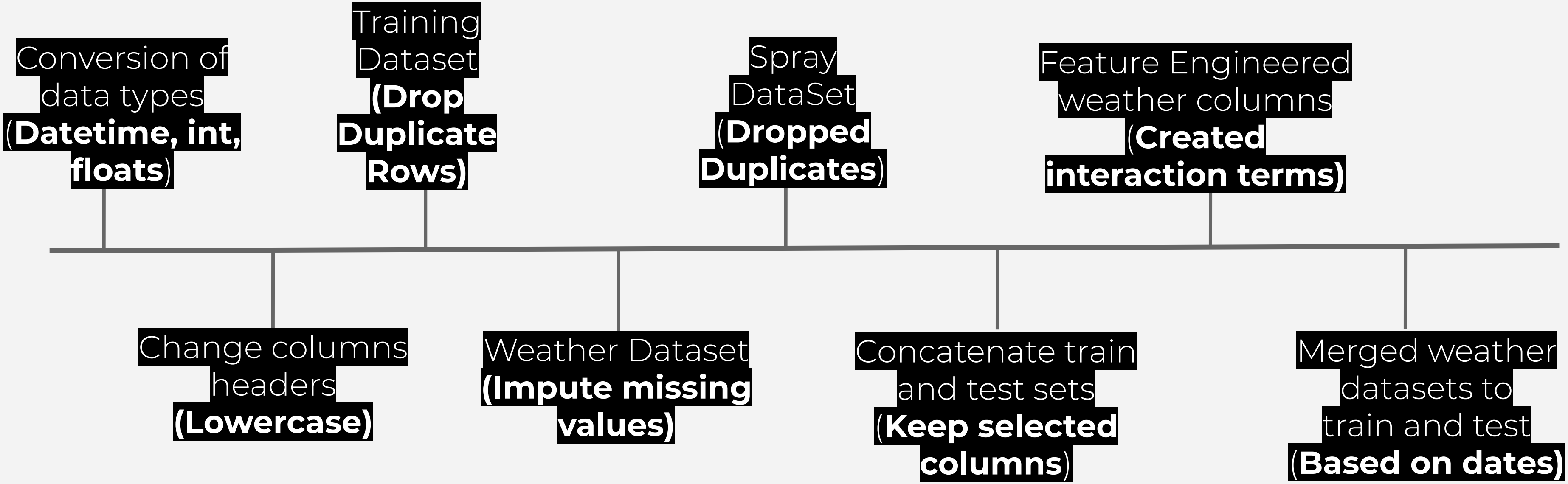
Project 4 (Group 2)

PREDICTING WEST NILE VIRUS IN CHICAGO

Problem Statement

- To build an **effective classifier** and make **predictions of outbreaks of the West Nile virus** in mosquitoes.
- People infected with the virus can develop **fever and serious neurological illnesses leading to death.**
- Success will be evaluated via **ROC-AUC** and **sensitivity**.

Data Cleaning & Preprocessing



Conversion of data types (**Datetime, int, floats**)

Change columns headers **(Lowercase)**

Training Dataset **(Drop Duplicate Rows)**

Weather Dataset **(Impute missing values)**

Spray DataSet (**Dropped Duplicates**)

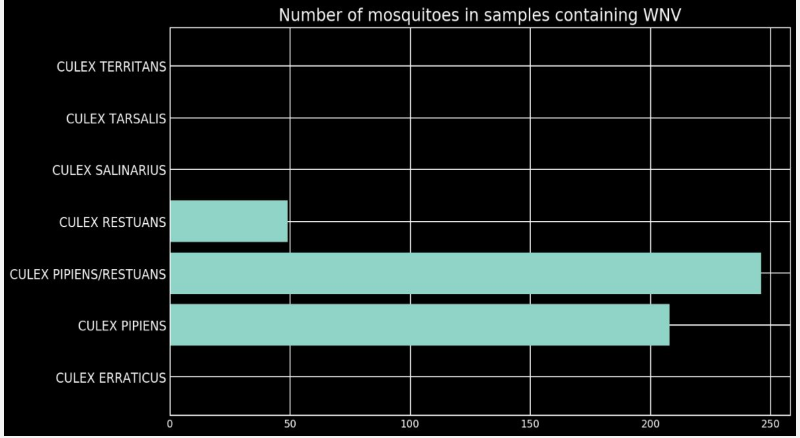
Concatenate train and test sets (**Keep selected columns**)

Feature Engineered weather columns (**Created interaction terms)**

Merged weather datasets to train and test (**Based on dates)**

EDA (Training Dataset)

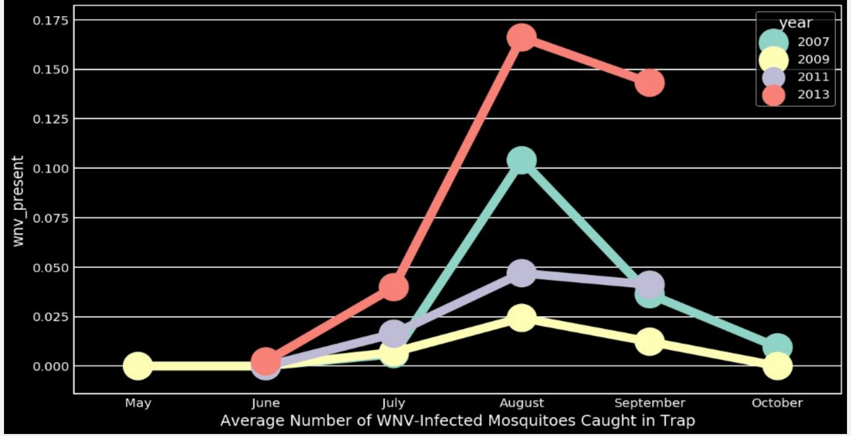
WNV by **species type**



CARRIERS OF VIRUS:

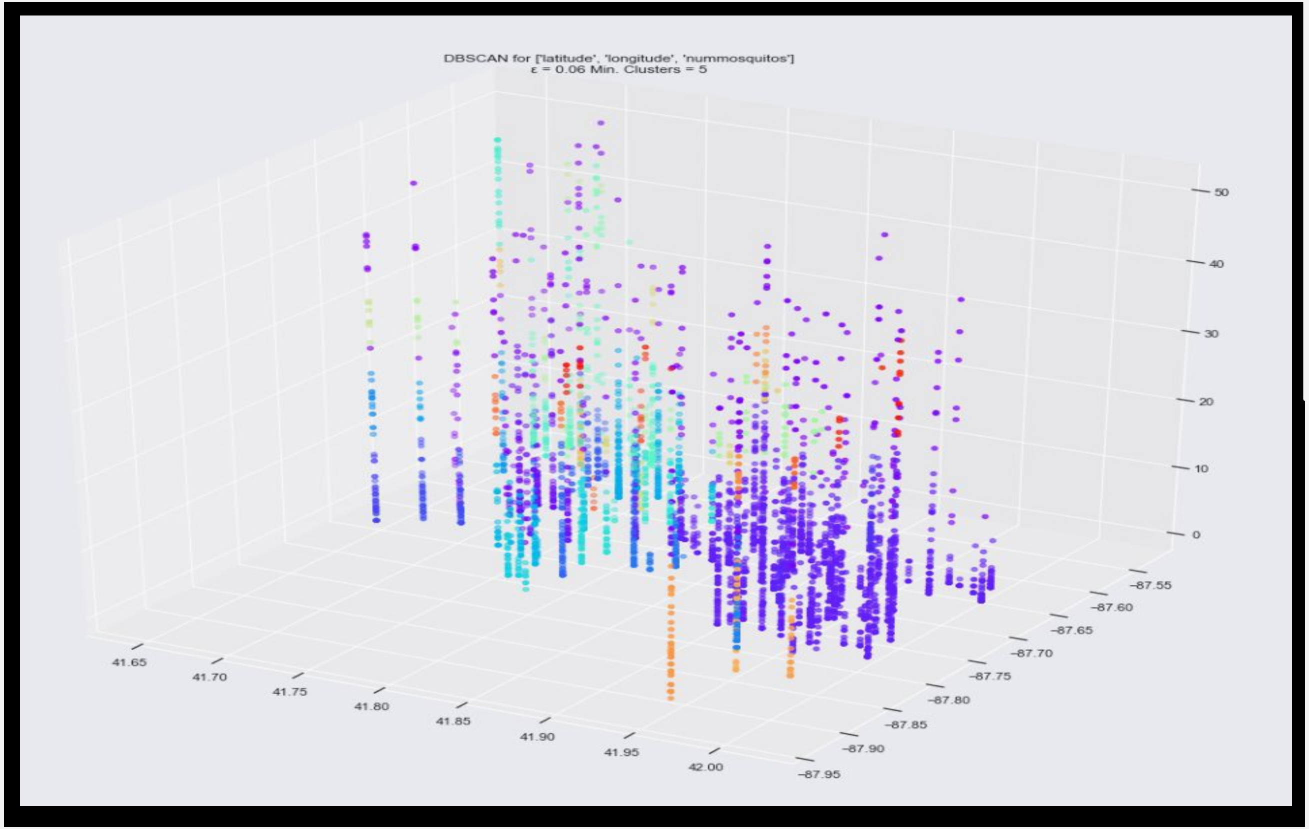
- **CULEX PIPIENS/RESTUANS**
- **CULEX RESTUANS**
- **CULEX PIPIENS**

Average number of WNV **mosquitoes caught** in trap



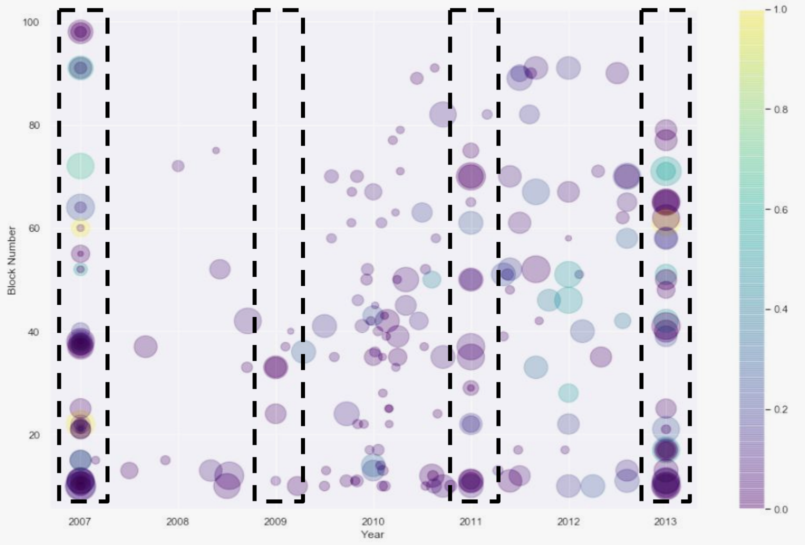
PEAKS IN AUGUST

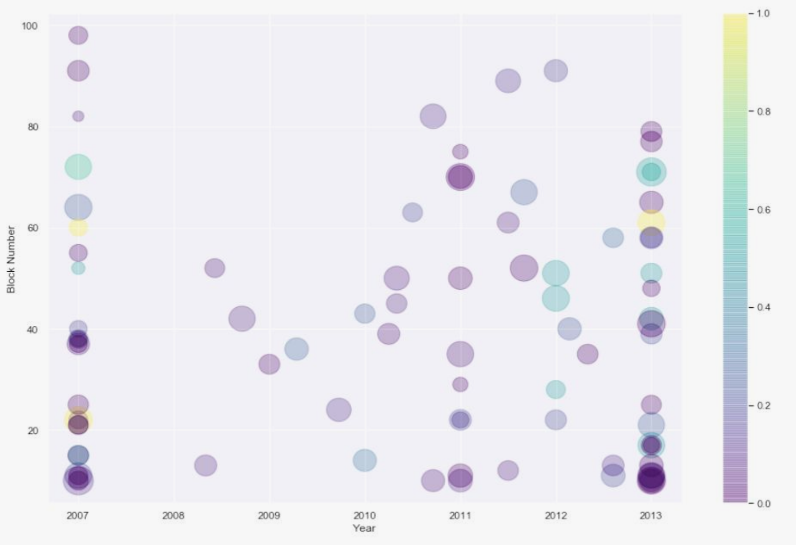
EDA (Training Dataset)



SILHOUETTE SCORE:	**-0.129**
NUMBER OF OUTLIERS:	**291**
NUMBER OF CLUSTERS:	**64**

EDA (Training Dataset)

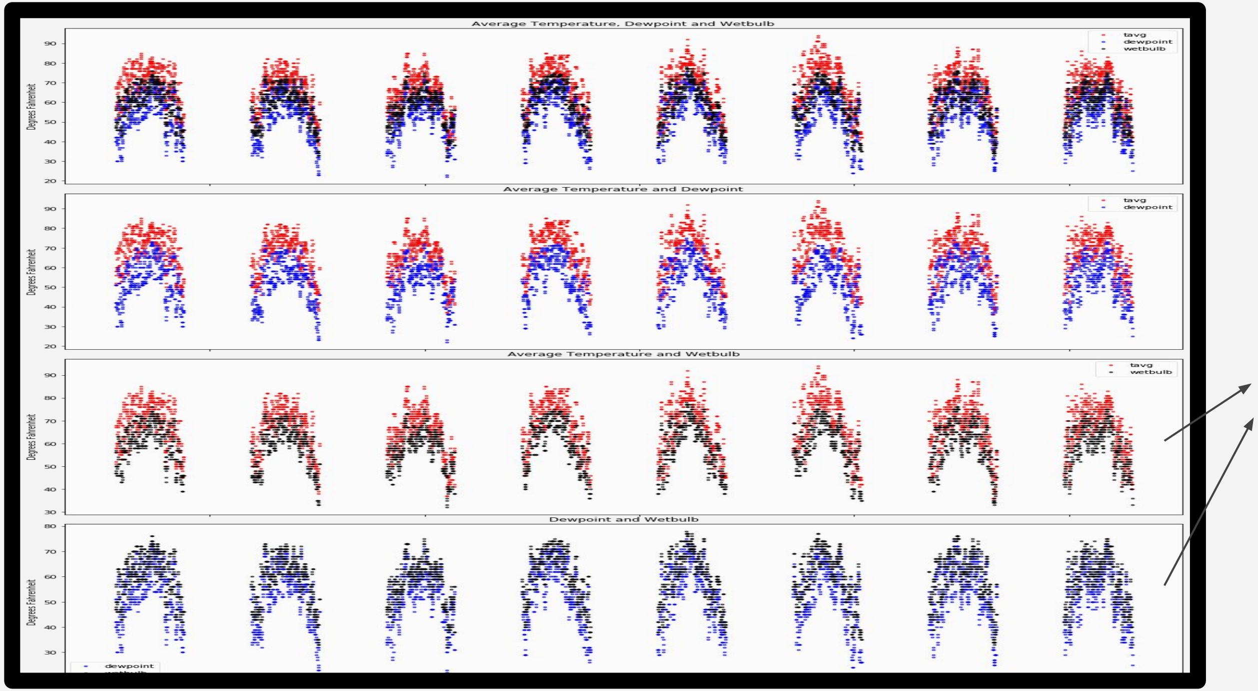




CLUSTERS

OUTLIERS

EDA (Weather Dataset)



● **tavg**

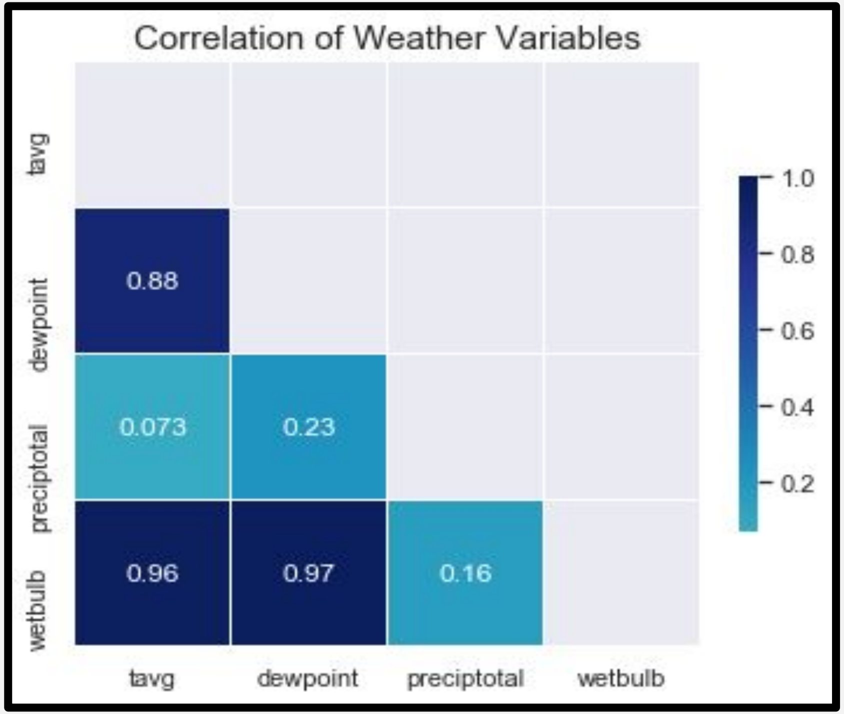
● **wetbulb**

● **dewpoint**

OVERLAPS BETWEEN:

- **wetbulb & dewpoint**
- **wetbulb & tavg**

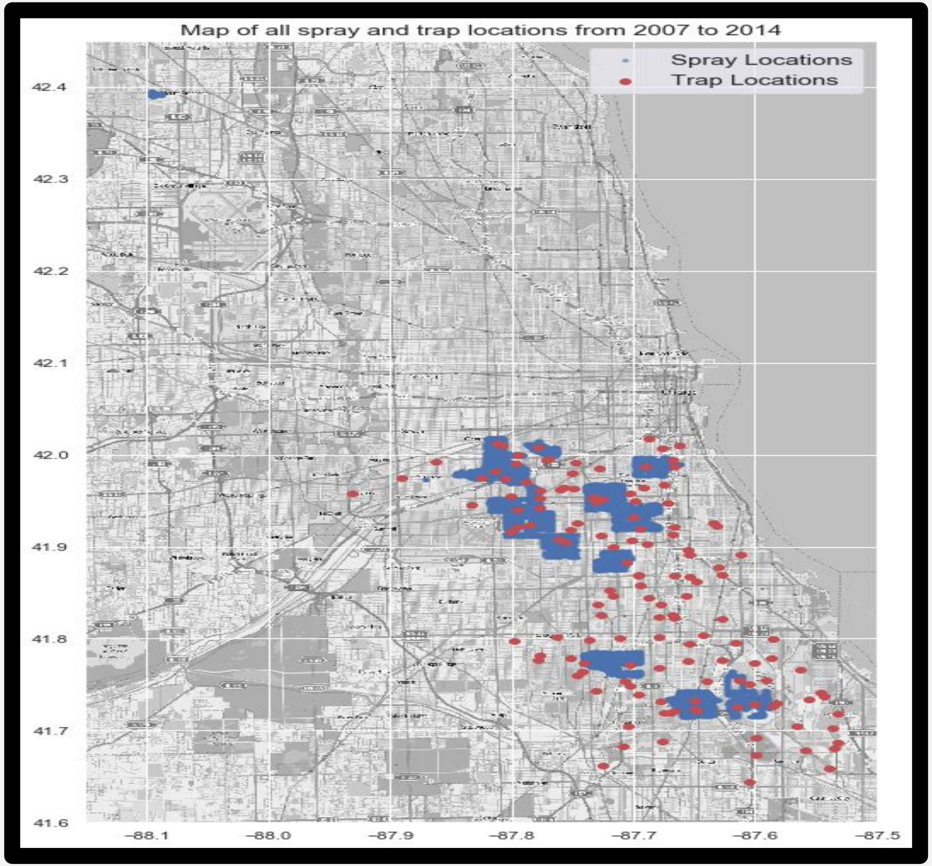
EDA (Weather Dataset)



HIGH CORRELATION BETWEEN:

- **wetbulb & dewpoint**
- **wetbulb & tavg**

EDA (Spray Dataset)



SPRAY AND TRAP LOCATIONS

Modelling & Predictions

Applied SMOTE to Imbalanced Data

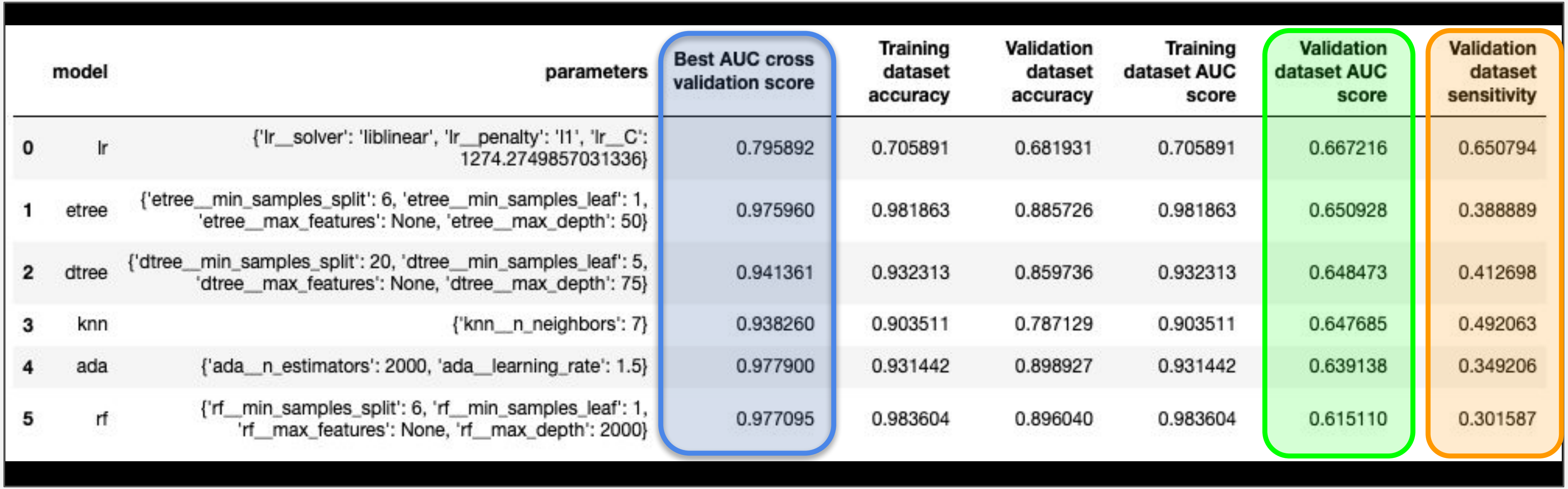
- K Nearest Neighbours
- Logistic Regression
- Extra Trees
- Random Forests
- Decision Trees
- ADA Boost

Chose Classifier Models

- Accuracy
- ROC AUC
- Sensitivity

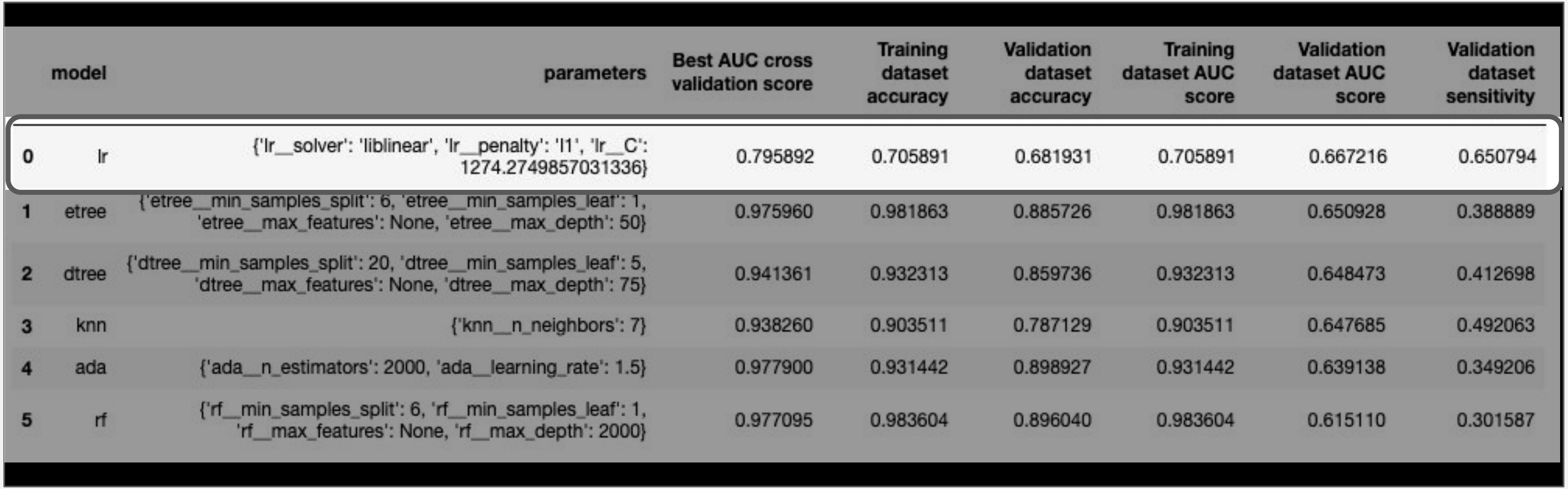
Evaluated Models

Modelling & Predictions



	model	parameters	Best AUC cross validation score	Training dataset accuracy	Validation dataset accuracy	Training dataset AUC score	Validation dataset AUC score	Validation dataset sensitivity
0	lr	{'lr__solver': 'liblinear', 'lr__penalty': 'l1', 'lr__C': 1274.2749857031336}	0.795892	0.705891	0.681931	0.705891	0.667216	0.650794
1	etree	{'etree__min_samples_split': 6, 'etree__min_samples_leaf': 1, 'etree__max_features': None, 'etree__max_depth': 50}	0.975960	0.981863	0.885726	0.981863	0.650928	0.388889
2	dtree	{'dtree__min_samples_split': 20, 'dtree__min_samples_leaf': 5, 'dtree__max_features': None, 'dtree__max_depth': 75}	0.941361	0.932313	0.859736	0.932313	0.648473	0.412698
3	knn	{'knn__n_neighbors': 7}	0.938260	0.903511	0.787129	0.903511	0.647685	0.492063
4	ada	{'ada__n_estimators': 2000, 'ada__learning_rate': 1.5}	0.977900	0.931442	0.898927	0.931442	0.639138	0.349206
5	rf	{'rf__min_samples_split': 6, 'rf__min_samples_leaf': 1, 'rf__max_features': None, 'rf__max_depth': 2000}	0.977095	0.983604	0.896040	0.983604	0.615110	0.301587

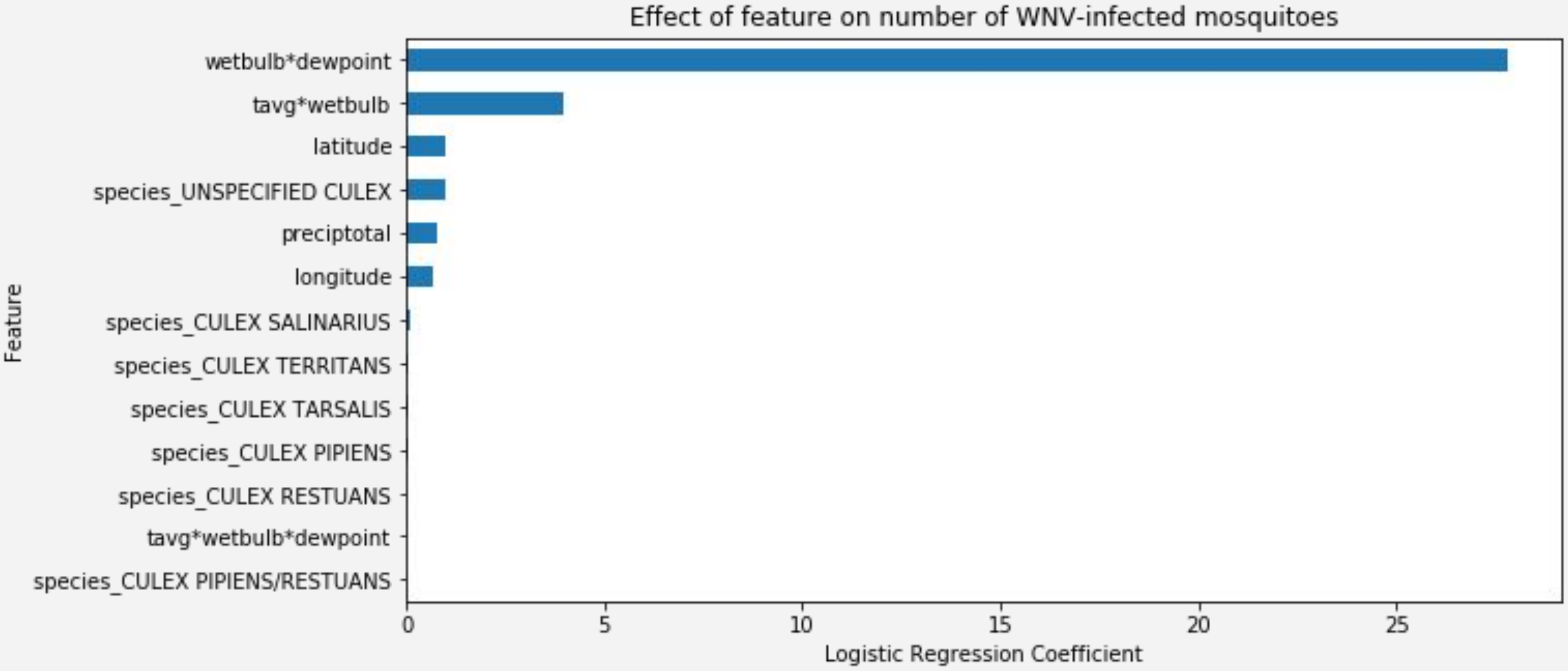
Modelling & Predictions



	model	parameters	Best AUC cross validation score	Training dataset accuracy	Validation dataset accuracy	Training dataset AUC score	Validation dataset AUC score	Validation dataset sensitivity
0	lr	{'lr__solver': 'liblinear', 'lr__penalty': 'l1', 'lr__C': 1274.2749857031336}	0.795892	0.705891	0.681931	0.705891	0.667216	0.650794
1	etree	{'etree__min_samples_split': 6, 'etree__min_samples_leaf': 1, 'etree__max_features': None, 'etree__max_depth': 50}	0.975960	0.981863	0.885726	0.981863	0.650928	0.388889
2	dtree	{'dtree__min_samples_split': 20, 'dtree__min_samples_leaf': 5, 'dtree__max_features': None, 'dtree__max_depth': 75}	0.941361	0.932313	0.859736	0.932313	0.648473	0.412698
3	knn	{'knn__n_neighbors': 7}	0.938260	0.903511	0.787129	0.903511	0.647685	0.492063
4	ada	{'ada__n_estimators': 2000, 'ada__learning_rate': 1.5}	0.977900	0.931442	0.898927	0.931442	0.639138	0.349206
5	rf	{'rf__min_samples_split': 6, 'rf__min_samples_leaf': 1, 'rf__max_features': None, 'rf__max_depth': 2000}	0.977095	0.983604	0.896040	0.983604	0.615110	0.301587

Our model achieved a Kaggle Greatness Index (KGI) score of **0.64734**

Modelling & Predictions



Cost-Benefit-Analysis

Quantitative **COST** Factors

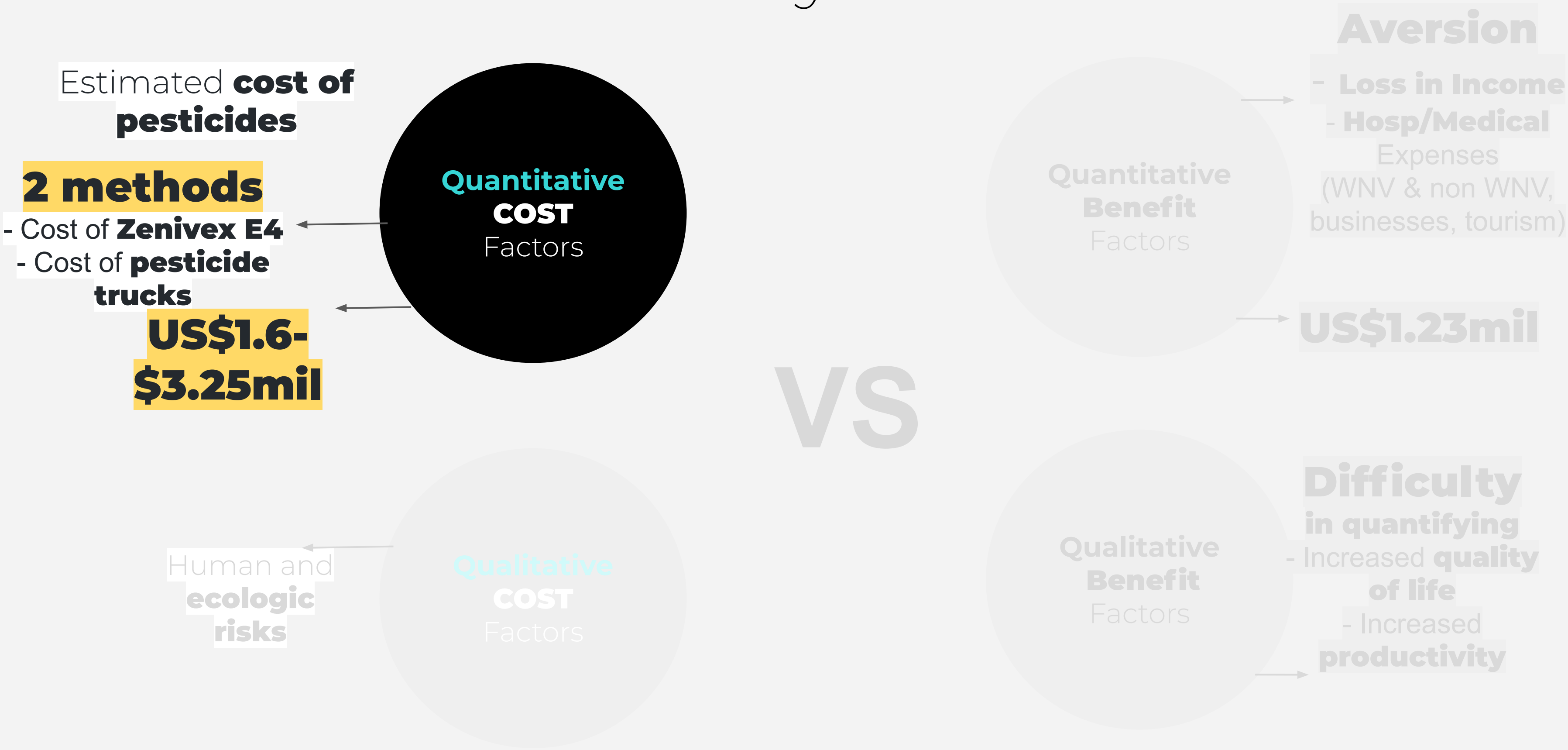
Qualitative **COST** Factors

VS

Quantitative **Benefit** Factors

Qualitative **Benefit** Factors

Cost-Benefit-Analysis



Estimated **cost of pesticides**

2 methods

- Cost of **Zenivex E4**
- Cost of **pesticide trucks**

US$1.6-$3.25mil

Quantitative **COST** Factors

VS

Quantitative **Benefit** Factors

Aversion

- **Loss in Income**
- **Hosp/Medical** Expenses (WNV & non WNV, businesses, tourism)

US$1.23mil

Qualitative **COST** Factors

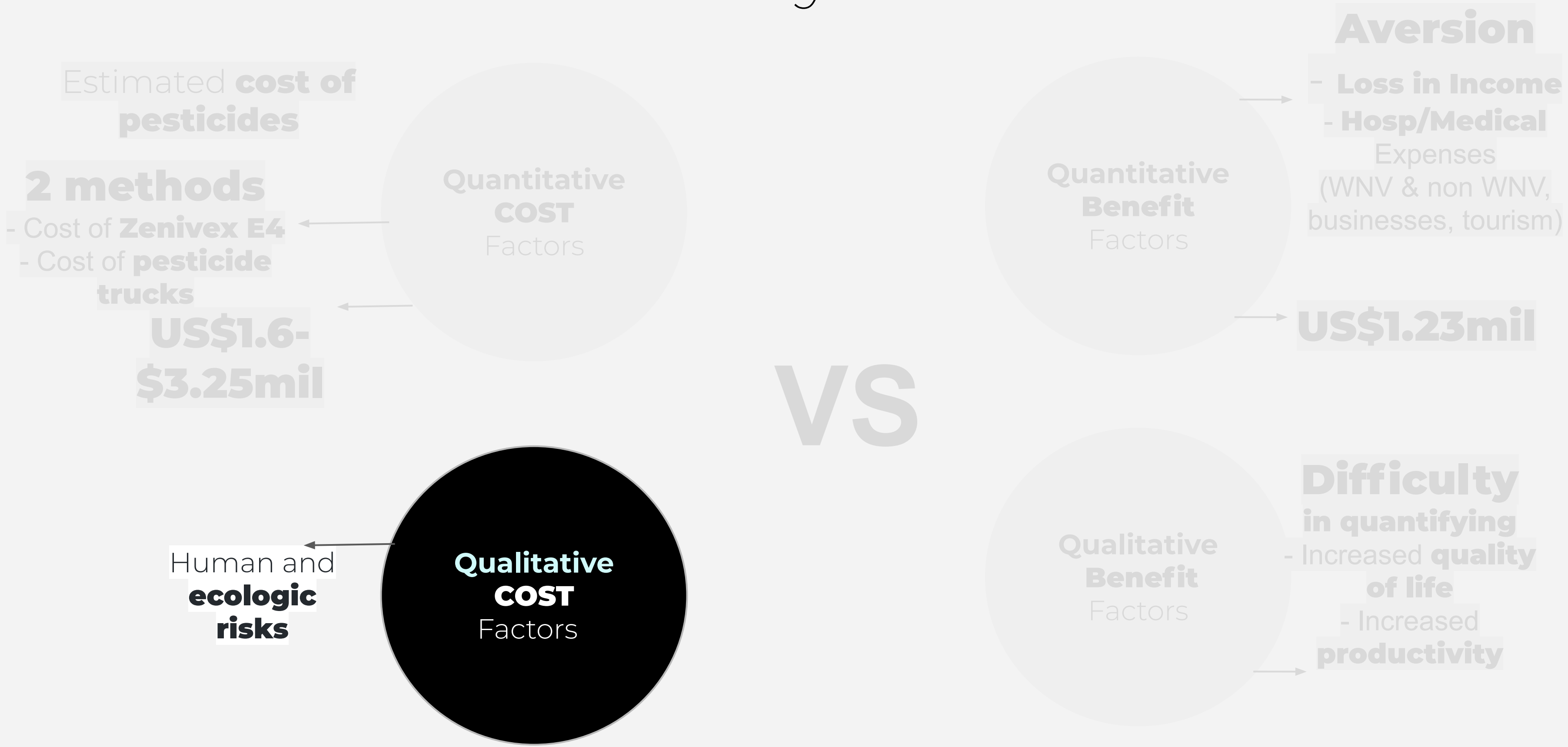
Human and **ecologic risks**

Qualitative **Benefit** Factors

Difficulty in quantifying

- Increased **quality of life**
- Increased **productivity**

Cost-Benefit-Analysis



Estimated **cost of pesticides**

2 methods
- Cost of **Zenivex E4**
- Cost of **pesticide trucks**

US$1.6-$3.25mil

Quantitative COST Factors

Human and **ecologic risks**

Qualitative COST Factors

VS

Quantitative Benefit Factors

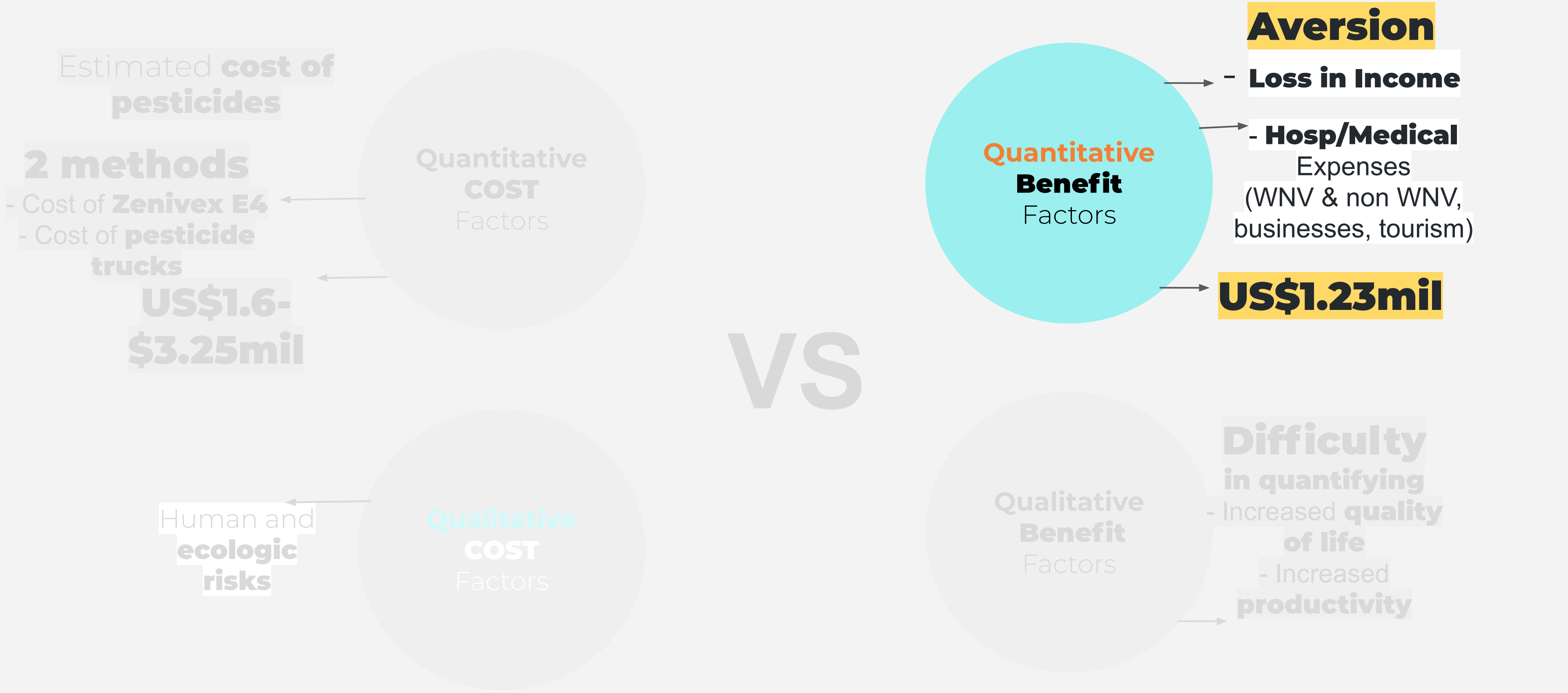
Aversion
- **Loss in Income**
- **Hosp/Medical** Expenses (WNV & non WNV, businesses, tourism)

US$1.23mil

Qualitative Benefit Factors

Difficulty in quantifying
- Increased **quality of life**
- Increased **productivity**

Cost-Benefit-Analysis



Estimated **cost of pesticides**

2 methods

- Cost of **Zenivex E4**

- Cost of **pesticide trucks**

US$1.6-$3.25mil

Quantitative
COST
Factors

VS

Quantitative
Benefit
Factors

Aversion

- **Loss in Income**

- **Hosp/Medical** Expenses (WNV & non WNV, businesses, tourism)

US$1.23mil

Human and **ecologic risks**

Qualitative
COST
Factors

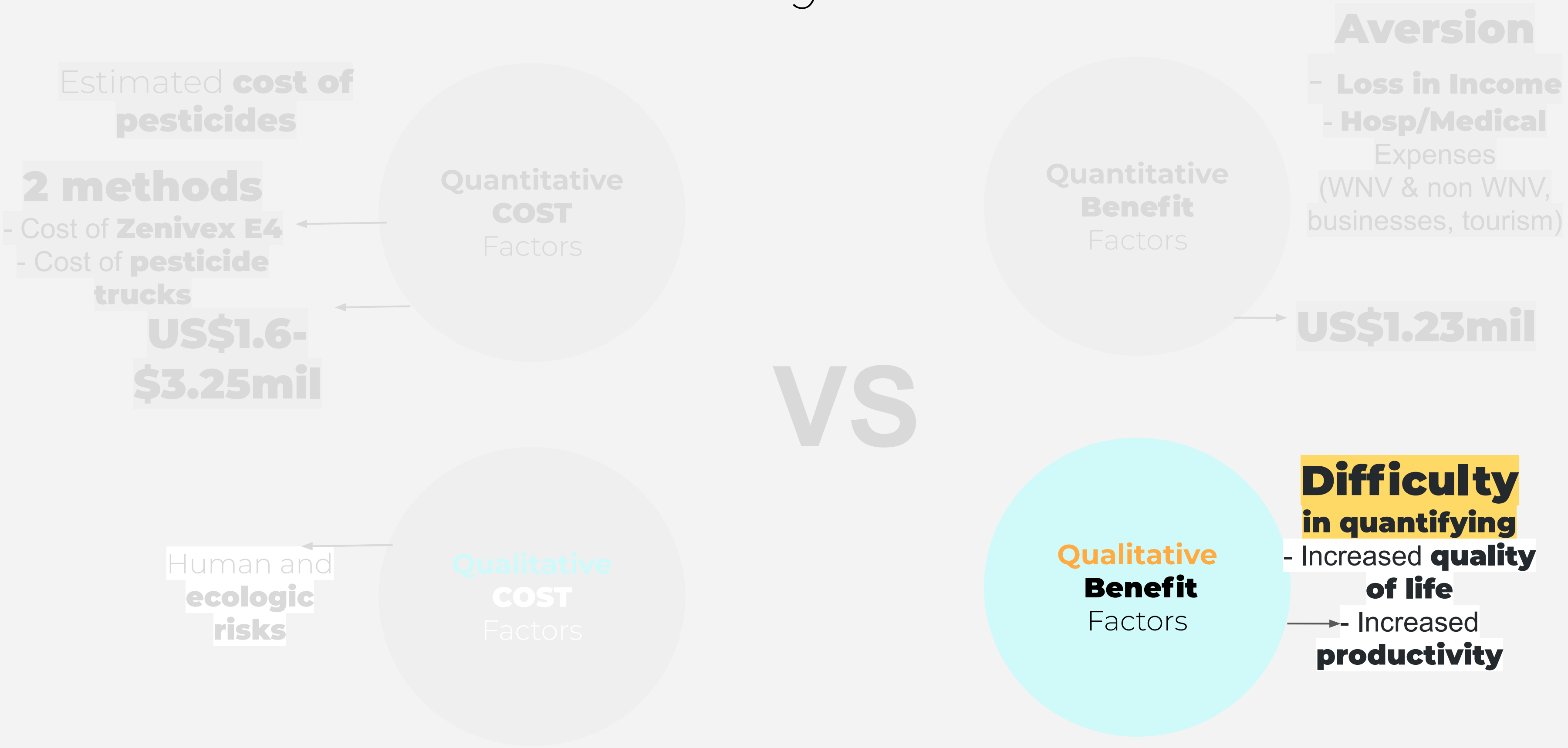
Qualitative
Benefit
Factors

Difficulty in quantifying

- Increased **quality of life**

- Increased **productivity**

Cost-Benefit-Analysis



Estimated **cost of pesticides**

2 methods
- Cost of **Zenivex E4**
- Cost of **pesticide trucks**

US$1.6-$3.25mil

Quantitative COST Factors

Quantitative Benefit Factors

Aversion
- **Loss in Income**
- **Hosp/Medical** Expenses (WNV & non WNV, businesses, tourism)

US$1.23mil

VS

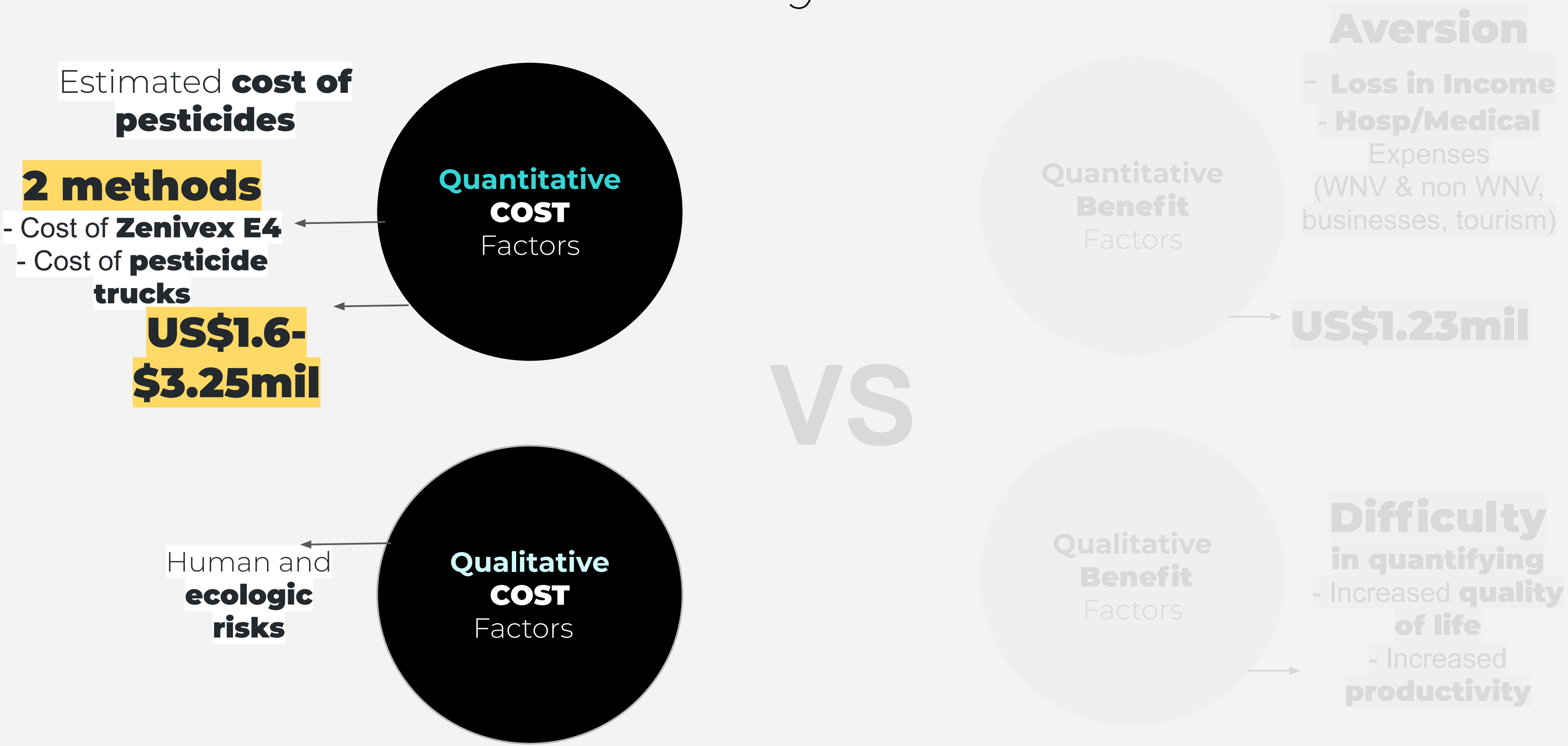
Human and **ecologic risks**

Qualitative COST Factors

Qualitative Benefit Factors

Difficulty in quantifying
- Increased **quality of life**
- Increased **productivity**

Cost-Benefit-Analysis



Estimated **cost of pesticides**

2 methods

- Cost of **Zenivex E4**
- Cost of **pesticide trucks**

US$1.6-$3.25mil

Quantitative **COST** Factors

Human and **ecologic risks**

Qualitative **COST** Factors

VS

Quantitative **Benefit** Factors

Aversion

- **Loss in Income**
- **Hosp/Medical** Expenses (WNV & non WNV, businesses, tourism)

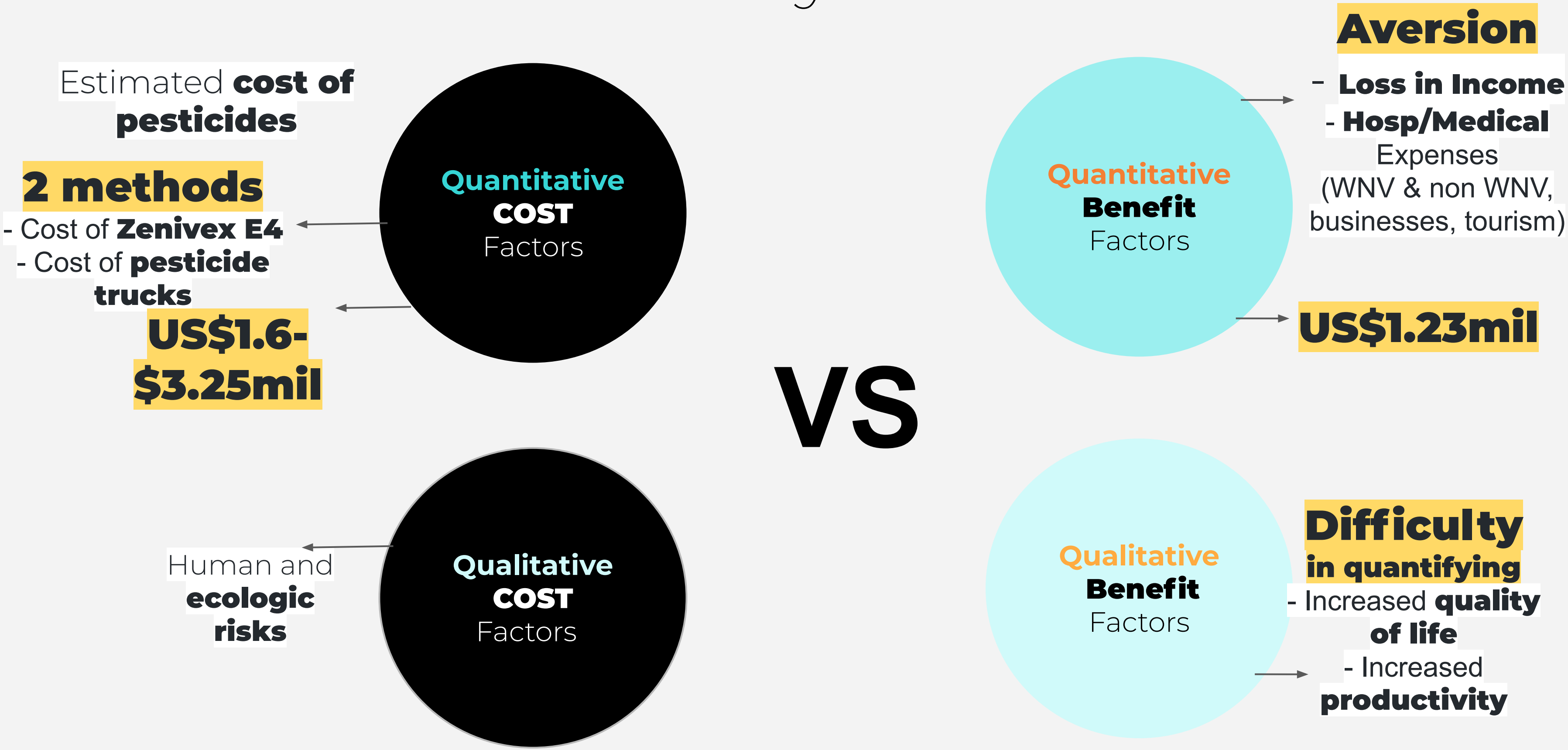
US$1.23mil

Qualitative **Benefit** Factors

Difficulty in quantifying

- Increased **quality of life**
- Increased **productivity**

Cost-Benefit-Analysis



Estimated **cost of pesticides**

2 methods

- Cost of **Zenivex E4**
- Cost of **pesticide trucks**

US$1.6-$3.25mil

Quantitative
COST
Factors

Qualitative
COST
Factors

Human and **ecologic risks**

VS

Quantitative
Benefit
Factors

Aversion

- **Loss in Income**
- **Hosp/Medical** Expenses (WNV & non WNV, businesses, tourism)

US$1.23mil

Qualitative
Benefit
Factors

Difficulty in quantifying

- Increased **quality of life**
- Increased **productivity**

Cost-Benefit-Analysis

US$1.6-$3.25mil

US$1.23mil

For every **1 case**

30 to 60

cases **go unreported.**

Conclusion

Logistic Regression WITH SENSITIVITY RATE OF 65%

Limitations:

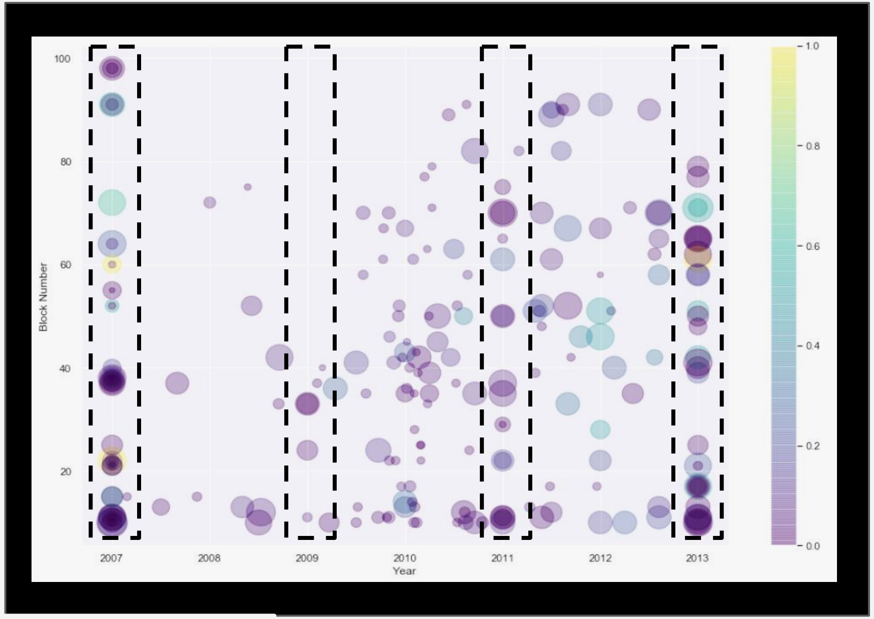
1. Use of time-series data.

2. Inclusion of all dates in the training and testing dataset.

Further Explorations:

1. The effect of number of mosquitoes on presence of WNV

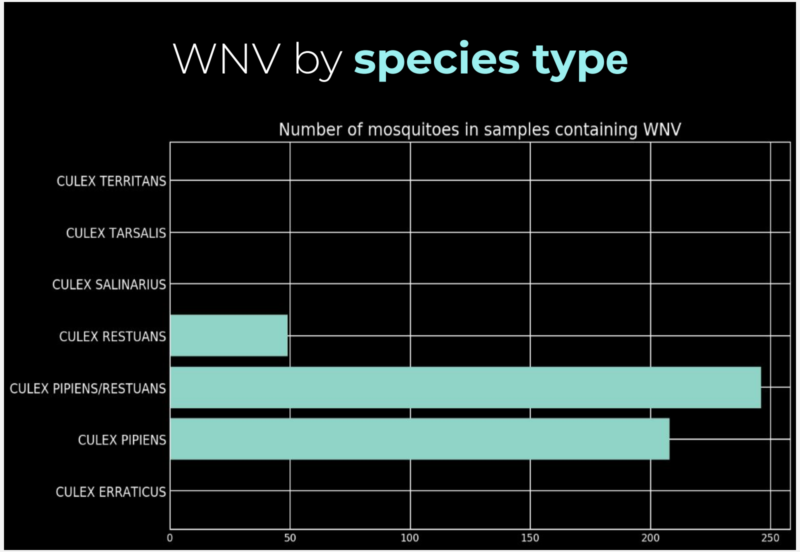
Recommendations



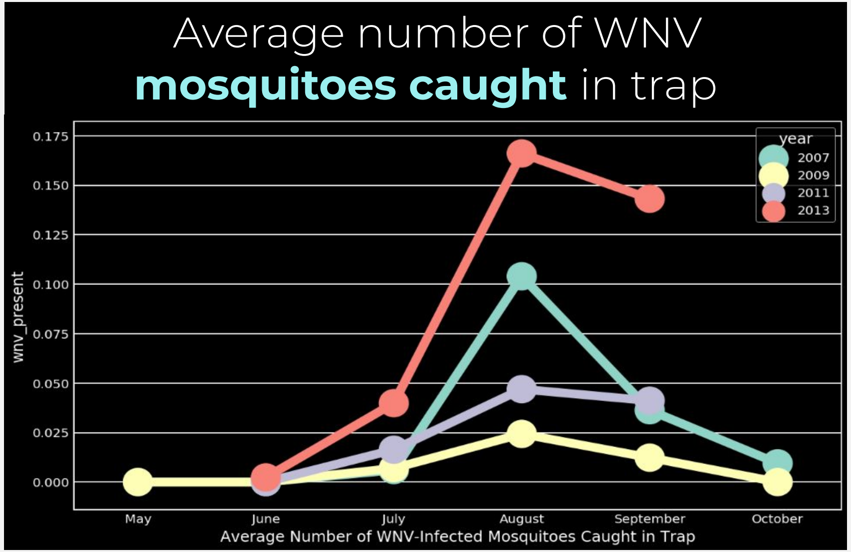
CLUSTERS

TARGET GEOGRAPHICAL CLUSTERS WITH WNV

Recommendations



WNV by **species type**



Average number of WNV **mosquitoes caught** in trap

TARGET THESE SPECIES:

- **CULEX PIPIENS/RESTUANS**
- **CULEX RESTUANS**
- **CULEX PIPIENS**

INCREASE SPRAYING FREQUENCY IN AUGUST

Recommendations

Daytime is the most dangerous
Mosquitoes that spread Zika are aggressive daytime biters. They can also bite at night.

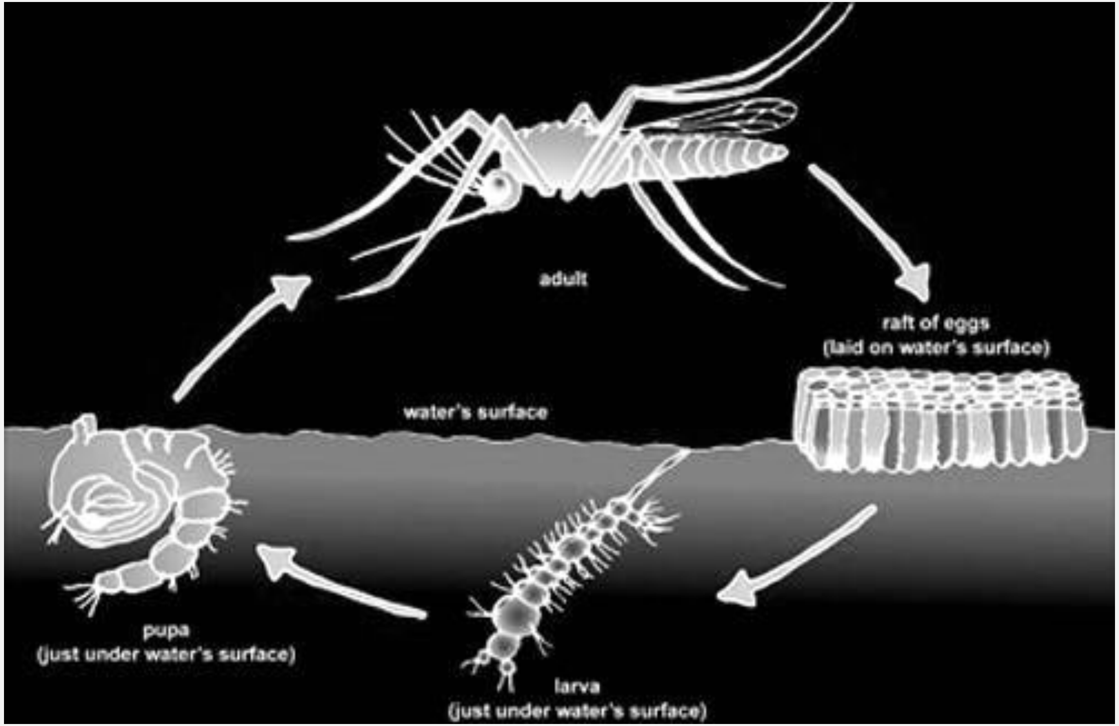
Use insect repellent
It works!
Look for the following active ingredients:
- DEET • PICARIDIN • IR3535
- OIL of LEMON EUCALYPTUS
- PARA-MENTHANE-DIOL

Wear protective clothes
Wear long-sleeved shirts and long pants or use insect repellent. For extra protection, treat clothing with permethrin.

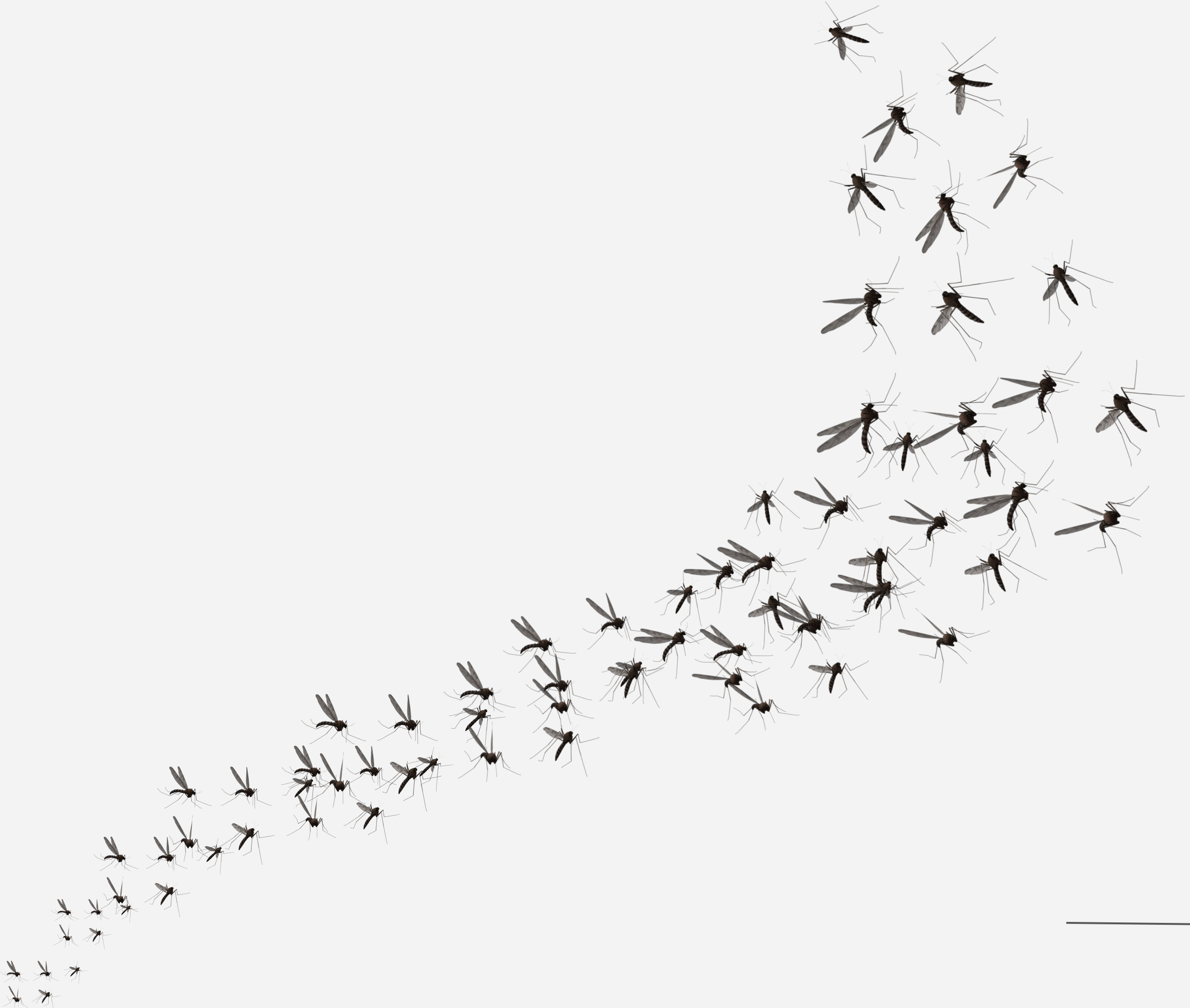
Mosquito-proof your home
Use screens on windows and doors. Use air conditioning when available. Keep mosquitoes from laying eggs near standing water.

EDUCATION

Recommendations


adult
raft of eggs
(laid on water's surface)
water's surface
pupa
(just under water's surface)
larva
(just under water's surface)

USE OF LARVICIDES VS ADULTICIDES



Thank You