Total mosquito count (by species)




Total mosquito count by Year-Month per Cluster
0
20
40
60
80
100
Number of Mosquitoes
0
2000
4000
6000
8000
10000
Year-Month
2007-05
2007-06
2007-07
2007-08
2007-09
2007-10
2009-05
2009-06
2009-07
2009-08
2009-09
2009-10
2011-06
2011-07
2011-08
2011-09
2013-06
2013-07
2013-08
2013-09

Cost Benefit Analysis

Costs (using pesticide Zenivex E4):

1 sprayer truck:
$844 - $1688 for area of 0.6 km^2

~1000 trucks to cover Cook County (606.1 km2)
Total Cost: $844 000 - $ 1 688 000
Total spend in 2013 was also about $800 000

Benefits:
Fewer people dying/falling ill —> increased workplace productivity and healthcare savings

120 more predicted WNV cases in 2013:
total income loss ~ -$20,000 from sickness
medical bill ~ -$6 300 per pax, $ 756 000 total

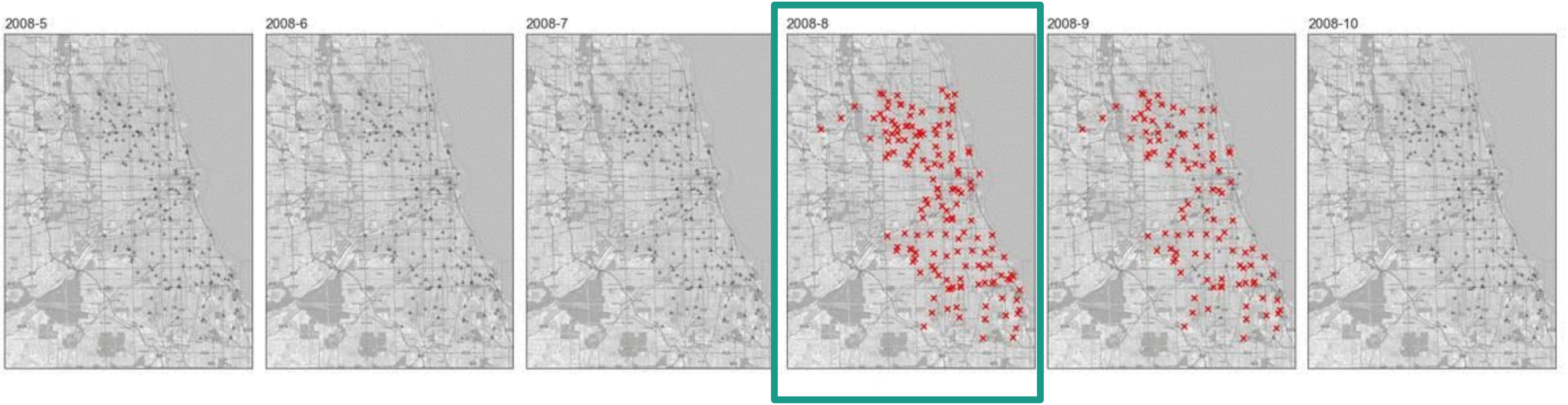
Critical: spray early to keep mosquito numbers less than 2000 per trap.
No sprays after September

Conclusion and Recommendations

Concentrate spraying efforts

- Time: July
- Locations: Identified locations in red

Spray periodically during seasonal month (May to September) to keep mosquitoes numbers low



2008-5

2008-6

2008-7

2008-8

2008-9

2008-10