Predicting West Nile Virus (WNV) Occurrences in North Dakota using Data Mining Techniques

1Mitch Campion, 1Calvin Bina, 1Martin Pozniak, 2Todd Hanson, 3Jeff Vaughan, 4Joseph Mehus, 5Scott Hanson, 6Laura Cronquist, 6Michelle Feist, 1Prakash Ranganathan, 1Naima Kaabouch, 7Mark Boetel
1Department of Electrical Engineering, University of North Dakota, Grand Forks, ND 58202
mitchell.campion@ndus.edu

Abstract—This paper discusses a model that predicts trap counts of Culex tarsalis, a female mosquito that is responsible for West Nile Virus (WNV) using machine-learning algorithms. Culex mosquitoes are the main transmission vectors for WNV infections. In this research, a Partial Least Square Regression (PLSR) has been deployed to predict mosquito trap counts of Culex tarsalis using historical meteorological and trap count data from 2005-2015. The associations between 10 years of mosquito capture data and the time lagged environmental quantities trap counts, rainfall, temperature, precipitation, and relative humidity were used to generate a predictive model for the population dynamics of this vector species. Statistical measure of Mean Absolute Error (MAE) is compared with other existing actual collected trap counts to analyze accuracy the predictive models. The paper also details the development of a user-friendly web-interface containing interactive web pages that allow users to visualize the North Dakota mosquito population, weather pattern, and WNV incidence data. The interface utilizes multi-layered Google Maps developed through Google Fusion Tables. An understanding of historical data and weather variables is essential for providing sufficient lead time to predict WNV occurrence, and for implementing disease control and prevention strategies such as spray period and hiring of seasonal mosquito workers. Further, an approach similar to the proposed approach of this paper, which involves the integration of data mining and data visualization techniques, brings novelty to vector control initiatives.

Keywords—Culex tarsalis; Partial Least Squares Regression (PLSR); West Nile Virus (WNV); Vector Control

I. INTRODUCTION

West Nile Virus (WNV) [1] is a potentially fatal mosquito-borne flavivirus and human, equine, and avian neuropathogen. Birds are natural carriers for the virus and contribute to its transmission. The transmission of this virus is delicately maintained by a mosquito-to-bird-to-mosquito transmission cycle that primarily involves the Culex species of mosquitoes. In North Dakota, the mosquito species Culex tarsalis is the primary vector that carries WNV [2],[3]. Humans can contract WNV through being bit by mosquitoes that are infected with the virus. Most human cases of WNV are not severe enough to differentiate from a common flu, so it is thought that many infections are not clinically treated. However, clinical infections have the ability to be quite severe and even fatal in causing meningoencephalitis [1],[4]. The risk of incidence of severe neuro-invasive WNV as well as death due to the virus increases with age. There is no existing WNV specific treatment or vaccine available. Current prevention protocols depend exclusively on organized and sustained vector/mosquito control and public education. These protocols include timely public spraying of insecticides as well as public awareness initiatives to educate and inform public of WNV related risk factors [1].

West Nile Virus is native to Africa, Asia, Europe, and Australia; but incidence of the disease in the United States has become of recent concern. The first known incidence of WNV occurred in 1999 in New York during an epidemic of meningoencephalitis [1],[4]. Since that time, the virus has spread throughout much of the rest of the USA and Western Hemisphere. The first discovery of WNV in North Dakota occurred near Grand Forks, ND, in June of 2002 [2].

Culex tarsalis is the species that is crucial to observe and control WNV in North Dakota. Larval production of Culex tarsalis is known to begin during the late spring and continues until early fall with several generations being produced in this time. Depending on weather conditions, Culex tarsalis can complete egg to adult development in as few as 4 to 10 days [3],[5]. The maximum adult population of these mosquitoes is generally observed during August or September. The females are known to be painful and persistent biters. Most of this biting occurs at dusk and after dark [3]. In addition to incidence of WNV, mosquito bites also pose a nuisance to the general human population. Many species of mosquito cause annoyance during outdoor activities as well as general discomfort from bites which cause itchy bumps on the skin. These itchy bumps, unlike WNV, are not exclusive to being caused by the Culex tarsalis mosquito in North Dakota, and can be observed after being bitten by a number of different mosquito species [2],[3],[5].

The goal of this project is to use insights made possible from data mining techniques to aid in public awareness and education of seasonal risks due to WNV. Specifically, the objective is to produce accurate predictions for weekly mosquito population in North Dakota based on historical

798-1-5090-4171-8/16/$31.00 ©2016 IEEE
mosquito data collected from years 2005-2015. Data containing corresponding climate and weather phenomena is used extensively as predictive variables. The relationship between mosquito population and weather phenomena is vaguely known, but varies depending upon compounding factors including species, location, and local ecosystems [6], [7], [8]. The combination of historical mosquito population data and robust historical weather data makes for a rather large amount of data. Aggregation and analysis of large amounts of data can be cumbersome and tedious. However, underlying trends in the data must be discovered in order know beforehand the expected incidence of West Nile Virus and mosquito populations based on the predictive influence of meteorological data. In order to discover these underlying trends, advanced techniques converging from fields of mathematics, computer science, and data science are necessary.

In addition to modeling mosquito trap count populations and WNV incidence, another objective is to create user friendly and interactive visualization tools that have the capability to showcase the trends and risks of West Nile Virus throughout the state. These interactive visualization methods will be showcased on a disease control website for use of industry persons as well as the general public. This specific objective requires conglomeration of data visualization and web development. Together, the use of data mining techniques will efficiently meet these objectives and will modernize the public awareness and vector control initiatives and provide an increased depth of insight into mosquito population and West Nile Virus incidence in North Dakota.

II. DATA MINING

A concise definition for the umbrella term “data mining” is: “the process of discovering and extracting useful patterns in large data sources [9].” Data mining exists as an intersection and aggregation of a large number of different fields including mathematics, statistics, computer science, artificial intelligence, and many others. Techniques of data mining are related to and serve as an umbrella to analytical tools including statistical computations, clustering algorithms [10], predictive modeling algorithms [11], and many more. The number of possible applications for data mining techniques is innumerable. Some notable applications of data mining that have had large influences on modern society include image recognition techniques used by popular social media applications such as Facebook [12] and Snapchat [13], pattern recognition algorithms used by online retailers to suggest items to shoppers [14], and is even used in the modern power grid [15]. Because data mining is such a large conglomeration of techniques and has vast application, an accepted definition of data mining varies slightly depending upon the source. The concept of data mining can have other aliases including “Big Data” or “Data Science.”

A. Predictive Modeling

The formal definition of predictive modeling given by [11] is “the process of developing a mathematical tool or model that generates an accurate prediction.” Predictive modeling can be used to quantify available information and make more informed decisions based upon new discovery of information. In scenarios with large amounts of information (data), modeling techniques afford computational abilities to make models with predictive inference that aid in practical decisions. Predictive modeling is strongly associated with data mining due to overlapping algorithms of discovering hidden knowledge in data. Predictive modeling was used extensively in this project in forming statistical univariate and multivariate regressions for determining trap counts based upon historical data as well as multi-variable climate data. There are a variety of predictive modeling regressions and algorithms. For purposes of this paper, a brief discussion of a relevant method is included.

1) Partial Least Squares Regression (PLSR)

A number of different multivariate regression models were tested for the modeling section of this project. After analysis of many algorithms Partial Least Squares Regression (PLSR) [11], [16] was determined to be an appropriate method to predict mosquito trap counts based on a number of weather variables. PLSR is a regression that is similar to Multiple Linear Regression [17], Ordinary Least Squares Regression [18], and Principal Components Regression [19]. PLSR is a useful method for finding underlying relationships among predictive variables that are correlated to the target response.

The PLSR algorithm produces linear combinations of predictor variables that maximally summarize the covariance with the response variable. These linear combinations are called components. In essence, PLSR iteratively finds linear combinations of the predictor variables that have maximum covariance with the target variable. Figure 1 shows a general block diagram of the PLSR algorithm. The algorithm does this by creating “n” and “m” dimensional vector spaces, where n is the number of predictor variables and m is the number of target (dependent) variables. The PLSR will find the multidimensional linear direction in the n-dimensional space that has maximal covariance with the target variables of the m dimensional target space. After it finds a component, it subtracts the component from the original variables and recomputes more components until 100% of variance is accounted for. Figure 2 shows a general illustration of these spaces and maximal covariance. PLSR was chosen because it works well when the number of predictor variables is high and/or there is correlation or collinearity among the predictor variables [11], [16], [19]. The meteorological datasets used in this project contain many variables and these variables show collinearity.

Figure 2 is a generalized diagram showing the
transformation of predictor and target variable spaces into a component spaces formed from linear combinations of these spaces. The “X” or predictor space is the n dimensional space formed by the predictor variables. The “Y”, or target variable space shown in the left side of figure 2 displays a general representation of an m dimensional target variable space. For our project, the target variable space has only one dimension, the targeted variable of Culex tarsalis trap count. The predictor space of this project is the space containing all of the predictor meteorological variables. The PLSR computes linear combinations of predictor variables into a transformation vector, “T.” This vector is computed such that the PLSR component maximally summarizes covariance with the predictor space vector.

Fig. 1. Block Diagram of partial least squares algorithm

2) Time-series Decomposition by Loess (STL)

Time series decomposition [20] is a statistical method that decomposes a time series dataset into a number of sub components. A time-series [21] is any dataset that consists of a sequence of data points representing successive measurements during a given time interval. In our work it was of interest to decompose the trap count time series data into trend, seasonality, and remainder components.

The seasonality component is a part of the data set that recurs seasonally. It is essentially a pattern within the data series that reoccurs during a certain season of the time series. The trend components represents long-term increase of decrease in the data. The remainder component is the difference between the observed data and the sum of the trend and the seasonality. A time-series decomposition with consistently small remainder component would indicate a very ordered time series and a very accurate decomposition. An example of a data series that exhibits trend and seasonality would be monthly electricity demand. Daily electricity demands follow a daily seasonality pattern based on electricity consumption during peak and non-peak hours. This series would also exhibit an overall increasing trend as the demand for electricity has increased with an increasing number of consumer appliances that run on electricity [20], [21]. This method is shown as a segmented plot of trap count as a time series decomposition in Figure 7 of the results section of this paper.

III. METHODS

A. Datasets

In order to model and visualize the trends of mosquito counts and West Nile Virus in North Dakota, a large amount of data must be collected and analyzed. The North Dakota Department of Health coordinates a state wide observation of mosquito trap counts. Each county’s public health department sets mosquito traps of multiple varieties in various locations of the county in North Dakota. The traps considered in this work are New Jersey light traps that are set in each county. The traps are harvested weekly, usually beginning with the last week of May or the first week of June until the first week of September. Each week the mosquitoes that are trapped are identified according to their species and the number of each species is recorded. This data is recorded and published to the North Dakota Department of Health Website [22]. Grand Forks county has a relatively advance vector control unit and also publishes daily trap count information and other public awareness material via their website [23].

Mosquito trap count (number of each species of mosquito caught in the traps) data has been collected since 2002. A number of species are identified in these datasets; however, for this work the primary focus is the Culex tarsalis species. Future work may also consider Culex pipiens, which is a main vector in transmitting WNV to the bird population. After much time and consideration in organizing the state-wide historical trap count data, a master trap count data set was formed. This master data set has, for each year: a weekly listing of mosquito trap count for each species, the county the population was trapped in, the location within the county, the start and end dates of the weekly collection period, and the numerical week of the summer. In order to make predictive inferences of mosquito trap count based on climate factors, a dataset containing climate variables is necessary. The ND Department of Health does not record climate data in addition to the trap count data so open source weather data from the internet was discovered, organized, and utilized.

The climate data was found using an open source database available through Weather Underground website [24]. This provides local and long range weather forecasts and reports for locations worldwide. The website has a feature that allows users to view historical meteorological data for specific locations. In addition to viewing historical weather, the site allows historical weather data to be extracted to a .csv file. The extracted .csv file contains daily weather metrics for a time period specified by the user. On a daily basis this data contains max, min, and average temperatures, precipitation, humidity, wind speed, and barometric pressure. Weather data from each county was extracted for each summer
corresponding to trap count data, organized, and appended to the mosquito trap count dataset.

1) Data Cleaning, Aggregating, and Pre-Processing

A majority of time allocated to this project related to data cleaning, aggregating, and processing. Because the data for this project is quite voluminous and arrives from multiple sources, the process of organizing and standardizing the data into a homogenous and useful format was quite extensive. The trap count data alone consists of data sheets from each of North Dakota’s 53 counties. Each week for the past 11 years has its own separate data sheet and entry of the data was not always standardized to a similar format.

Data entries for each county have been recorded weekly since 2005. This results in a mass of data to organize. It is worth noting, that the trap data was not originally collected or organized with this type of analysis in mind. Clever techniques were employed to organize this data in a meaningful way, and were quite time consuming. Additionally, the aggregation of meteorological data sets was equally cumbersome. Proper analysis and visualization is impossible without effective techniques of pre-processing and organizing these data sets. The results of this work are largely dependent upon the effective organization of these datasets.

B. R Software Programming Language

Data cleaning, organization, statistical analysis, predictive modeling, and data visualization for this project was conducted using R software. R [25] is an open source software programming language with specific statistical functionalities. The R environment is an integrated suite for calculation, graphical display, statistics, visualization techniques, and data manipulation. R software framework and the R-studio interface were used extensively for this project. R was initially written by Robert Gentleman and Ross Ihaka of the statistics department at the University of Auckland. Today, R is a result of contributions from users all over the world. In addition to users creating new functions there are thousands of developed packages available with countless functions and capabilities available on the Comprehensive R Archive Network (CRAN).

C. Visualization

The end goal of this work is to create interactive methods available to industry and the general public that can be of assistance in understanding mosquito population and WNV risk factors. Preliminary work has focused on development of interactive tools for a website. Current web development showcases a user-friendly and interactive interface that allows a user to effectively visualize and analyze the North Dakota Mosquito count and West Nile Virus incidence data. The website accomplishes this by the use of interactive maps, charts, and mathematical models. The interface utilizes layered interactive Google maps developed using the Google Maps API as well as interactive charts developed through Google Fusion Tables.

The layered Google maps are geocoded with clear representation for the counties of North Dakota. The maps offer “on-click” functionalities that allow a user to interactively click on their county on the map and mosquito population data as well as climatology data are displayed. In addition to layered geocoded maps, “heat maps” showcasing population density have also been developed. These heat maps allow for intuitive visualization of areas in the state where population density of mosquitoes is highest. Examples of a layered google maps sorted by species trap count are shown in figures 3 and 4. An example of a heat map showcasing population density of Culex tarsalis is shown in figure 5. Continued visualization work will involve developing these visualization maps to automatically update to correspond with collected trap data as well as developed prediction models. This will allow for real time interactive use for the public and also aid units in vector control initiatives.
the trend for that given year since the seasonality component is consistent to each year. In general, the trend component is the component most related to outside factors including climate variables.

Because of the independence of the seasonal component and the dependency of the trend component, the predictive modeling approach involved predicting the level of the trend at which the seasonality would occur. Discussion with subject matter experts [23], [26], [27] as well as literature review show that specific weather variables effect the population of *Culex tarsalis* [6]–[8], [27], [28]. In addition to the information provided, principal components and regression analysis confirmed the influence of temperature and rainfall variables.
Preliminary modeling utilizes Partial Least Squares Regression considering the influence of average of monthly maximum temperatures for months March, April, and May as well as cumulative rainfall amounts for April and May in an initial 13-week prediction for the 2016 summer. The general modeling process can be explained by the flow chart shown in figure 6. The initial PLSR model is shown in figures 8 and 9. This model considers weather data accumulated in the spring to predict an initial trend component for a summer prediction. Additional variables are considered and updated for more accurate predictions throughout the summer. Weeks 2-13 additionally consider rainfall, mean max temperature, mean maximum humidity, and *Culex tarsalis* trap count of the previous week. Weeks 3-13 consider these same variables with the addition of their influence from two weeks prior. With the consideration of variables lagging 1 and 2 weeks the initial prediction is updated each week. This update modifies trap count of week ahead and two week ahead predictions from the initial model.

**TABLE I.** Tabular display of different prediction algorithms compared with the actual trap count for 2015 and the corresponding MAE of each prediction. PLSR shows the least amount of prediction error.

<table>
<thead>
<tr>
<th>Week</th>
<th>2015 Actual</th>
<th>STL+Random Walk</th>
<th>Holt-Winters</th>
<th>ARIMA</th>
<th>PLSR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0</td>
<td>12</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>1</td>
<td>16</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>26</td>
<td>16</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>15</td>
<td>8</td>
<td>2</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td>40</td>
<td>15</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>

**MAE**  

| 5.62 | 4.5 | 5.3 | 3.3 |
IV. RESULTS DISCUSSION

Figures 8 and 9 show PLSR trap count prediction models for Grand Forks County for the current summer 2016. Grand Forks County Public Health Department sets a number of traps throughout the county during the summer months. The models produced in this work predict the average trap counts of the “New Jersey light” traps that are reported to the North Dakota State Department of Health and are published publicly. The PLSR predictive modeling schemes with spring and summer meteorological variables were tested against actual data from 2015 and 2016 to test model accuracy. Data recorded up to 7/24/2016 shows a Mean Absolute Error (MAE) of 3.63 for the initial PLSR model and the same MAE of 3.63 for the PLSR model that is updated weekly with weekly meteorological data. Even though these models predict different values for a given week, as of 7/24/16 they have shown the same accuracy. The 2015 test of the PLSR method is shown in figures 9 and 10. Table 1 shows this test from figure 10 in addition to other predictive modeling algorithms that were employed. As shown by Table 1, PLSR performed most accurately in predicting the average trap count for Grand Forks County in 2015 compared to other algorithms. The 2015 PLSR test prediction shows an (MAE) of 3.3 trap counts. This accuracy in prediction justified the use of this prediction scheme for the current 2016 summer.

V. CONCLUSION

The mosquito population depends on several variables such as weather patterns, mosquito species type, public awareness and mosquito control initiatives. Preliminary work has shown great potential to be of assistance mosquito control initiatives and public awareness campaigns. Visualization methods developed through Google fusion tables offer an intuitive, familiar, and modern interface for control groups and general public interaction. In conjunction with predictive modeling approaches these visualization techniques will also be helpful in informing and quantifying public health departments of mosquito populations so that resources can be accurately allocated to effectively control WNV vector populations.

The modeling of mosquito trap count based upon climate data is a daunting task. There are known factors to mosquito population that are difficult to quantify such as existence of standing water in residential areas. Initial model results lead to known statistical accuracy of MAE=3.3. A continuous and frequent update of the model (weekly/bi-weekly) is needed to improve the accuracy of prediction. Otherwise, prediction can be a perilous task.

VI. FUTURE WORK

In addition to building predictive models, more analysis is needed in quantifying specific variables that correlate strongly with mosquito population as well as WNV incidence. The effect of weather phenomena on mosquito population is generally known in literature, but there has been no known published work that specifically quantifies these relationships in North Dakota. Continued work will focus on improving prediction model accuracy and further development of interactive visualization techniques that correspond to mosquito population modeling. More accurate predictions are desired to aid in mosquito control efforts. Additional modeling attempts may include improving current PLSR framework with new predictive data, combining PLSR framework with other methods for a hybrid modeling approach, use of other existing algorithms, or development of a novel prediction algorithm. Additionally, specific predictive models will be created to predict not only mosquito population, but also risk levels of human WNV contraction. Mosquito population and risk of WNV contraction are correlated, but specific attempts for incidence risk modeling are currently in progress and aim to be implemented for the 2017 summer.

An integrated data mining, predictive modeling, and data visualization web framework is the ultimate goal of this work. This approach involving data mining and enhanced visualization techniques brings novelty to current mosquito control initiatives. A similar approach could theoretically be instituted in managing vector populations for other species of mosquito, including those species that are primary vectors for Zika virus. A comprehensive approach similar to the method proposed by this paper, one that involves data mining technologies, shows promise to be used for the development of future technologies that will aid in the control vector populations and public awareness initiatives.

ACKNOWLEDGMENT

The authors acknowledge the support of Subject Matter Experts (SME) on the project that include: Todd Hanson, Mosquito Control Program Manager, Grand Forks City, Scott Hanson, Mosquito Research Expert, Turtle Mountain Reservation, Jeff Vaughan, Department of Biology at UND, and Joseph Mehus at Mayville State University. Also acknowledged are Laura Cronquist and Michele Feist of the North Dakota Department of Health for data collection and aggregation. Resources for this work were provided by NASA ND EPSCoR Grant #0020769.

REFERENCES


