**West Nile Virus Data**

The data in this project was based on 2 main sources:

1. Dataset of mosquitos catch and presence of West Nile Virus (WNV table).

2. Dataset of weather observations from 2 local climatological data stations.

The first table was consisted of the following features:

Date, Address of the Trap, Species of mosquitos, Block, Street name, Trap number, Full Address, Latitude, Longitude, Address Accuracy, number of Mosquitos, number of mosquitoes caught in this trap and WnvPresent (presence of WNV).

Each observation was dedicated to mosquito species caught in a certain trap on a specific date and whether or not it was positive to the presence of WNV.

The second Table was consisted of different weather observation. Some are quantitative, like temperature, barometric pressure, precipitation and a variable of different weather descriptive phenomena such as fog or thunderstorm.

Data Processing

At first, the two tables were treated separately. In WNV table, from the date-variable: A month, month in year, and a season variables were driven. Number of mosquitos-caught per date and per trap was united, while keeping the differentiation between Mosquitos species.

The outcome variable was defined as: whether or not (1 / 0) a WNV was tested positive among the mosquitos-caught in a trap in a given date.

The weather data contain observation from 2 separate weather station, the data was divided by those station and the two dataset were treated in the exact the same manner. The variable of the descriptive weather phenomena was transformed to a binary indicator variable for each phenomenon.

The three tables were then joined based on the dates; in this stage, variables without differential distribution were removed, redundancy of weather variables (within and between the two stations) were tested by Pearson correlation test (table 1) and pairs- plot visualization (see fig.1). Variables found to be highly correlated were removed.

Table 1: a correlation matrix using Pearson correlation test of the weather variables (temperature related mainly).

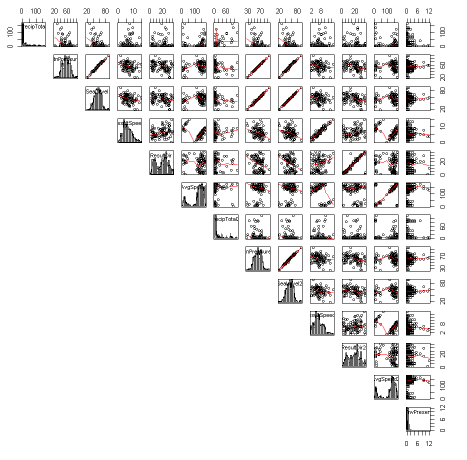


Fig 1: Pairs plot of weather variables (temperature related mainly) from station 1 & 2, variables with identical distribution can be identify (nb “ffWNV2 data exploration R”).

Data exploration

Two significant trends revealed by data exploration:

1. Two mosquito’s types (*Culex Pipiens* and *Culex Pipiens* or *Restuans*) were correlated to the presence of WNV (table 2 & fig.2).

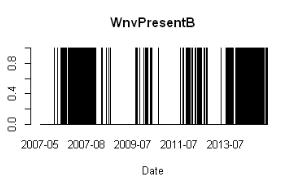
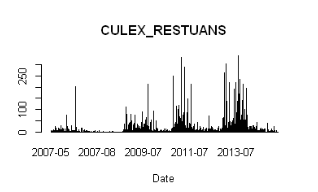
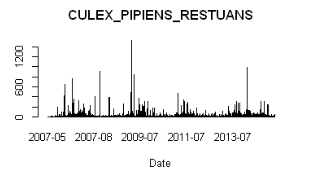


Fig. 2 Presence of mosquito types Culex Pipiens and Culex Pipiens or Restuans is visually and statistically correlated to positive tested WNV

1. Elevated temperatures were positively correlated to WNV presence (table 2 & fig.3).

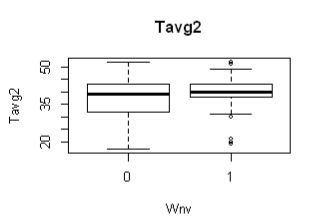
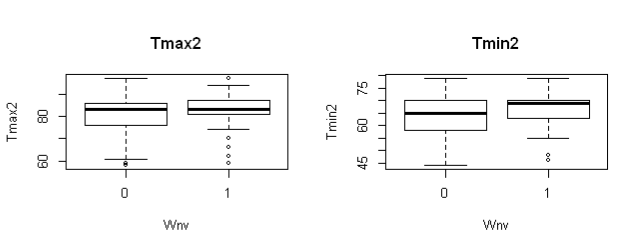
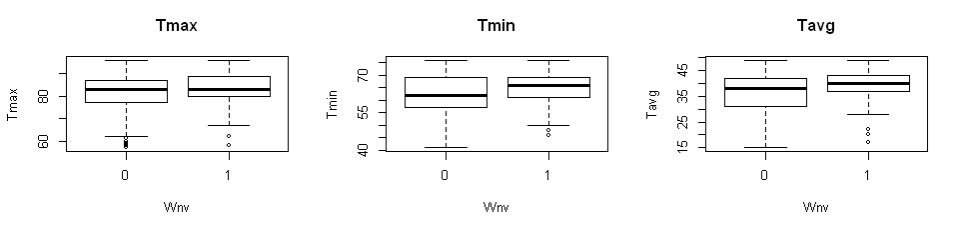


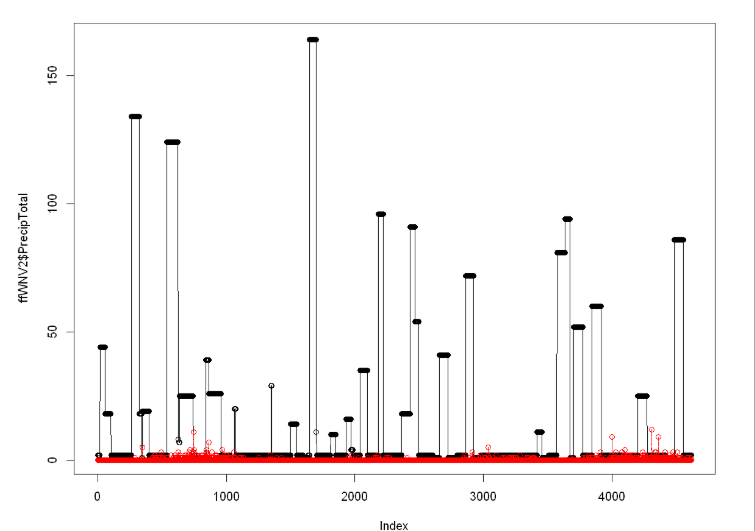
Fig. 3: Maximum, minimum and average temperatures in relation to WNV presence. The positive correlation was visible mainly in the minimum temperature (Tmin), though in all 3 temperature variables when WNV was positive the temperature was statistically significantly higher.

Table 2: “Tableone” analysis containing variables and categories in which p.v.<0.05



* Full data exploration can be found in notebook: “ffWNV2 EDA 030219”, “ffWNV5 EDA 110219 R”, “ffWNV2 data exploration R”.
* Table one and can be found in notebook: “ffWNV5 - Table1 & exploreData”.
* Visual exploratory of the data can be found in PDF files: Exploratory Data Analysis weaq, Exploratory Data Analysis wead, Exploratory Data Analysis Other, Exploratory Data Analysis Newvar, Exploratory Data Analysis MOS.

Additional interesting feature can be seen when the precipitation variable was exhibited versus the presence of WNV (fig.4). It appears that in the days post major precipitation-events, an increase in WNV presence occurred. This observation is supported by previous knowledge and is treated in the variables engineering (see below).



WNV

PrecipTotal

Fig.4: In days following precipitation events (>50) the presence of WNV is increasing. The x axis is index but it is in direct relationship to time (precipitation-black dots, WNV-red circles). \*Please note the WNV in this fig. is semi-quantitative and not binary as found in the rest of the data analyses (the semi-quantitative phase of WNV was a result of data transformation and is used only to demonstrate this phenomenon).

Variables engineering

1. Two variables added based on hospitalization data in Cook County. The data is originally displayed as percentage from total number of hospitalization per Quarter. In aim to adjust to the project dataset, the given data was divided to 3 and for each observation the variable stand as hospitalization per month (<http://www.idph.state.il.us/emsrpt/form-hospitalization.asp>).
2. Percentage of parasitic infection hospitalizations from total number of hospitalizations in Cook County. Taken from Cause of Hospitalization characterized as “Infection/Parasitic – Other” code 010-139 (exclude 038) by the ICD-9 codes for the principal diagnosis associated with the hospitalization that include West Nile Virus (ParasiticIn.).
3. Percentage of pneumonia/influenza hospitalizations from total number of hospitalizations in in Cook county, Taken from Cause of Hospitalization characterized as Pneumonia/Influenza code 480-4883 by the ICD-9, due to the similarity in symptoms of WNV illness to influenza (Influenza.).
4. Four additional variables were engineered based on the fact that 2 weeks (and up to 2 months) after an area received more rainfall than usual an increase may occur in the number of virus-spreading mosquitos:
5. For each of the weather stations, the amount of rain events was counted 45 days backwards (RA45i and RA45\_2i).
6. For each of the weather stations, a binary indicator variable was added based on the original precipitation variable: Events of precipitation that were higher the 50 were marked as 1 if occurred in the last 45 days (PT45\_gt50i and PT45\_gt50\_2i).

Outliers determination and treatment

Weather data was tested for outliers, most of the data appear to be uniform and no outliers that need to be treated were identified (See nb: ffWNV4 R Data Cleansing).

Missing values

PrecipTotal2 was the only variable that originally had some missing observations (and was not filtered out due to redundancy). 35 subsequent observations were missing in this variable.

The engineered variables (RA45i , RA45\_2i, PT45\_gt50i and PT45\_gt50\_2i) that calculated based on precipitation and events of rain were missing 58 observations resulting from their calculation method.

The test for Missingness Generation Mechanism resulted with inconclusive outcome for all 5 variables.

For PrecipTotal2 variable it can be assumed that the missingness was a result of measurement failure (MCAR), thus allowed imputation to PrecipTotal2.

As for the engineered variables, the missingness is clearly a result of the variable generation method (MAR) and thus imputation is allowed as well (nb: ffWNV4 R Data Cleansing).