Predicting West Nile virus in mosquitos across the city of Chicago

Author

address

Abstract: West Nile virus is most commonly spread to humans through infected mosquitos. About 20% people who are infected 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 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 goal of this research is to predict when and where different species of mosquitos will test positive for West Nile virus given weather, testing trap location and spraying data. A more accurate method of predicting outbreaks of West Nile virus in mosquitos will help CPHD more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.

Keywords: Data Mining, Data Science, Public Health Care, Applied Data Science, West Nile virus, Machine Learning

Reference to this paper should be made as follows: Author. (xxxx) 'Title', *Int. J. xxxxxxxxxx xxxxxxxxxx*,

1 Introduction

West Nile virus is an arthropod-borne virus (arbovirus) most commonly spread by infected mosquitoes. It is a member of the *flavivirus* genus and belongs to the Japanese encephalitis antigenic complex of the family *Flaviviridae*. Virus was first isolated in a woman in the West Nile district of Uganda in 1937. It was identified in birds (crows and columbiformes) in Nile delta region in 1953. It was first detected in North America in 1999 producing a large and dramatic outbreak that spread throughout the continental United States of America (USA) in the following years. It has resulted in the deaths of more than 450 people and tens of thousands of birds, horses, and other animals in North America.

West Nile Virus is maintained in nature in a cycle involving transmission between birds and mosquitoes. Humans, horses and other mammals can be infected. Humans get infected with West Nile virus by the bite of an infected mosquito. Most people (70-80%) who become infected with West Nile virus do not develop any symptoms at all. About 1 in 5 people who are infected will develop a fever with other symptoms such as headache, body aches, joint pains, vomiting, diarrhea, or rash. Most people with this type of West Nile virus disease recover completely, but fatigue and weakness can last for weeks or

months. Less than 1% of people who are infected will develop a serious neurologic illness such as encephalitis or meningitis (inflammation of the brain or surrounding tissues). The symptoms of neurologic illness can include headache, high fever, neck stiffness, disorientation, coma, tremors, seizures, or paralysis. Recovery from severe disease may take several weeks or months. Some of the neurologic effects may be permanent. About 10% of people who develop neurologic infection due to West Nile virus will die.

West Nile virus disease cases have been reported from all 48 lower states. The only states that have not reported cases are Alaska and Hawaii. Seasonal outbreaks often occur in local areas that can vary from year to year. The weather, numbers of birds that maintain the virus, numbers of mosquitoes that spread the virus, and human behaviour are all factors that can influence when and where outbreaks occur.

The data set that has been analysed is based on city of Chicago. Every year from late-May to early-October, public health workers in Chicago setup mosquito traps scattered across the city. Every week from Monday through Wednesday, these traps collect mosquitos, and the mosquitos are tested for the presence of West Nile virus before the end of the week. The location of the traps is described by the block number and street name. For the convenience of the analysis these attributes were mapped in to longitude and latitude. The test results include the number of mosquitos, the mosquito's species, and whether or not West Nile virus is present in the cohort. Also, please note that some traps are "satellite traps". These are traps that are set up near (usually within 6 blocks) an established trap to enhance surveillance efforts. Only the data that has been collected in 2007, 2009, 2011 and 2013 will be considered to the analysis. It is believed that hot and dry conditions are more favourable for West Nile virus than cold and wet. Dataset from National Oceanic and Atmospheric Administration of the weather conditions of 2007 to 2014 in Chicago, during the months of the tests were also taken in to consideration for the analysis. There were two weather stations in Chicago.

1. CHICAGO O'HARE INTERNATIONAL AIRPORT Lat: 41.995 Lon: -87.933 Elev: 662 ft. above sea level
2. CHICAGO MIDWAY INTL ARPT Lat: 41.786 Lon: -87.752 Elev: 612 ft. above sea level

The next section will discuss the patterns identified in the data sets and the 3rd section will contain the research about machine learning algorithms that are used to predict the West Nile Virus presence. A more accurate method of predicting outbreaks of West Nile virus in mosquitos will help the City of Chicago and CPHD more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.

2 Data Analysis

2.1 Trap Location and West Nile Virus presence

Title

Figure 1 Heat map on West Nile Virus Detection

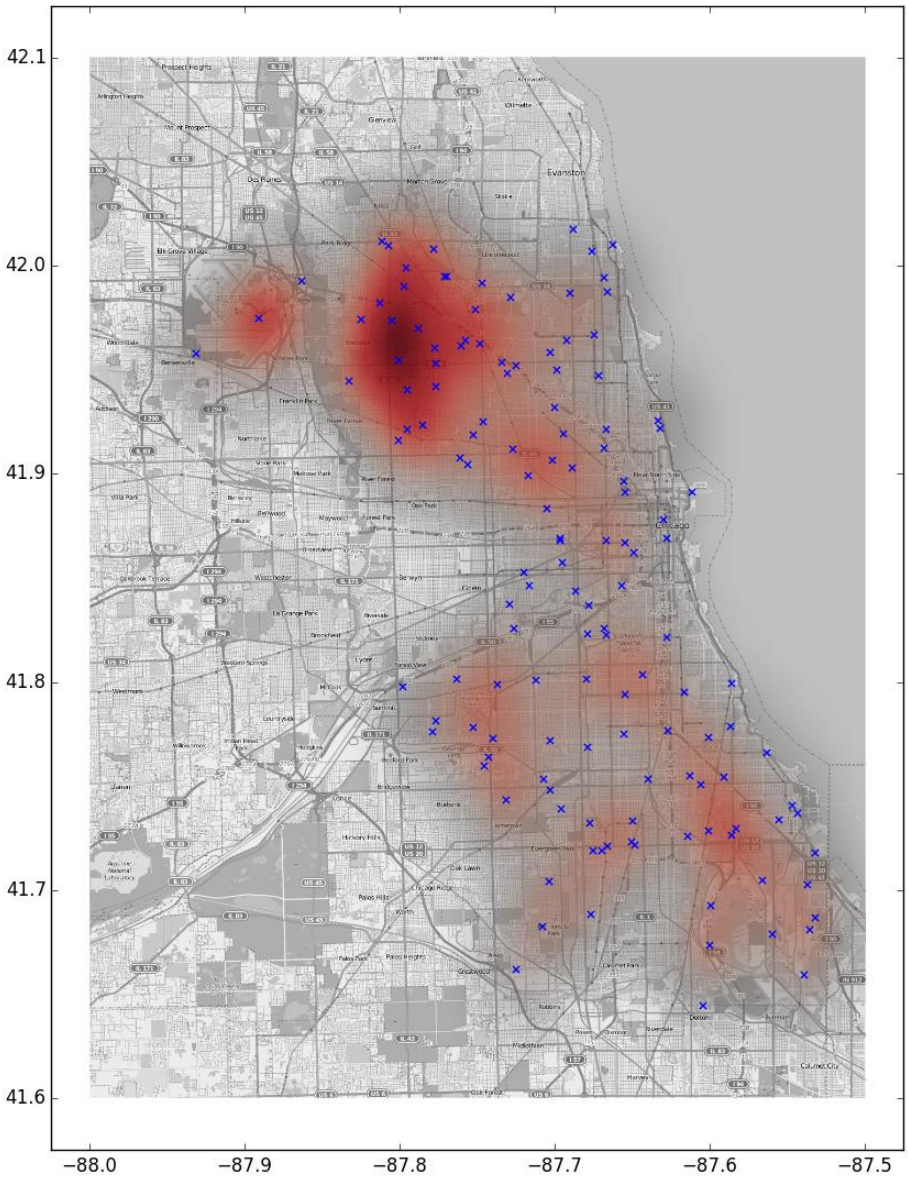
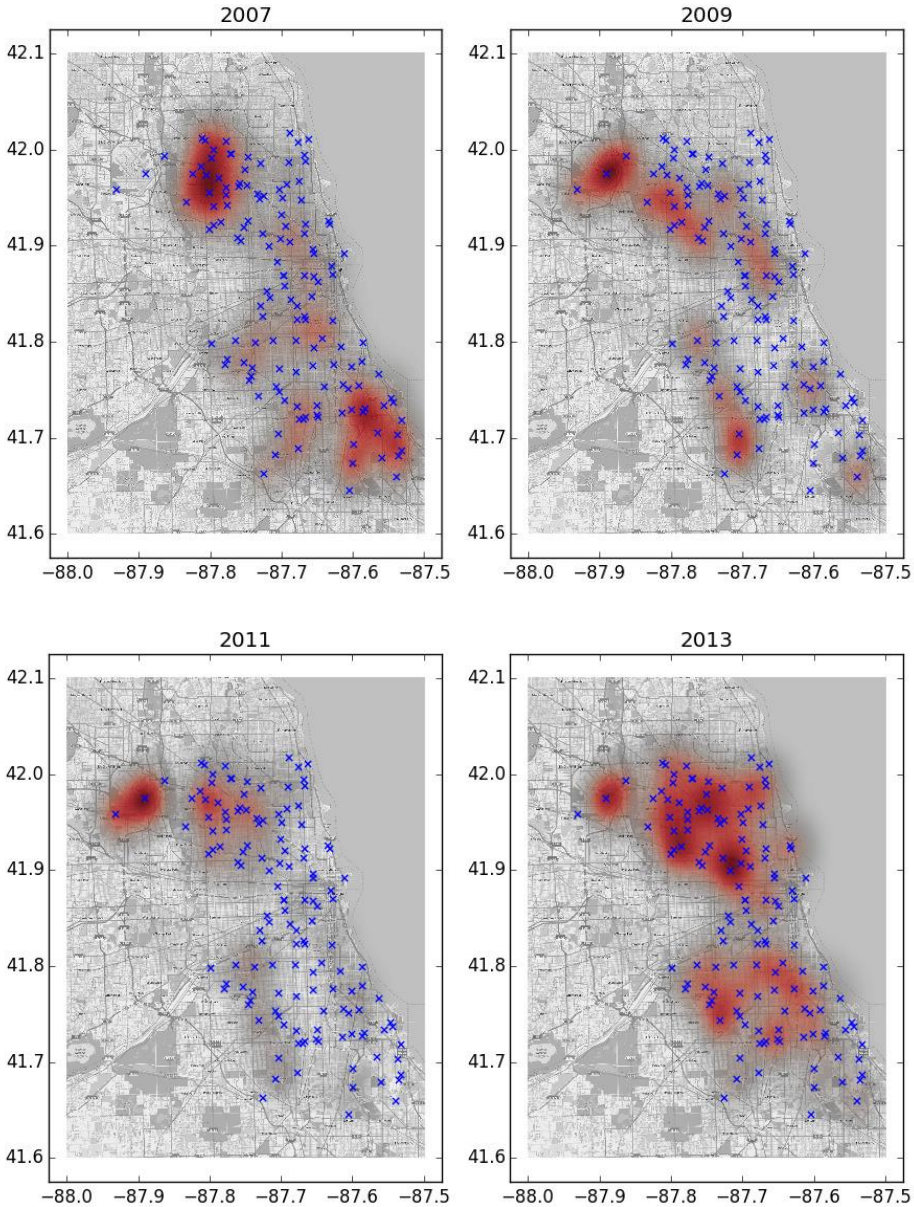


Figure 1 is a map of Chicago city. Blue crosses imply the mosquito traps that were placed to test West Nile virus. Red coloured areas imply the detection of West Nile virus. kernel density estimation with 0.02 bandwidth was used for the data smoothing of the visualization. As the image shows there is a concerned area in the top half of the city where the virus can be detected commonly. After this analysis data was split based on the tested year to see whether the same area is common for West Nile virus continuously. The same kernel density estimation was used for this visualization also.

Author

Figure 2 Heat Map of West Nile virus each year

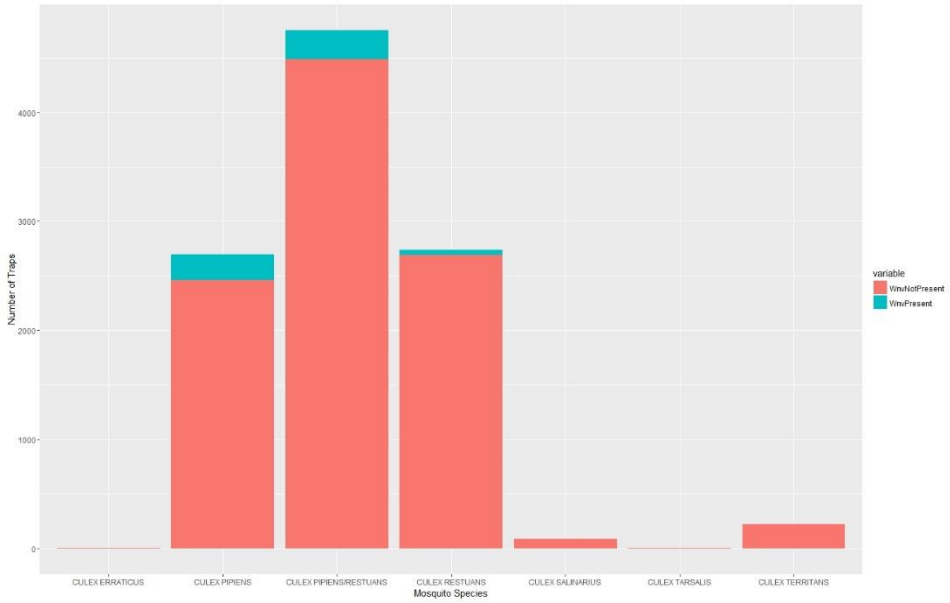


As the Figure 2 implies the area we discussed in the figure 1 has been continuously affected by the virus throughout the years. So, that area should be given more priority when allocating resources to prevent this virus. In 2013, the virus has been common in the city of Chicago. Heat map of 2013 shows that infected mosquitos have been detected in many areas which haven't detected infected mosquitos in previous years.

2.2 Mosquito Species and West Nile virus presence

The next attempt of the research was to figure out whether there is a relationship between mosquito species and West Nile virus presence.

Figure 3 Mosquito Species and West Nile virus presence



6 mosquito species were found in the traps. Among them *CULEX ERRATICUS, CULEX SALINARIUS, CULEX TARSALIS, CULEX TERRITANS* were not common and didn't have West Nile virus presence in them. But Centres for Disease Control and Prevention reported that all of these species have been tested positive for the virus between 1999 to 2012 in USA. As the image implies out of the 2 most common mosquito species in Chicago *CULEX PIPIENS* showed a higher probability than *CULEX RESTUANS*. The most common situation in the traps was a combination of *CULEX PIPIENS and CULEX RESTUANS* where 5.51% tested positive for the virus.

2.3 Time of the year and West Nile virus presence

Medical reports show that most people are infected with the West Nile virus from June through September. Dataset was analysed to observe whether there is a peak time for the virus between June to September. So, the West Nile virus detection count for a week was plotted against the week number for all the years as in Figure 4.

Since the variability is too high from year to year, West Nile virus count was scaled per year using a min-max scaler and the graph was plotted again in Figure 5.

Author

Figure 4 West Nile virus detection count vs Week number

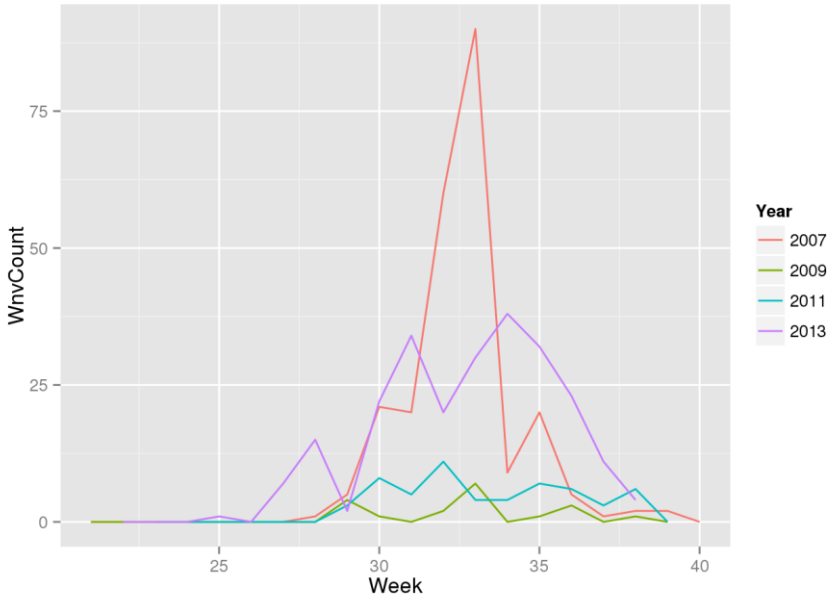


Figure 5 Scaled West Nile virus count vs Week number

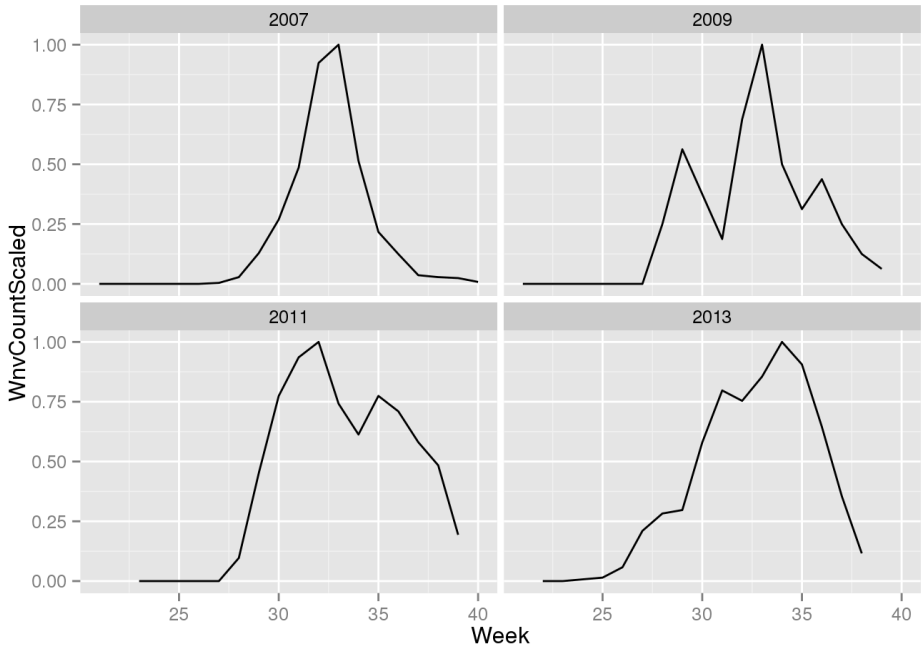


Figure 5 is better for analysing than Figure 4. As the figure implies the peak of each year happens between week 30 and week 35. This data can be fitted in to a Gaussian Distribution.

Equation 1: Gaussian Distribution for West Nile virus detection count

$$P = height \times e^{-(Week\ number - center)^2 / width}$$

The fitted curve can be visualized as in Figure 6

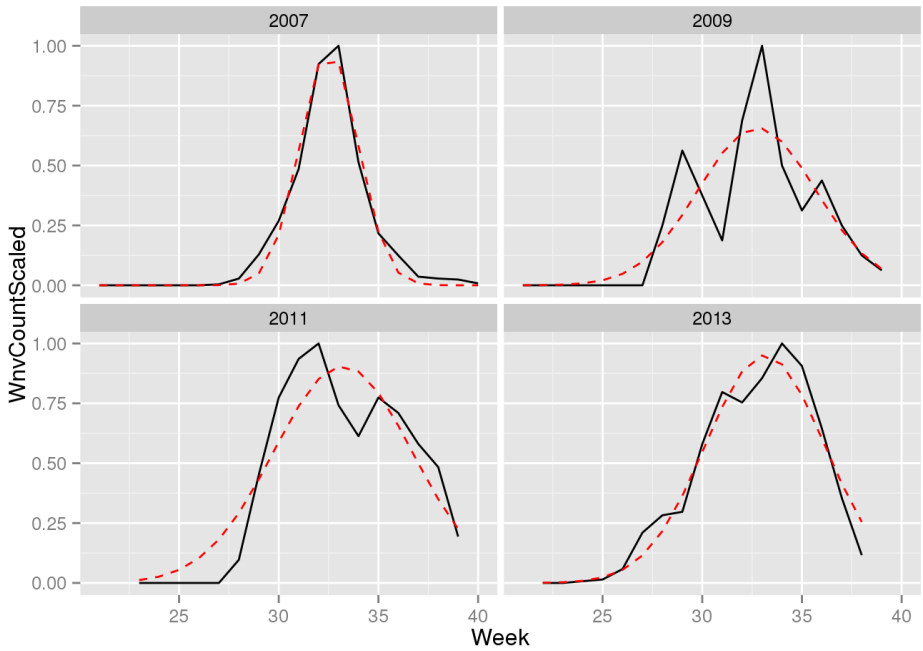


Figure 6: Gaussian Distribution fitted to Data

The curve fits well specially in 2007 data. The curve has a strong predictive ability too. By fitting a curve to the data for the current year, you can get an estimate of how the epidemic will progress during the remainder of the year; when it will peak, when it will end and how large it will be.

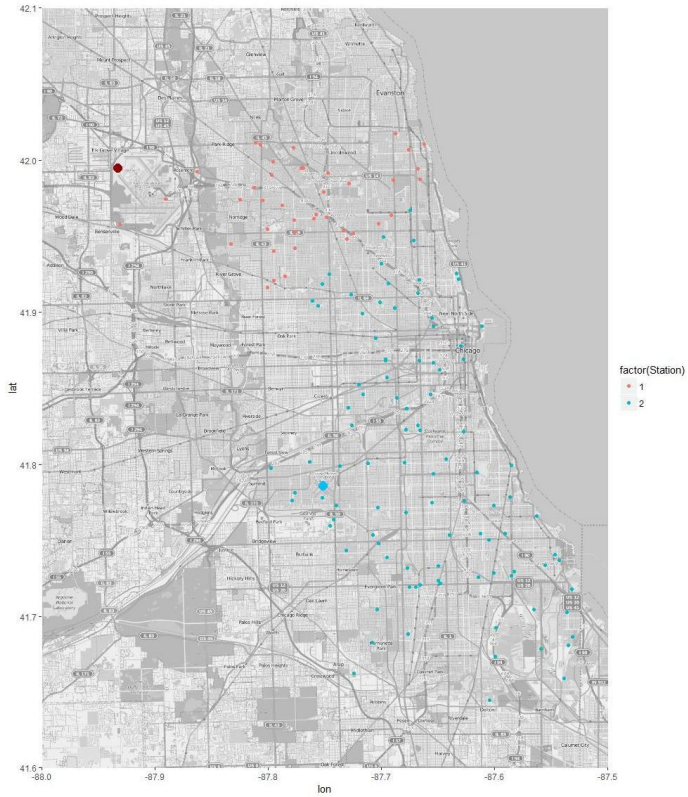
2.4 Weather information and West Nile virus presence

The next and final part of data analysing was to see a co-relation between weather information and West Nile virus detection. For that weather information from National Oceanic and Atmospheric Administration were used. There were two weather stations used for their reporting. To get what weather information is most suitable for a certain mosquito trap we needed to find the closet weather station for that mosquito trap. To that we used Haversine distance between the trap and the weather stations and got the weather station with minimal Haversine distance as the closet weather station.

After colouring the traps with different colours depending on the closet weather station, the map of the mosquito trap would look like Figure 7.

Author

Figure 7 Closet weather station with Haversine distance



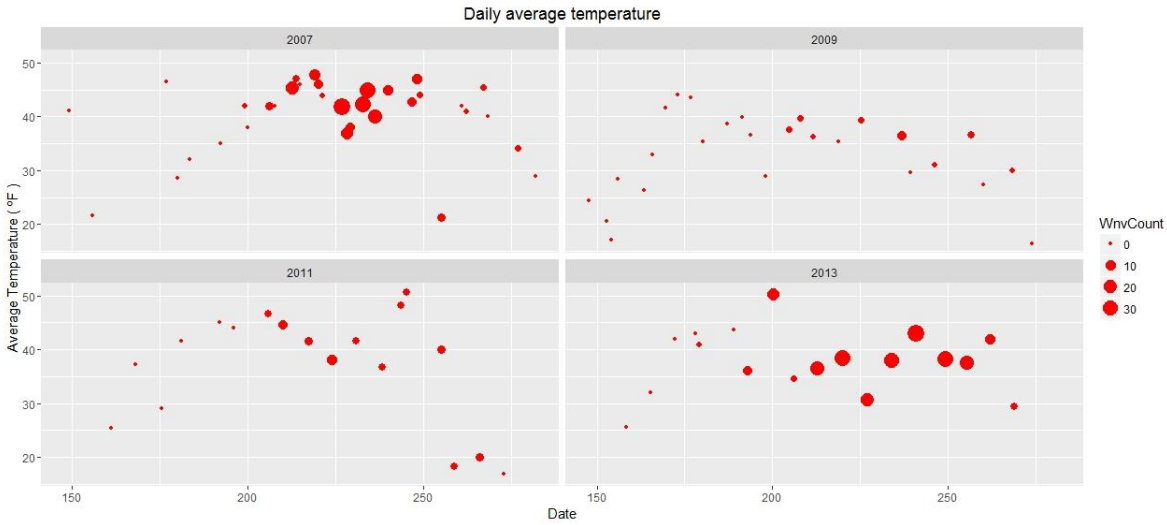
With the weather information from the closet weather station following graph was drawn to identify a relationship between average temperature and West Nile virus detection. Average temperature was plotted against the day number of each year indicating West Nile virus count for that particular day from the point size.

As Figure 8 implies each year West Nile virus detection count has increased when the temperature is in a high level – specifically when more than 35° F. So the temperature can be considered as an important factor when predicting West Nile virus.

Several other weather factors like wind, snow, rain were taken in to consideration when building the machine learning models.

Title

Figure 8 Day number of the year vs Average temperature



3 Machine Learning Research

The final prediction was based on an ensemble of two machine learning algorithms – Extreme gradient boosting and Regularized Greedy Forest. Following features that were identified in data exploration were used for the classifiers.

- Basic Features – Mosquito Species, Block, Latitude, Longitude, Address Accuracy, Month, Year
- Weather Features – Average Temperature, Dew Point, Heat level, Cool level, Precipitation Total, Average Wind Speed, Snow Fall, Sun Rise, Sun Set\

The weather information from the closet weather station were used for a trap as I have stated in Data Analysis Section. Also, the weather features have been smoothed inside the year with local polynomial regression fitting with span = 0.4. List parameters used in each classifier have been displayed in Table 1 and Table 2.

Table 1: Parameters for Extreme Gradient Boosting Classifier

Parameter	Value
n_estimators	1000
learning_rate	0.0035
loss	deviance

Author

max_depth	7
subsample	1

Table 2: Parameters for Regularized Greedy Forest

Parameter	Value
reg_L2	0.2
reg_sL2	0.07
algorithm	RGF
loss	Log
num_iteration_opt	7
num_tree_search	1
min_pop	8
opt_stepsize	0.7
max_leaf_forest	1400

It is worth to notice that we tried a lot of other methods (such as Random Forest Classifier, generalized linear methods, ExtraTrees). Some of them showed good results, but, unfortunately, did not improve the final predictions. Both Extreme Gradient Boosting classifier and Regularized Greedy Forest models were given a same weight in ensemble. Finally, we applied smoothing for the predictions based on closet traps and dates. Closet traps were obtained with the Haversine distance.

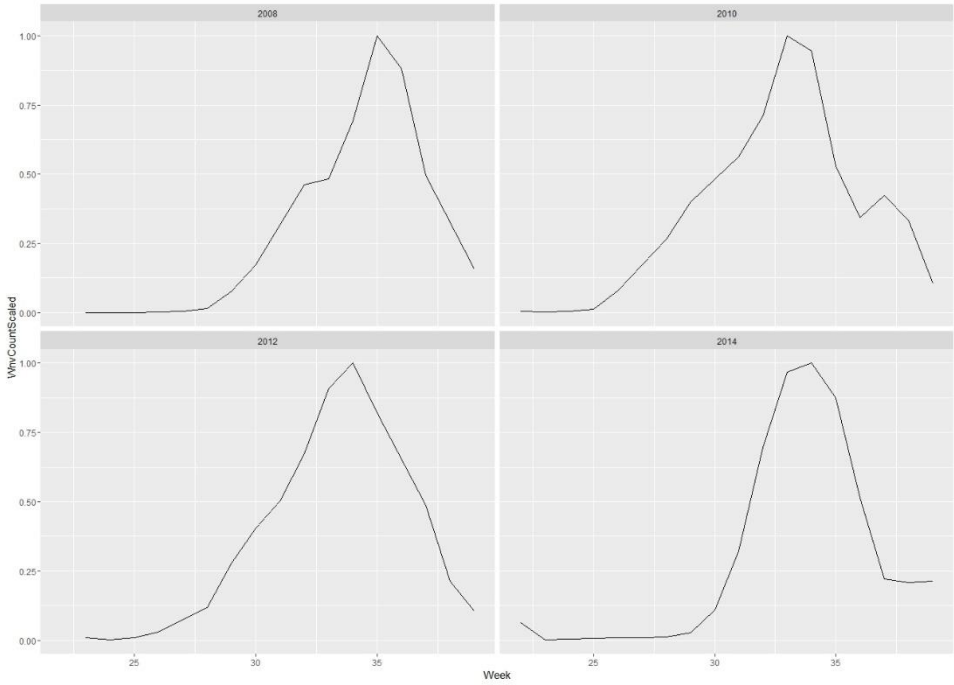
Using the above machine learning model, West Nile virus for 2008, 2010, 2012 and 2014 were predicted. The AUC score for the model was 0.79044. The predictions over time can be visualized below.

4 Conclusion and Future work

West Nile virus spread depends heavily on the mosquito species, time of the year, location and weather features. With the research discussed the virus can be predicted for a good accuracy to help the City of Chicago and CPHD more efficiently and effectively allocate resources towards preventing transmission of the virus.

Since there is no vaccine available to West Nile virus, the only way to control the virus is through spraying. The City of Chicago also does spraying to kill mosquitos. Spraying can reduce the number of mosquitos in the area, and therefore might eliminate the appearance of West Nile virus. If the spraying times and locations can be used to machine learning models there is a higher probability that the accuracy of the models will be better than this.

Figure 9 Prediction for 2008,2010,2012 and 2014



References

John Hart Jr, Gail Tillman, Michael A Kraut, Hsueh-Sheng Chiang, Jeremy F Strain, Yufeng Li, Amy G Agrawal, Penny Jester, John W GnannJr, Richard J Whitley and the NIAID Collaborative Antiviral Study Group West Nile Virus 210 Protocol Team. *West Nile virus neuroinvasive disease: neurological manifestations and prospective longitudinal outcomes*

W. S. Cleveland, E. Grosse and W. M. Shyu (1992). *Local regression models. Chapter 8 of Statistical Models in Seeds* J.M. Chambers and T.J. Hastie, Wadsworth & Brooks/Cole.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, pp. 2825-2830.

R. Johnson and T. Zhang (2011). *Learning nonlinear functions using regularized greedy forest.* Technical report, Tech Report: arXiv:1109.0887.

L. Breiman (2001). *Random Forests.* Machine Learning, pp.5-32.
J. Friedman, T. Hastie and R. Tibshirani (2008). *Regularization paths for generalized linear models via coordinate descent.* Journal of Statistical Software, Vol. 33(1), 1-22.

Author

J. Simm and I. Magrans de Abril (2013). *Package for ExtraTrees method for classification and regression.*

R.W. Sinnott (1984). *Virtues of the Haversine.* Sky and Telescope, 68 (2), 159.