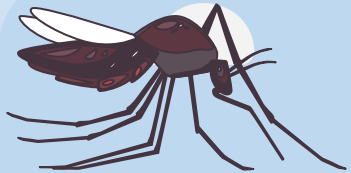
Project 4: West Nile Virus Prediction



Chan Haosheng Timothy
Kho Guan Guo
Soh Sze Ron

Disease & Treatment Agency

Societal Cures in Epidemiology &
New Creative Engineering

Table of contents

01 Background & Problem Statement

02 Data Cleaning

03 EDA

04 Modelling

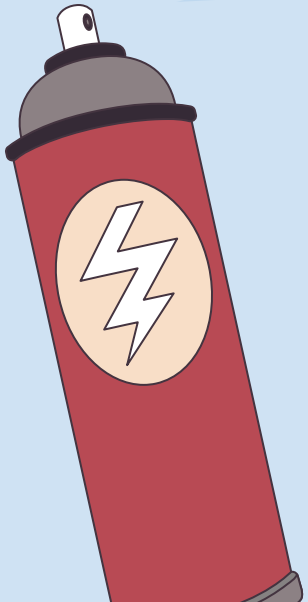
05 Cost-Benefit Analysis

06 Conclusion & Recommendations

Background

- In 2002, the first human cases of West Nile virus were reported in Chicago.
- West Nile virus, is typically spread by mosquitoes.
- Cost of treatment for severe symptoms developed from WNV can be staggeringly high.
- Preventative measures like spraying of pesticides, can also be costly and potentially harmful to the environment.
- It is thus essential to manage the spread of WNV efficiently and effectively.

Problem Statement



- Team of data scientists at DTA
- Understand key features for making accurate predictions
- Predict areas where WNV mosquitoes can be detected
- Optimise pesticide spraying effort
- Formulate plan to deploy pesticides throughout the city

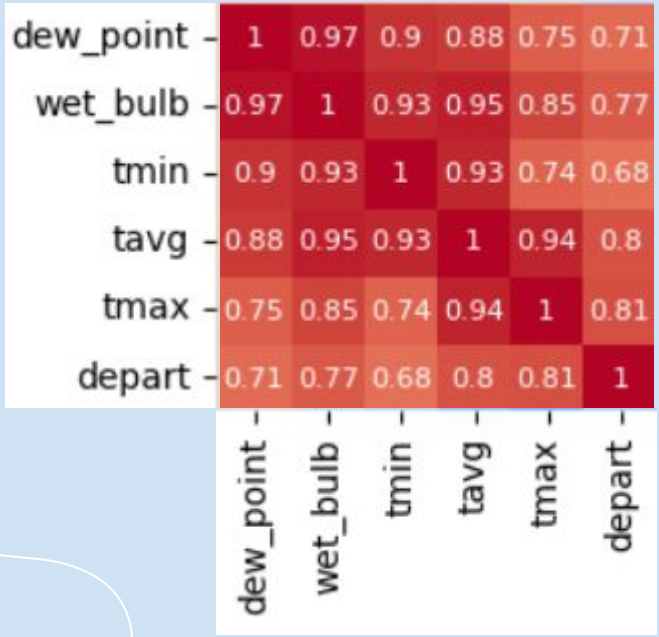
Data Cleaning

- Missing data: Imputing weather data by deriving values from available data.
- Duplicates: Number of mosquitos, dedupe < 50 and aggregate remaining.
- Duplicates: Spray, remove
- Columns: Similar categories/features not useful dropped

EDA

- Correlation
- Periods
- Further investigation of periods (Average Monthly Temperature by Year)
- Species
- Traps
- Barplots of various features
- Location

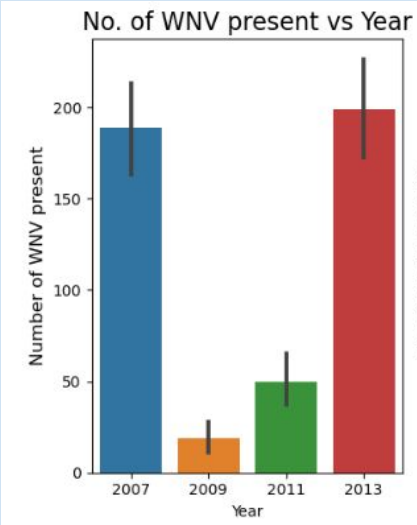
Correlation

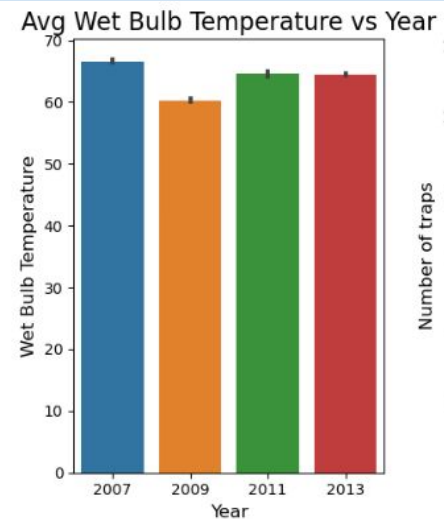


- Heatmap: Spearman Correlation*
- Temperature related all highly correlated to each other. To only choose one of them as a feature.
- High linear correlation between temperature features.

* Refer to Notebook for full visuals

Periods

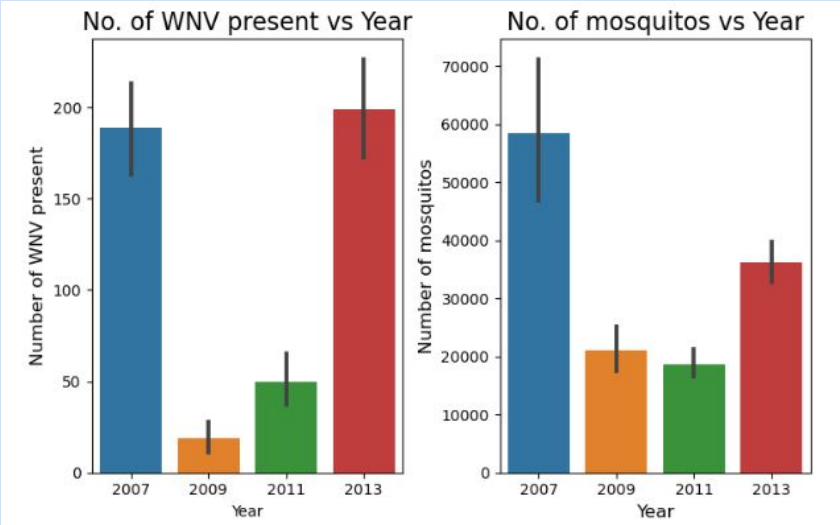




- Mosquitoes are most active at 80°F, become lethargic at 60°F, and cannot function below 50°F.**
- Wetbulb temperature was highest in 2007, but relatively high as well in 2011 and 2013.
- Wetbulb temperature was lowest in 2009 which corresponds to the lowest WNV presence as well.

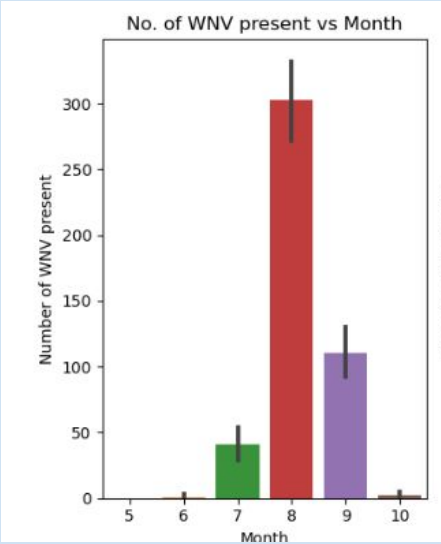
**Source:
https://www.beyondpesticides.org/resources/mosquitos-and-insect-borne-diseases/documents/the-truth-about-mosquitoes,-pesticides-and-west-nile-virus

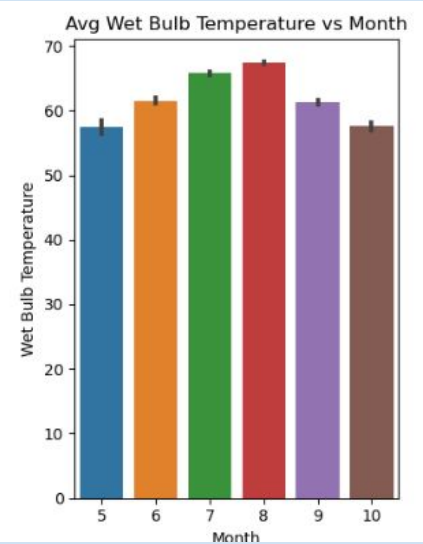
Periods



- Higher WNV present in 2007 and 2013.
- Same is true for number of mosquitos.

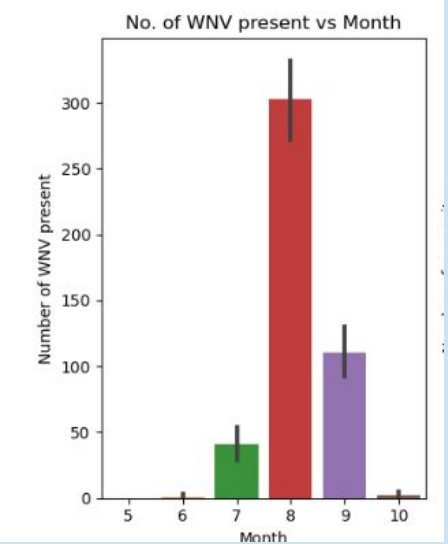
Periods

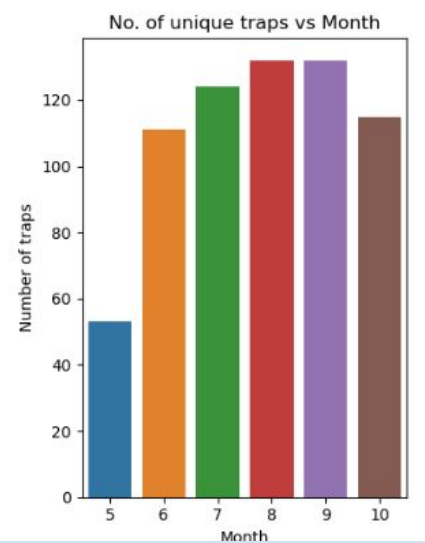




- Number of mosquitoes and WNV cases peak in Aug (highest wet bulb temperature).

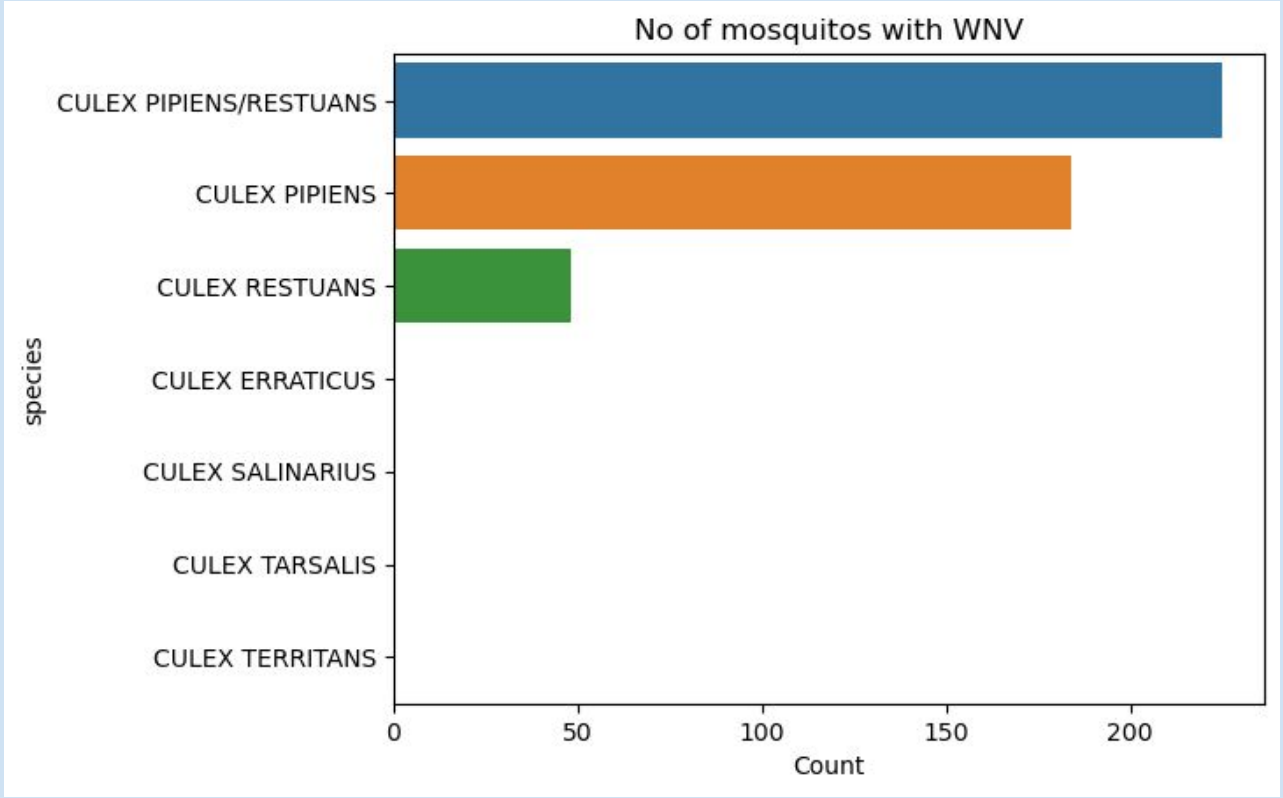
Periods



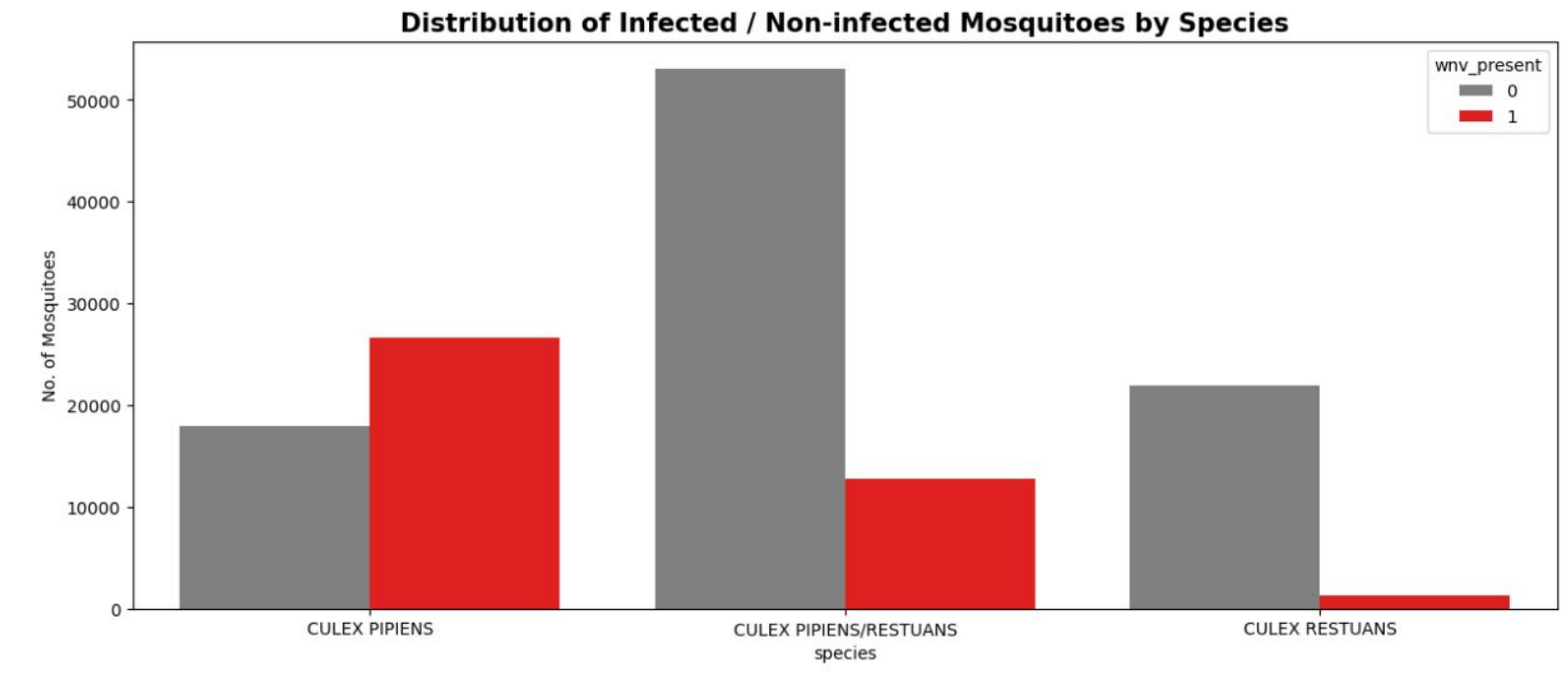


- Most number of traps in Aug and Sep, corresponding to the peak WNV cases.

Species

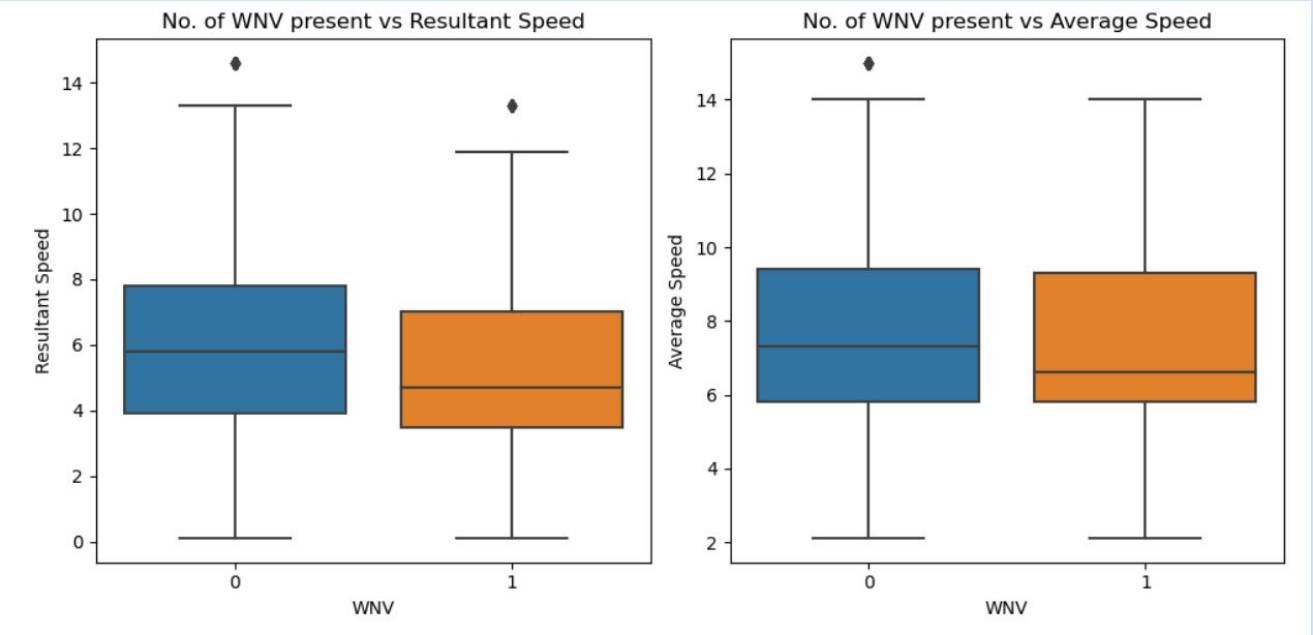

No of mosquitos with WNV
CULEX PIPIENS/RESTUANS
CULEX PIPIENS
CULEX RESTUANS
CULEX ERRATICUS
CULEX SALINARIUS
CULEX TARSALIS
CULEX TERRITANS
species
0
50
100
150
200
Count

Species



- The Culux Pipiens has a higher ratio of WNV present compared to Culux Restuans.

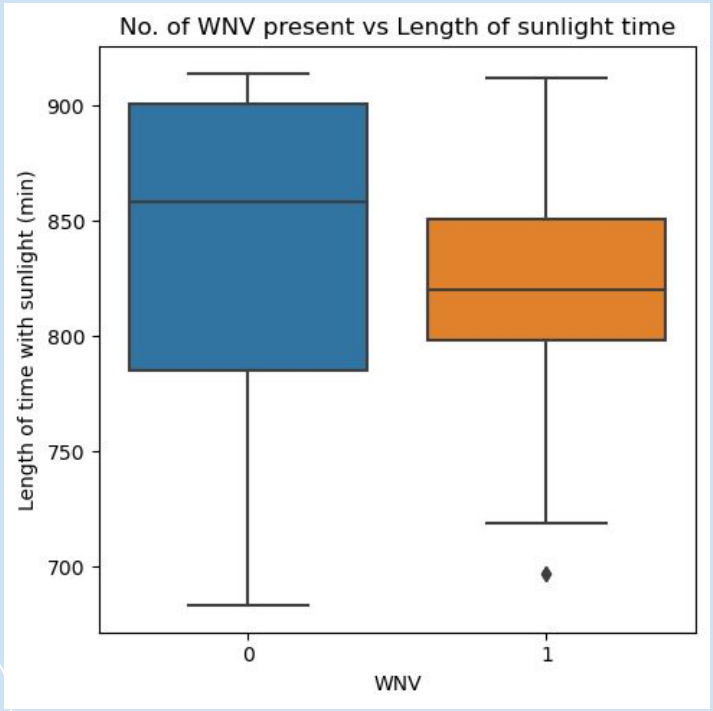
Wind Speed



- Lower wind speed means more WNV. This is because mosquitoes are generally are not strong flyers.**

**Source:
https://www.orkin.com/pests/mosquitoes/when-are-mosquitoes-most-active

Daylight Time



- Less sunlight duration means more cases. This could be because the Culux Pipiens is normally active at night.**

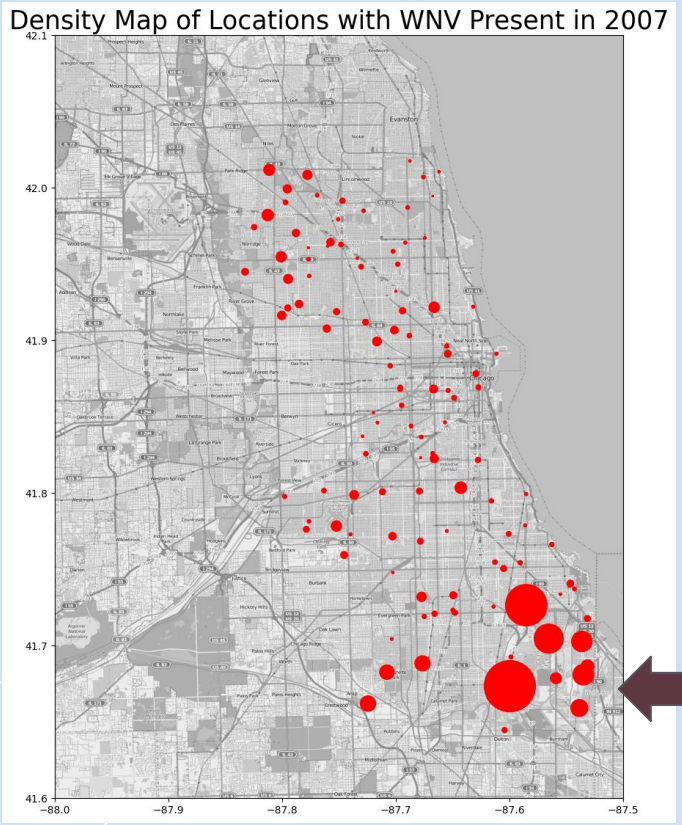
**Source: https://www.mosquitomagnet.com/advice/mosquito-info/biting-insect-library/culex-pipiens-mosquito

Location

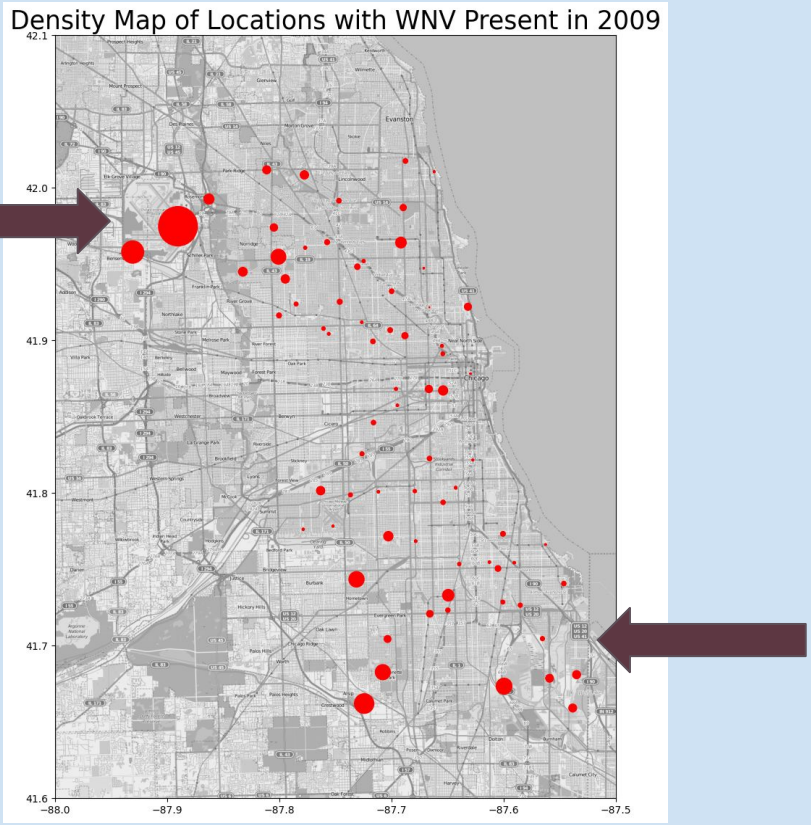
- Areas near the airport and near water sources are likely to be better environments for mosquitoes to thrive. Locations with high number of cases vary from year to year.
- Ohare Airport, Lake Camulet and Wolf Lake
- Density of WNV found by location shows higher concentration in the general North West and South East regions.

Location

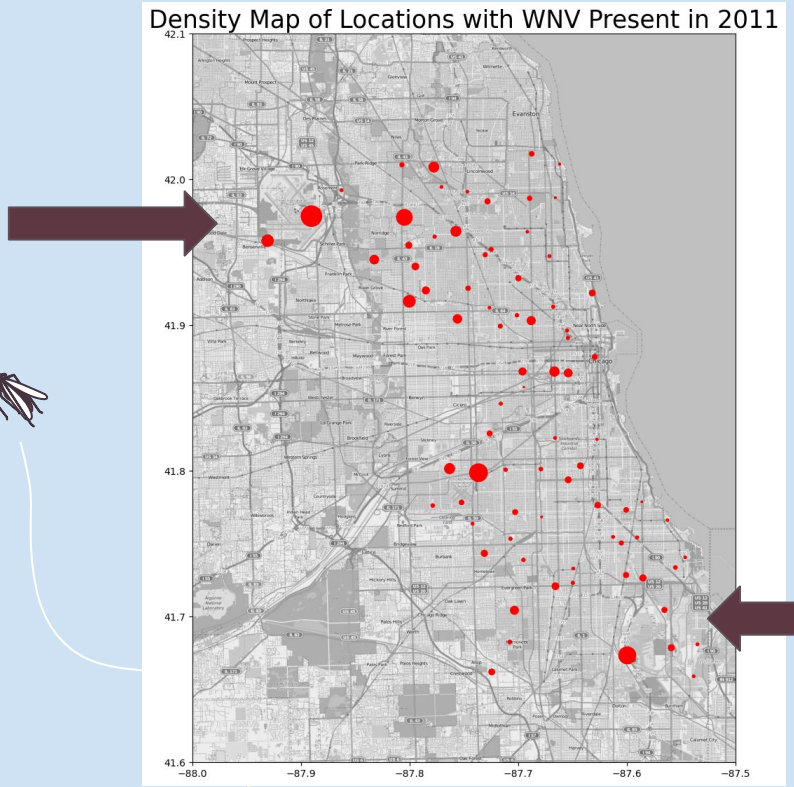
Density Map of Locations with WNV Present in 2007

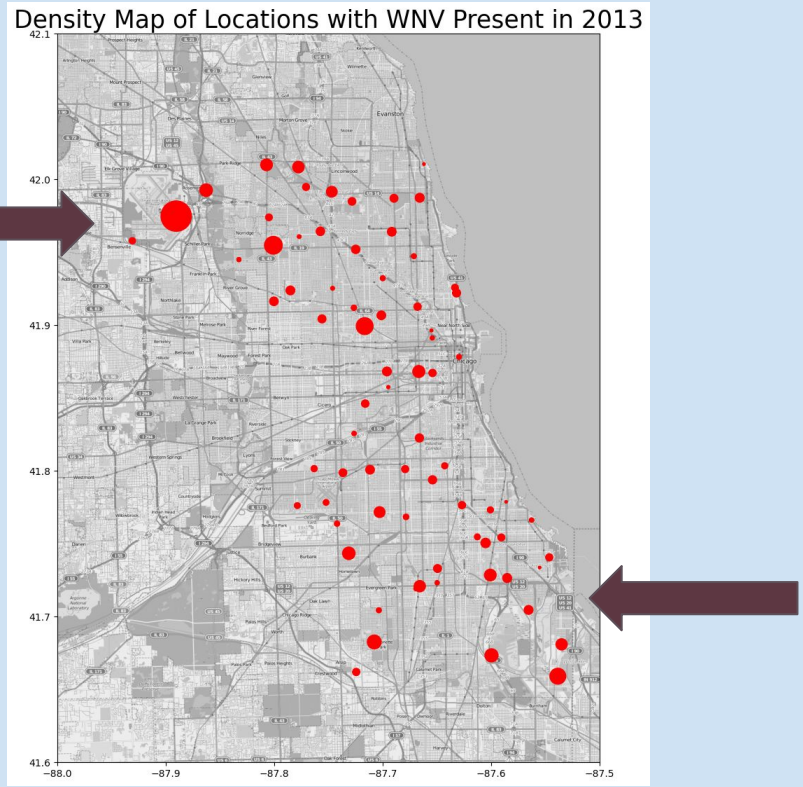


Density Map of Locations with WNV Present in 2009

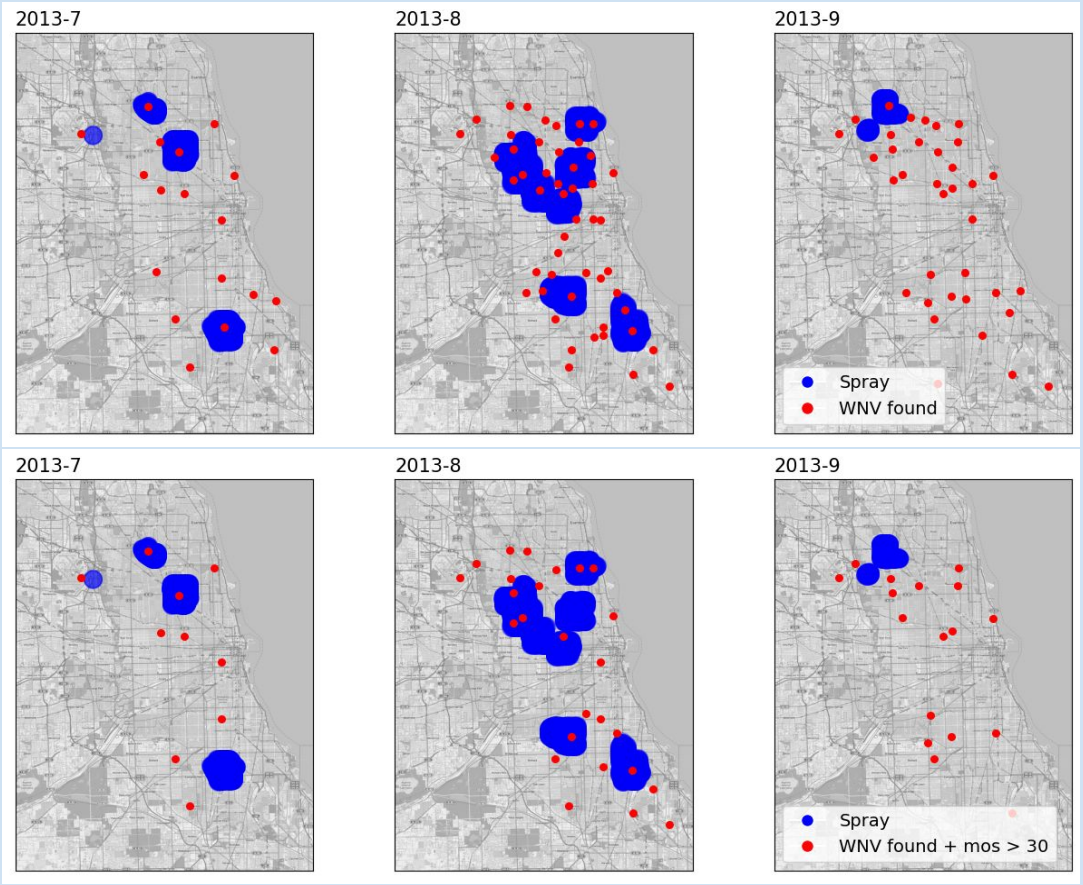


Location


Density Map of Locations with WNV Present in 2011
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

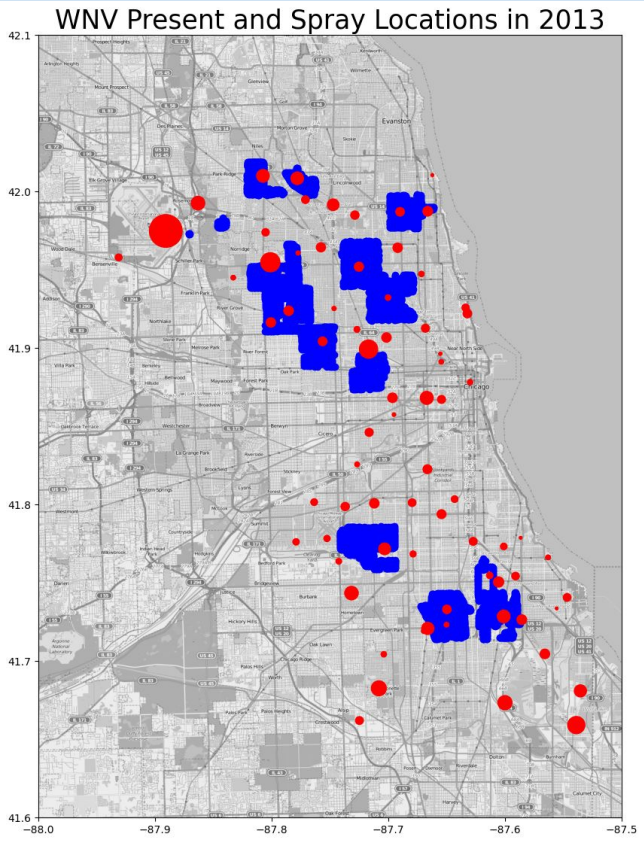

Density Map of Locations with WNV Present in 2013
42.1
42.0
41.9
41.8
41.7
41.6
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5

Spray



- Spray + WNV present
- Spray + WNV present and > 30 Mosquitos
- We noted that number of cases (the red spots) decreased in Sep.
- Number of mosquitoes decreased naturally after the peak in Aug.

Spray



- Most of the concentration seems to be in 2 broad range hot spots in the North West and the South East.
- The spray areas do not necessarily cover all areas with WNV presence.

Modelling

Evaluation Metrics

- Use **Area under the curve (AUC)**
  - better at handling class imbalance
  - captures the trade-off between the TPR and FPR
- Precision-Recall Trade-off:
  - Recall: over-spraying areas with low number of WNV present
  - Precision: under-spraying areas with high number of WNV present
  - Recall more important: target > 0.70 recall score

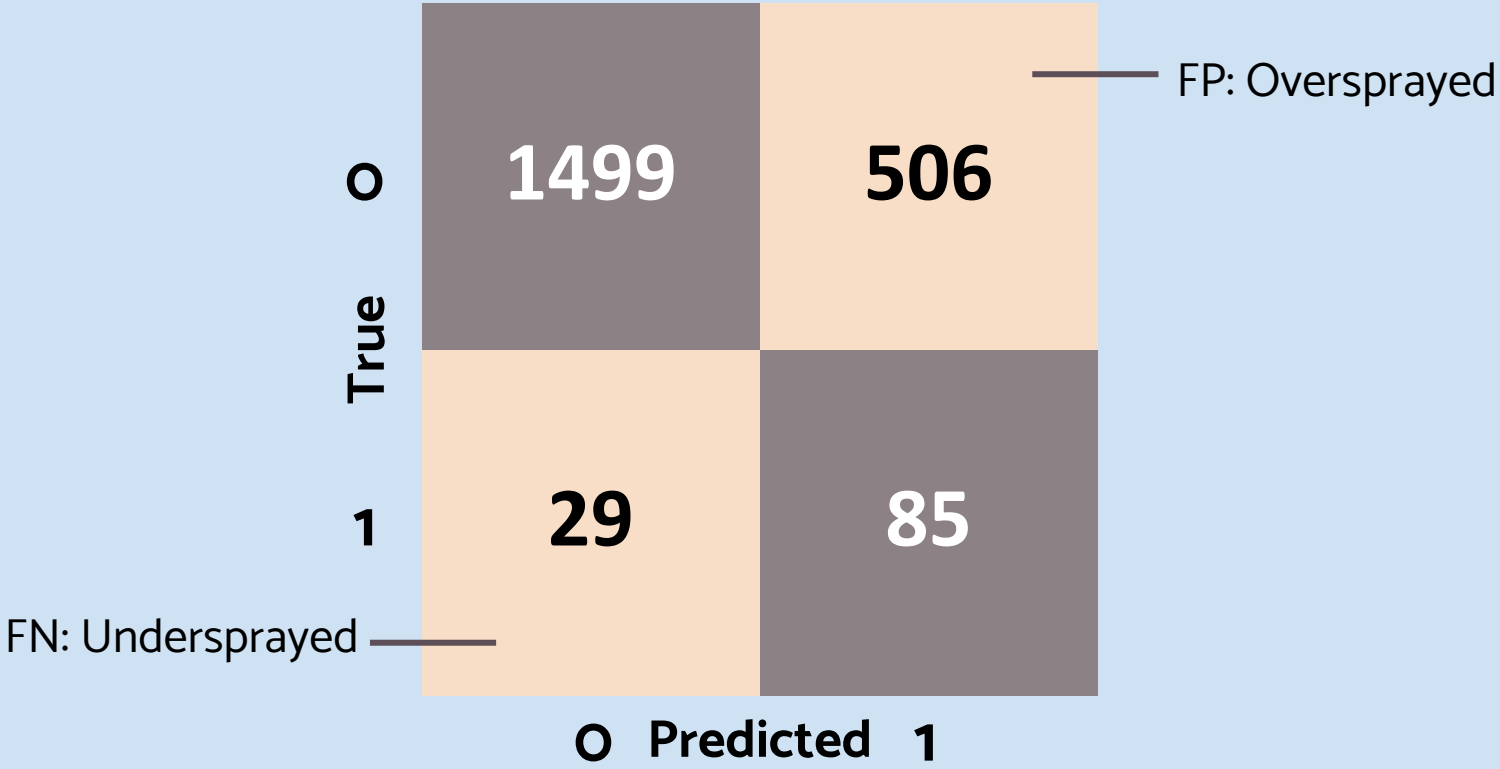
Feature selection

- Daylight duration
- Week number
- Wet bulb temperature
- Species (culex pipiens, culex pipiens/restuans and culex restuans)
- Resultant wind speed
- Location (clustering of coordinates using DBscan)

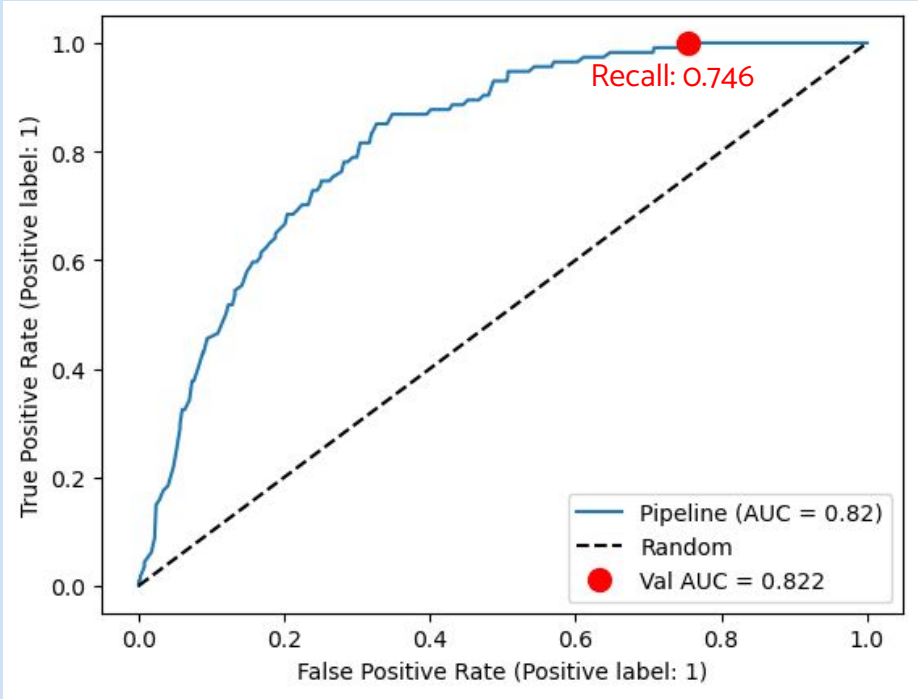
Models

Model	Classifier	AUC (Val)	AUC (Kaggle)	Recall (Val)
1	Logistic Regression	0.791	0.656	0.781
2	Logistic Regression (less features)	0.748	0.582	0.772
3	Multinomial Naive Bayes	0.708	N/A	0.737
4	Random Forest	0.838	0.609	0.702
5	Random Forest (less features)	0.822	0.680	0.693
6	XGBoost	0.844	0.579	0.544
7	XGBoost (less features)	0.832	0.600	0.693
8	**Random Forest (less features) & fine-tuned**	**0.822**	**0.701**	**0.746**

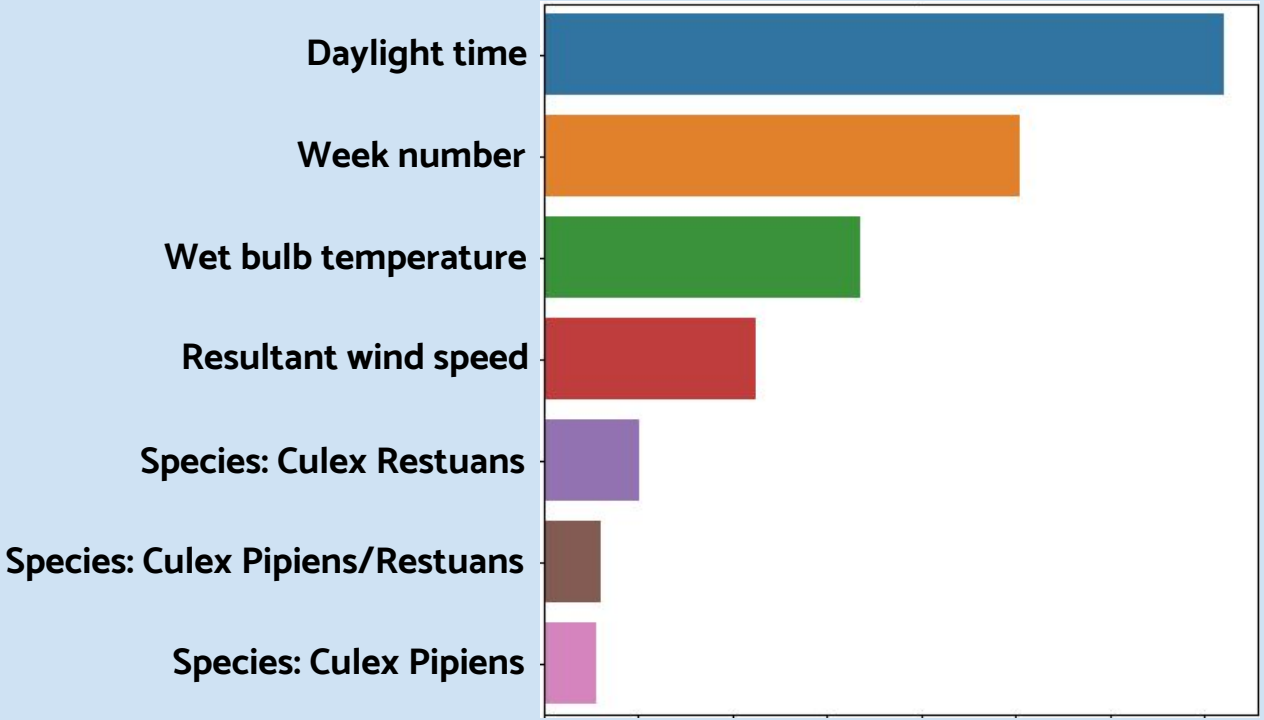
Confusion Matrix



AUC curve


Recall: 0.746
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
Pipeline (AUC = 0.82)
Random
Val AUC = 0.822

Feature Importance



Possible model improvement

- Validation design to avoid overfitting e.g.
  - train-test split by years
    - Train data: 2007, 2009, 2011, 2013
    - Test data: 2008, 2010, 2012, 2014
  - perform training and validation on different years

Cost-Benefit Analysis

Cost-Benefit Analysis

Costs

- Pesticides
- Manpower

Benefits

- Savings in medical expenses
- Savings in economic costs

Goal

- Find the optimal balance between total costs and level of benefit achieved

Costs

- Spray data: 2011: 2x in 2 weeks,
  2013: 8x in 8 weeks
- Chicago Department of Public Health (CDPH) uses Zenivex E4 insecticide for spraying citywide
- Applied at 1.5 fl oz / acre at 10mph
- Assumptions: 50 trucks available, spray radius of 100sqft, labour cost @ $20/hr
- 68h to cover entire Chicago area

Benefits

Savings in medical costs

- ASTMH: annual medical costs of WNV in U.S = $57M
- Annual medical costs of WNV in Chicago (interpolated) = $1.83M

Savings in economic costs

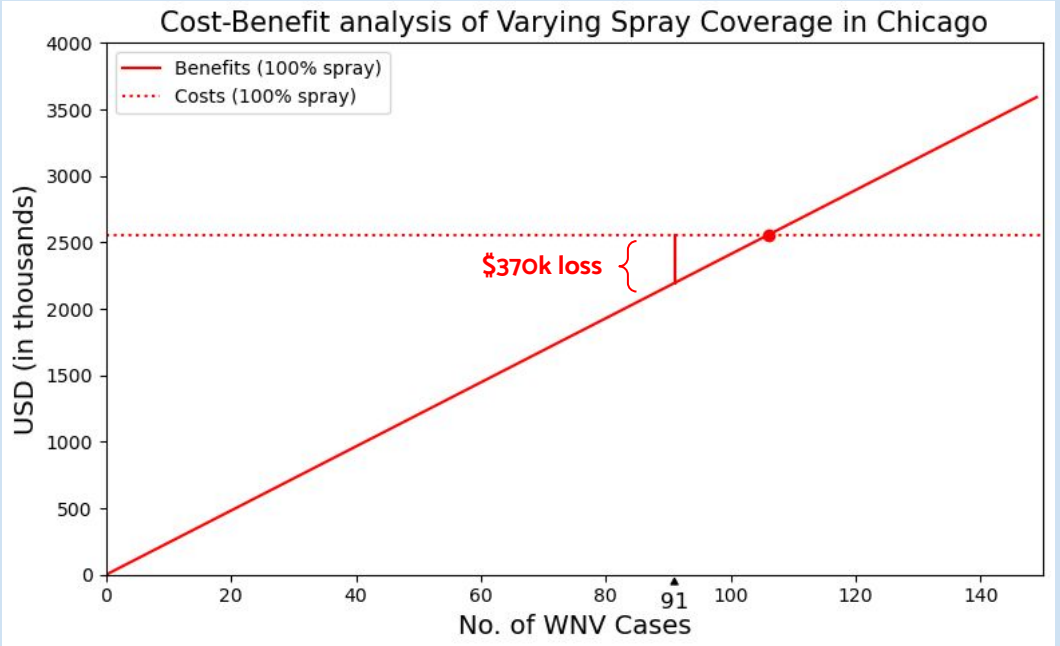
- NLM: productivity loss for those < 60 = $955 daily
  > 60 = $625 daily
- Symptoms last 3-6 days -> productivity loss per individual = $4000

Base Scenario

- According to CDC, average annual Chicago cases = 91
- 12 applications across 12 weeks (Jul - Sep)
- 100% Coverage
- Costs of spray: $1.75M
- Manpower costs: $810k
- **Total Costs: $2.56M***

- Savings in medical expenses = $ 1.83M
- Savings in economic costs = $364k
- **Total Benefits: $2.19M***

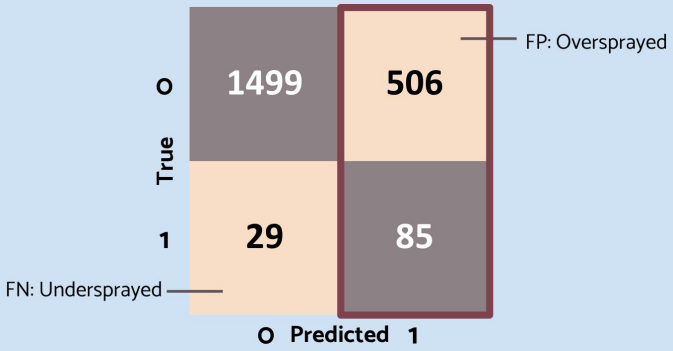
- **Net Loss: $370k***

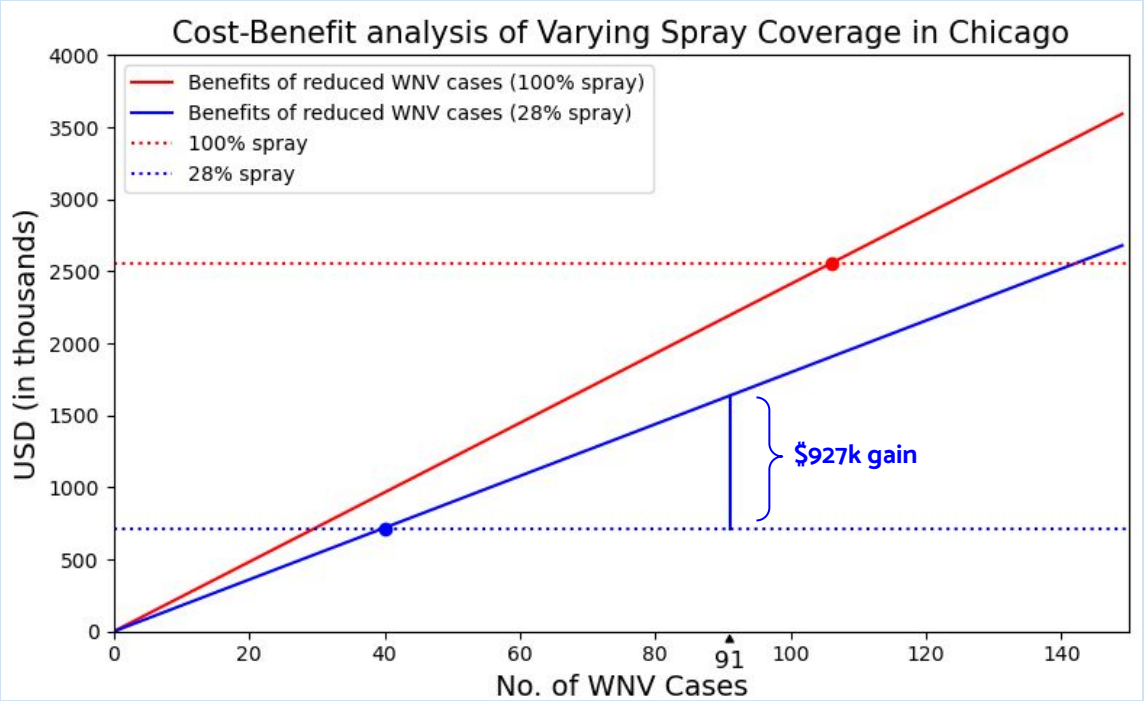


* Based on average annual cases in Chicago (91)

Scenario 1: Targeted Spray Coverage

- % Spray Coverage in Chicago: 28% (based on Predicted / Total Score)
- **Total Costs: $713k***
- **Total Benefits: $1.64M***
- **Net Gain: $927k***

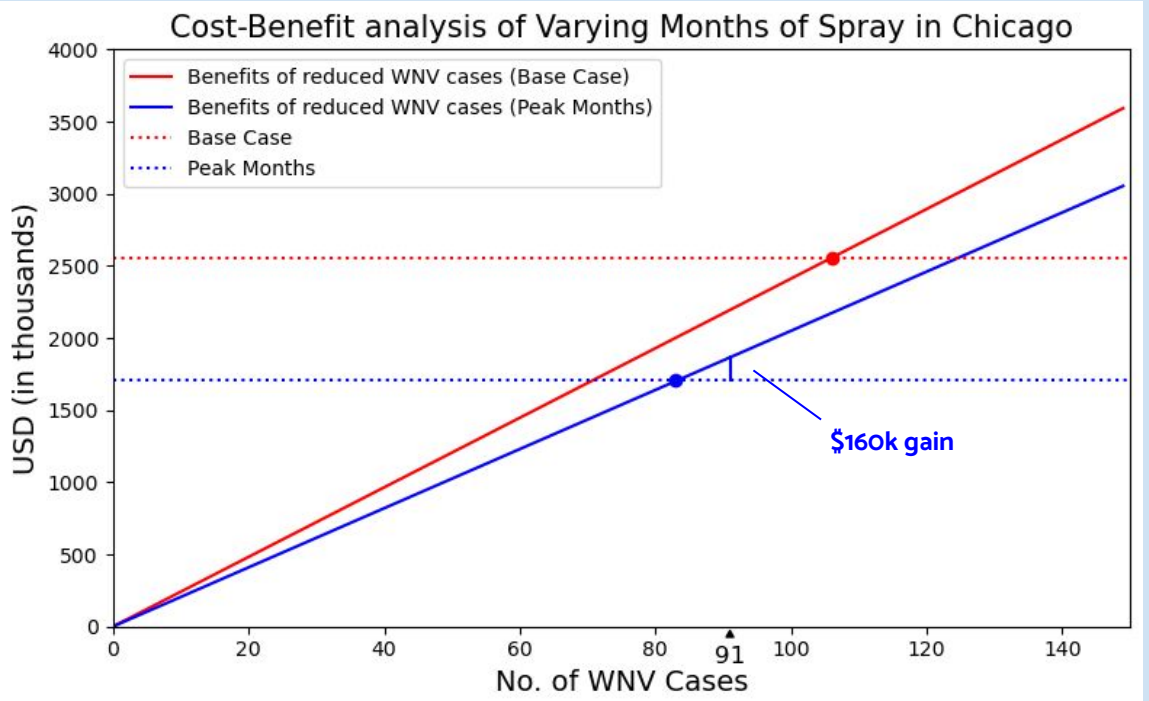




* Based on average annual cases in Chicago (91)

Scenario 2: Spray in Peak Months (Aug/Sep)

- Based on 20-year historical average, the number of WNV cases in USA during WNV season are as follows:
  - July: 7,217
  - Aug: 21,791
  - Sep: 18,080
- Spray in Aug/Sep as 85% cases
- **Total Costs: $1.71M***
- **Total Benefits: $1.87M***
- **Net Gain: $160k***



* Based on average annual cases in Chicago (91)

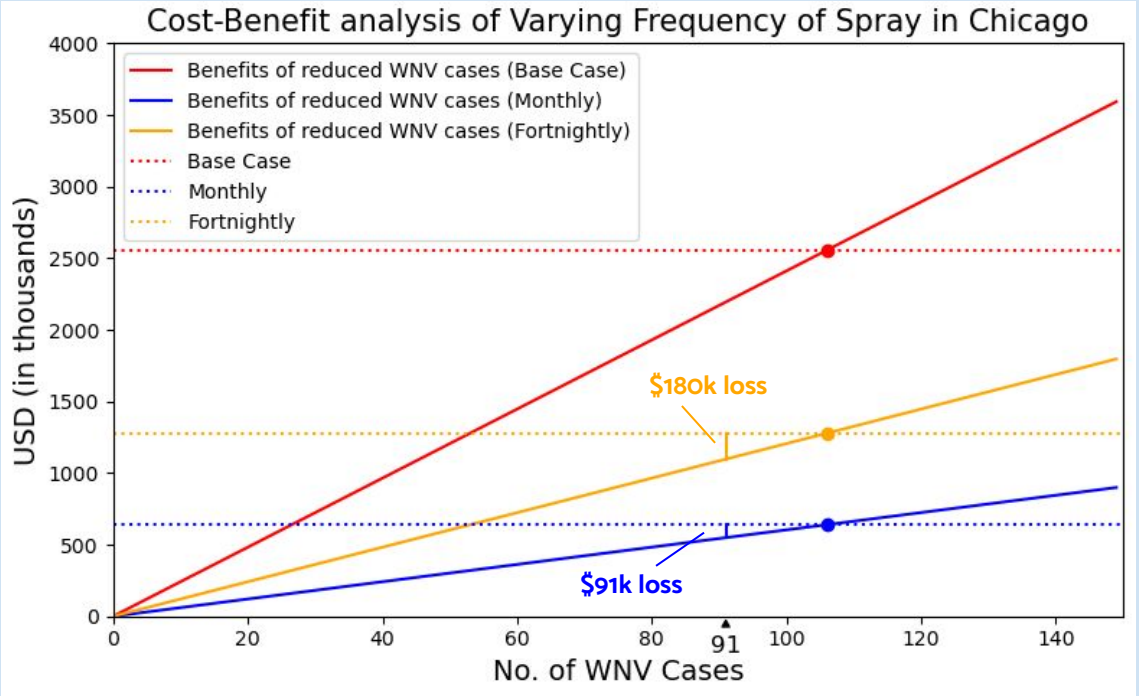
Scenario 3: Vary Frequency (Monthly/Fortnightly)

Monthly basis

- **Total Costs: $640k***
- **Total Benefits: $549k***
- **Net Loss: $91k***

Fortnightly basis

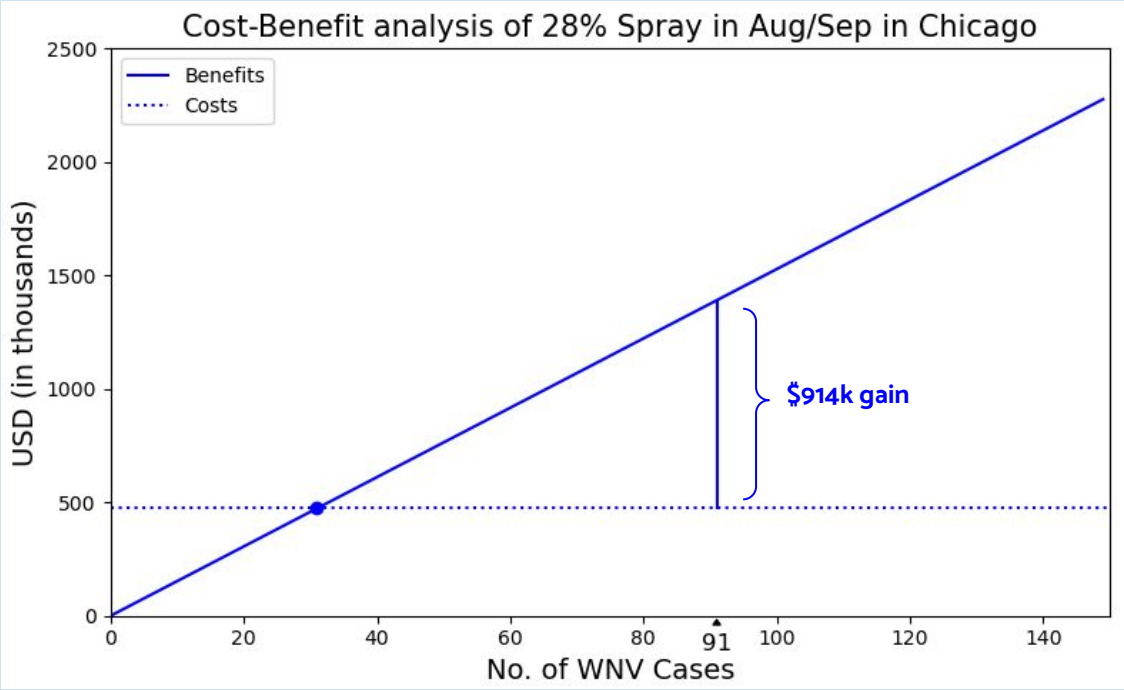
- **Total Costs: $1.28M***
- **Total Benefits: $1.1M***
- **Net Loss: $180k***



* Based on average annual cases in Chicago (91)

Hybrid approach - 28% Targeted Spray + Peak Months (Aug/Sep)

- Total Costs: $476k
- Total Benefits: $1.39M
- Net Gain: $914k



* Based on average annual cases in Chicago (91)

Results of analysis

	Coverage	Total Costs* ($)	Total Benefits* ($)	Breakeven Cases	Net Gain / Loss* ($)
Base Scenario	100%	2.56M	2.19M	106	-370k
28% Spray Coverage	28%	713k	1.64M	40	927k
Spray in Peak Months (Aug/Sep)	100%	1.71M	1.87M	83	160k
Monthly Basis	100%	640k	549k	106	-91k
Fortnightly Basis	100%	1.28M	1.1M	106	-180k
28% Coverage + Peak Months	**28%**	**476k**	**1.39M**	**31**	**914k**

* Based on average annual cases in Chicago (91)

Recommendations

Subsidising costs of pesticide

Motivate people to take action

Agricultural Drones

- Cheaper
- Greater accessibility

Lab-grown mosquitoes

Wolbachia-carrying mosquitoes

Conclusion

- Key factors that impact presence and spread of WNV:
  - Longer daylight hours
  - Period of the year
  - Higher temperatures
- Spraying efforts should target high-risk areas during peak months
  - Targeted approach relies on having a good predictive model

Thank you