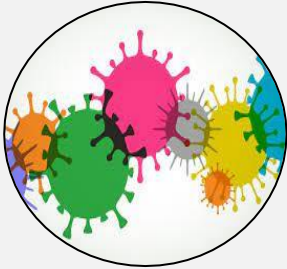


PROJECT 4: WEST NILE VIRUS PREDICTION

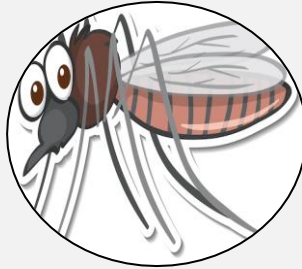
By:

Samuel Koh
Benjamin Yen
Tan Jun Jie

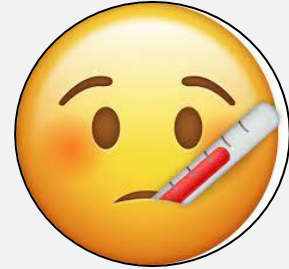
WEST NILE VIRUS



Originated in Africa, and it is named after the West Nile region of Uganda



Mosquito:
Culex Species



Illness:
Fever, headache, body aches, and fatigue

CONTROL MEASURES



Vaccines is still developing

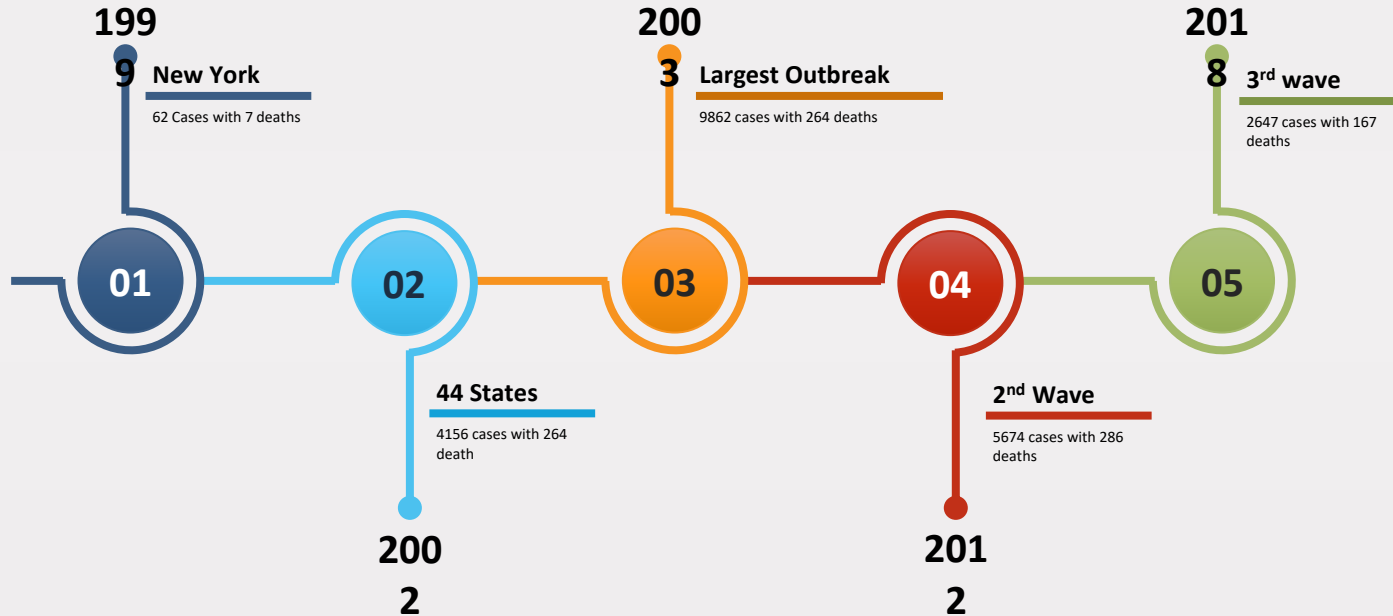


Wearing of PPE during mosquito feeding time



Sprays & Traps

Timeline of the Outbreak

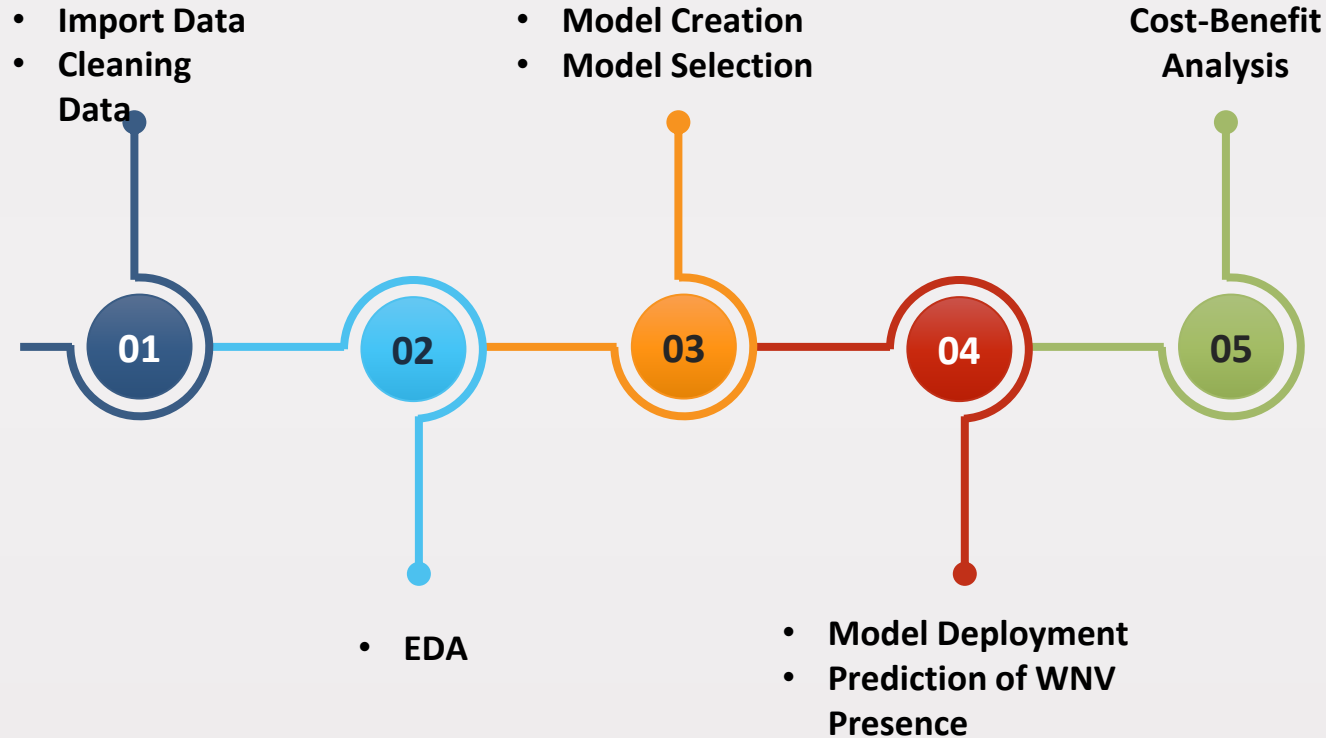


PROBLEM STATEMENT

West Nile virus is most commonly spread to humans through infected mosquitos. Around 20% of people who become infected with the virus develop symptoms ranging from a persistent fever, to serious neurological illnesses that can result in death.

The City of Chicago and CPHD needs a model that can help them to predict more efficiently and effectively when and where WNV+ mosquitos is found within the city.

Workflow





A diagram illustrating five data cleaning steps, each represented by an orange circle. The circles are arranged in a loose cluster. The text inside each circle is as follows:

- Top-left circle: **Dropping Duplicates**
- Top-middle circle: **Dropping columns**
- Top-right circle: **Merging**
- Bottom-left circle: **Creation of Date-Time Columns**
- Bottom-right circle: **Combine records for traps**

At the bottom center of the image, the text **DATA CLEANING (TRAINING)** is displayed in a bold, black, sans-serif font.

Dropping Duplicates

Dropping columns

Merging

**Creation of
Date-Time
Columns**

Combine records
for traps

DATA CLEANING (TRAINING)



A diagram illustrating data cleaning techniques. It features four orange circles arranged in a diamond pattern around a central point. Each circle contains a text label. At the bottom center, the text 'DATA CLEANING (Spray)' is displayed in a bold, black, sans-serif font.

**Dropping
Duplicates**

**Dropping
rows**

**Creation of
Date-Time
Columns**

Impute of values

DATA CLEANING (Spray)



The diagram consists of three orange circles on a light gray background. The top-left circle is the largest and contains the text 'Convert dates to DT format'. The top-right circle is smaller and contains the text 'Impute of values'. The bottom-center circle is also smaller and contains the text 'Dropping rows from null values'. The circles are arranged in a triangular pattern, suggesting a sequence of steps in a data cleaning process.

**Convert dates
to DT format**

Impute of values

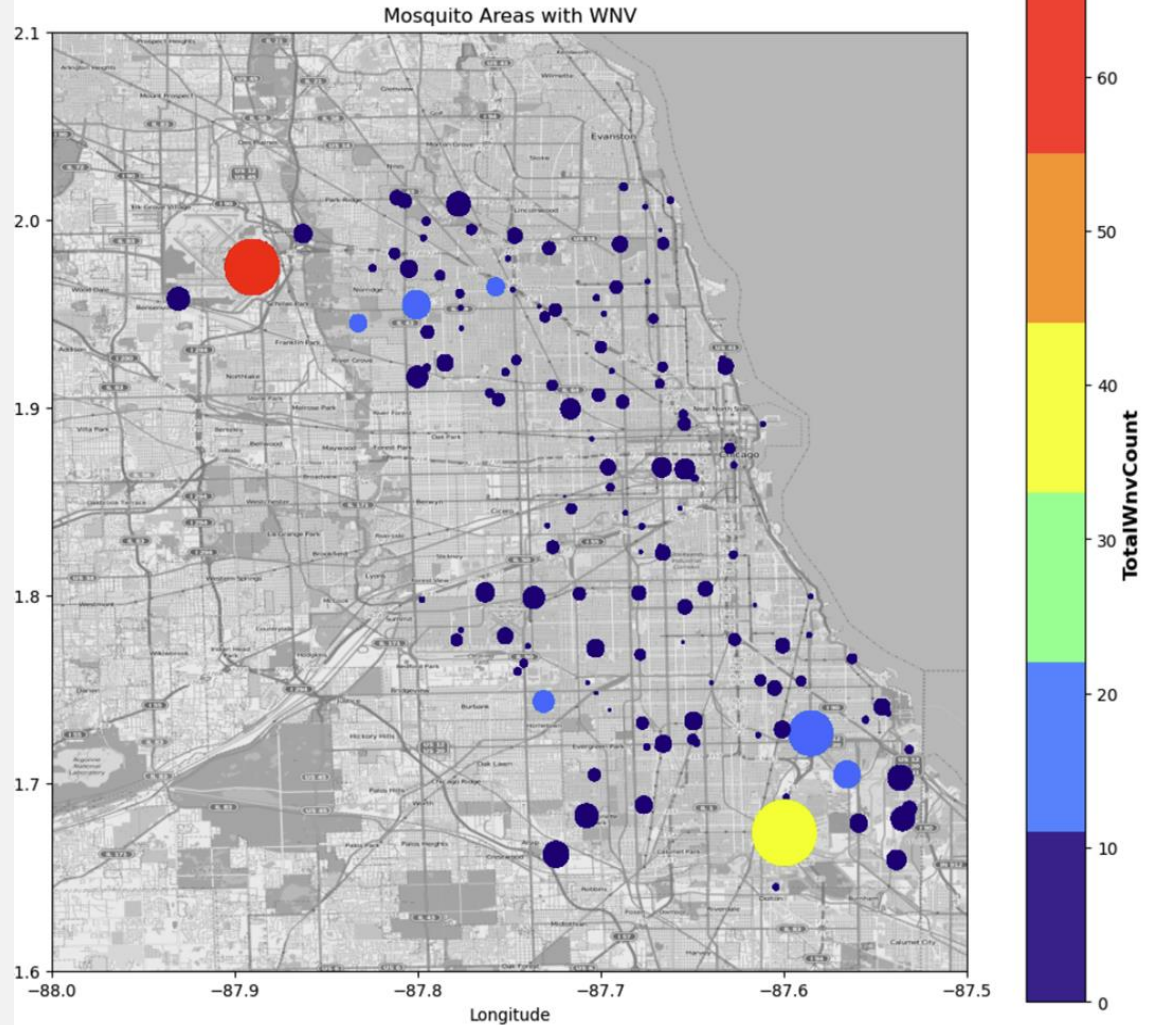
**Dropping
rows from
null values**

D A T A C L E A N I N G (Weather)

EDA

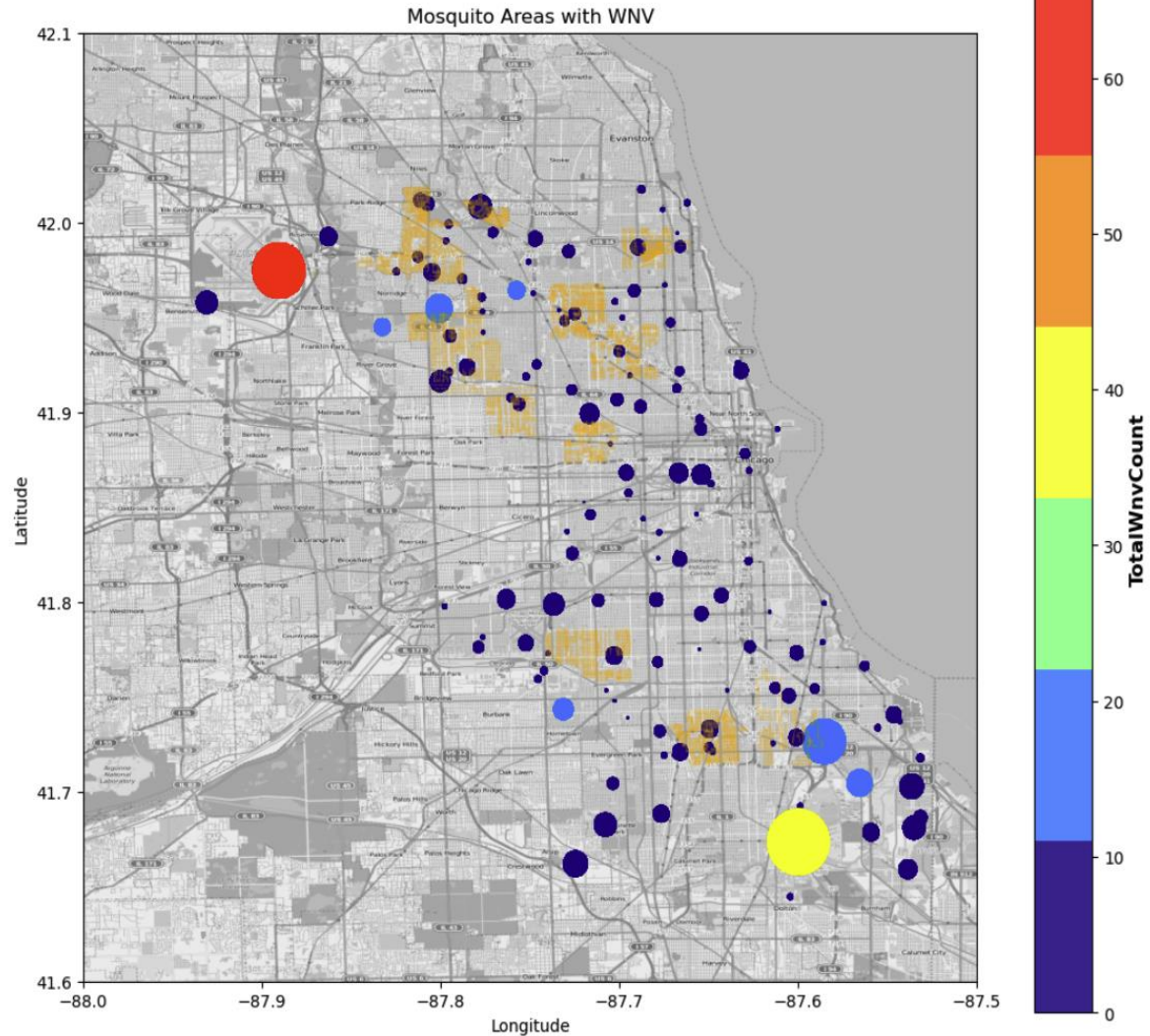
AREAS WITH WNV

- Airport – 66 Cases
- Port of Chicago - 44 Cases

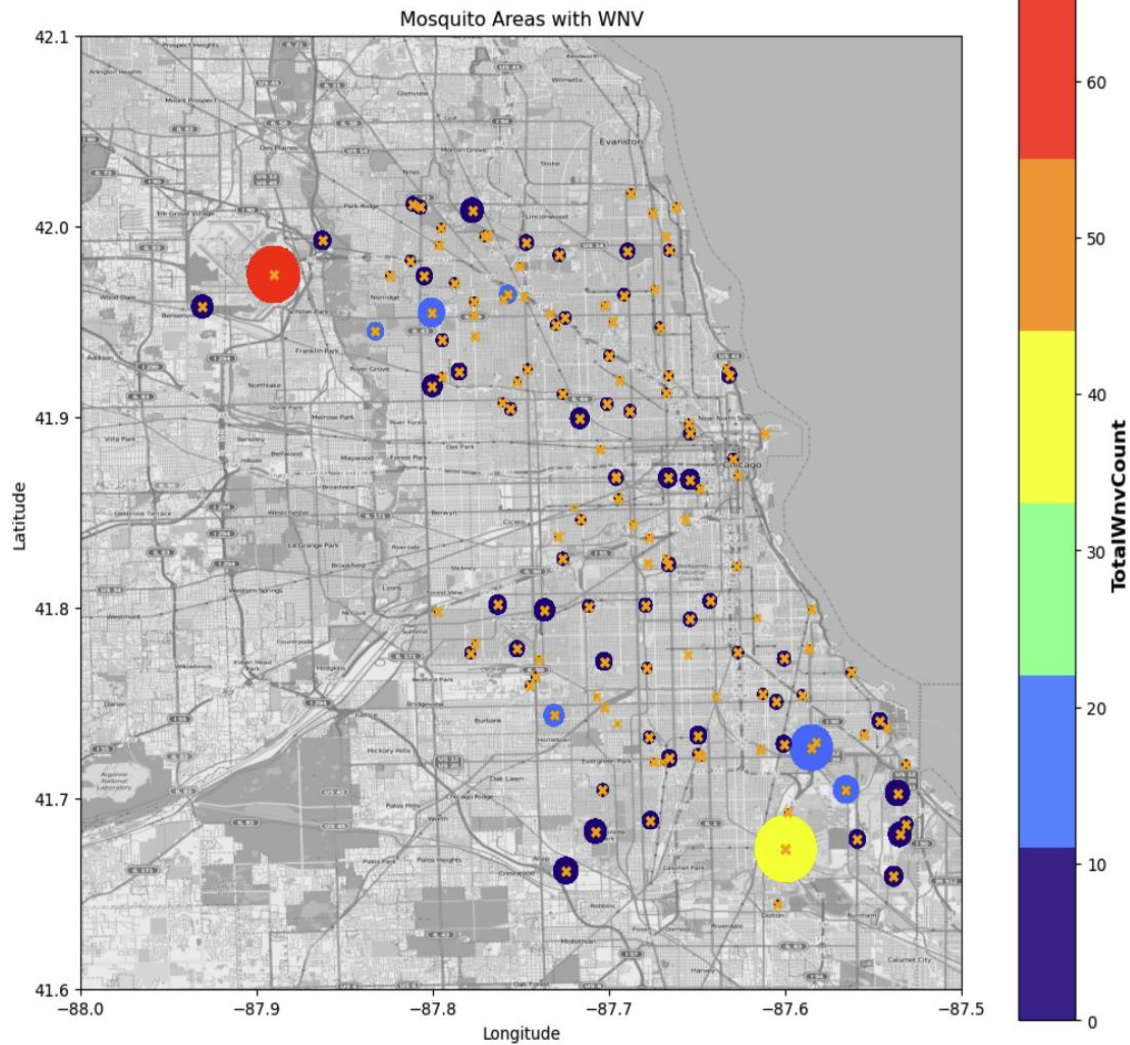


EFFECTIVENESS OF SPRAY

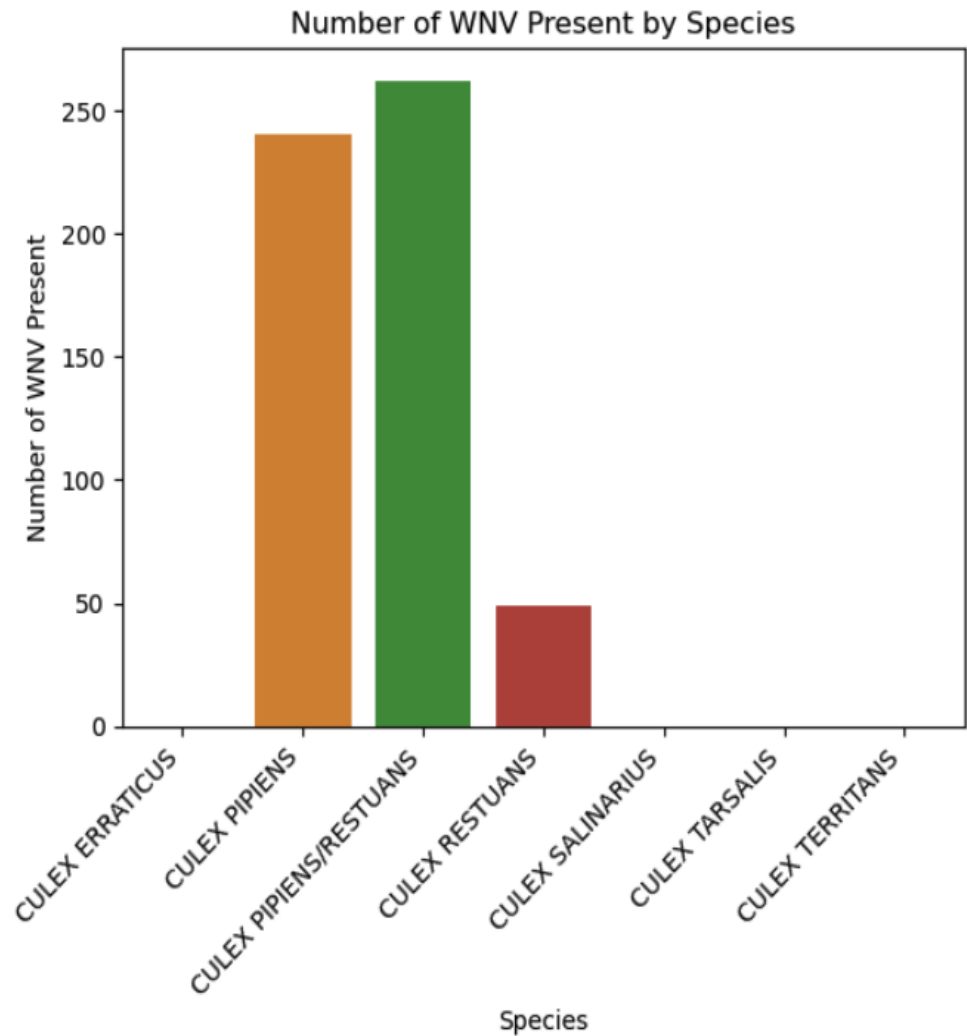
- Spray is the are the light orange areas
- More analysis is needed



TRAPS WITH MOSQUITO CLUSTERS AND WNV PRESENT

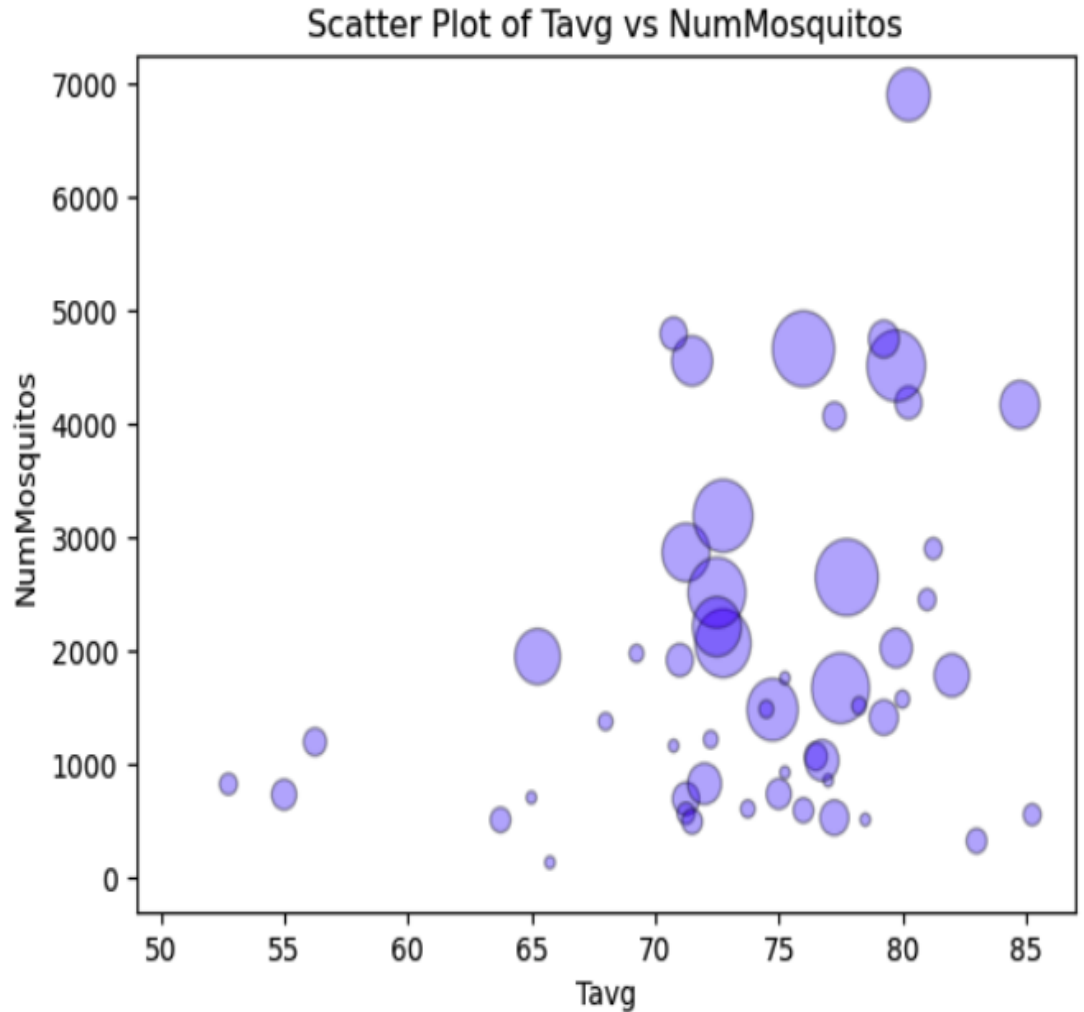


MOSQUITO SPECIES

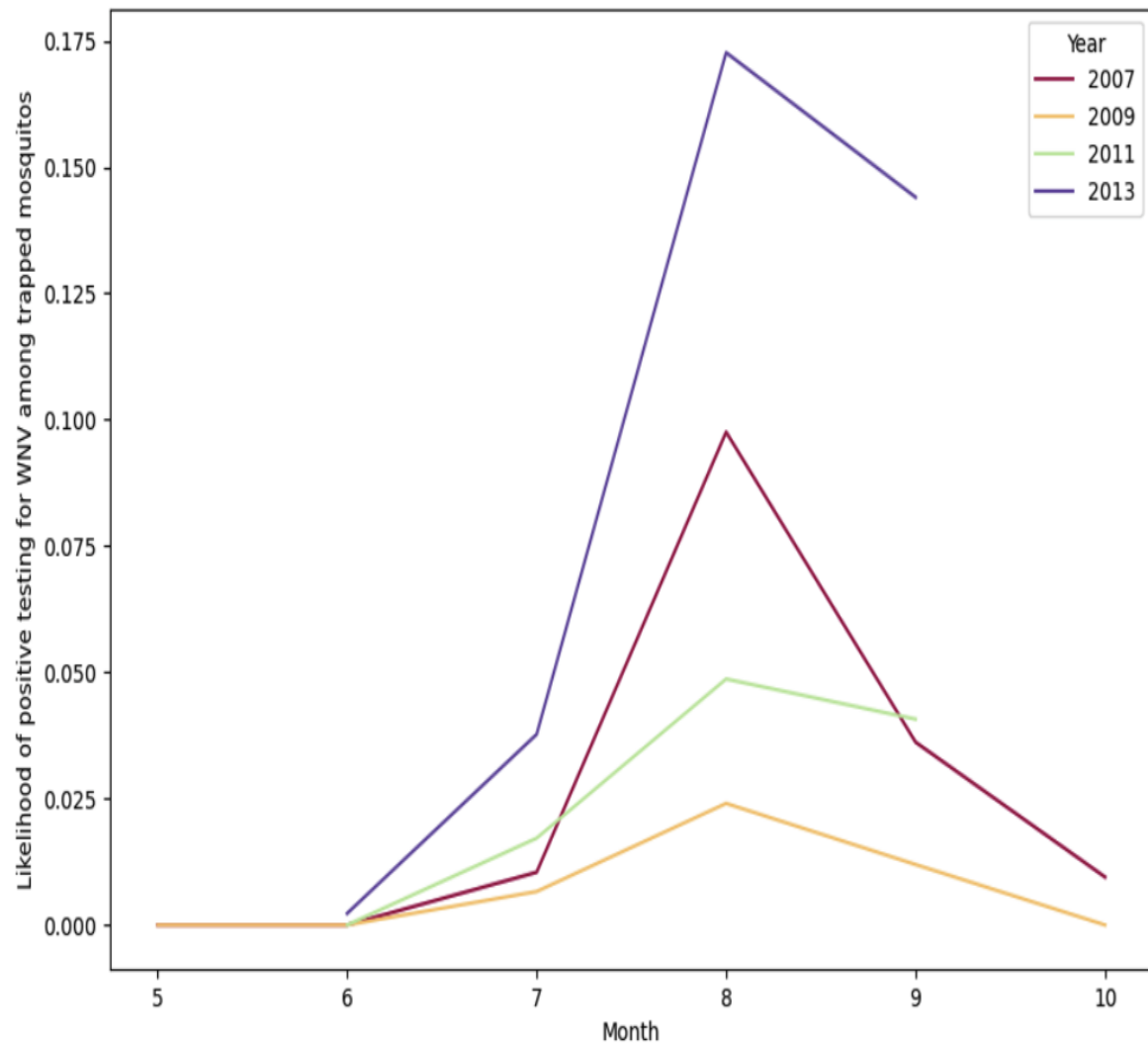


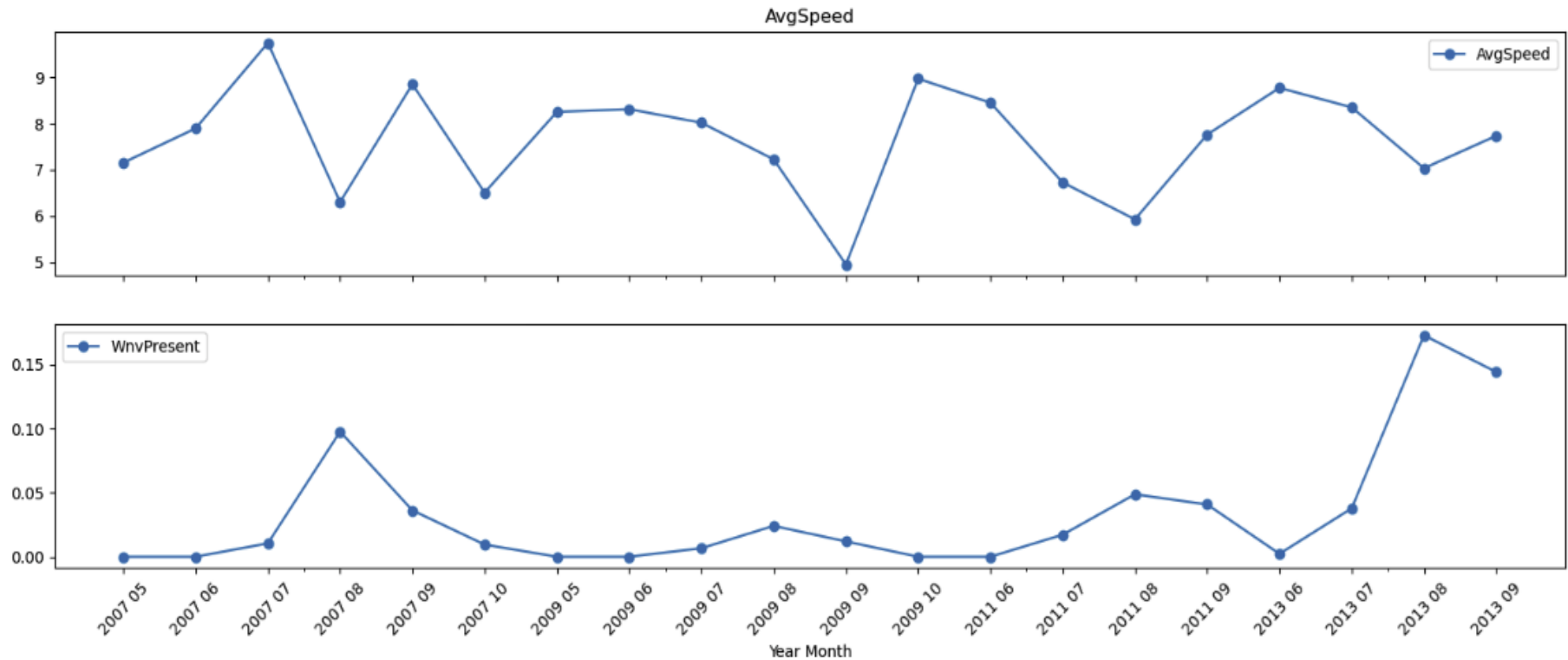
Effect temperature have on the mosquito/WNV

- Mosquito thrive between 70 - 80F
- Most of the WNV positive mosquito are active around this period

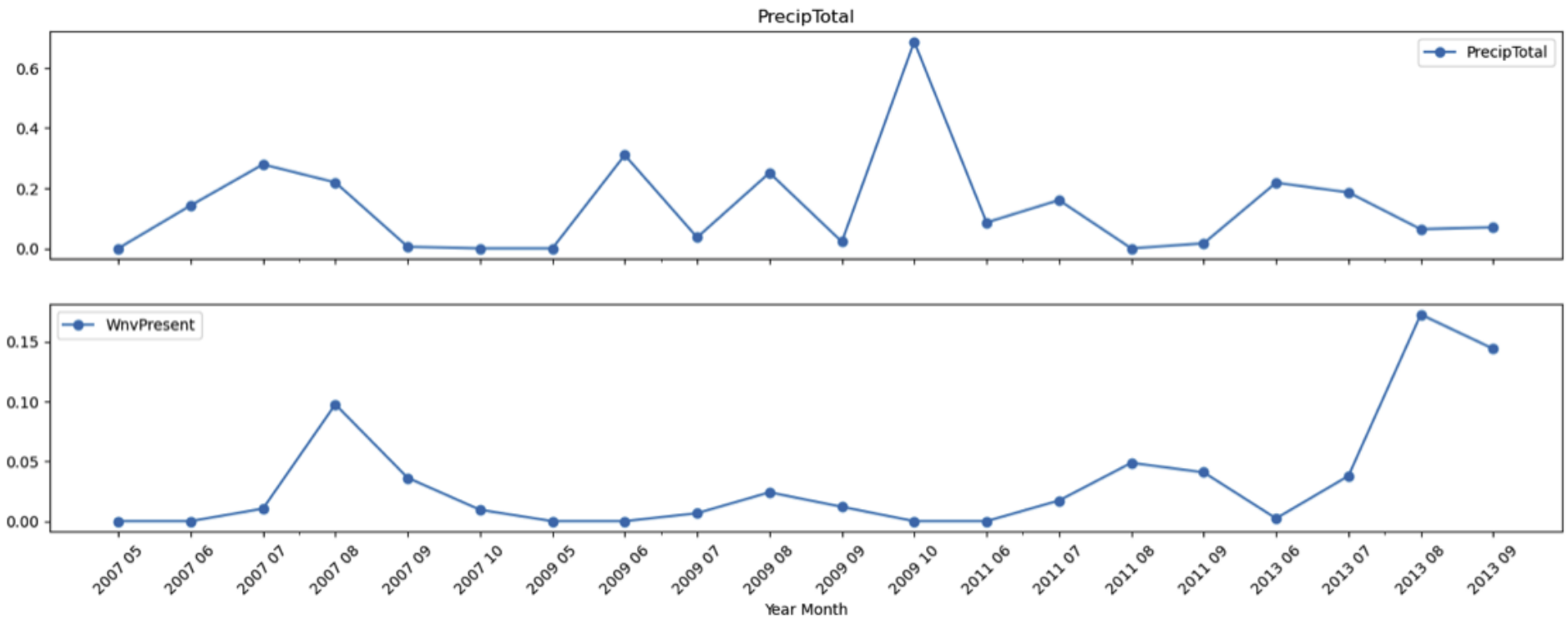


- The highest WNV cases which also has the highest mosquito counts happens to be in Aug.

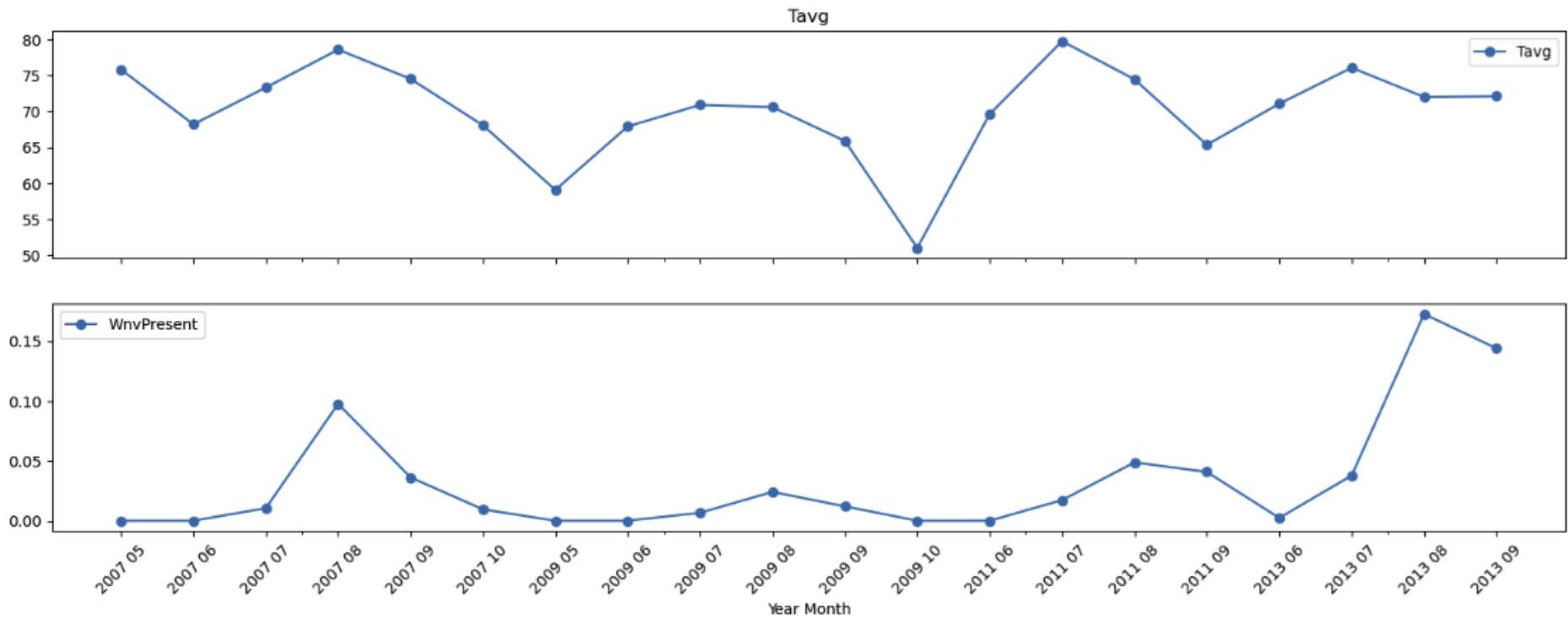




AVG SPEED WITH WNV PRESENT

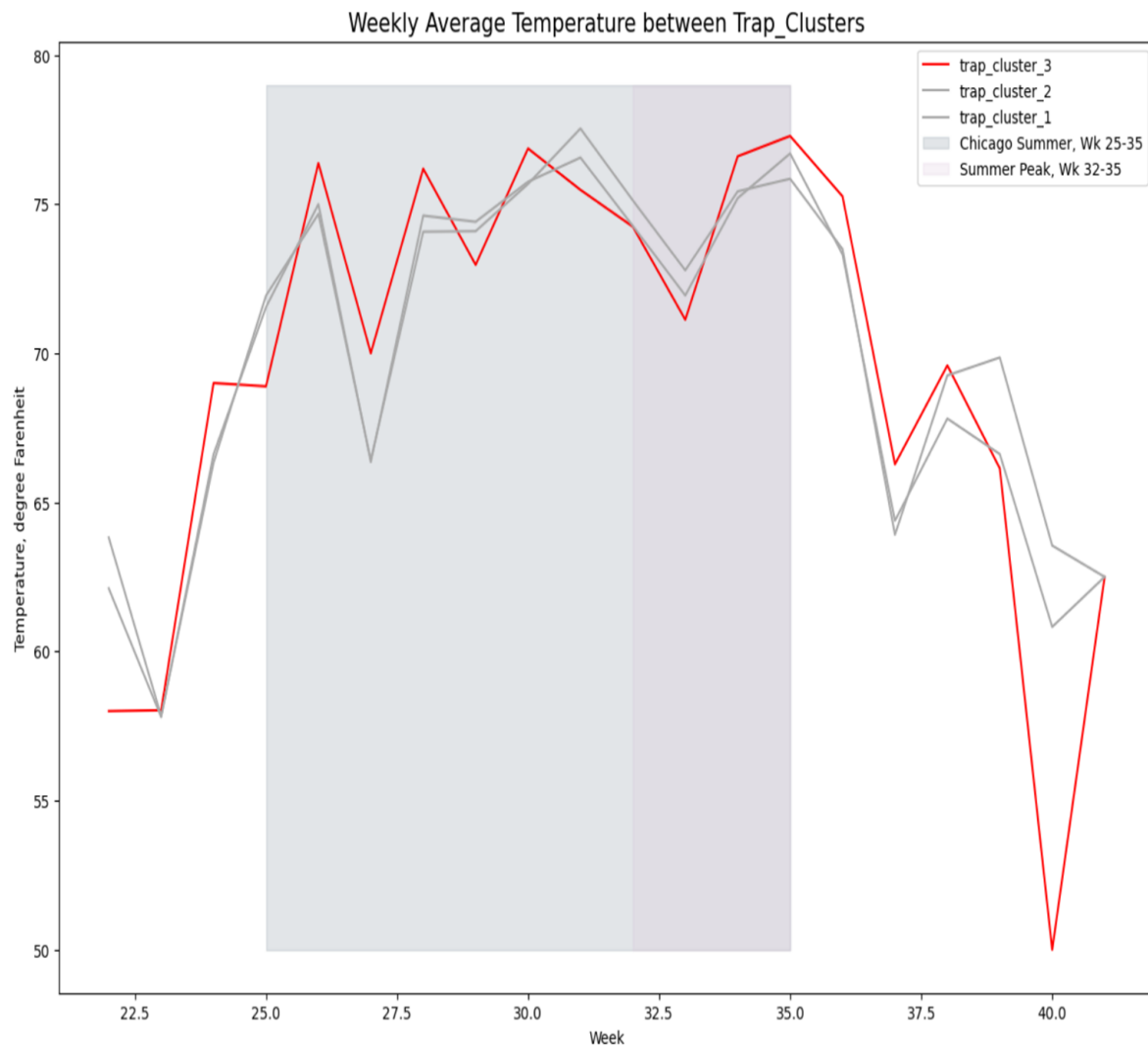


TOTAL PRECIPITATION WITH WNV PRESENT



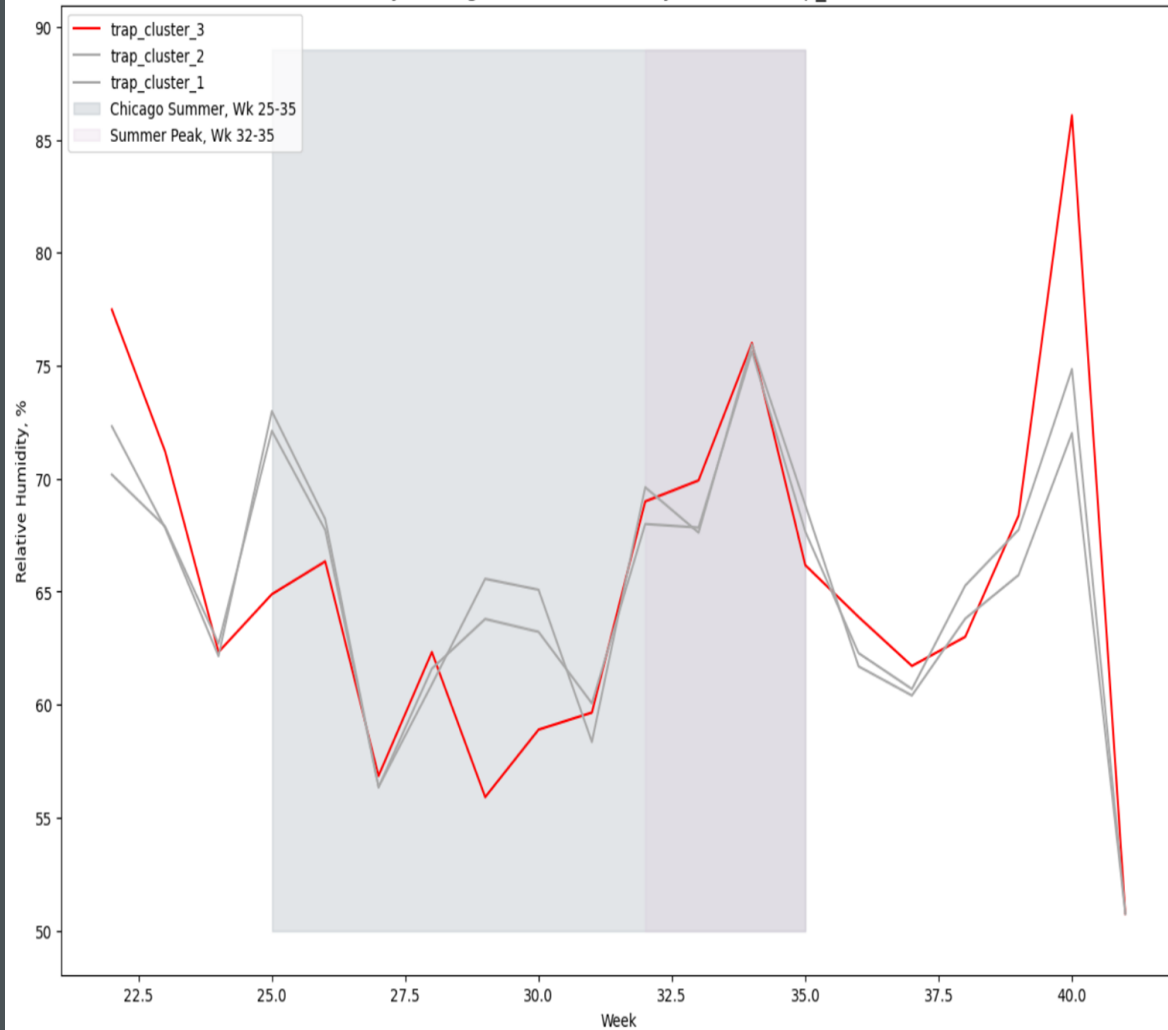
TEMPERATURE WITH WNV PRESENT

- The temperature range for trap cluster 3 is distinct from 1 and 2.
- Tavg tends to stay above 70F consistently for trap cluster 3 which causes more mosquitos to be hatched and result in bigger WNV clusters.

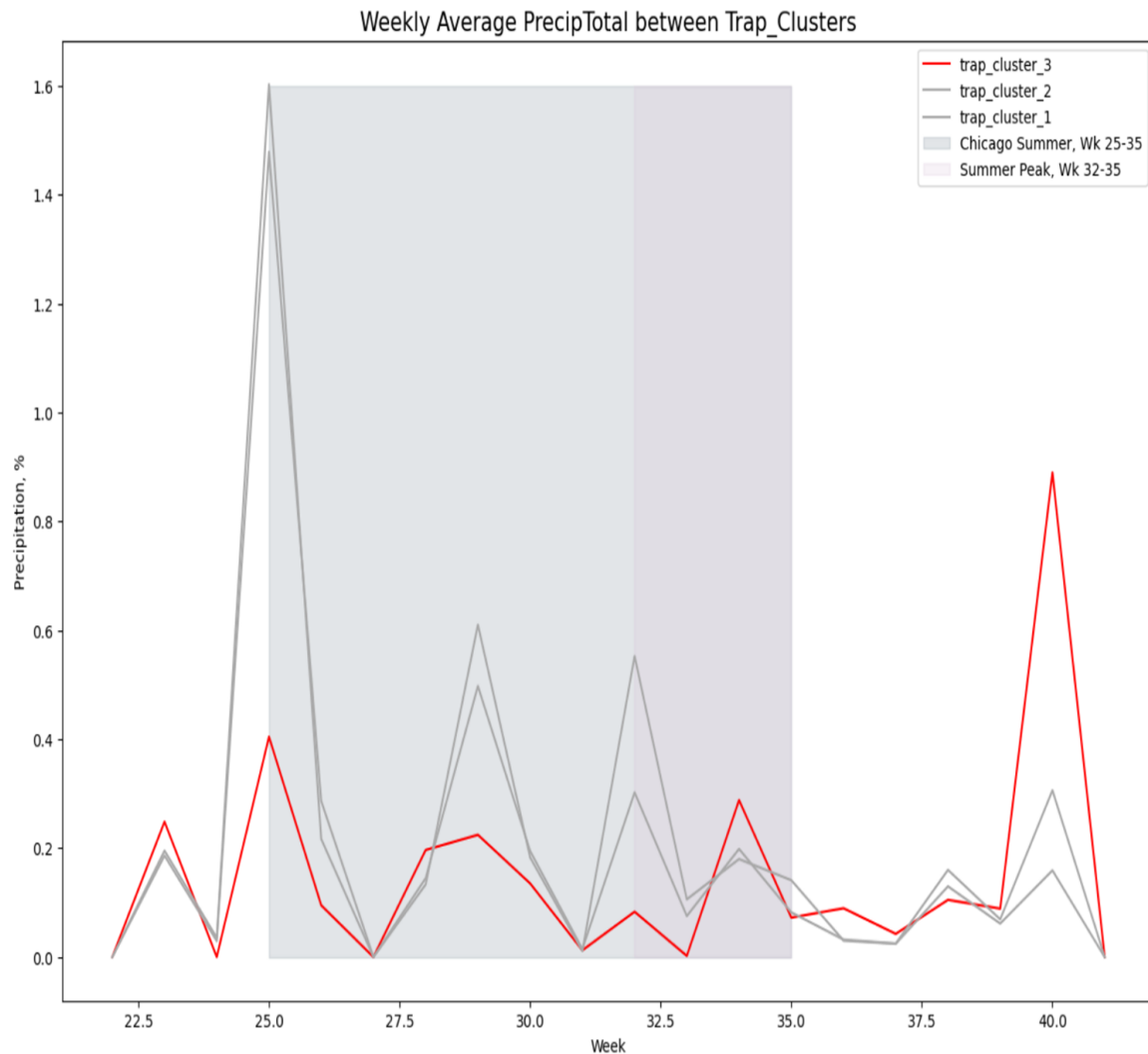


Weekly Average Relative Humidity between Trap_Clusters

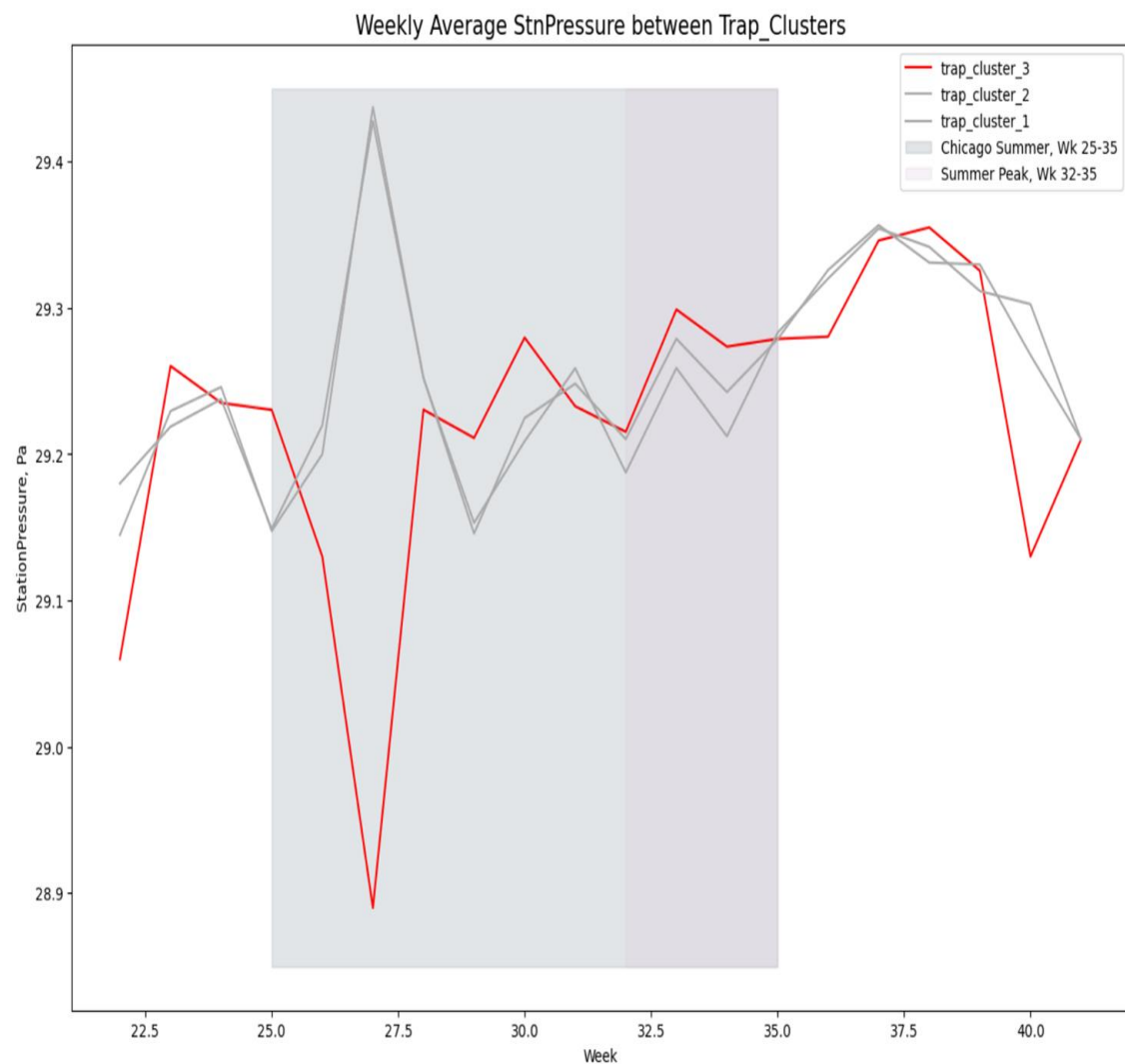
- Strong differences in lagged Relative Humidity between trap_cluster_3 and others leading up to the Summer Peak
- Trap_cluster_3 tends to stay consistently drier for longer in the weeks leading up to Summer Peak.



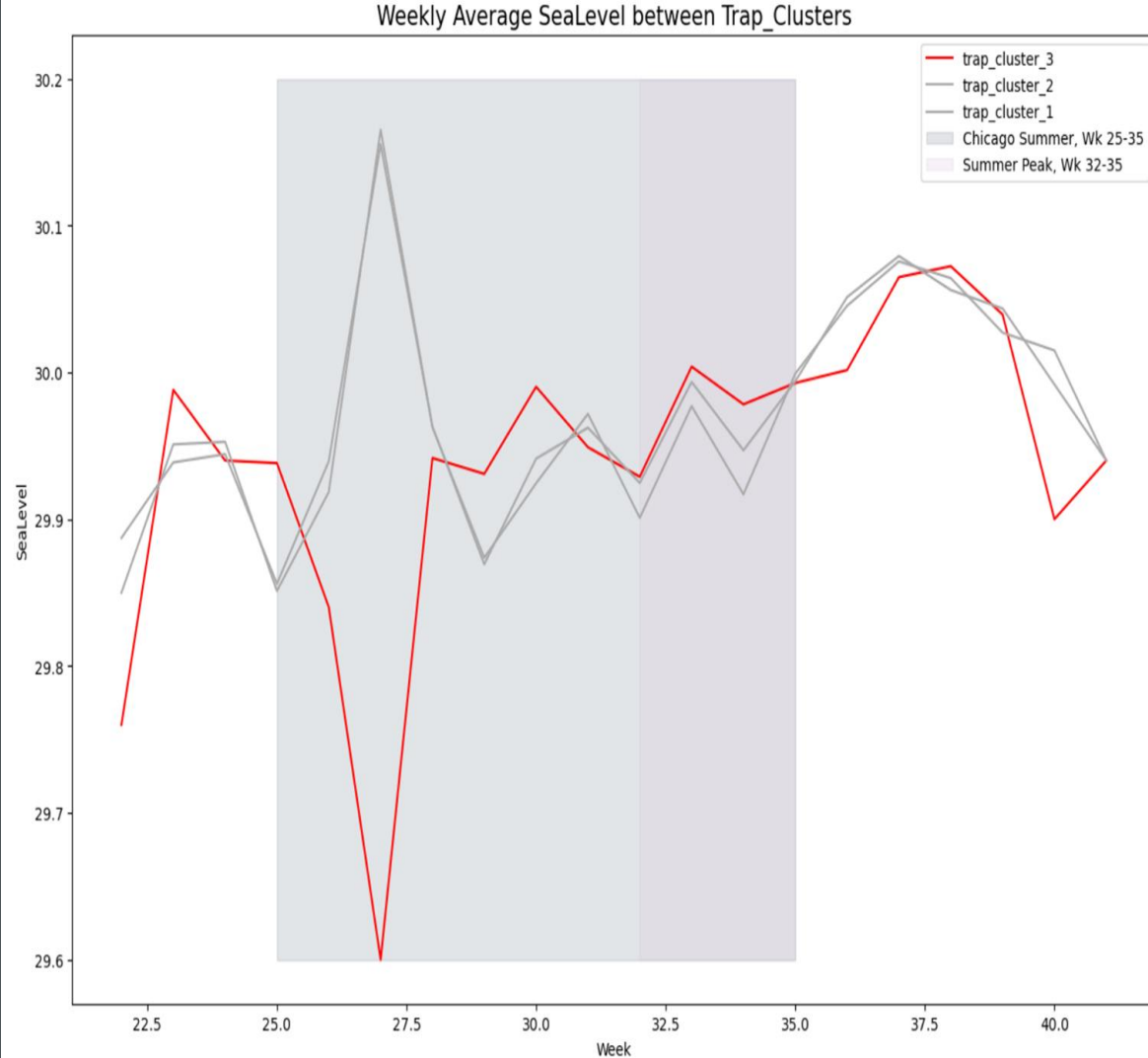
- Very distinct differences in the weekly lags between trap_cluster_3 and the rest leading up to Summer Peak.
- Precipitation conditions tend to be consistently drier leading up to Summer Peak
-



- `weekly_StnPressure_lag_2,3,4` are very important features leading up to the Summer Peak that influences whether `WnvPresent` is going to be in `trap_cluster_1,2,3`.
- 5 weeks before Summer Peak, the `StnPressure` always plummet resulting in `trap_cluster_3`, hence we used weekly lag 4 and 5 for `StnPressure`.



- 5 weeks before Summer Peak, the Sea Level always plummet resulting in trap_cluster_3, hence we used weekly lag 4 and 5 for Sea Level.
- Trap_cluster_3 tends to stay consistently drier for longer in the weeks leading up to Summer Peak.



FACTORS FOR FEATURES ENGINEERING

TRAINING DATASET

- Year/Month/Day
- Week

WEATHER DATASET

Standard Pressure

Humidity

Lagged Weather Duration

```
graph TD; TD[TRAINING DATASET] --> M((Modelling)); WD[WEATHER DATASET] --> M;
```

Modelling

Model Evaluation

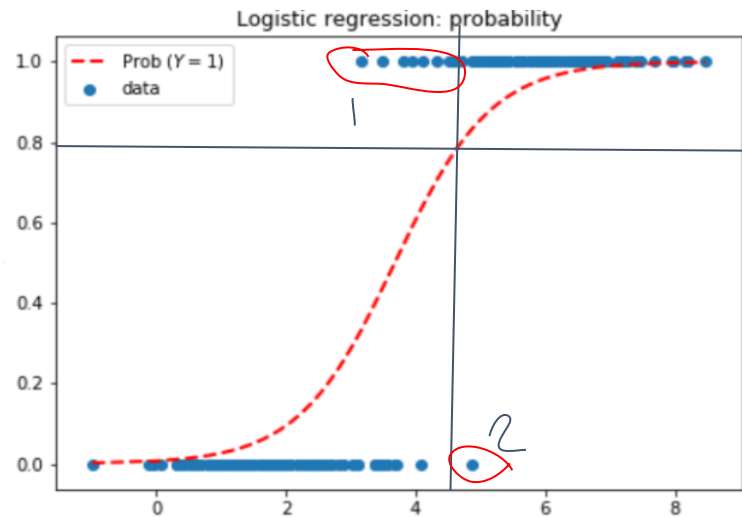
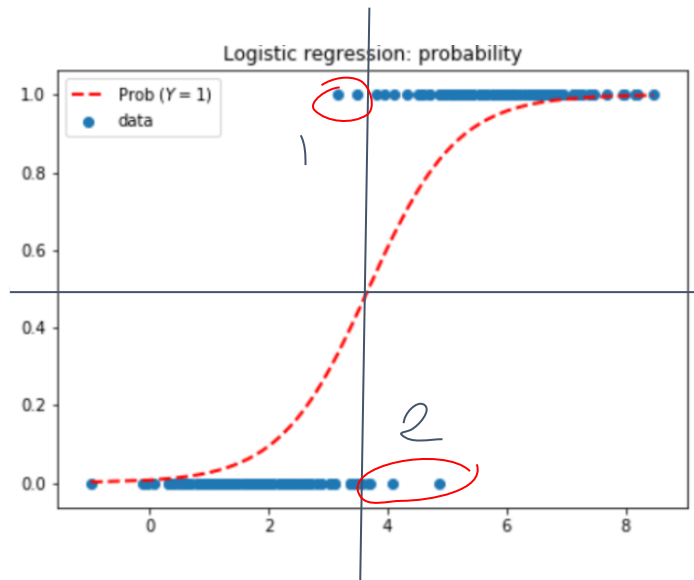
1. Model: Logistic Regression
2. Metric: AUC (kaggle score 0.70)
3. Coefficients

Model Evaluation: Logistic Regression

- Odds: $\frac{p}{1-p}$, where $p = P(y = 1)$
- $\log(\text{odds})$: $\log\left(\frac{p}{1-p}\right)$
- $\log\left(\frac{P(y=1)}{1-P(y=1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$
- $$\frac{P(y=1)}{1-P(y=1)} = e^{\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p\right)}$$
$$= e^{\beta_0} e^{\beta_1 x_1} e^{\beta_2 x_2} \dots e^{\beta_p x_p}$$

Model Evaluation: Logistic Regression

- Default threshold: $p = 0.5$



Model Evaluation: What is AUC

- Metric used: AUC
- Area under the curve (AUC) of the Receiver Operator Characteristic (ROC)
- Measure of the True Positive Rate (TPR) vs False Positive Rate (FPR)
- In the case of a Logistic Regression model, each pair (FPR, TPR) corresponds to a choice of the decision threshold of a binary decision

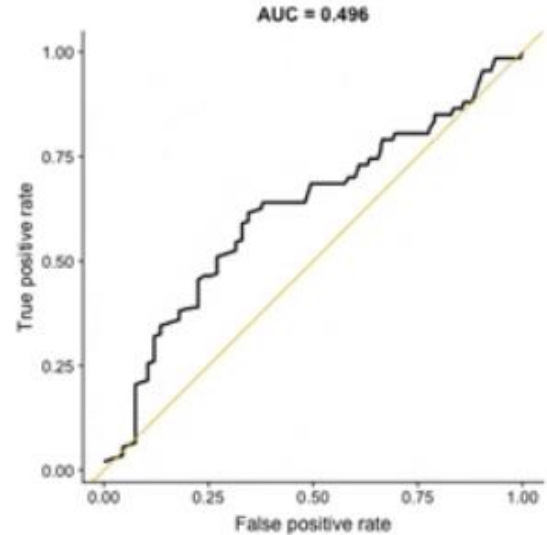
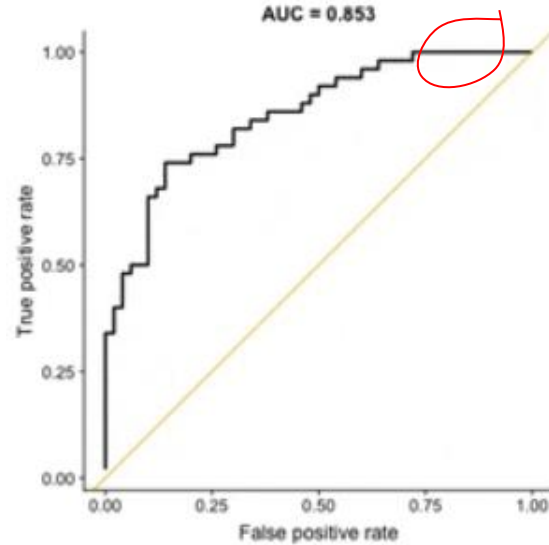
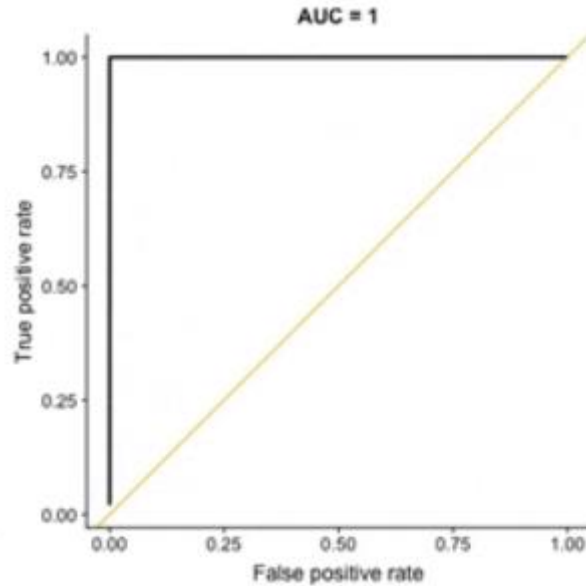
Model Evaluation: TPR/FPR

- **TPR (Sensitivity):**
$$\text{TPR} = \frac{\text{TP}}{P} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

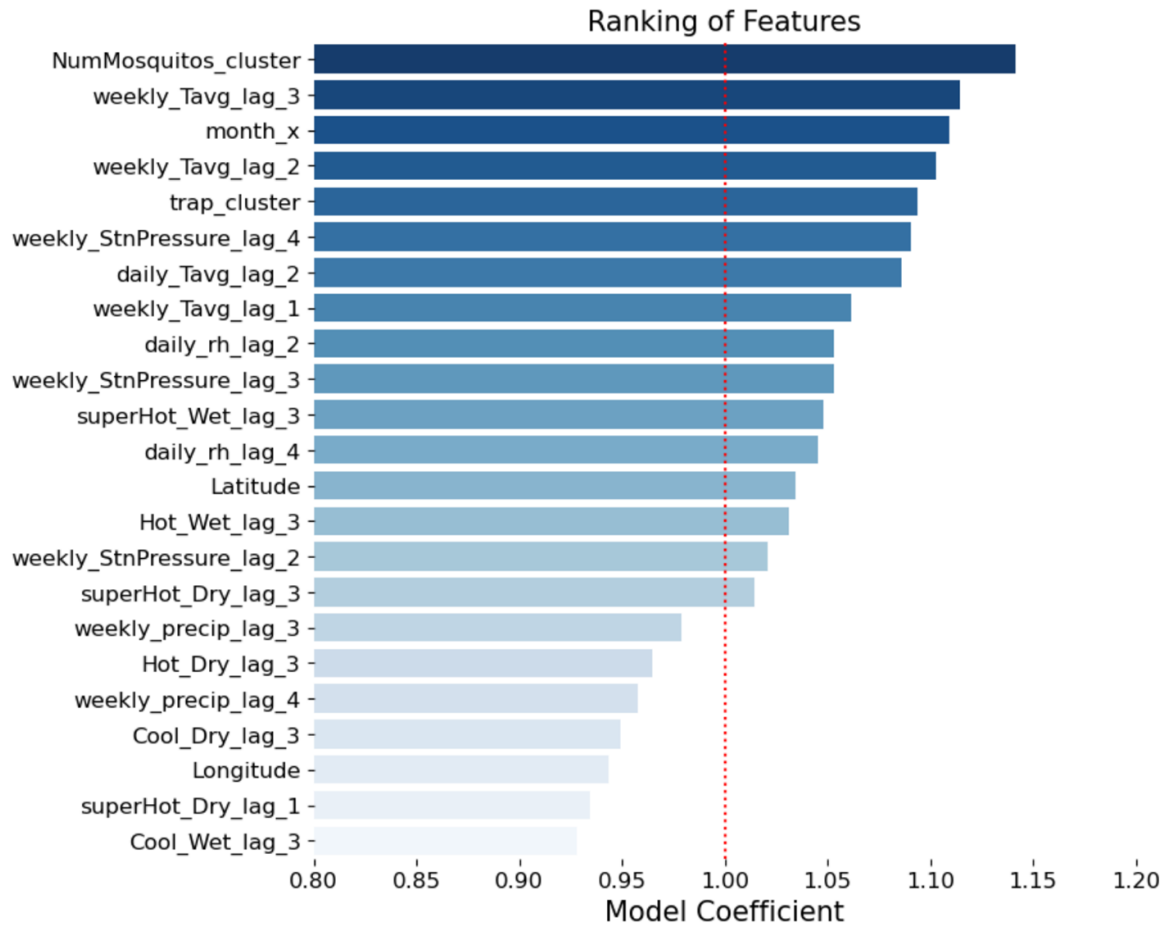
- **FPR (1 - Specificity):**
$$\text{FPR} = \frac{\text{FP}}{N} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

		Predicted condition	
Total population = P + N		Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

Model Evaluation: Visualisation of ROC and AUC



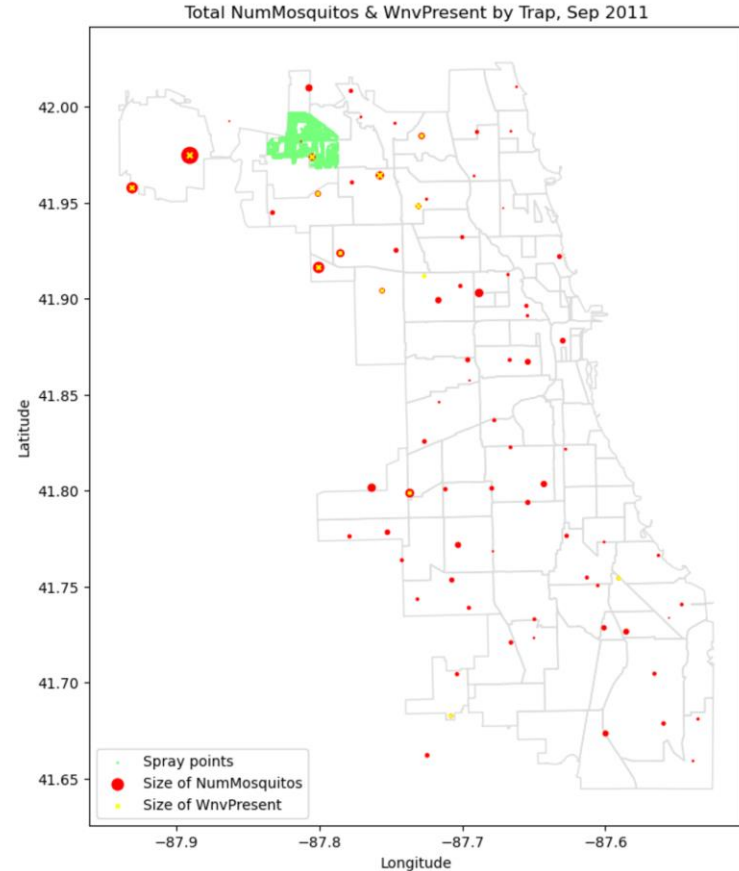
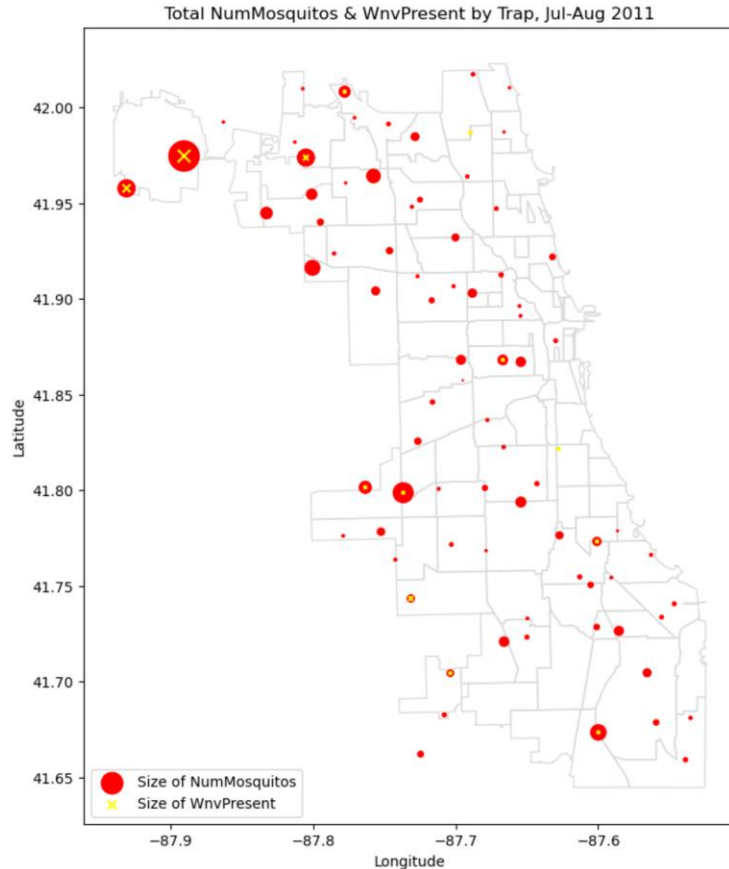
Model Evaluation: Coefficients



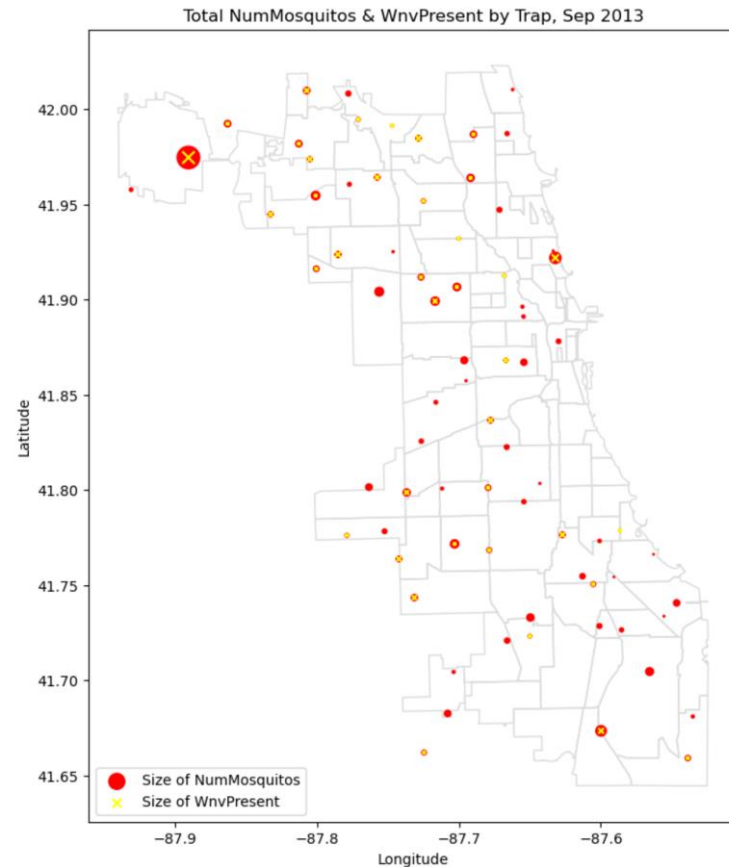
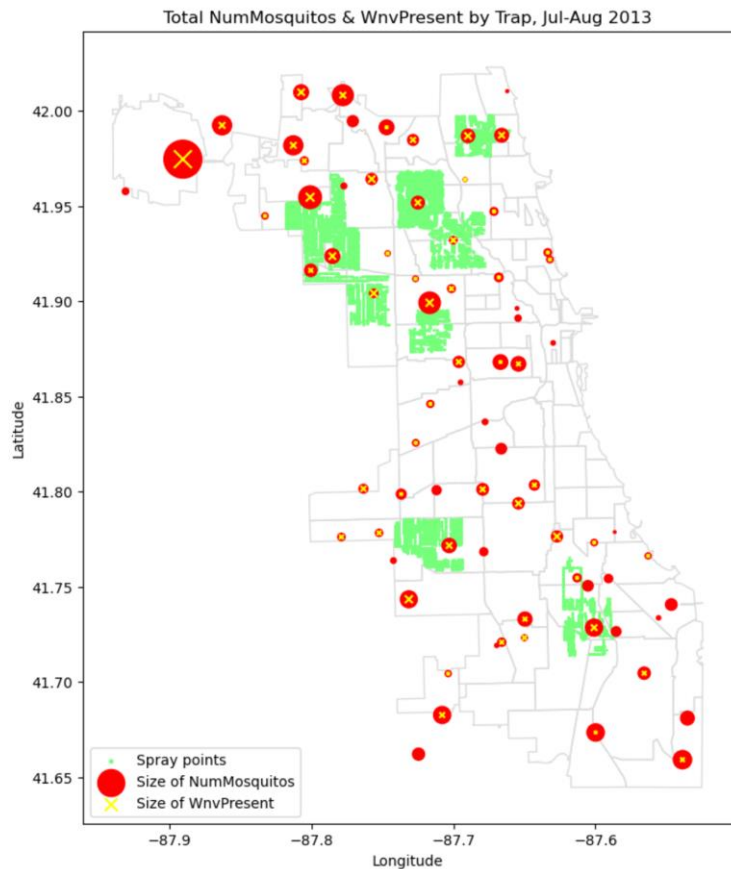
Cost-Benefit Analysis

Chicago's Spray Program vs Our Model

Chicago failed to spray in time despite finding strong clusters of NumMosquitos and WnvPresent in Jul-Aug 2011. In Sep 2011, they sprayed a small area but it was not the hotspot for WnvPresent.



In Jul-Aug 2013, Chicago sprayed extensively but inaccurately as BIG NumMosquitos and WnvPresent clusters have been detected from their Traps in those months. And in Sep 2013, they did not spray at all despite big clusters still being discovered in the most Northwest region.



Chicago's WNV Spray Program Post-Mortem



MISSED THE MARK!



MISSED THE TIMING!

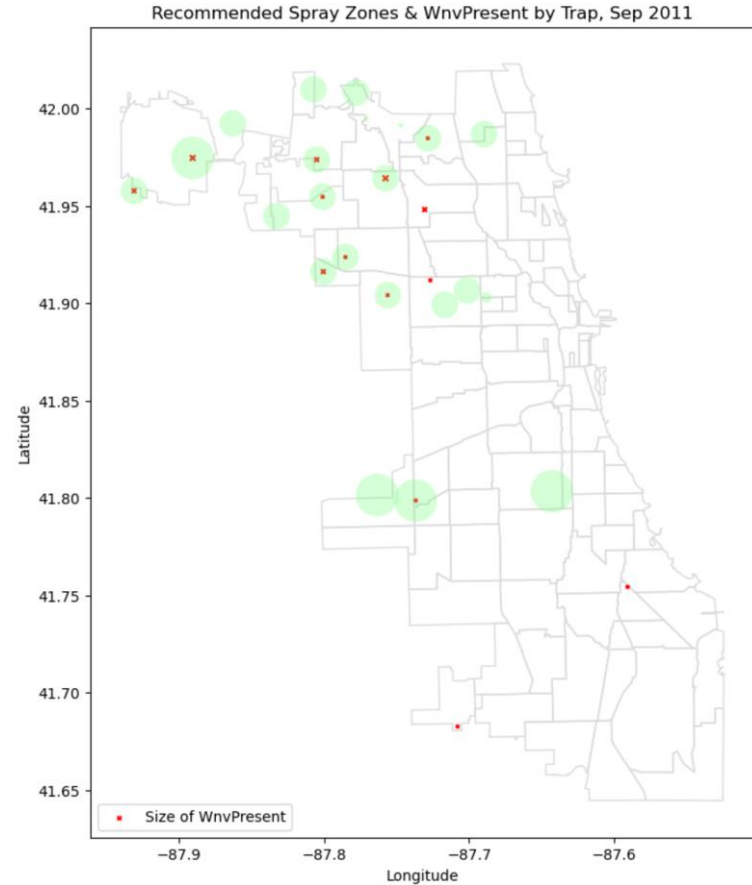
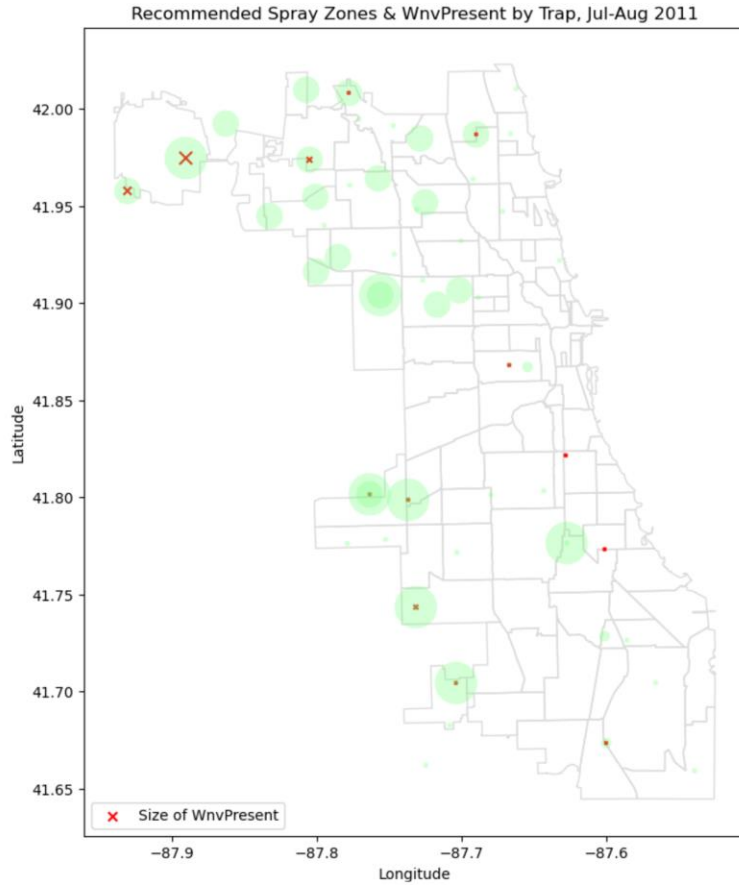


REACTIVE SPRAYING IS FUTILE!

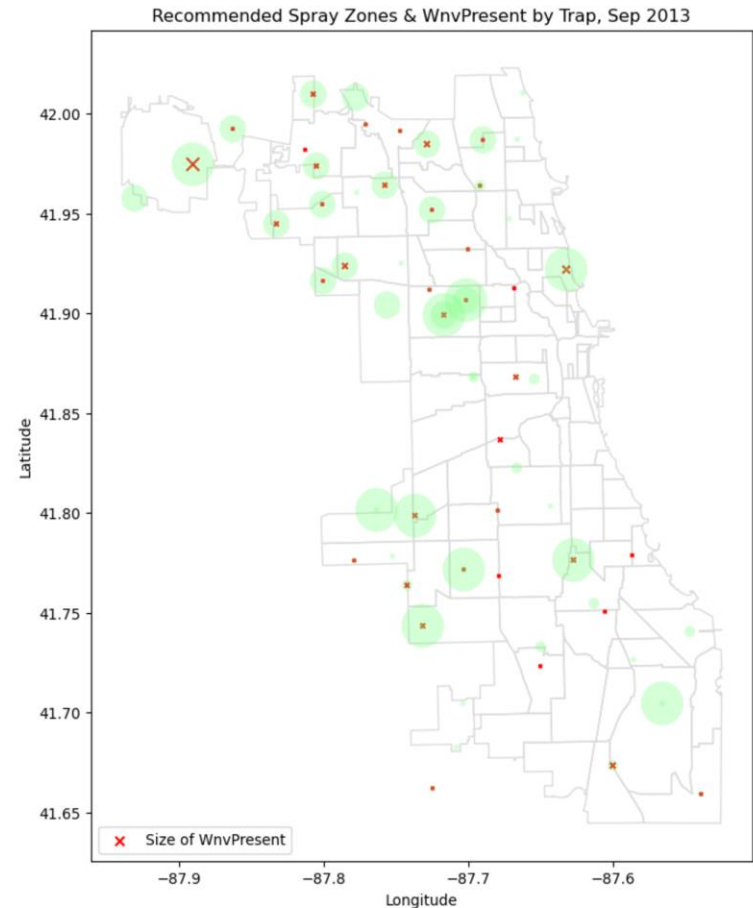
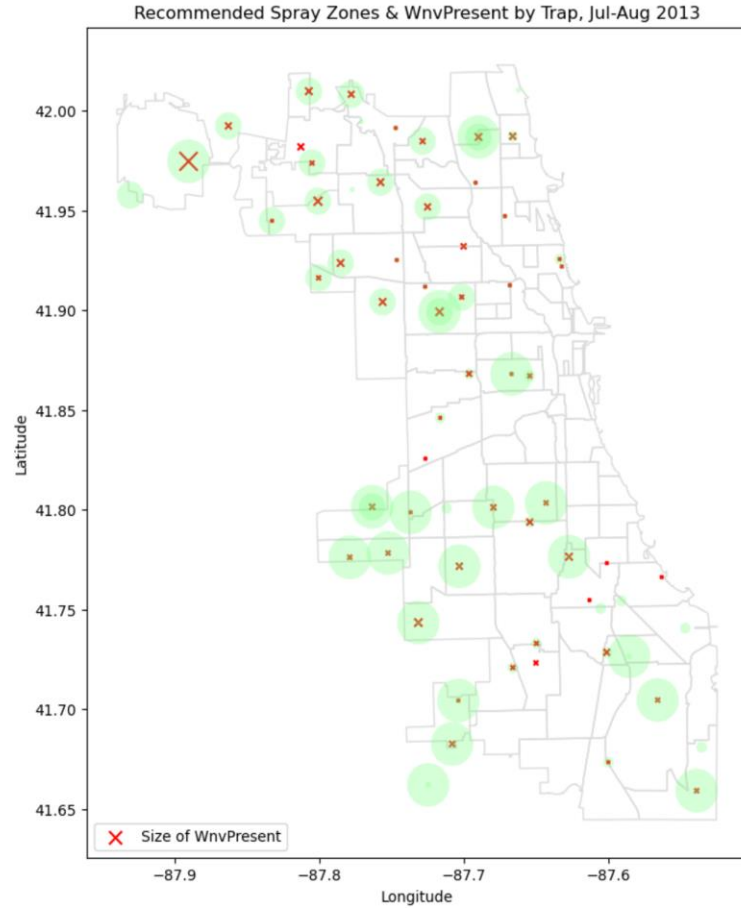
Let our Model recommend when, where, and how much to spray

- High Recall Score (“True positive rate”) : 0.78
- Contains only Leading Indicators
 - Most indicators’ data are available weeks in advance
 - Only three late indicators but still 2-4 days in advance
 - Helps in anticipating and planning weeks and days in advance
- Informs the necessary coverage of sprays around the Traps
- Spray **PROACTIVELY** and **ACCURATELY!**

Most recommended spray zones correspond well to the actual location and size of WnvPresent clusters.



Even more accurate in 2013! **Less False Positives!**



Case-study: Cost-Benefit Analysis for 2013

Table 2



Annual human WNV cases, average seasonal mosquito infection rate (MIR), and mosquito testing from 2005 to 2016 in Cook and DuPage counties.

Year	Number of human cases	Average MIR	Number of pools tested	Number of positive pools	Total number of mosquitoes tested
2005	181	5.33	7,165	1,939	271,235
2006	129	5.35	9,428	1,984	318,386
2007	43	2.65	12,131	1,259	375,520
2008	10	1.91	9,024	587	298,995
2009	1	1.14	9,450	298	311,220
2010	47	5.19	11,491	2,086	393,279
2011	24	3.10	8,911	939	287,774
2012	229	7.35	10,162	3,182	323,497
2013	66	4.26	11,078	1,967	407,326
2014	31	2.97	9,273	990	333,489
2015	36	3.57	7,725	1,046	314,363
2016	108	6.34	6,144	1,687	219,909

[Open in a separate window](#)

MIR = Mosquito infection rate; WNV = West Nile virus

Case-study: Cost-Benefit Analysis for 2013

	Chicago's Spraying	With our Model
Number of Infected Human Cases	66	15*
Average Medical Burden Cost per case (US\$) ¹	21,000	21,000
Total Medical burden costs (US\$)	1,386,000	304,920
Average Spray Cost per acre (US\$) ²	1.60	1.60
Total Spray(ed) Area, acres	60,234 [^]	338,081
Total Spray Costs (US\$)	96,374	540,929
Total Costs (US\$)³	1,483,000 	846,000 

List of Assumptions:

* We assumed that the actual infected human cases could have been reduced proportionately by a factor of our Model's recall rate of 0.78: $[(1-\text{Recall rate}) \times \text{Actual cases}]$

¹ According to a study published in the Journal of Infectious Diseases in 2014

² Assumed spray used is *Larvicide* which is less harmful to humans and environment and has a longer duration of 1-28 days depending on sunlight levels

³ Total Costs=Total Medical Burden costs + Total Spray Costs

[^] Actual sprayed area is based on the spray data provided, added with an effective radial zone of 100meters from each spray point

Conclusion & Recommendations

- Change their reactive spray model to a preventive and anticipatory model
- Use Traps' samplings to test the efficacy of their spray program, not use it to decide whether to spray or not
 - Better able to quantify the cost-benefit of their spray program
- Spray **PROACTIVELY**, and **ACCURATELY**!

