West Nile Virus Prediction

Team Members: Sarah Lim, Daiyu, Yew Tong, Yuan

Problem Statement

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.

In 2002, the first human cases of West Nile virus were reported in Chicago. By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program that is still in effect today.

Every week from late spring through the fall, mosquitos in traps across the city are tested for the virus. The results of these tests influence when and where the city will spray airborne pesticides to control adult mosquito populations.

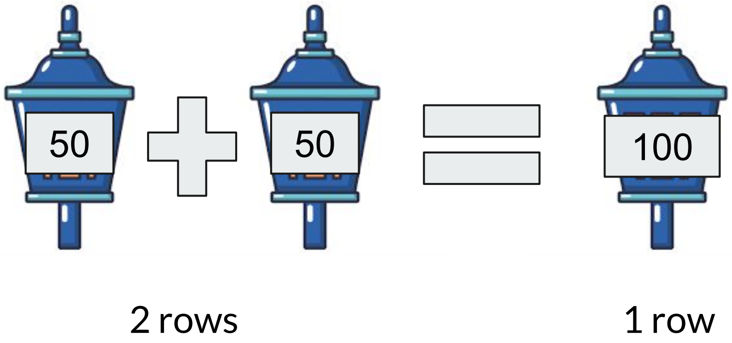
Representing CDPH, we will develop a model to predict outbreak of West Nile virus in mosquitoes so that the City of Chicago can more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.

Preprocessing and Analysis

Data Reshaping

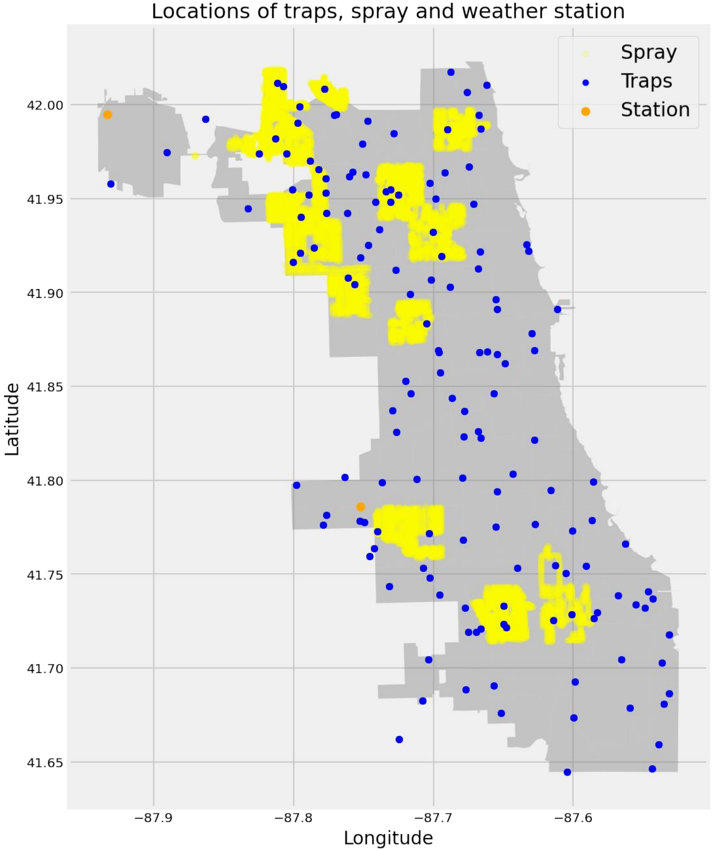
- Traps only recorded for >=1 mosquito
- Data is implicitly missing, so we added these rows back in

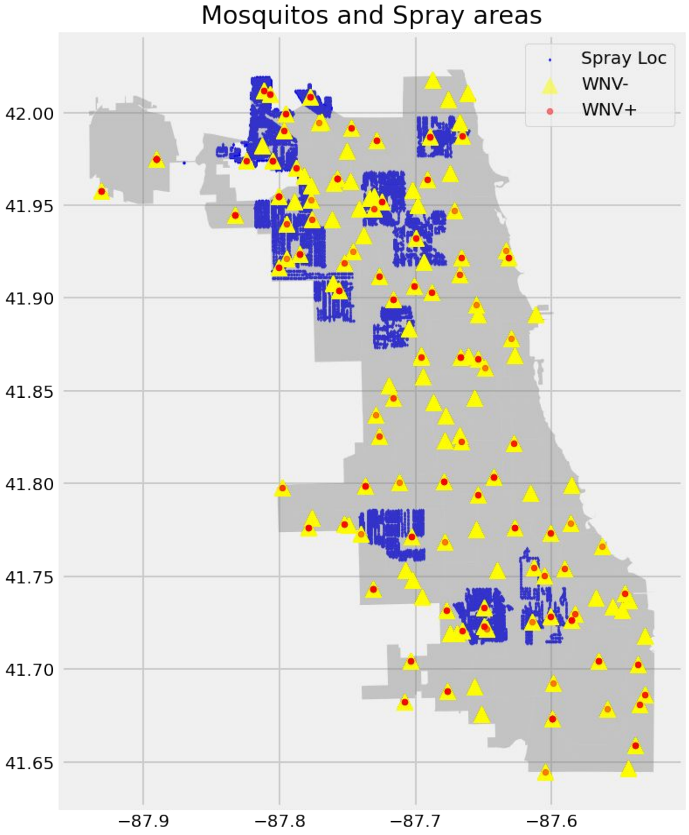
- Traps with >50 mosquitoes divided into multiple rows
- Combined these rows



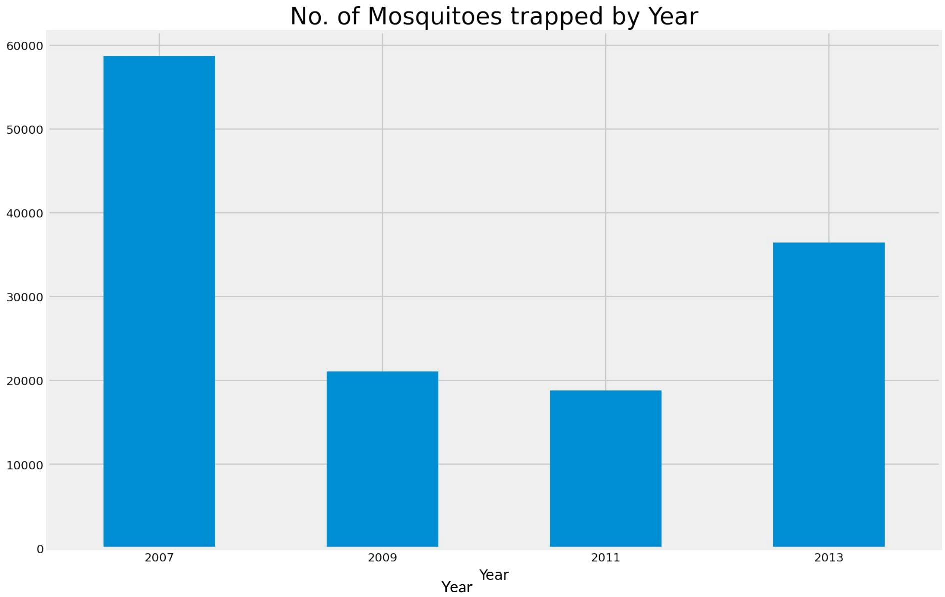
Note: this was done **before** the EDA

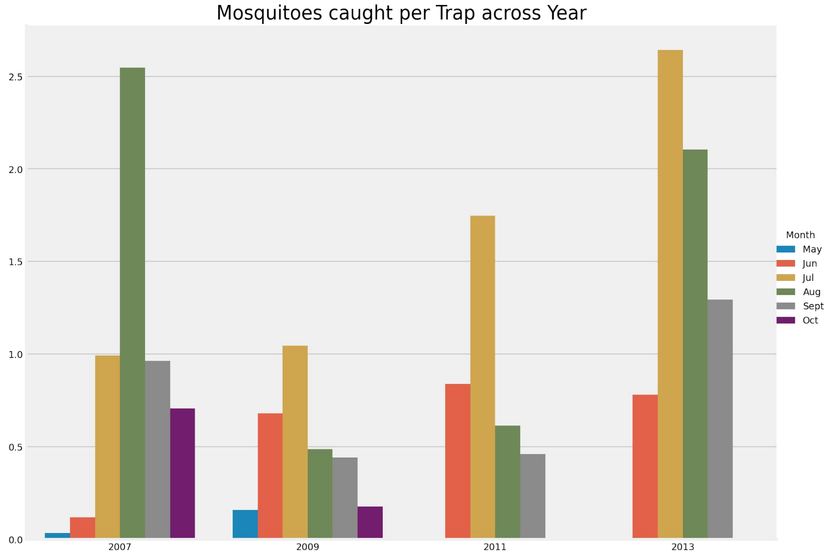
EDA


Locations of traps, spray and weather station
Spray
Traps
Station
Latitude
Longitude
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.9
−87.8
−87.7
−87.6

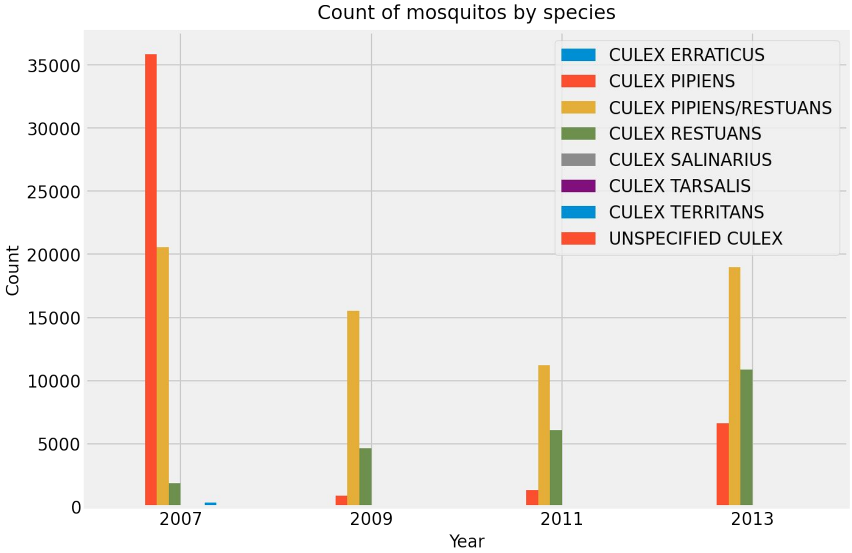

Mosquitos and Spray areas
Spray Loc
WNV-
WNV+
42.00
41.95
41.90
41.85
41.80
41.75
41.70
41.65
−87.9
−87.8
−87.7
−87.6

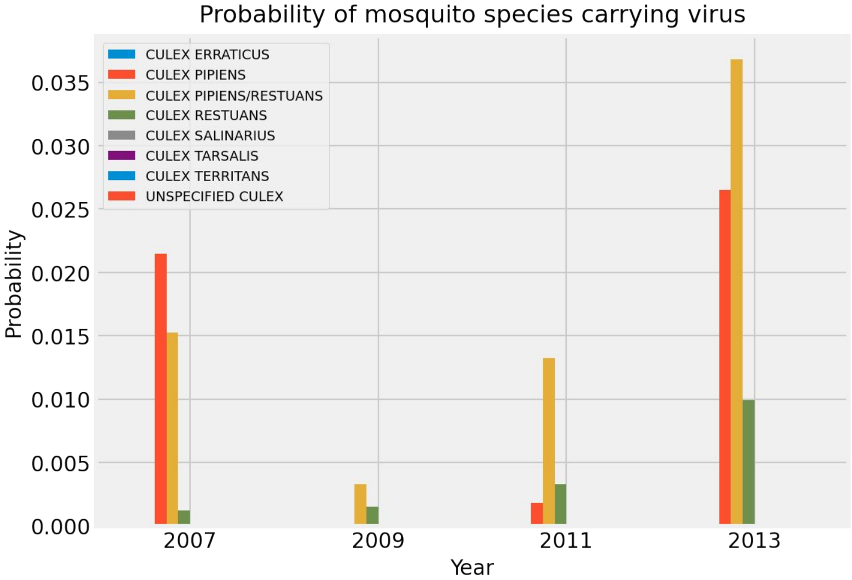
EDA


No. of Mosquitoes trapped by Year
60000
50000
40000
30000
20000
10000
0
2007
2009
2011
2013
Year
Year


Mosquitoes caught per Trap across Year
2.5
2.0
1.5
1.0
0.5
0.0
2007
2009
2011
2013
Month
May
Jun
Jul
Aug
Sept
Oct

EDA


Count of mosquitos by species
Count
Year
35000
30000
25000
20000
15000
10000
5000
0
2007
2009
2011
2013
CULEX ERRATICUS
CULEX PIPIENS
CULEX PIPIENS/RESTUANS
CULEX RESTUANS
CULEX SALINARIUS
CULEX TARSALIS
CULEX TERRITANS
UNSPECIFIED CULEX


Probability of mosquito species carrying virus
Probability
Year
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
2007
2009
2011
2013
CULEX ERRATICUS
CULEX PIPIENS
CULEX PIPIENS/RESTUANS
CULEX RESTUANS
CULEX SALINARIUS
CULEX TARSALIS
CULEX TERRITANS
UNSPECIFIED CULEX

Modelling

Baseline Model - Before EDA

- 99.6% of traps did NOT have West Nile Virus - this includes traps with no mosquitoes
- Models scored using AUC ROC, which has a baseline score of 0.5 (larger is better).

We made a model using only the provided data (no feature engineering) and got about 0.87 AUC ROC:

- Used logistic regression
- Adjusted the threshold for positive prediction (not 0.5)
- Not an accurate number for how well the model generalizes, more on this later

Modelling after EDA

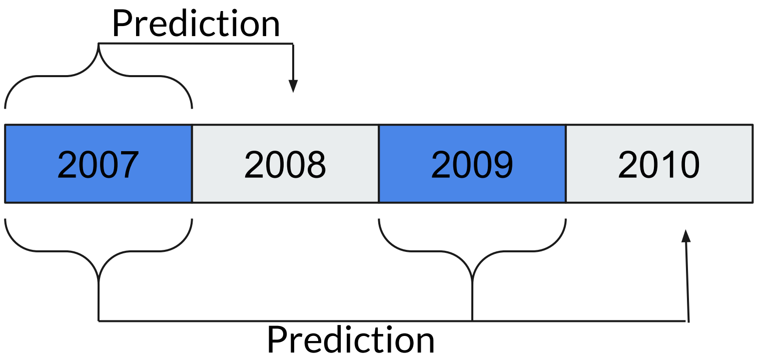
- Added features from EDA
  - Weather
  - Max number of mosquitoes is in Aug, we added features showing the time difference from Aug, to try and achieve linearity
  - Encoded species/trap as dummy variables
- Removed some measurements
  - For traps that were measured after spraying (for that year) we removed these measurements
  - They do not provide an accurate picture of whether WNV is present

Modelling - Data used

To simulate actual predictions:

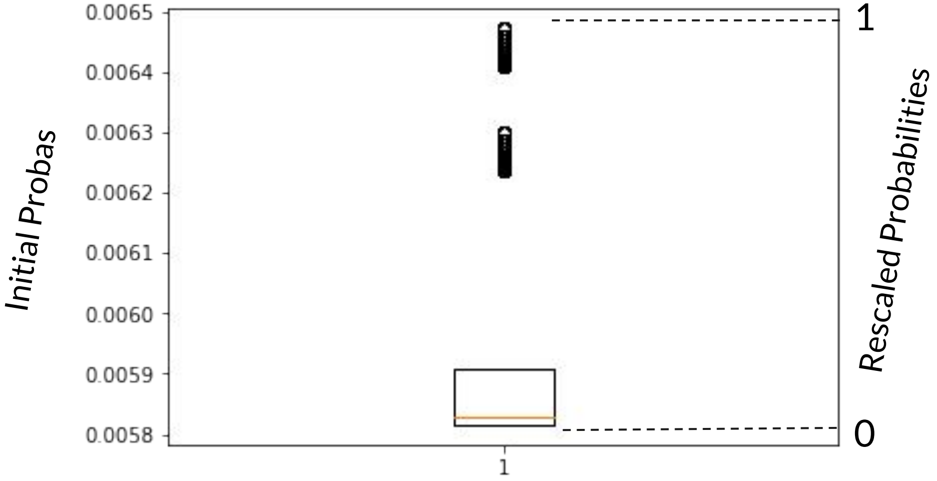
- 2008 predicted using only 2007 data
- 2010 predicted using only 2007 and 2009 data
- .etc

* Model selection process also used this method, so for one class of model (e.g. Logistic Regression) we needed to test 4 models



Models - Logistic Regression

- 4 models (one for each year)
- Predicted probabilities are very low (near 0)
- Rescaled probabilities to [0,1] range, then used 0.5 as the threshold
- About 0.97 AUC ROC but this is because the data is overfit to the years that we have.

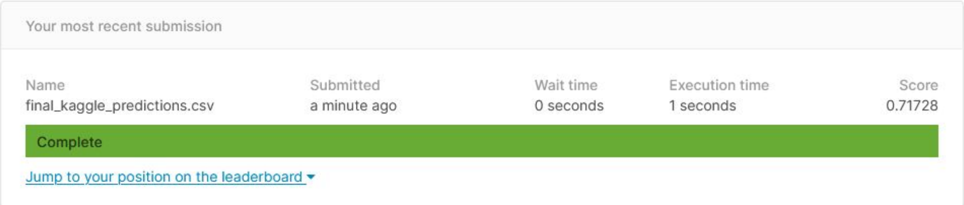


Models - Support Vector Classifier

- Model only predicts majority class
- Used bootstrapping to get balanced classes and avoid this issue
  - SVCs don’t provide probabilities so we couldn’t use the same approach as before
- AUC ROC is ~0.6-0.7

Final Model - Logistic Regression

- Extracted coefficients to see the relative importance of features (next slides)
- Used this to predict on Kaggle
- AUC ROC of 0.717 from Kaggle submission
- AUC ROC is not as good as training set because of several reasons
  - Mosquito population varies greatly between years
  - We have zero data for any test years so we can’t compare this
  - Model is likely overfitted to the years in which we had data.



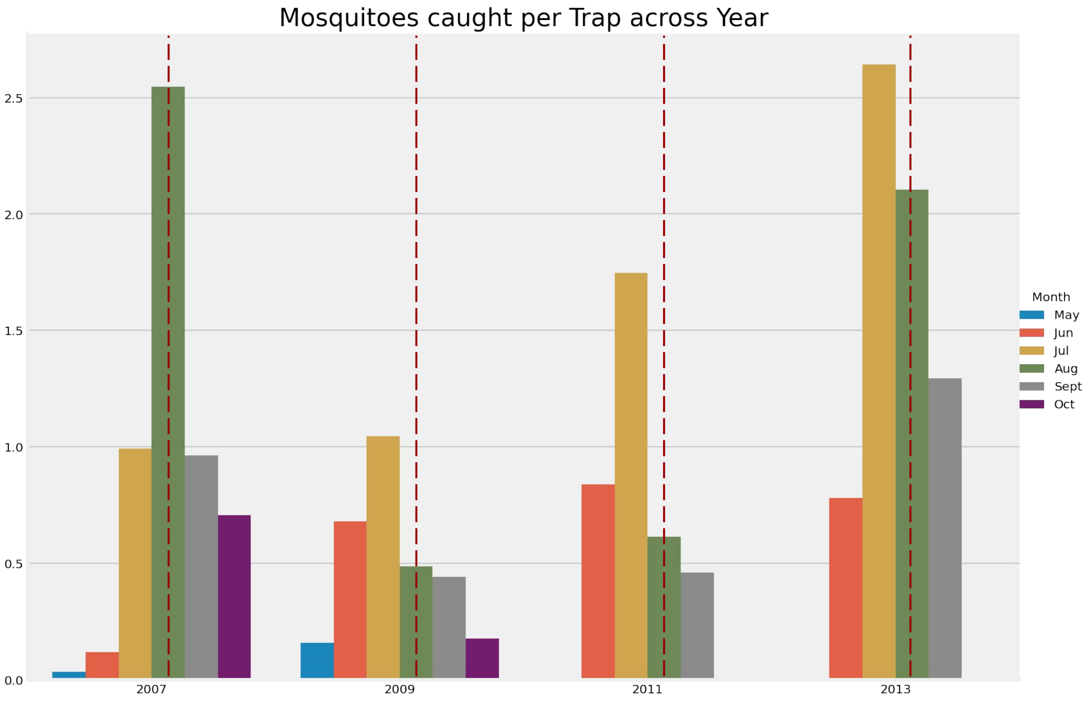
Your most recent submission

Name	Submitted	Wait time	Execution time	Score
final_kaggle_predictions.csv	a minute ago	0 seconds	1 seconds	0.71728

Conclusions

Conclusions

Relative time from August was the most important feature.



Conclusions

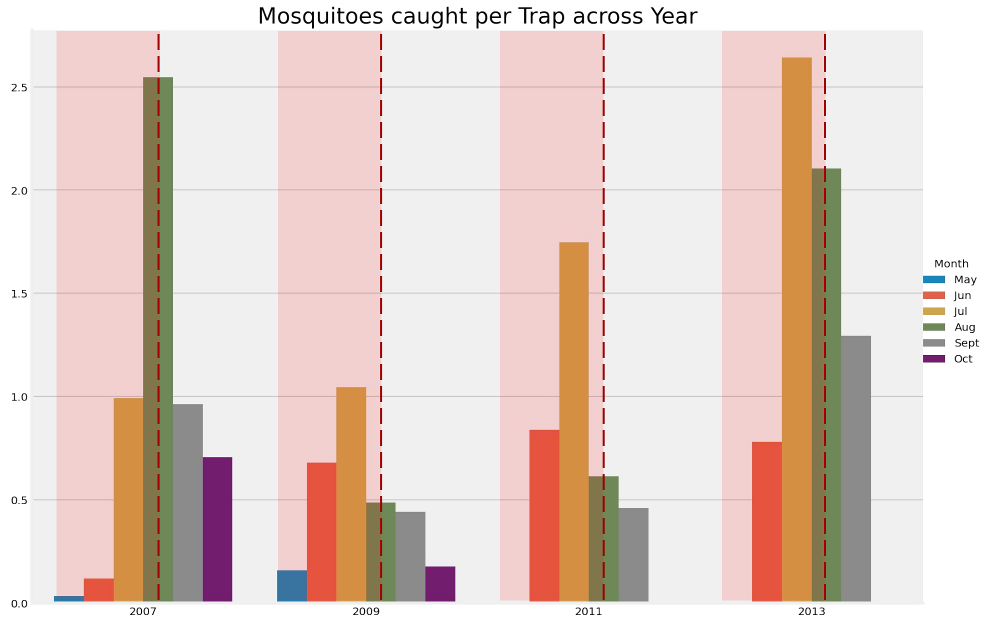
Mosquito Species is an important indicator:

Culex Pipiens

Culex Restuans

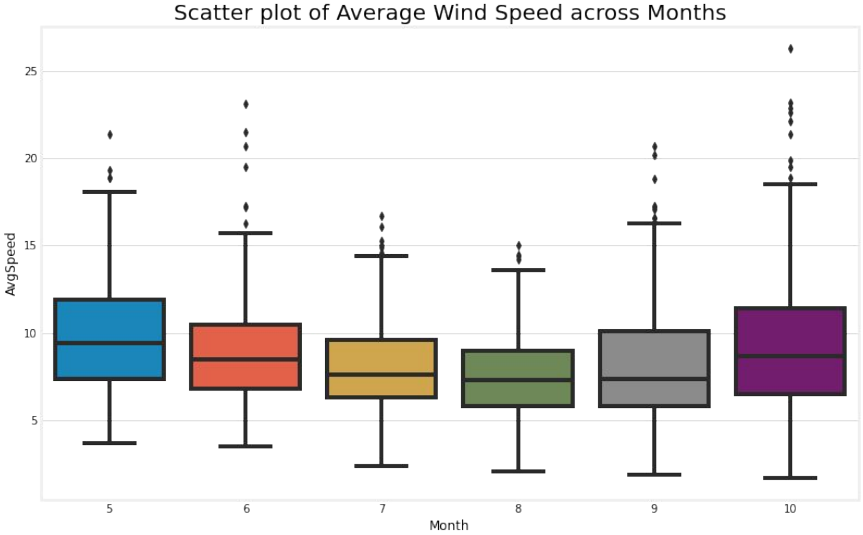
Recommendations

- Spraying should be done before and up till August each year.
- Research how to more effectively target Culex Pipiens and Restuans.



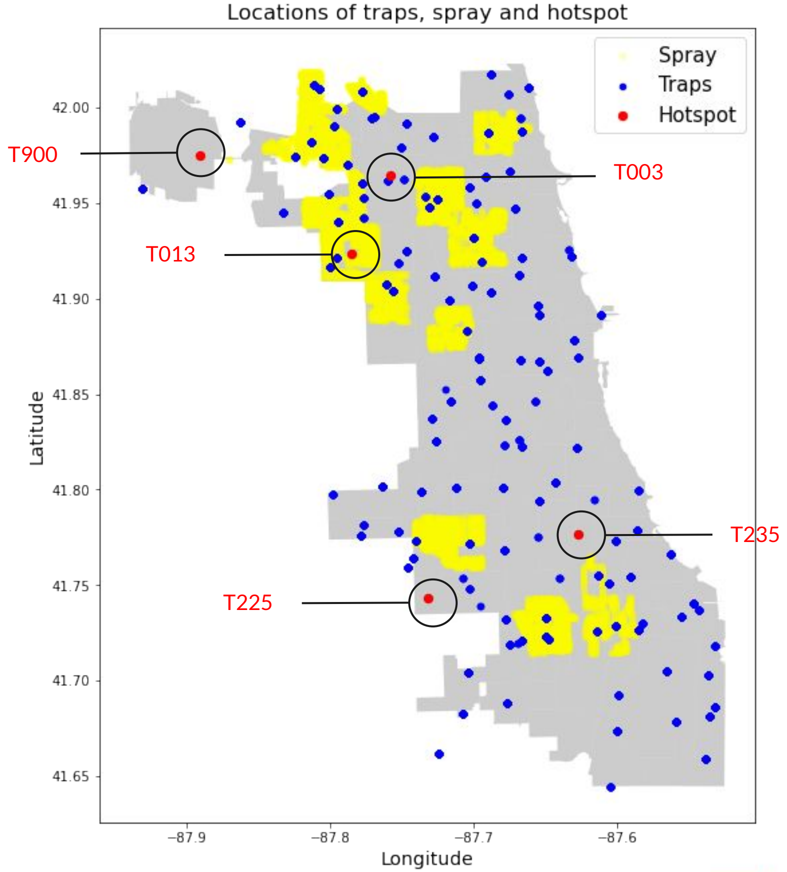
Recommendations

- Save budget for spraying when snowfall is expected.
- Intervention should be considered for windy periods or periods with lower humidity.



Recommendations

- Areas in the vicinity of traps that best predict for WNV should be sprayed or studied as case studies in greater detail.



Questions?