West Nile Virus prediction in Chicago

Hanpu Yao
MUSA capstone
Professor: Jonathan Tannen

Since the breakout of COVID-19, public health is drawing everybody's attention. When people around the world are concerned about this never-ending pandemic, we might as well reflect on another public health crisis that broke out in 1999 and hopefully get some insights from disease control, even though they are different. In this article, I analyzed the previous trend of West Nile virus in Chicago and built a model to predict future possible presence of the virus.

Introduction

West Nile virus (WNV) is a single-stranded RNA virus that causes West Nile fever, that often breaks out in tropical and temperate regions. It primarily infects birds, but it also infects humans, horses, cats, skunks, squirrels, and domestic rabbits. It is most spread to people by the bite of an infected mosquito.

West Nile virus was first discovered in Uganda in 1937 and is transmitted by domestic mosquitoes.[1] A major epidemic of West Nile virus spread in the population occurred in Israel in 1950. After the Romanian outbreak in the mid-1990s, there were subsequent small outbreaks in Morocco (1996), Tunisia (1997), Italy (1998), and Israel (1998). In particular, the 1998 outbreak in Israel that was fatal to geese and storks was the only flavivirus with the potential to be lethal to poultry.

In the hot summer of 1999 in New York City, several unusual phenomena are gradually being linked together. Crow deaths have begun in large numbers around the city; the eastern end of Long Island has begun to see an unusual outbreak of equine encephalitis; and the illness and death of Chilean flamingos and snowy owls at the Bronx Zoo. Then, encephalitis outbreak among people in Queens. The public health response to the outbreak cannot wait for the results of virus isolation verification, and the mayor of New York City has directed an aggressive, multi-faceted public health intervention in Queens. Mosquito repellent is dispensed, and trucks are sprayed with mosquito repellent at the US Open. But new cases have since emerged in Brooklyn, the Bronx and Manhattan. St. Louis encephalitis, which is common in the Americas, makes little connection between morbidity in humans and mortality in birds. Finally, CDC obtained tissue specimens of dead birds from the USDA NVSL, determined that the virus causing the dead birds was WNV (NY99), and ruled out the possibility of St. Louis encephalitis virus.[2]

---

[1] "West Nile Virus: An Historical Overview" https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3111838/
[2] "Exotic Virus Is Identified In 3 Deaths" https://www.nytimes.com/1999/09/26/nyregion/exotic-virus-is-identified-in-3-deaths.html

Since then, the virus has spread throughout the US, and it is now the leading cause of mosquito-borne disease in the continental United States.

The Chicago Department of Public Health maintains an environmental surveillance program to track the citywide trend of West Nile Virus (WNV). This program includes the collection of mosquitoes from specific traps located throughout the city; the identification and sorting of mosquitoes collected from these traps; and the testing of specific species of mosquitoes for WNV.

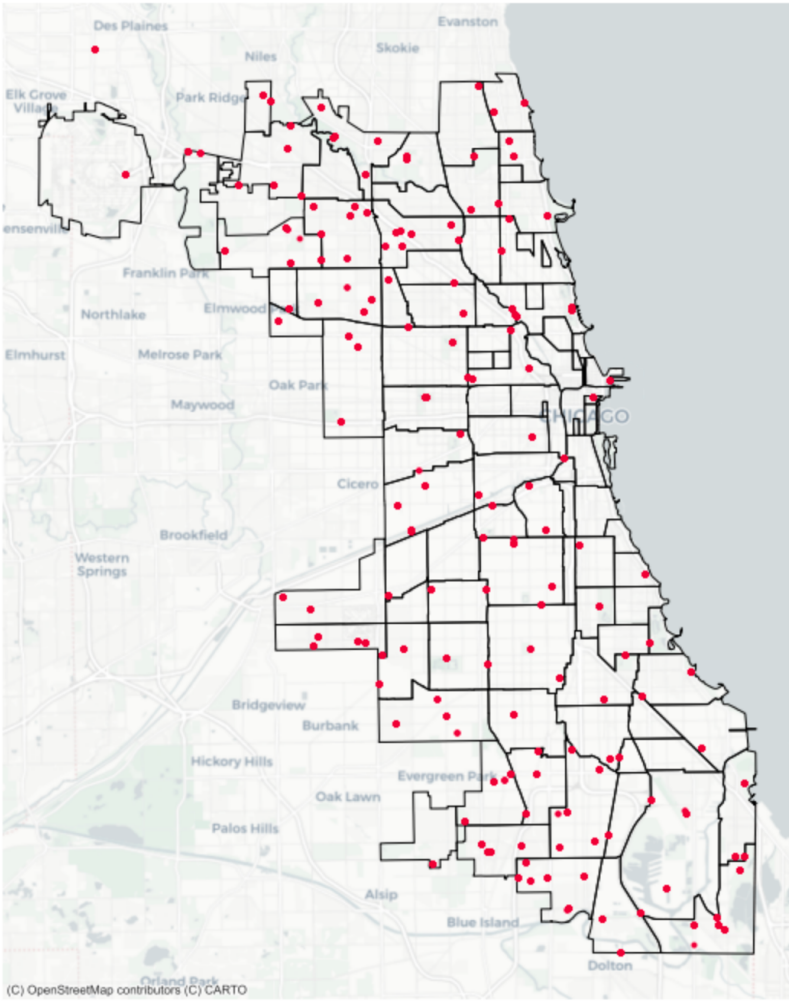

Figure 1 All mosquito traps in Chicago city (2007-2021)

Question

How to predict the new presence of WNV in Chicago?

Literature Review

Virology and epidemiology

According to Murray (2011), the most important transmission factors may include: (i) competent mosquito vectors and susceptible hosts on which they depend, (ii) well-suited climatic conditions, and (iii) a reservoir population of flavivirus-naive birds to ensure efficient virus dispersal.[3]

Besides, Moser (2015) demonstrated that mosquito saliva acts in a dose-dependent manner to enhance virus levels in the blood.[4] Based on these researches, we can assume that most WNV infections in human is results from mosquitoes bites.

Weather

Paz's (2015) study found the following: "As predictions show that the current trends are expected to continue, for better preparedness, any assessment of future transmission of WNV should take into consideration the impacts of climate change."[5]

The study show that climate is an important factor in assessing of transmission of WNV. And different aspects of weather patterns contribute to the presence of WNV in complicated ways. For example, wind patterns are also relevant to virus spread by carrying mosquitoes with air flows.[6] And in terms of precipitation, as its impact is more indirect in WNV transmission, the findings regarding North America are inconsistent especially when the analyses include different vectors.[5]

My paper builds on these studies about weather's impact on virus transmission by wrangling different weather factors.

Data sources:

1. West Nile Virus (WNV) Mosquito Test Results[7]

Time range: week 21, 2007 – week 39, 2021

---

[3] MURRAY, K., WALKER, C., & GOULD, E. (2011). The virology, epidemiology, and clinical impact of West Nile virus: A decade of advancements in research since its introduction into the Western Hemisphere. Epidemiology and Infection, 139(6), 807-817. doi:10.1017/S0950268811000185

[4] Moser, Lindsey & Lim, Pei-Yin & Styer, Linda & Kramer, Laura & Bernard, Kristen. (2015). Parameters of Mosquito-Enhanced West Nile Virus Infection. Journal of virology. 90. 10.1128/JVI.02280-15.

[5] Paz S. (2015). Climate change impacts on West Nile virus transmission in a global context. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 370(1665), 20130561. https://doi.org/10.1098/rstb.2013.0561

[6] Mackenzie JS, Gubler DJ, Petersen LR. 2004. Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses. Nat. Med. 10, S98–S109. (10.1038/nm1144)

[7] https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s

This is the main dataset of this analysis. It is a list of information collected from every trap throughout Chicago. Each observation is a piece of data collected per trap per week. Specifically, there are following features:

Feature Name	Description
season_year	Year of collection
week	Week of the year
test_id	Test ID
block	Address of the trap
trap	Trap ID
trap_type	Trap type
test_date	Test date of the sample
number_of_mosquitoes	Number of mosquitoes in a certain trap (test pool)
result	Pooled test results of a trap. This is the dependent variable of our analysis, which means whether WNV is presence in a certain trap in a certain week.
species	Species of mosquitoes in a certain trap (test pool)
latitude	Latitude of the trap
longitude	Longitude of the trap
location	Location of the trap

Table 1 Data description of traps dataset

2. Weather dataset

Time range: Mar 1, 2007 – Dec 31, 2021

This dataset is provided by NOAA (National Oceanic and Atmospheric Administration). It is the daily weather summary data of a station in Chicago midway airport (GHCND: USW00014819)

Type	Code	Description
Temperature	TMAX	Maximum temperature
	TMIN	Minimum temperature
	TOBS	Temperature at the time of observation
Precipitation	PRCP	Precipitation
Wind	AWND	Average wind speed
	WDF2	Direction of fastest 2-minute wind
	WSF2	Fastest 2-minute wind speed
Weather Type	WT01	Fog, ice fog, or freezing fog (may include heavy fog)
	WT02	Heavy fog or heaving freezing fog (not always distinguished from fog)

	WT03	Thunder
	WT04	Ice pellets, sleet, snow pellets, or small hail
	WT05	Hail (may include small hail)
	WT06	Glaze or rime
	WT08	Smoke or haze
	WT09	Blowing or drifting snow
	WT10	Tornado, waterspout, or funnel cloud"

Table 2 Data description of weather dataset

3. Locations of water bodies
4. Locations of forests
5. Location of parks
6. 311 request – sanitation violation

My method

The research process of my paper consists of following steps:

1. Data Wrangling: To put my features into a model, I synchronize all datasets into a single data frame, with columns of features and rows of one observation per trap per week.
2. Spatial lag features: Create a new spatial lag feature indicating the impact of blowing from nearby traps from previous week
3. Temporal lag features: Create new spatial lag features indicating the impact of positivity of previous 1 and 2 weeks

Exploratory Analysis

Temporal trends

The Figure 2 shows the monthly trends of number of mosquitoes in a trap in different years. Though in each year the available time ranges vary, the missing months are in cold weather in which I assume there is so few mosquitoes and WNV cases that it can be ignored.

There is an obvious trend that the number of mosquitoes peaks in July or August, which are the hottest months in a year. This is consistent with a common sense that mosquitoes are almost only active in warm weather. Noticeable, in 2007 the peak in August is relatively higher than other years.

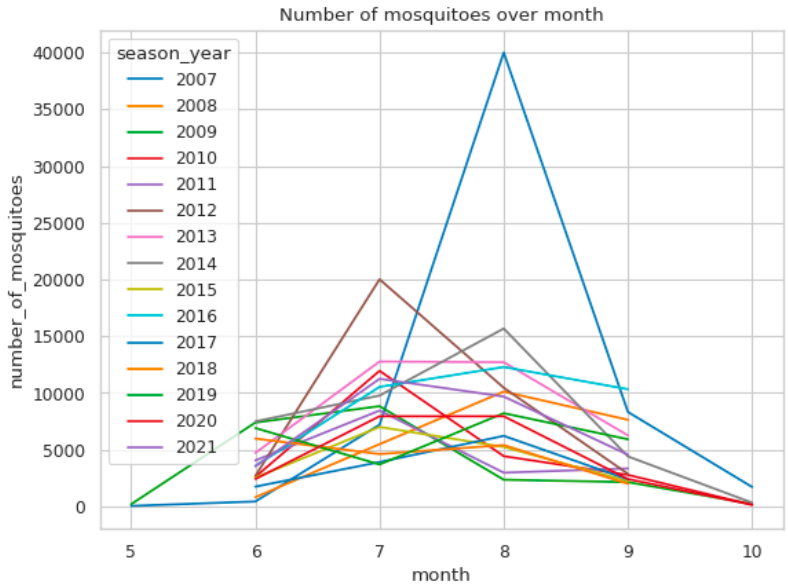

Figure 2 Number of mosquitoes over month

Figure 3 shows the trend of positivity. This is highly similar with number of positivity pools, but with a greater rise in almost every August.

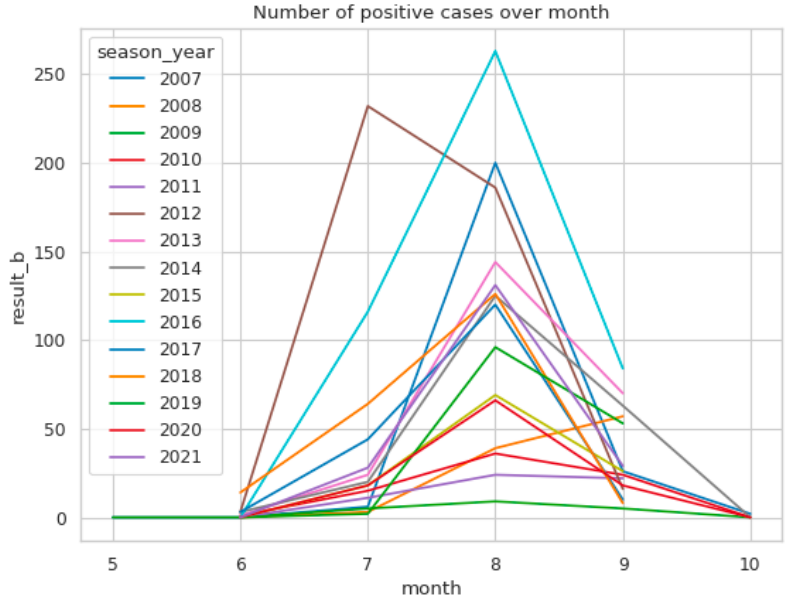

Figure 3 Number of positive pools over month

Figure 4 shows the overall situation of positive pools in different years. The blue line is positivity while orange is the 7-day moving average of it. Despite of the monthly pattern that I mention above, there is hardly a yearly trend: it begins with a peak in 2007, and decreases dramatically until the beginning of 2011; then the positivity also witnesses another peak in 2012 which is almost double the one in 2007, following by years without an obvious rising or falling

trends. What's more, in terms of the smoother moving average line, the peak in 2012 is a watershed after which the positivity maintains on a higher platform.

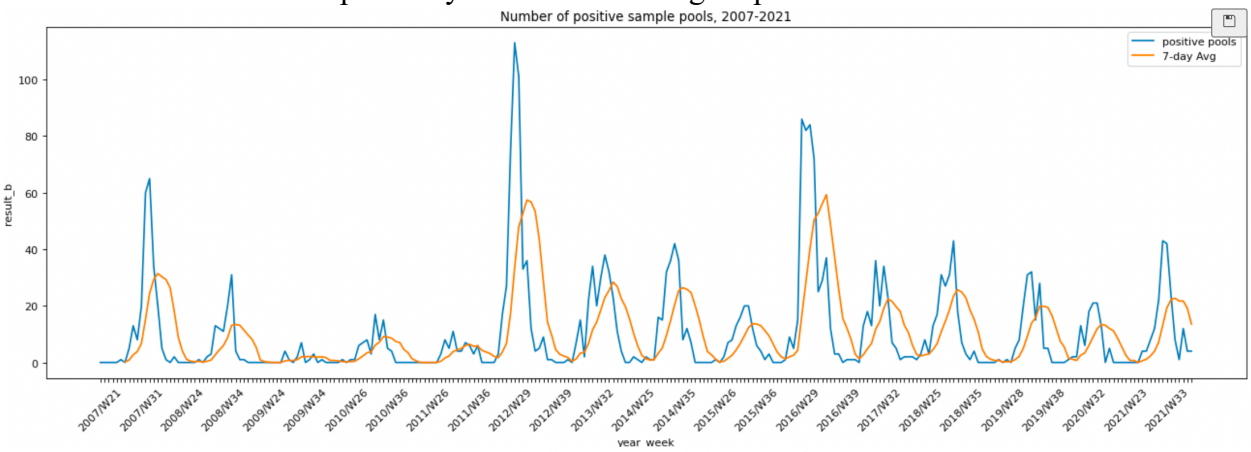

Figure 4 Overall positive pools by week of years

Sampling bias

However, do the charts above really represent the real situation in Chicago? In this chapter, I will demonstrate the sampling bias resulted from traps.

Firstly, the inconsistent number of traps in different years bring temporal bias. The mosquitoes are only the sample counted in traps, since it is unrealistic to know the population of all mosquitoes in the city. If there are some issues with traps, the data collection tool, the dataset would be problematic because of systematic error or bias.

Figure 5 shows the total number of traps of different years. The total number of traps is not consistent throughout the years and there are the most traps in 2007 which somehow explains the peak of mosquitoes' number also in 2007 in Figure 2. Of course, if there are more traps, it is expected to capture more mosquitoes.

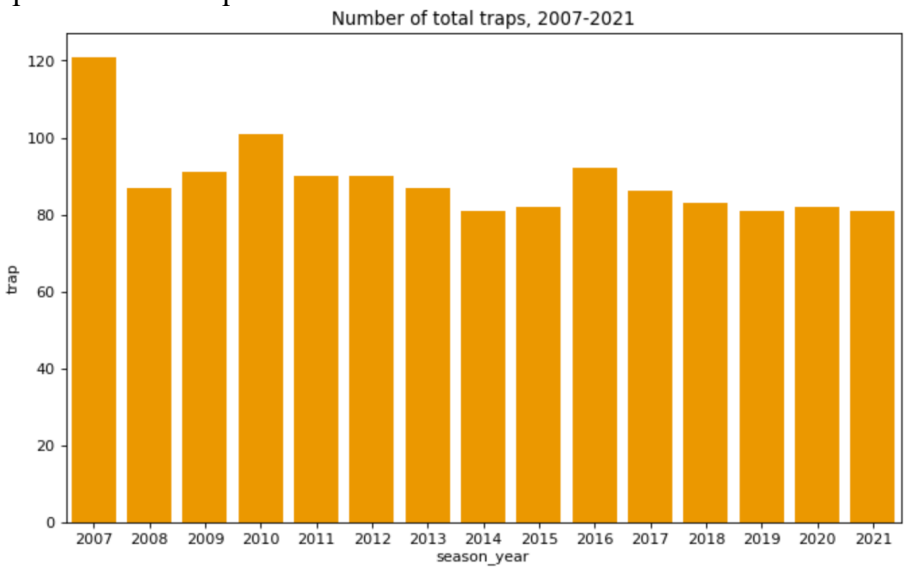

Figure 5 Number of total traps

Second, different types of traps bring spatial bias. The number of traps is broken down into 4 different types in Figure 6: CDC, GRAVID, OVI and SENTINEL. While there is always relatively large number of GRAVID trap, SENTINEL traps are replacing CDC traps in recent years, and there is only one OVI trap in 2007.

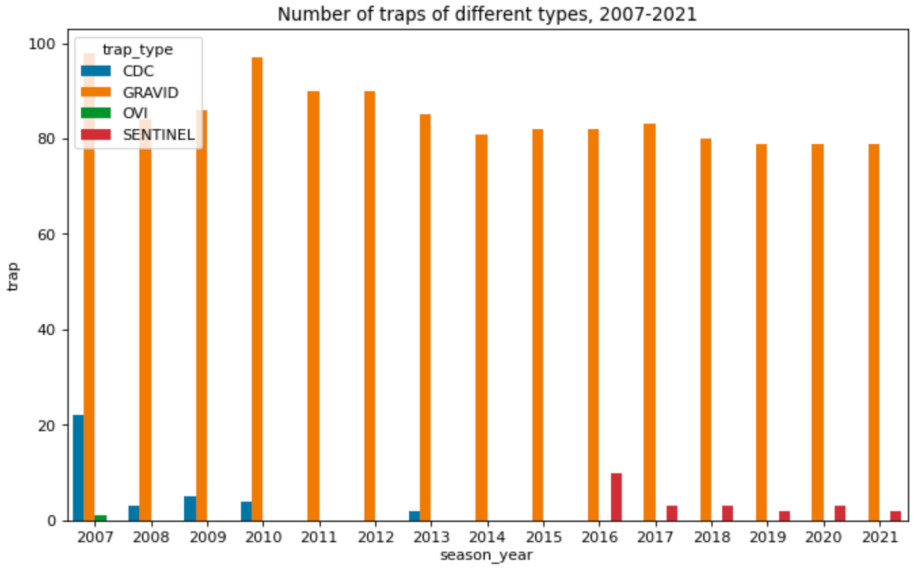

Figure 6 Number of traps of different types

This pattern is also reflected in Figure 7, the number of positivity is related to trap types over years to a extent. And SENTINEL traps capture more positivity pools than GRAVID traps, even though there are much fewer SENTINEL traps. This indicates they are essentially different in their ability to capture mosquito samples.

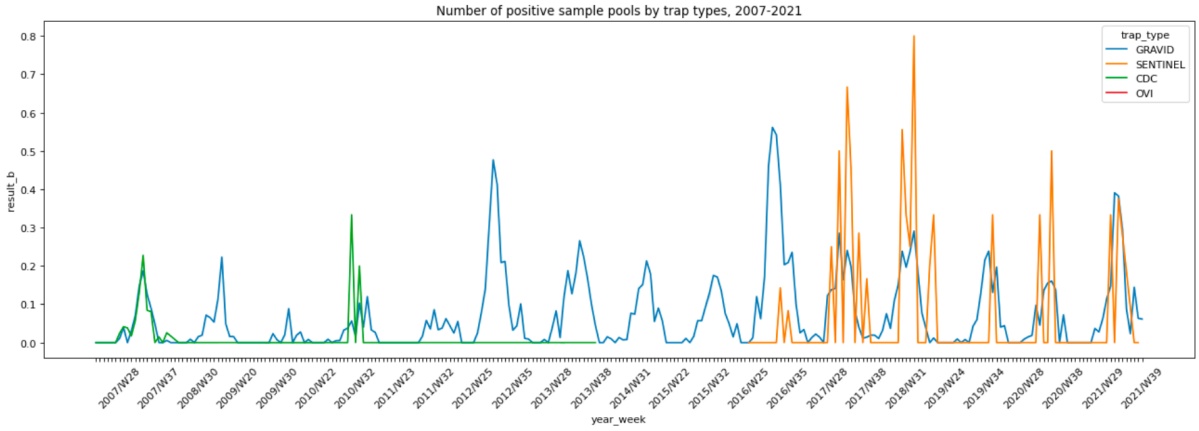

Figure 7 Number of positive sample pools by trap types

Table 3 shows different type of traps' performance in capturing mosquitoes and positivity samples. The tools data collection are clearly different that CDC traps might capture more

mosquitoes and SENTINEL traps are more likely to capture more positive mosquitoes. This is a result of their mechanism of attracting mosquitoes.[8]

Trap type	Total positivity	Total number of mosquitoes	Total observations	Positivity / Observations	Mosquitos / observations
CDC	79	35334	1256	0.062898	28.132166
GRAVID	2651	372014	31278	0.084756	11.893791
OVI	0	1	1	0	1
SENTINEL	48	6805	343	0.139942	19.83965

Table 3 Ability of different trap types

Finally, the location of traps is also a resource of selection bias. Whether the researcher in public health department sets a trap in a park or in the CBD would be very different. With different and uneven spatial features, selection bias from the trap locations is the most complicated.

Feature engineering

Temporal feature
The test results of each observation of 1- and 2-week lag are also two new features in the data frame.

Spatial feature
The first is the distance to spatial features. I collect some spatial features in Chicago city which are the geometries of forests, parks, water bodies and 311 requests of sanitation violation. To wrangle them into the dataset, I calculate the nearest distance from each trap to each type of these spatial features and create some features like "distance to forests", "distance to water bodies" and so on. Lake Michigan is taken out from water bodies feature and it is wrangled into a separate feature because it is too huge compared to other water bodies.

The second is the spatial lag of traps, which is taking the nearby traps' impact into consideration. Since mosquitoes are easily carried by wind blowing, I use wind as a factor to create spatial lag. I assume that wind direction and speed in one week before would impact the transmission of WNV. The function is as follow:

$$SpatialLag_{ij} = \sum_{k=0}^{n} Positivity_{j-1} * Number\ of\ mosquitoes_{j-1} * Speed_{j-1} * \frac{cos\theta}{distance_{ik}}$$

Here i is the ID of a trap, j is the week of the collection, and n is the number of traps last week and k is the ID of a trap in last week. This means I time the positivity (0 or 1), the number of

[8] https://www.clarke.com/blog/adult-mosquito-surveillance-equipment/

mosquitoes, wind speed and azimuth decomposed to the direction between two traps, and all this data is from 1 week before.

After the feature engineering, Figure 8 is the final data set.

```
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   geometry_str                   15868 non-null  object
 1   year_week                      15868 non-null  object
 2   season_year                    15868 non-null  int64
 3   week                           15868 non-null  int64
 4   test_id                        15868 non-null  object
 5   block                          15868 non-null  object
 6   trap                           15868 non-null  object
 7   trap_type                      15868 non-null  object
 8   test_date                      15868 non-null  object
 9   number_of_mosquitoes           15868 non-null  int64
 10  result                         15868 non-null  object
 11  species                        15868 non-null  object
 12  month                          15868 non-null  int64
 13  result_b                       15868 non-null  int64
 14  geometry                       15868 non-null  geometry
 15  distance_to_water              15868 non-null  float64
 16  distance_to_michLake           15868 non-null  float64
 17  distance_to_parks              15868 non-null  float64
 18  AWND                           15868 non-null  float64
 19  PRCP                           15868 non-null  float64
 20  tavg                           15868 non-null  float64
 21  WDF2                           15868 non-null  float64
 22  WT01                           15868 non-null  float64
 23  WT02                           15868 non-null  float64
 24  WT03                           15868 non-null  float64
 25  WT04                           15868 non-null  float64
 26  WT05                           15868 non-null  float64
 27  WT06                           15868 non-null  float64
 28  WT08                           15868 non-null  float64
 29  WT09                           15868 non-null  float64
 30  WT10                           15868 non-null  float64
 31  distance_to_bad_santination    15868 non-null  float64
 32  trap_index                     15868 non-null  int64
 33  wind_lag                       15868 non-null  float64
 34  result_lag_1                   15868 non-null  float64
 35  result_lag_2                   15868 non-null  float64
 36  species_CULEX ERRATICUS        15868 non-null  uint8
 37  species_CULEX PIPIENS          15868 non-null  uint8
 38  species_CULEX PIPIENS/RESTUANS 15868 non-null  uint8
 39  species_CULEX RESTUANS         15868 non-null  uint8
 40  species_CULEX SALINARIUS       15868 non-null  uint8
 41  species_CULEX TARSALIS         15868 non-null  uint8
 42  species_CULEX TERRITANS        15868 non-null  uint8
 43  species_UNSPECIFIED CULEX      15868 non-null  uint8
dtypes: float64(20), geometry(1), int64(6), object(9), uint8(8)
```

Figure 8 final dataset information

Prediction Models

There is no much multicollinearity in the dataset according to Figure 9 correlation matrix.

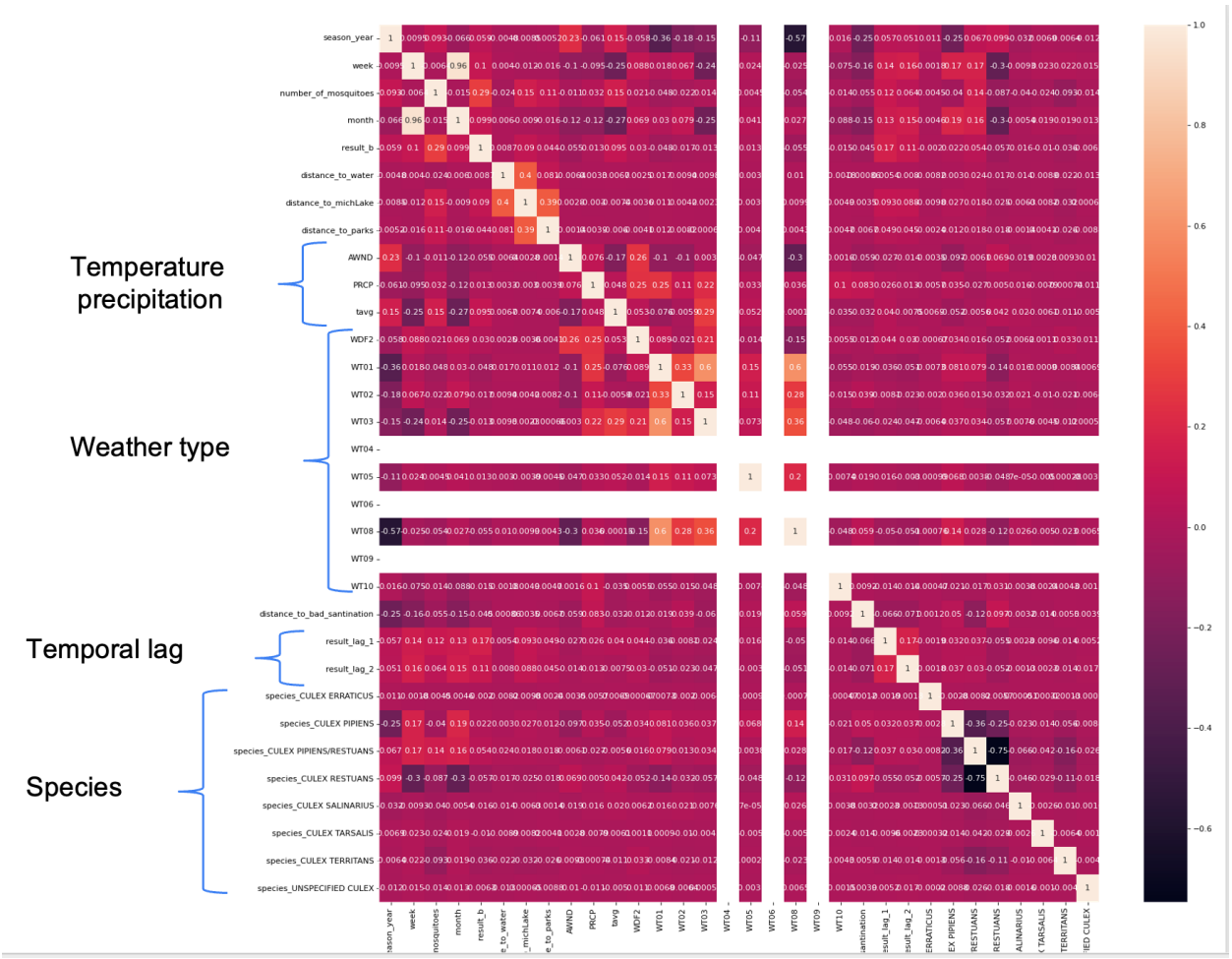

Figure 9 correlation matrix

Logistic regression

After standardization of all predictors I run a logistic regression. The result is as followed.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                result_b   No. Observations:                15868
Model:                           Logit   Df Residuals:                    15840
Method:                            MLE   Df Model:                           27
Date:                 Thu, 28 Apr 2022   Pseudo R-squ.:                 0.2664
Time:                         20:15:34   Log-Likelihood:                -2595.1
converged:                       False   LL-Null:                       -3537.7
Covariance Type:             nonrobust   LLR p-value:                     0.000
=================================================================================================
                                    coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
season_year                       0.0308      0.012      2.522      0.012       0.007       0.055
week                              0.0522      0.016      3.298      0.001       0.021       0.083
number_of_mosquitoes              0.0576      0.002     23.828      0.000       0.053       0.062
distance_to_water                -3.1483      3.592     -0.876      0.381     -10.189       3.893
distance_to_michLake              6.2772      0.888      7.070      0.000       4.537       8.017
distance_to_parks                -1.6588      3.049     -0.544      0.586      -7.634       4.317
distance_to_bad_santination   -5.705e-05      0.005     -0.011      0.991      -0.010       0.010
AWND                              -0.1377      0.035     -3.954      0.000      -0.206      -0.069
PRCP                               0.1959      0.254      0.770      0.441      -0.303       0.694
tavg                               0.0375      0.011      3.515      0.000       0.017       0.058
WDF2                               0.0013      0.001      1.299      0.194      -0.001       0.003
```

WT01	-0.0701	0.042	-1.661	0.097	-0.153	0.013
WT02	-0.1393	0.191	-0.730	0.465	-0.513	0.235
WT03	0.0191	0.047	0.406	0.685	-0.073	0.111
WT05	0.3768	0.285	1.321	0.187	-0.182	0.936
WT08	-0.0247	0.045	-0.552	0.581	-0.112	0.063
WT10	-8.7232	183.961	-0.047	0.962	-369.279	351.833
result_lag_1	0.4503	0.109	4.146	0.000	0.237	0.663
result_lag_2	0.2297	0.120	1.917	0.055	-0.005	0.465
species_CULEX ERRATICUS	-85.8398	9861.920	-0.009	0.993	-1.94e+04	1.92e+04
species_CULEX PIPIENS	-71.8010	24.489	-2.932	0.003	-119.799	-23.803
species_CULEX PIPIENS/RESTUANS	-71.8812	24.513	-2.932	0.003	-119.925	-23.837
species_CULEX RESTUANS	-71.8389	24.512	-2.931	0.003	-119.881	-23.797
species_CULEX SALINARIUS	-154.0229	4.53e+17	-3.4e-16	1.000	-8.87e+17	8.87e+17
species_CULEX TARSALIS	-87.9549	3210.119	-0.027	0.978	-6379.673	6203.764
species_CULEX TERRITANS	-73.7517	24.517	-3.008	0.003	-121.804	-25.700
species_UNSPECIFIED CULEX	-96.7378	3.2e+05	-0.000	1.000	-6.26e+05	6.26e+05
wind_lag_log	0.8170	0.054	15.206	0.000	0.712	0.922

Table 4 Logistic regression result

The statistically significant features are week of the year, number of mosquitoes, distance to Lake Michigan, average wind speed, average temperature, temporal lag of 1 week, some certain species and spatial lag with wind.

Machine learning models

To prepare data to train machine learning models, I do a max-min standardization to the predictors. Also, since there are too few positive observations in the dataset and the positive and negative ration is only 0.06, I upsample the positive label in training set.

1. **Support vetor machine (SVM)**

After tuning the model, the best result is 0.705 accuracy, 0.744 sensitivity, 0.134 precision and 0.726 AUC (area under curve).

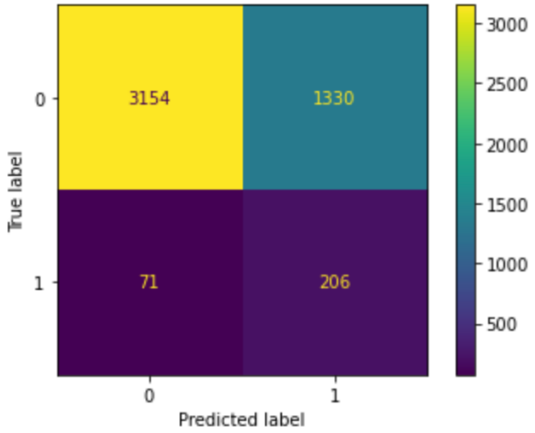

Figure 10 Confusion matrix of SVM model result

2. **Random forest**

The best result of random forest is 0.942 accuracy, 0.209 sensitivity, 0.514 precision and 0.598 AUC.

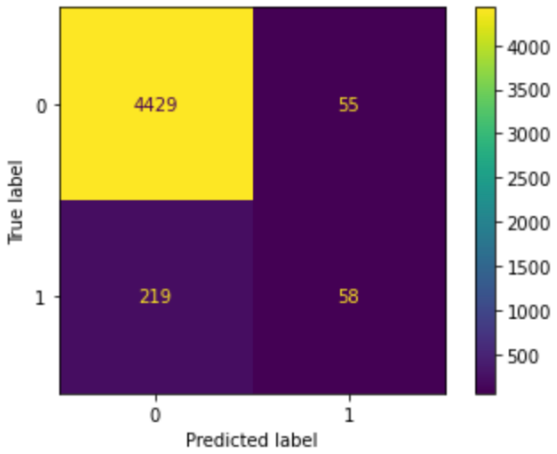

Figure 11 Confusion matrix of Random Forest model result

3. Comparison

In terms of overall accuracy, SVM model performs better than Random Forest model. However, my objective is to only predict the positive risk and I don't focus on how accurate the prediction of negative pool is. The high accuracy is mainly contributed by the success in predicting negative pools.

In terms of sensitivity, SVM gets a better result of 0.744, which means that this model can find out 74.4% positive pools of the entire positive samples.

In terms of precision, Random Forest gets a better result of 0.514, which means that 51.4% of predicted positive traps are truly positive.

Conclusion

Although upsampling solves some problem from labels imbalance, the overall prediction quality is not satisfied. However, the trade-off between SVM and Random Forest models worth discussing.

My goal of this article is to train a model to predict WNV presence, so the further application of this model is my 'user demand'. If this model is utilized by public health department to spray pesticide to places where the model predicts high risky, then the trade-off mainly depends on the effect of pesticide. If the pesticide is expensive, complicated, or harmful to other creatures, then the precision of where to spray it is the most important. In this case the Random Forest is more suitable because it won't waste pesticide or cause side-effect from excessive spraying. Vice versa, if the WNV is a issue that need urgent action to control and fault tolerance is high, then the SVM model is a better choice because it can block as many of WNV transmission as possible.

Reference

1. “West Nile Virus: An Historical Overview” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3111838/
2. “Exotic Virus Is Identified In 3 Deaths” https://www.nytimes.com/1999/09/26/nyregion/exotic-virus-is-identified-in-3-deaths.html
3. MURRAY, K., WALKER, C., & GOULD, E. (2011). The virology, epidemiology, and clinical impact of West Nile virus: A decade of advancements in research since its introduction into the Western Hemisphere. Epidemiology and Infection, 139(6), 807-817. doi:10.1017/S0950268811000185
4. Moser, Lindsey & Lim, Pei-Yin & Styer, Linda & Kramer, Laura & Bernard, Kristen. (2015). Parameters of Mosquito-Enhanced West Nile Virus Infection. Journal of virology. 90. 10.1128/JVI.02280-15.
5. Paz S. (2015). Climate change impacts on West Nile virus transmission in a global context. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 370(1665), 20130561. https://doi.org/10.1098/rstb.2013.0561
6. Mackenzie JS, Gubler DJ, Petersen LR. 2004. Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses. Nat. Med. 10, S98–S109. (10.1038/nm1144)
7. https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s
8. GUIDE TO ADULT MOSQUITO SURVEILLANCE EQUIPMENT: FINDING THE RIGHT TRAPS FOR YOUR MOSQUITO CONTROL PROGRAM https://www.clarke.com/blog/adult-mosquito-surveillance-equipment/