Table of Contents

Harnessing Computational Tools and Complex Biological Data for Environmental and Health Applications

I. Hypothesis

Interdisciplinary bioinformatics approaches can elucidate the complex interactions between environmental factors, vector biology, and pathogen dynamics, contributing to comprehensive public health strategies for disease prevention and control.

Advanced computational methods offer critical tools for dissecting the multifaceted influences on West Nile virus transmission dynamics. This research proposes to delve into the subtleties of these relationships, acknowledging the observed minimal direct impacts from environmental and demographic factors. By refining predictive models to incorporate a broader spectrum of data, including nuanced interactions and indirect effects, this work aims to uncover overlooked aspects of disease ecology that may be important for public health interventions against vector-borne diseases.

The integration of environmental, biological, and socio-demographic data using advanced bioinformatics techniques will elucidate previously unrecognized patterns that significantly contribute to human health outcomes and the efficacy of disease prevention measures.

II. Specific Aims

- Aim 1: To identify and characterize genetic variations in *Culex tarsalis* populations that may confer environmental adaptations.

- Aim 2: To develop and validate a national-scale predictive model for annual West Nile virus transmission, integrating environmental and demographic data. This model aims to enhance early warning systems for public health preparedness at a country-wide level.

- Aim 3: To create and assess a regional predictive model for monthly West Nile virus cases, focusing on county-level data. This model will incorporate detailed environmental and demographic information to support more localized public health interventions and vector control strategies.

III. Background and Significance

Mosquito-borne diseases, such as West Nile Virus (WNV), present formidable public health challenges worldwide. Since its initial detection in the United States in 1999, WNV has inflicted a significant toll, with over 56,000 documented cases and 2,700 fatalities. Approximately 20% of those infected develop West Nile fever, with a critical 1% suffering from severe neurological illnesses [1]. Its pervasive spread has firmly established WNV as the foremost mosquito-borne disease in the country, warranting meticulous attention from public health authorities. This emergence underscores the

intricate interplay of environmental conditions, geographical features, and vector biology in shaping disease transmission dynamics.

WNV transmission dynamics are intricately influenced by a multitude of factors. Environmental conditions, including temperature, humidity, and precipitation, significantly impact mosquito populations and virus replication rates [2]. Moreover, geographical features such as land use patterns and water bodies play crucial roles in shaping vector habitats and breeding grounds [3]. Population density and human behavior also contribute to WNV transmission, as urbanization and human activities can create favorable conditions for mosquito proliferation and contact with infected reservoir hosts [4,5] . Additionally, In the continental U.S., here are about 12 distinct mosquito species which can transmit diseases to humans, but not all of them have comprehensive genetic resources. *Cx. tarsalis* is a major vector for WNV in the United States [6], and is a predominant vector of the disease in the most severely impacted states in the West and Midwest [7]. Despite its significance, comprehensive genetic resources for *Cx. tarsalis* remains scarce, impeding our ability to elucidate its population dynamics and adaptability. Interestingly, our study of population genetics in *Cx. tarsalis* reveals a pattern of genetic differentiation that suggests a potential role for selection in addition to genetic drift. This pattern hints at environmental adaptations driving population divergence in *Cx. tarsalis*, suggesting that identifying the environmental factors and genetic determinants under selection is vital for predicting the spread of *Cx. tarsalis* and, by extension, WNV outbreaks.

Recent research endeavors have endeavored to predict WNV transmission dynamics through diverse methodological approaches across varied geographical contexts. Holcomb et al. [8] highlighted the significance of historical disease incidences and population density over climate anomalies in the U.S., while José-María et al. [9] found climatic variables, human-related factors, and topo-hydrographic features to be key in Europe. Additionally, John M. Humphreys et al. [10] pointed out the role of drought in amplifying virus transmission within the U.S. Despite these varied approaches, the overarching theme from these studies suggests that while predictive models can identify potential risk factors and outbreak patterns, the actual predictive power for specific outbreak events remains limited. These findings underscore the complex and multifaceted nature of WNV transmission dynamics, where single factors or models may not capture the full spectrum of variables influencing disease spread, indicating a need for more sophisticated and integrative predictive frameworks.

In summary, the collective research efforts across different regions and disciplines highlight the multifaceted nature of mosquito-borne disease transmission and the critical role of genetic, environmental, and ecological factors in shaping disease dynamics. Understanding the genetic diversity and population structure of vectors like *Cx. tarsalis*, alongside environmental and host factors, is paramount in developing comprehensive strategies to mitigate the impact of diseases like WNV and protect public health.

IV. Previous Work in Metabolomics & Machine Learning

The rapid expansion of metabolomics studies, particularly in databases like Metabolomics Workbench [11] and Metabolights [12], underscores the need for sophisticated tools capable of efficiently managing and analyzing mass spectrometry data. My research focused on refining the ADAP-KDB algorithm [13], crucial for efficiently processing mass spectrometry data, including identifying and prioritizing spectra of both known and unknown compounds. This enhancement is key to effectively analyzing extensive metabolomics data, enabling the identification of distinct metabolic signatures. In parallel, I developed a pipeline to discern these signatures through untargeted metabolomics studies. This dual approach aims to reveal robust metabolic patterns, crucial for understanding disease mechanisms and enhancing diagnostic and treatment strategies in biomedical and public health research.

1. Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreening *[14]*

The original ADAP-KDB spectral search algorithm (Figure 1), which used a relational database for storing and querying spectral data, became increasingly slow with the growth of the spectral database. This slowdown was primarily due to the method used to calculate spectral similarity between a query spectrum and library spectra, which involved sqrt-cosine similarity calculations for each comparison. Each query spectrum needs to be compared with an increasingly large number of library spectra, a process that is computationally intensive. Additionally, using a general-purpose relational database like MySQL, while memory-efficient and cost-effective, may not be the fastest for spectral search, particularly when dealing with vast data. As the database grows, the time taken to match a query spectrum to all the library spectra in the database increases, leading to slower overall performance.

```
SELECT COUNT(*) AS Common, TempTable.Id FROM(
    SELECT Id from Spectrum where ((ABS(TopMz1 - mz1) < 0.1) OR (ABS(TopMz2 - mz1) < 0.1)
        OR (ABS(TopMz3 - mz1) < 0.1) OR (ABS(TopMz4 - mz1) < 0.1) OR (ABS(TopMz5 - mz1) < 0.1)
        OR (ABS(TopMz6 - mz1) < 0.1) OR (ABS(TopMz7 - mz1) < 0.1) OR (ABS(TopMz8 - mz1) < 0.1))
    UNION ALL
    SELECT Id from Spectrum where ((ABS(TopMz1 - mz2) < 0.1) OR (ABS(TopMz2 - mz2) < 0.1)
        OR (ABS(TopMz3 - mz2) < 0.1) OR (ABS(TopMz4 - mz2) < 0.1) OR (ABS(TopMz5 - mz2) < 0.1)
        OR (ABS(TopMz6 - mz2) < 0.1) OR (ABS(TopMz7 - mz2) < 0.1) OR (ABS(TopMz8 - mz2) < 0.1)
        OR (ABS(TopMz9 - mz2) < 0.1))
    ...
    UNION ALL
    SELECT Id from Spectrum where ((ABS(TopMz1 - mz8) < 0.1) OR (ABS(TopMz2 - mz8) < 0.1)
        OR (ABS(TopMz3 - mz8) < 0.1) OR (ABS(TopMz4 - mz8) < 0.1) OR (ABS(TopMz5 - mz8) < 0.1)
        OR (ABS(TopMz6 - mz8) < 0.1) OR (ABS(TopMz7 - mz8) < 0.1) OR (ABS(TopMz8 - mz8) < 0.1)
        OR (ABS(TopMz9 - mz8) < 0.1) OR (ABS(TopMz10 - mz8) < 0.1) OR (ABS(TopMz11 - mz8) < 0.1)
        OR (ABS(TopMz12 - mz8) < 0.1) OR (ABS(TopMz13 - mz8) < 0.1) OR (ABS(TopMz14 - mz8) < 0.1)
        OR (ABS(TopMz15 - mz8) < 0.1))
) AS TempTable
JOIN Spectrum ON Spectrum.Id = TempTable.Id
GROUP BY Id
ORDER BY Common DESC
```

Figure 1: Pseudo SQL query for calculating similarity scores between a query spectrum and all library spectra.

- Improved algorithm

The new algorithm in ADAP-KDB is faster due to the implementation of a prescreening search step (Figure 2). This step allows the algorithm to quickly identify candidate spectra, reducing the need to calculate similarity scores for every library spectrum. For this speed improvement to be effective, three conditions must be met: (1) the pre-screening algorithm is much faster than the main search algorithm, (2) it returns a relatively small number of candidate spectra, and (3) the returned candidate spectra include the correct match to the query spectrum.

In this new algorithm, the process involves pre-calculating m/z values of the largest peaks in all library and query spectra. Then, the m/z value of the largest peak in the query spectrum is matched to m/z values of n largest peaks in the library spectra, and this process continues until the m/z value of the n-th largest peak in the query spectrum is matched to m/z values of m largest peaks in the library spectra, as illustrated in Figure 2 where (n = 8, m = 15). The library spectra are ranked based on the number of matched peaks. Candidate spectra are determined based on certain criteria, such as the number of matched peaks and a threshold value R. Spectral similarities between the query spectrum and candidate spectra are then calculated, and the candidate spectra with the highest scores are returned to the user.

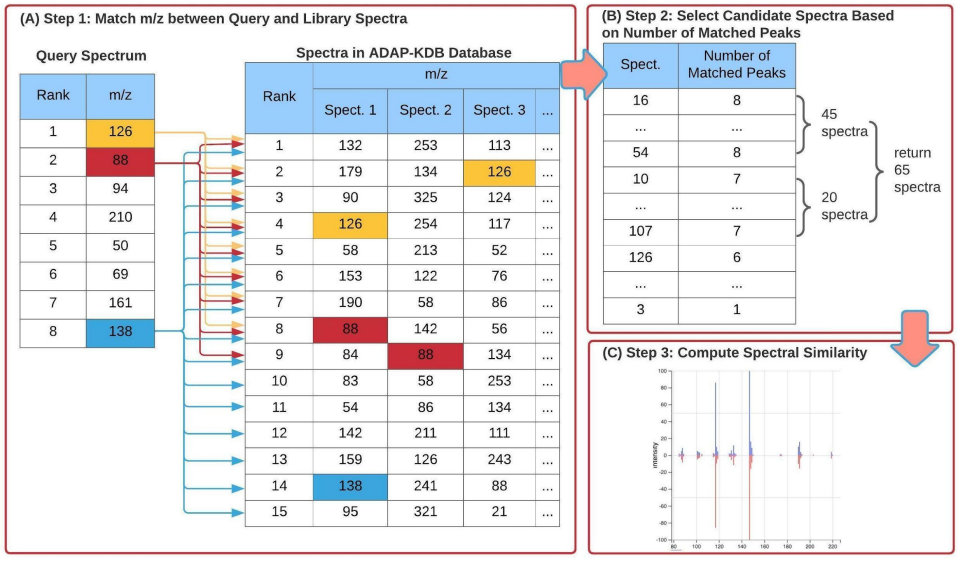


Figure 2: New ADAP-KDB spectral search algorithm with the prescreening search: (A) m/z values of 8 largest peaks in the query spectrum are matched to m/z values of 15 largest peaks in every library spectrum; (B) all library spectra are ranked based on the number of matched m/z values, and top 50+ candidate spectra are returned by the prescreening search; (C) the similarity score is calculated for each candidate spectrum.

Figure 3 shows the comparison of execution time between the new library search algorithm and the original algorithm. The result shows that performance of the library search with prescreening is about four-times faster for the low-mass-resolution spectra and very similar for the high-mass-resolution spectra. Moreover, the execution time stays

about the same for all pairs of parameters n and m, while pairs (n = 4, m = 7), (n = 6, m = 11), and (n = 8, m = 15) demonstrate slightly better inclusion rate than pairs (n = 4, m = 4), (n = 6, m = 6), and (n = 8, m = 8), respectively. Based on these results, comparing eight largest peaks in the query spectrum to 15 largest peaks in the library spectra seems to be the optimal approach for the preliminary search. Selecting threshold R is based on the tradeoff between the execution time and the inclusion rate. Based on the comparison results, value R = 50 seems to be optimal for keeping high inclusion rate and low execution time for both low-mass-resolution and high-mass-resolution spectra. Therefore, (n = 8, m = 15 and R = 50) were selected as optimal for the prescreening search of the new spectral search algorithm.

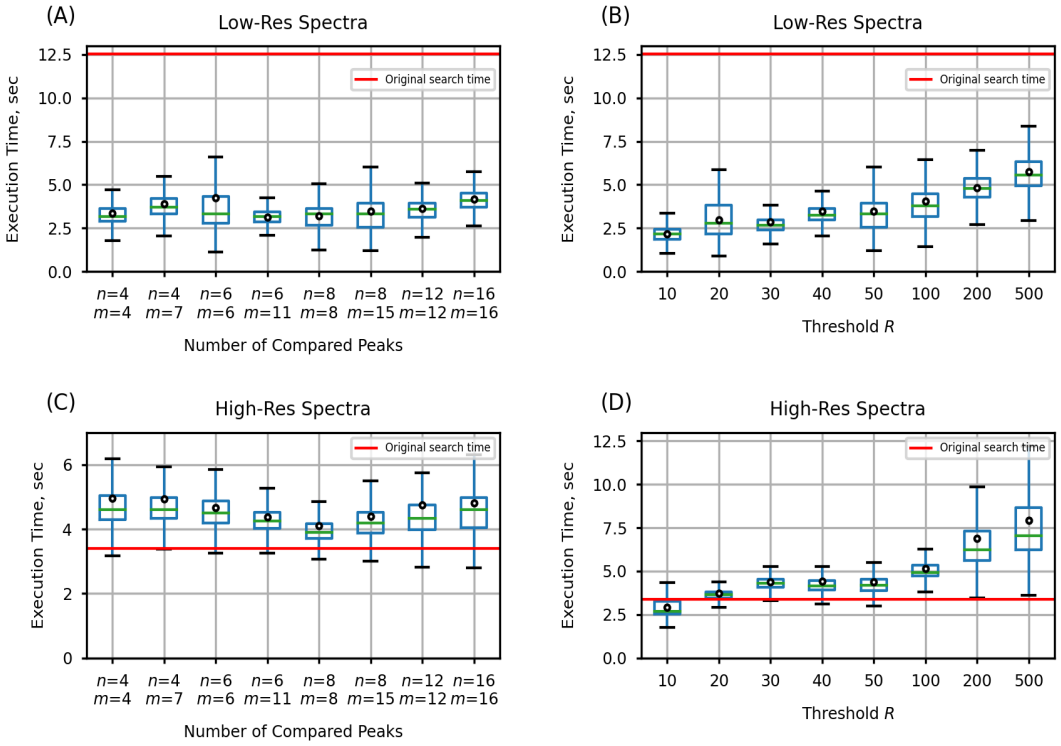


Figure 3: Execution times of the new search algorithm, estimated for prescreening parameters (n, m, R), where n is the number of query spectrum peaks participating in the prescreening search, m the number of library spectrum peaks participating in the prescreening search, and R determines the number of candidate spectra returned by the prescreening search: (A,B) search against low-resolution spectra; (C,D) search against high-resolution spectra.

2. Metabolic Signature Discovery

Figure 4 illustrates the workflow for discovering metabolic signatures. Raw data from public metabolomics studies undergo preprocessing using ADAPBIG, a software for processing untargeted mass spectrometry data [15]. This step includes formatting the data for compatibility with machine learning models. Subsequently, the data undergo cleaning, imputation, scaling, and target creation using a dummy matrix. The processed data is then fed into Partial Least Squares Discriminant Analysis (PLS-DA), a supervised learning algorithm used for identifying metabolite patterns that differentiate sample

groups. VIP scores are computed in PLS-DA to quantify the contribution of each metabolite to group separation, with a threshold typically set at 1. Metabolites surpassing this threshold are considered metabolic signatures. Finally, PLS-DA performance is validated through cross-validation, and the results are visualized.

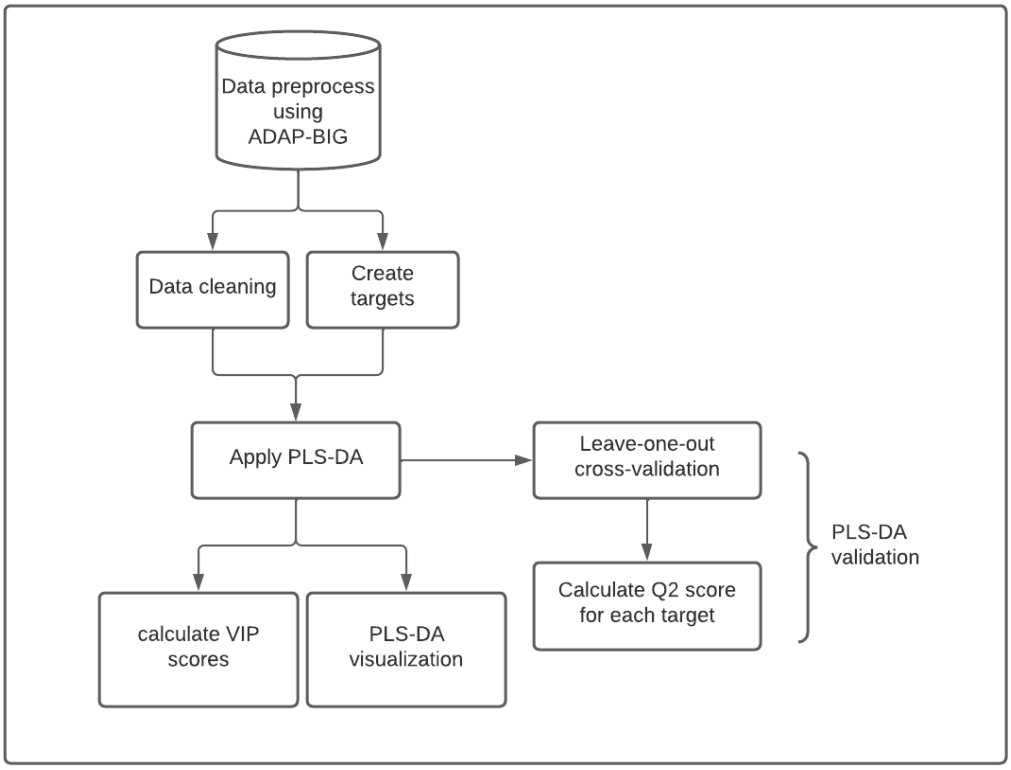


Figure 4: Metabolic Signatures Discover Pipeline

Study ST000058 [16] from the Metabolomics Workbench [11] investigates alterations in metabolite levels associated with methionine stress sensitivity in cancer using GC TOF MS analysis. The study comprises seven groups, each containing four samples. Group 1 serves as the control, receiving methionine treatment, while groups 2 to 7 undergo homocysteine treatment for varying durations from 2 hours to 48 hours. This design allows for the examination of metabolite responses under different stress conditions, offering valuable insights into cancer metabolism.

Figure 5 presents the PLS-DA visualization of results across six different pairs of groups, demonstrating clear separation between each group. The predictive capability of the model is assessed through Q2 scores after leave-one-out cross-validation, with the highest Q2 observed between the control group and the 48-hour treatment group (0.76), and the lowest between the control group and the 8-hour treatment group (0.36).

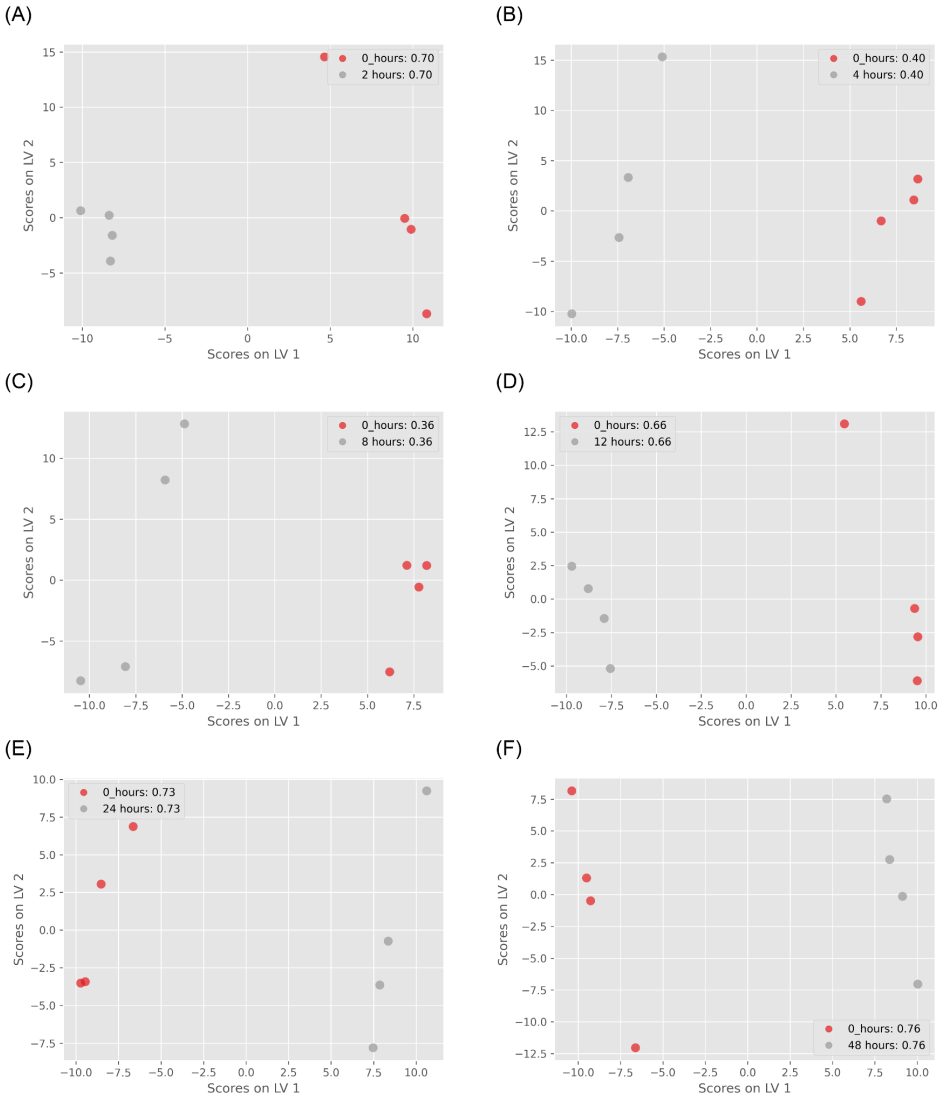


Figure 5: ST000085 PLS-DA results: (A) control group vs. 2 hours treatment group (Q^2=0.7); (B) control group vs. 4 hours treatment group (Q^2=0.4); (C) control group vs. 8 hours treatment group (Q^2=0.36); (D) control group vs. 12 hours treatment group (Q^2=0.66); (E) control group vs. 24 hours treatment group (Q^2=0.73); (F) control group vs. 48 hours treatment group (Q^2=0.76);

To identify metabolic signatures, I conducted PLS-DA analysis on control groups and 48-hour treatment groups, utilizing VIP cutoffs of 1.3 and 1.5 to select candidate metabolites. Additionally, I performed ANOVA tests on the same group pair to compare multivariate and univariate algorithms, correcting the findings with 5% and 10% FDR thresholds. Figure 6 presents the comparison between PLS-DA and ANOVA results. A total of 26 metabolites were identified as candidates by both methods, with 19 uniquely identified by PLS-DA (using a VIP score cutoff of 1.3), and all metabolites from ANOVA were found in PLS-DA results.

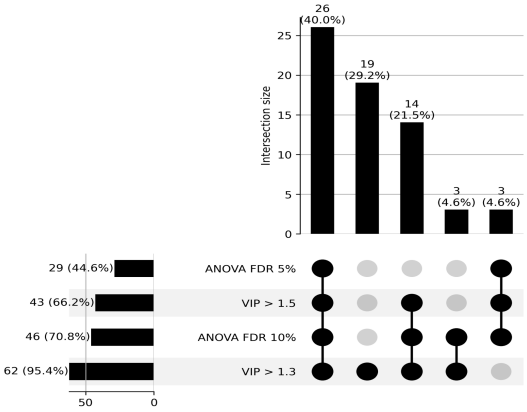


Figure 6: Upset plot between PLS-DA and ANOVA test. VIP cutoff is 1.3 and 1.5 in PLSDA; FDR threshold is 5% and 10% in ANOVA test.

In refining the ADAP-KDB algorithm for metabolomics data analysis, I significantly enhanced the efficiency of processing mass spectrometry datasets, enabling rapid identification of metabolic signatures. This experience honed my skills in algorithm optimization, data preprocessing, and machine learning, equipping me with the tools necessary for tackling complex datasets in genetic and disease prediction research.

Applying these skills, I am poised to contribute to *Cx. tarsalis* population genetics by efficiently analyzing genetic variations and to West Nile Virus prediction by developing models that integrate multifaceted data. The foundation laid in computational data analysis and pattern recognition prepares me for advancing these projects, ultimately aiming to inform public health strategies.

V. Climate Adaptation and Genetic Differentiation in the Mosquito Species *Culex tarsalis*

Aim 1: To identify and characterize genetic variations in *Cx. tarsalis* populations that may confer environmental adaptations.

To advance our understanding of population structure and identify alleles linked to local adaptation in *Cx. tarsalis*, we first assembled and annotated a *de novo* reference genome and generated Restriction-Site Associated DNA sequencing (RAD-seq) data for over 300 individuals from 28 diverse geographic locations. We analyzed these RAD-seq markers through a comprehensive landscape genetics framework to explore how various environmental variables influence population differentiation and to identify alleles associated with adaptation to these conditions. By leveraging a broad spectrum of environmental variables, we assessed the adaptive responses of populations to their local environments, enabling the identification of critical genetic-environment associations. This approach reveals how specific climate variables and genetic variants underpin local adaptation strategies across 28 representatives *Cx. tarsalis* collection sites. The findings

enrich our understanding of the complex interactions between genetics and environment, providing crucial insights into the ecological dynamics of this mosquito species.

ADMIXTURE [17] and PCA analysis were conducted to investigate the population structure of *Cx. tarsalis.* The ADMIXTURE analysis indicated a strong signature of population structure among the collected samples, with the optimal number of population assignments occurring at K=4. The genetic clusters corresponded to four different broad geographic regions: (1) California/the West Coast, (2) the Southwest, (3) the Northwest, and (4) the Midwest (Figure 7).

The PCA results confirmed this pattern (Figure 7C), while also showing evidence of some sub-structure among the West Coast and Northwest populations.

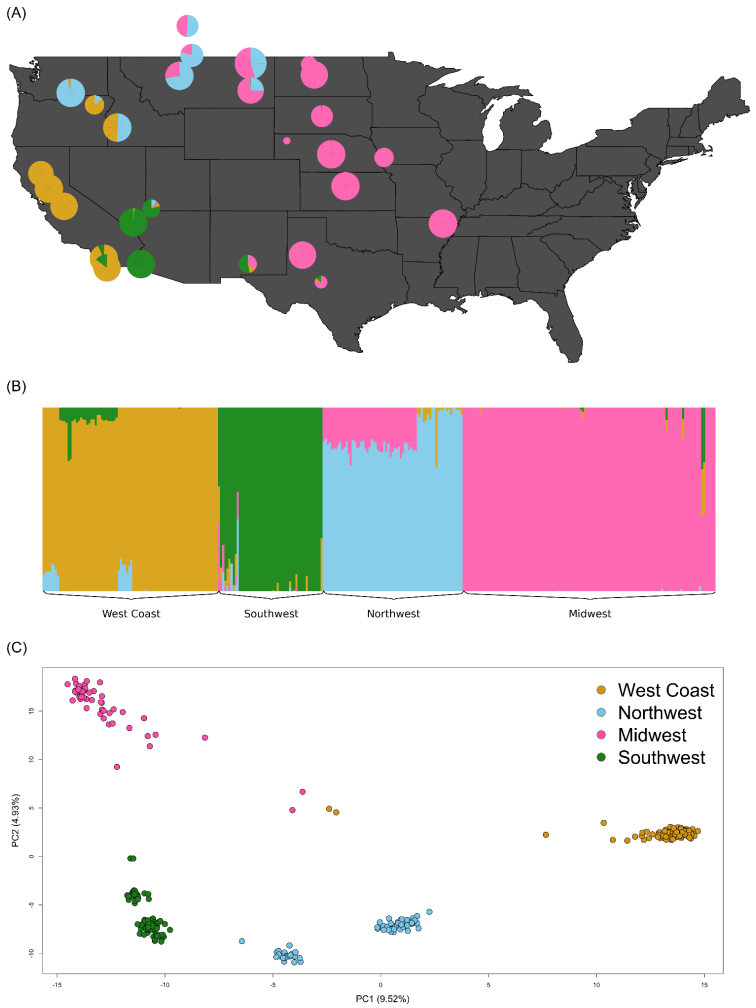


Figure 7: Population Structure of *Cx. tarsalis*: (A) Floating pie charts of the admixture proportions in *Cx. tarsalis* populations sampled across the Western and Midwestern U.S and parts of Canada. Pie chart sizes are proportional to the sample size at each collection site. (B) ADMIXTURE results for K=4. Labels along the x-axis indicate sampling locations and colors correspond to the admixture proportion for each of the 4 clusters. (C) PCA results for the top 2 principal components, with points colored by the 4 geographic regions identified by ADMIXTURE.

To analyze isolation by distance (IBD) [18] and isolation by environment (IBE) [19,20] patterns within the 28 mosquito populations, genetic distances, derived from Weir & Cockerham F_{ST} estimations [21] based on SNP data, were compared with geographic and environmental distances computed based on latitude and longitude coordinates. A mixed model was also utilized to analyze the relationships between genetic distance, geographic distance, and environmental distance.

From the results, both geographic distance and environmental distance showed a statistically significant relationship with genetic distance (Figure 8). The correlation with geographic distance (IBD) was much stronger, with a Mantel test statistic of 0.5661 (p=0.001), while the correlation with environmental distance (IBE) was weak but still significant, with a Mantel test statistic of 0.1866 (p=0.0115). The mixed model results indicate that considering both geographic and environmental factors simultaneously provides a more comprehensive explanation of genetic distance variations than either factor alone (Table 1). The moderate to strong correlation observed in IBD and the weaker correlation in IBE are echoed in the mixed model, where the combined effect of both distances is most significant in explaining genetic differences.

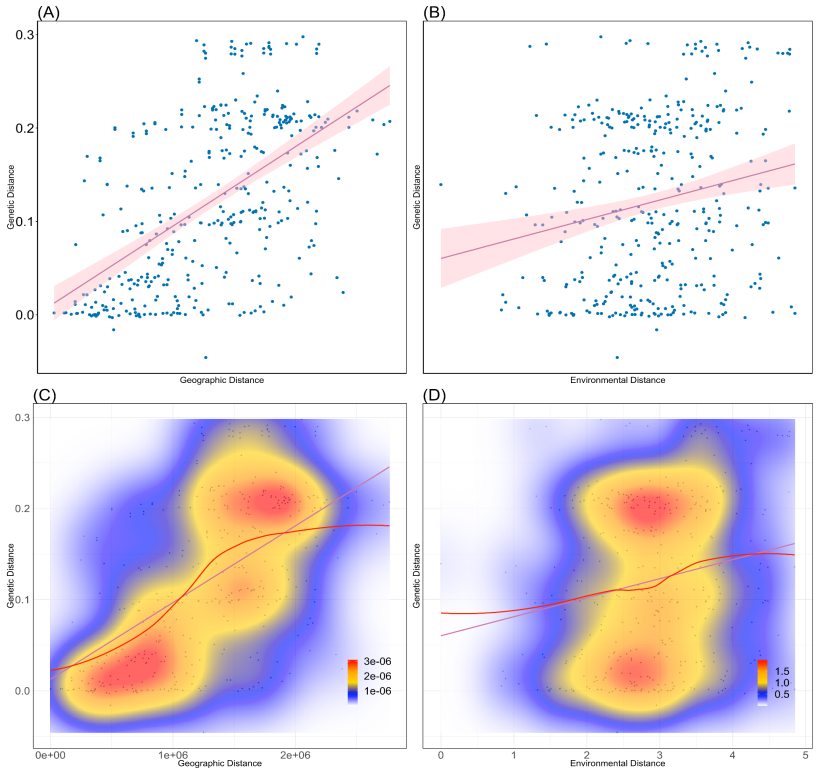


Figure 8: Isolation-by-Distance and Isolation-by-Environment. (A) Pairwise geographic distance versus genetic distance (F_{ST}) with best fit linear regression model (red line: $\boldsymbol{y = 0.012 + 8.4 \times 10^{-8}x, R^2 = 0.3}$). (B) Pairwise environmental distance versus genetic distance with best-fit linear regression mode (red line: $\boldsymbol{y = 0.06 + 0.021x, R^2 = 0.04}$). (C) Kernel density plot with best fit spline for geographic distance versus genetic distance. Areas of high, intermediate, and low density are represented by red, yellow, and blue colors, respectively. (D) Kernel density plot with best-fit spline for environmental distance versus genetic distance.

Table 1: Mixed model results for IBD and IBE

Models	AIC	BIC	k	AICc	AICcmin	BICew
Full	689.701	709.375	5	689.862	0.959	0.771
Distance	696.061	711.801	4	696.169	0.041	0.229
Environment	1001.974	1017.713	4	1002.081	0.000	0.000

After performing PCA analysis on the environmental data, the first PC is used as the predictors in the Latent Factor Mixed Model (LFMM) 92 candidate SNPs were identified as being significantly associated with the environment (FDR < 0.1). And Redundancy analysis (RDA) was performed to compare to the LFMM results. The RDA model reveals that environmental variables explain 10.42% of the genetic variance (Constrained), leaving 89.58% of the variance unexplained and presumably accounted for by geographical distance as supported by IBD and IBE tests, the environmental contribution to genetic variance nonetheless provides valuable insights into adaptive processes. From the RDA results (Figure 9), the clear regional clustering depicted within this biplot aligns with the four geographic regions (Midwest, Northwest, Southwest, and West Coast), underscoring a significant regional influence on the portion of genetic variation shaped by environmental factors. The alignment of populations with vectors for the annual average temperature at 2 meters (avg_t2m), the annual average total evaporation (avg_e) and annual average leaf area index for low vegetation (avg_lai_lv) signifies the role of temperature, humidity and vegetation density in this context.

For detecting SNPs under selection, PCAdapt and Bayescan are used to find the outliers. The BayeScan outlier analysis revealed 1,836 loci potentially under diversifying selection, 10,166 potentially under balancing selection, and 5,237 neutral loci. Of the 1,836 loci putatively undergoing selection for local adaptation, 1501 of these are found within 824 genes. Of these 824 genes, 24 also overlapped with genes that were identified by both the LFMM and RDA as significantly associated with environmental variables. We also independently used PCAdapt to search for outlier loci with notable allele frequency differences across populations that are potentially the result of natural selection and uncovered 173 SNPs in the top 1% of extreme p-values. Looking across all 4 of our analyses examining significant environmental associations and evidence of natural selection, we found 20 common genes that contain the candidate SNPs that were significant in every instance (Figure 10). Among the 20 common genes, Ct.00g04290 encodes a PERIOD CIRCADIAN PROTEIN stands out for the essential of regulating circadian rhythms, aligning life cycle events with environmental cues [22–24].

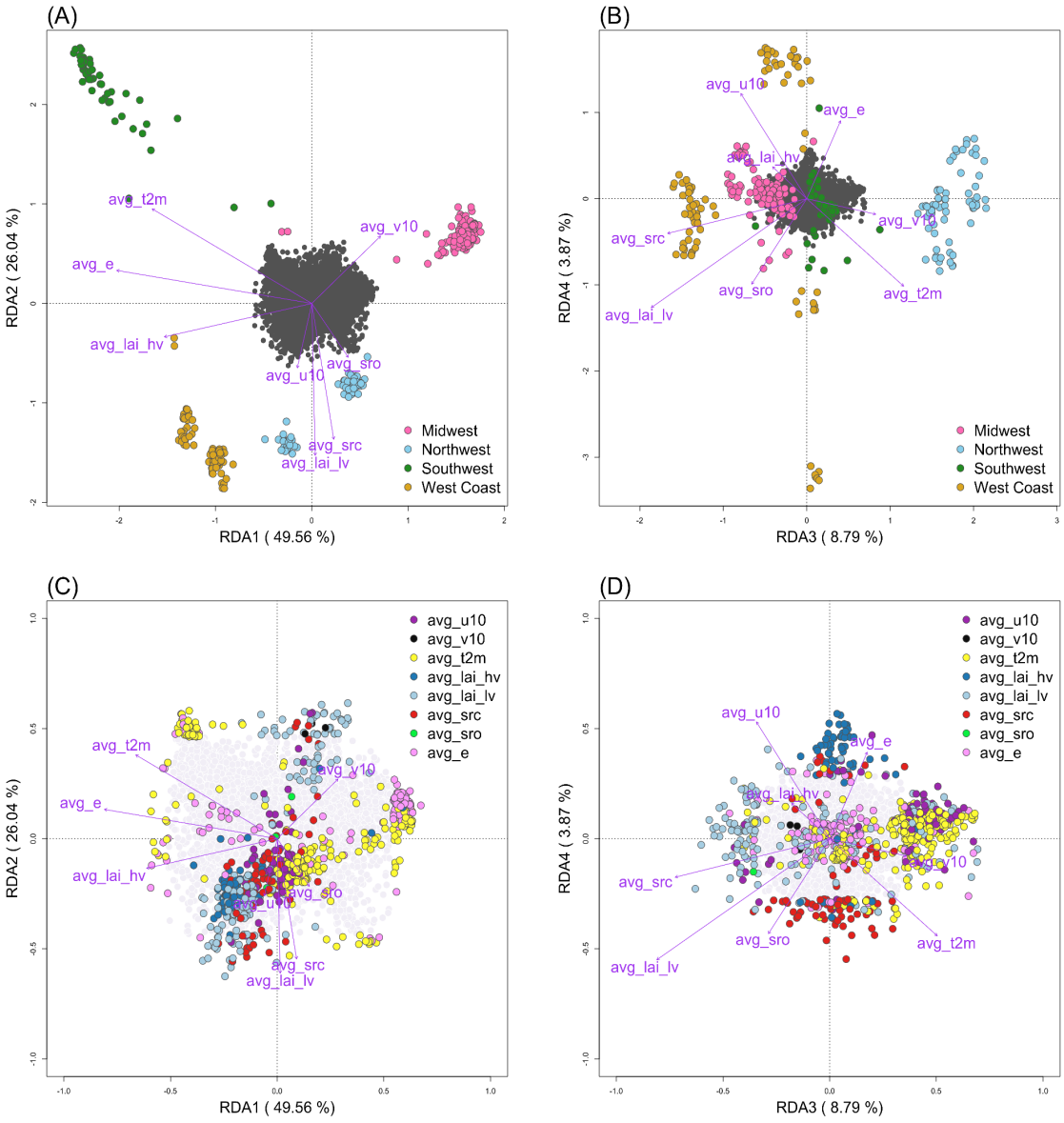


Figure 9. Environmental Correlates of Genetic Variation in *Cx. tarsalis* Across Diverse North American Regions. Panels A and B display the relationship between environmental factors and the distribution of *Cx. tarsalis*, using RDA to illustrate how regional differences affect genetic variation. In these panels, the position of each circle (representing an individual mosquito) and color (indicating regional groupings from ADMIXTURE results) reflects their association with environmental variables, shown as purple vectors. The first plot (A) focuses on RDA1 and RDA2, the primary axes explaining the most variance, while the second (B) explores more subtle influences in RDA3 and RDA4. Panels C and D shift focus to SNPs within the genetic data, colored to signify their strongest environmental association. These latter plots further detail the genetic-environment relationship, with the SNPs' distribution providing insight into the adaptive landscape of these populations.

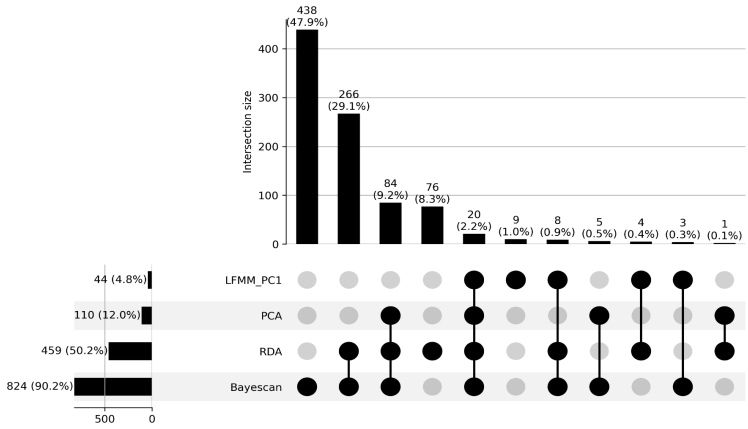


Figure 10: Upset Plot of within Genes Annotated from Candidate SNPs Identified for Local Adaptation and Environmental Association Across LFMM, RDA, Pcadapt, and BayeScan Analysis in *Cx. tarsalis*

Our analysis uncovered a significant link between environmental variables and genetic variation, particularly showing that temperature, evaporation rates, and vegetation density are critical environmental factors with strong associations to genetic differentiation within *Cx. tarsalis* populations. Among these, the identification of 53 SNPs with strong evidence of diversifying selection suggests they are likely involved in local adaptation processes. These SNPs, linked to crucial biological functions such as circadian rhythms, reproductive success, feeding habits and fat metabolism, and lay the groundwork for a detailed exploration of the genetic mechanisms driving adaptation in diverse environmental conditions.

VI. West Nile Virus Prediction

Aim 2: To develop and validate a national-scale predictive model for annual West Nile virus transmission, integrating environmental and demographic data. This model aims to enhance early warning systems for public health preparedness at a country-wide level.

Aim 3: To create and assess a regional predictive model for monthly West Nile virus cases, focusing on county-level data. This model will incorporate detailed environmental and demographic information to support more localized public health interventions and vector control strategies.

The prediction of West Nile Virus (WNV) transmission is a critical component in public health strategies for disease prevention and control. Given its impact on human health and the complexity of its transmission dynamics, there is a pressing need for accurate predictive models. These models can inform timely interventions and guide resource allocation for vector control. My research aims to develop a comprehensive predictive model integrating environmental, demographic, and possibly genetic data, to enhance our understanding and forecasting of WNV outbreaks.

1. Data source

1.1. Clinical Disease Surveillance Data

The disease data used in this study are from a few different sources. The first one is from ArboNET, which is a national arboviral surveillance system managed by CDC and state health departments. ArboNET collects data on arboviral infections among people, veterinary animals, mosquitoes, dead birds, and sentinel animals [25]. This dataset contains yearly counts of WNV patients for each county from 1999 to 2023. This dataset also contains a monthly count for horses, dead birds (separated by corvid and non-corvid) and mosquitoes tested with WNV for each county from 2000 to 2021.

The second dataset is from California Department of Public Health weekly reports from 2003 to 2023 [26]. This dataset contains human cases weekly reports with clear mention either is neuro-invasive, west nile fever or from blood donors in 2004, 2005, 2006, 2010, 2011 and 2012 from each county in California. Other years besides these 6 years are the same granularity as in the CDC dataset which is the yearly sum for each county and without mention the disease type. This dataset also contains weekly counts of the non-human counts the same as CDC dataset.

The third dataset is from the Illinois Department of Public Health, which only contains the weekly report for dead birds, mosquitoes and horses.

1.2. Climate Data

Same as in the *Culex tarsalis* project environmental data section, utilizing ERA5-Land monthly averaged data (1950-present) from the Copernicus Climate Data Store [27] to extract annual average and monthly average climate data.

1.3. Land Use Data

The land use information is extracted from the global 1-km Consensus Land Cover dataset [28]. This dataset comprises twelve data layers, each representing the prevalence of a distinct land-cover class, with values ranging from 0 to 100 to indicate the percentage of prevalence for each class. The spatial coverage of the dataset extends from 90ºN to 56ºS latitude and from 180ºW to 180ºE longitude, providing a global perspective at a resolution of approximately 1 km per pixel at the equator [29].

1.4. Birds Data

Sourced from eBird [30], this comprehensive dataset includes observations of common North American bird species. Data is aggregated by species, county, month, and year. Additionally, avian phylodiversity data based on 2019 observations is included.

1.5. Demographic Data

Compiled from various sources including the U.S. Census Bureau (population, poverty estimates, and land area data) and Federal Communication Commission (county data). The focus is on data from 2000 to 2021, with manual adjustments for changes such as county name alterations. Climate data representation for each county is based on county seat coordinates.

2. Methodology

2.1. Data Preprocessing

2.1.1. Integration of multi-sources data

Harmonizing diverse datasets for machine learning is challenging, especially with the inconsistencies in county identifiers. I use FIPS codes as standard identifiers, yet not all sources include these, and county names often differ between datasets. This requires a thorough manual examination to ensure accurate FIPS code allocation, a task complicated by occasional county renames and FIPS updates. Meticulous effort is essential for successful data integration.

2.1.2. Climate data processing

For each data row representing a specific geographic location and year within our study, I employed the xarray library [31] in Python to process the ERA5-Land monthly averaged climatic data, which covers the period from 1950 to the present. From this dataset, I extracted the maximum, minimum, and mean values for each climate variable on an annual basis. This approach ensures that I accurately represent the climatic variability and trends for each location over the course of a year, providing a robust foundation for analyzing the influence of climate on West Nile Virus transmission dynamics.

For the monthly report dataset, I extracted the monthly average values for each climate variable within the monthly on the corresponding county. To better understand the past climate variables effect. The average monthly data of the previous one month, two months is also extracted for each location and climate variable.

2.1.3. Integration of land use data

Additionally, I will incorporate land use variables from the global 1-km Consensus Land Cover dataset [29], employing the OpenCV library [32] and xarray library [31] in Python to extract consensus land use information pertinent to the county seat coordinates.

2.1.4. Standardization of data granularity

2.1.4.1. Annual Based Dataset

In the context of the data processing for my study, the human West Nile Virus (WNV) data is aggregated on an annual basis at the county level, whereas the non-human WNV data is reported with greater frequency, on a weekly basis, also by county. To harmonize the granularity for comparative and predictive analysis, I standardized all data to an annual county-level scale. This alignment allows for a consistent approach in analyzing the spread and impact of WNV across human and non-human populations, ensuring that the annual cycle of the disease's prevalence is accurately captured, and that the data is comparable across the different categories of subjects affected.

In preparing the dataset for predicting WNV cases, a series of data manipulation and imputation tasks are conducted to ensure data integrity and usability for predictive modeling. Initially, I filtered human infection records from 2004 to 2023 and extracted essential county information, removing duplicates and entries with missing FIPS or State data. Subsequently, I constructed a comprehensive data frame reflecting each unique FIPS code's yearly data, augmented with corresponding state, county, and geolocation details. To address the challenge of missing data, I employed a two-fold imputation strategy: setting case counts to zero in the absence of recorded activity and replacing missing human case reports with the average number of cases over the 20-year span for each county.

2.1.4.2. Monthly Based Dataset

For the monthly human reports in California, the standardization for monthly based Dataset is like annual dataset. First, I constructed a comprehensive data frame reflecting each unique FIPS and month in California in 2004, 2005, 2006, 2010, 2011 and 2012. First, fill in the data frame with already known data. For any missing data on human case report, if there is no report on any non-human cases in the same county and month, a value of 0 will be given. Otherwise, fill the missing value with the annual average number of human cases in the corresponding year and county for the missing value. One exception here is for the spatial generalized linear mixed models (sglmm) applied later, with assumed negative binomial distribution, I need to round up the counts to integer numbers.

2.2. Correlation Analysis using the Mantel Test

Before delving into machine learning, understanding the relationships between environmental variables and the target variable is essential. For this purpose, I'll utilize Kendall's tau correlation, a robust non-parametric test developed by Maurice Kendall in 1938 [33]. Unlike parametric tests, Kendall's tau doesn't assume a specific data distribution, making it ideal for analyzing ecological and environmental datasets.

2.3. Machine Learning / Statistical Models

The study employs various models to predict West Nile virus occurrences, considering the complex interplay between environmental, clinical, and land use factors. Each model

is selected based on its strengths in capturing non-linear relationships, temporal patterns and incorporating spatial dependencies. Criteria for model selection include their ability to handle high-dimensional data, robustness against missing information, and capacity to capture spatial and temporal dynamics intrinsic to disease spread. This selection process, informed by a literature review [8–10,34–36], underscores the value of utilizing a combination of traditional machine learning techniques and advanced models to deepen our understanding of disease determinants and improve predictive accuracy.

Table 2 summarizes key aspects of each model used in this study, including their descriptions, rationales for selection, as well as pros and cons. Understanding the strengths and limitations of each model is crucial for selecting the most appropriate approach based on the specific requirements and characteristics of the dataset and prediction task.

Table 2: Comparative Overview of Predictive Models for West Nile Virus Prediction

Model	Description	Rationale for Selection	Pros	Cons
Random Forest	Ensemble learning, multiple decision trees	Robustness, handles non-linear data and missing data	Robust, handles missing values	Complex, reduced interpretability
SVM	Finds hyperplane separating classes in feature space	High-dimensional effectiveness	Performs well in high-dimensional spaces	Complex parameter tuning
Neural Network	Computing system for pattern recognition	Learns complex relationships	Exceptional pattern recognition	Proneness to overfitting
Ensemble Model	Combines Random Forest with Autoregression	Leverages both strengths	Integrates spatial and temporal analysis	Constraints on temporal data structure
SGLMMs	Extend GLMs with spatially correlated random effects	Models' spatial correlation	Accurate spatial dependencies modeling	Complex interpretation

2.4. Evaluation metrics

To comprehensively assess our model's predictive capability for West Nile virus cases, I apply a few evaluation metrics on the testing dataset, including Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Squared Logarithmic Error (MSLE), and predictive coefficient of determination (Q^2). Each metric offers a unique lens through which to view model performance.

Besides the evaluation metrics I mentioned above, I also add a baseline model to compare with each model which is a model that will always predict the result with the mean of the disease count.

3. Preliminary Data

Figure 9 features Kendall's tau correlation heatmaps, analyzing the national CDC dataset on annual human West Nile Virus (WNV) cases in relation to environmental and demographic factors. The heatmaps reveal significant correlations with variables such as minimum surface net solar radiation (min_ssr), as well as land use classifications—specifically, Cultivated/Managed and Urban/Built zones, and avian phylogenetic diversity (PD). However, most of the aggregated environmental variables exhibit minimal correlation with annual WNV case counts. This suggests that macro-level environmental metrics may fail to adequately reflect the intricate dynamics that drive disease transmission.

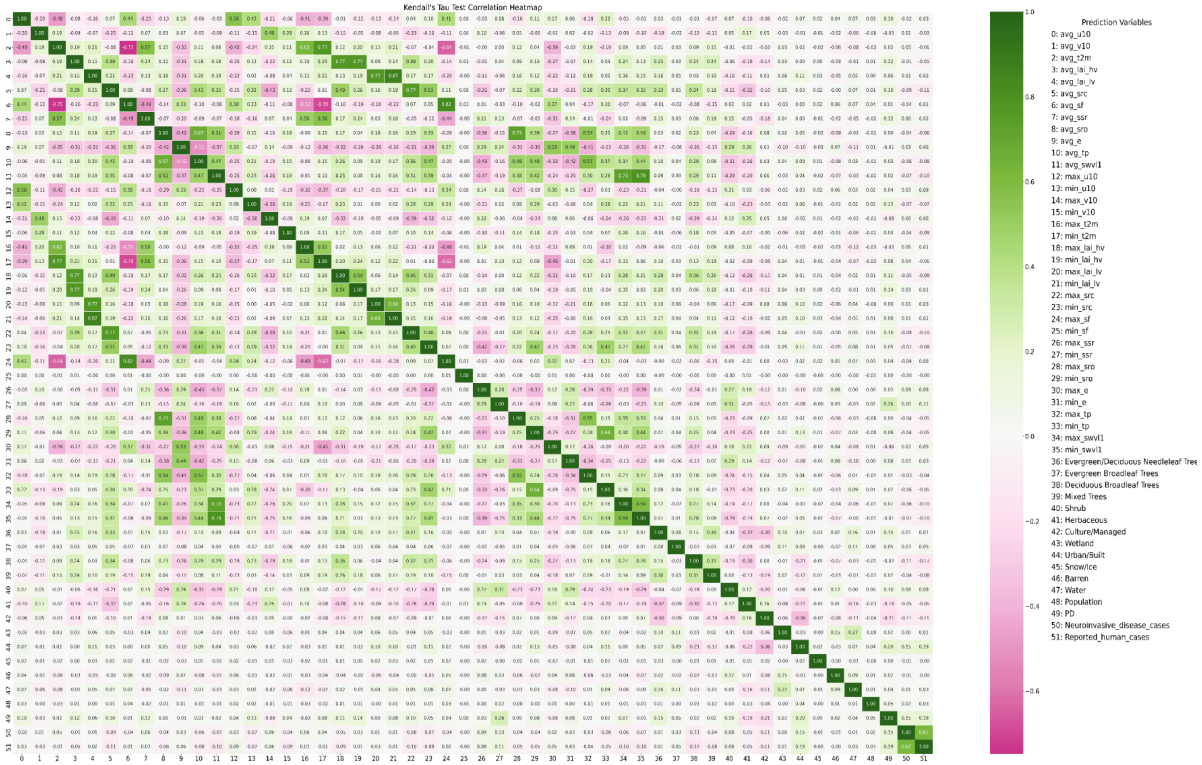


Figure 9: Kendall's Tau Correlation Heatmap for National CDC Dataset: Annual Human Cases (Both WNND and total WNV disease cases) vs. Environmental Variables

The initial approach aimed to predict nationwide WNND cases utilizing several models. The outcomes, as illustrated in Table 3, suggest that most models perform comparably, implying that predictive capabilities are generally modest. Among these, the SVM model emerges as the most effective, with an MSE of 80.5 and a Q2 score of 0.09, indicating that while performance is limited, SVM may offer marginal predictive advantages.

Table 3: Performance of Different Predict Models on National Model

Model	MSE	Q^2
Linear Regression	82.82	0.07
Random Forest *	84.28	0.05
SVM	80.5	0.09
Neural Network	86.00	0.04
SGLMM **	92.65	0.015

* The MSE and Q^2 for RF is average results of 100 runs
** Simplify the variables only use avg_src + min_ssr + Urban_Built + PD

While tuning the prediction models for West Nile Virus (WNV) incidence, several significant data-related challenges have emerged:

- **Inconsistency in Data Collection**: The collection of WNV case reports varies considerably across states, presenting substantial hurdles for modeling efforts at a national level. This inconsistency can lead to skewed data representations, which may compromise the accuracy and reliability of predictive analytics.

- **Significant Class Imbalance**: The imputation of missing values exacerbates the existing imbalance between reports of zero cases and those documenting one or more instances of WNV. Such disproportionality poses a risk of biasing the model towards predicting the majority class (0 disease case) and necessitates the adoption of specialized techniques to achieve a balanced representation of case occurrences.

- **Data Granularity**: An additional complexity encountered in predicting annual WNV cases is the granularity of environmental data. When forecasting yearly disease counts, environmental variables must be aggregated—typically as annual averages, maximums, or minimums. This approach, while necessary for aligning temporal scales, may obscure finer temporal correlations between environmental conditions and WNV transmission patterns. For instance, the use of yearly averages or extremes of temperature does not capture the subtleties of seasonal fluctuations that could be critical in understanding and predicting WNV outbreaks. The nuances of how and when these environmental factors influence mosquito populations and virus transmission may be lost, potentially diminishing the predictive power of our models.

Each of these issues requires tailored methodological adjustments to ensure the robustness of the predictive models. The following sections detail the specific approaches undertaken to mitigate these challenges and discuss the results yielded by these refined modeling strategies.

3.1. Inconsistency in Data Collection

Given the variability in data collection and reporting of West Nile Virus (WNV) cases among states, we adopted a state-centric modeling approach. Figure 10 presents the outcomes of employing Support Vector Machine (SVM) algorithms to predict West Nile Neuroinvasive Disease (WNND) occurrences for each state. The displayed models, limited to those achieving a positive Q2 score, indicate a disparity in performance across states. For instance, California, Illinois, Michigan, and New York demonstrate superior model efficacy compared to those of Pennsylvania, North Dakota, and Washington, suggesting that the robustness of local WNV surveillance systems may significantly influence predictive accuracy.

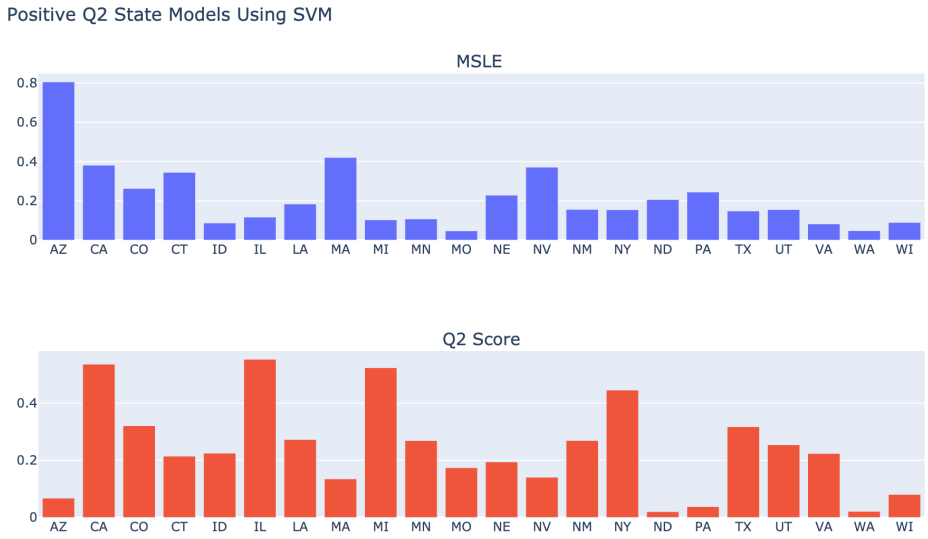


Figure 10: Comparative Analysis of SVM Models for WNND Across States

As part of an integrated strategy, I shifted my analytical focus from the quantitative assessment of West Nile Virus (WNV) incidences to a binary determination of WNV presence within individual counties. This pivot required reconceptualizing the issue from a regression framework to a classification paradigm. The newly adopted classification approach was applied to data from New York, California, and Illinois, which were specifically chosen due to their robust WNV case reporting systems that include both human and non-human instances. The efficacy of the SVM-based classification models is detailed in Table 4, where they are evaluated against a national model. These comparisons shed light on the models' capacity to accurately predict binary outcomes indicative of WNV presence.

To enhance model performance, I explored a range of probability thresholds from 0 to 1 to distinguish between zero cases (0) and non-zero cases (non-0). The optimal threshold was selected based on its ability to maximize the F1 score, a harmonic mean of precision and recall that balances the trade-off between false positives and false negatives. Among the state models, California's demonstrated the most robust performance. Conversely, the national model's high accuracy, coupled with a lower F1 score, highlights the persistent challenge posed by class imbalance within the dataset.

Table 4: SVM Predictive Outcomes for WNV Surveillance in State models (New York, California, Illinois) and National Model

State	Threshold	F1	Accuracy	AUC-ROC
New York	0.58	0.694	0.846	0.82
California	0.86	0.862	0.835	0.87
Illinois	0	0.552	0.381	0.60
National	0.3	0.496	0.748	0.74

3.2. Significant Class Imbalance

Figure 11 displays the distribution of West Nile Virus Neuroinvasive Disease (WNND) counts nationally and by state. The data reveal a predominant number of zero-case instances compared to non-zero cases. Such an imbalance skews prediction models towards forecasting zero cases, resulting in overly conservative estimates. To counteract this, a subsampling technique was applied to down sample the majority class (zero cases) to align with the minority class (non-zero cases), aiming for a balanced dataset that would enable more accurate predictions.

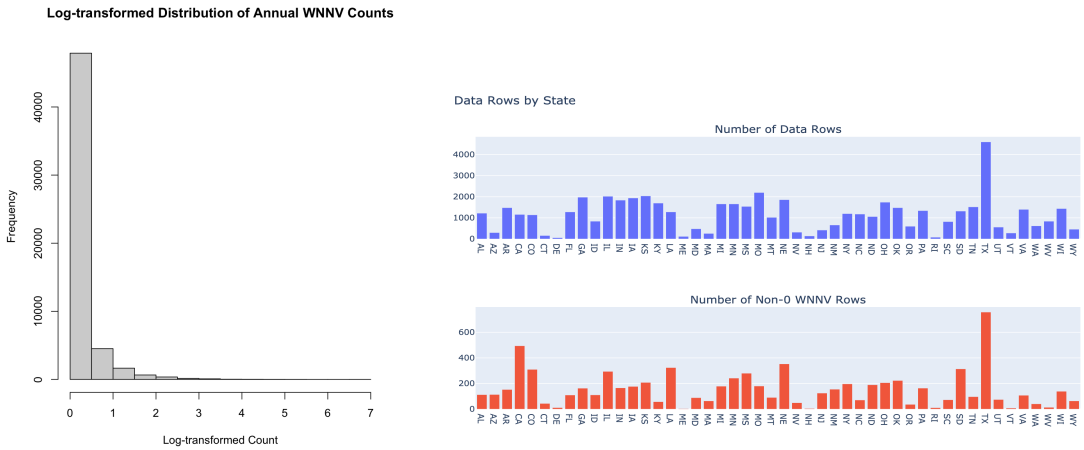


Figure 11 WNND count distribution across the nation (left) and between each state (right). On the right plot, the blue represents the total number of rows in each state, where the red represents the total number of rows with non-0 WNND in each state.

Table 5 delineates the performance of three different models in predicting national annual WNND counts. When compared with the results in Table 3, we observe that the Q^2 scores for each model remain relatively consistent. However, there is a notable increase in Mean Squared Error (MSE), suggesting that the models are now less conservative and more likely to predict higher numbers of cases. Similarly, Figure 12 showcases the performance of state specific SVM models in predicting annual WNND counts. The models exhibit Q^2 scores comparable to those of the original models depicted in Figure 10, yet with increased Mean Squared Logarithmic Error (MSLE), further indicating a tendency towards predicting higher case counts.

Table 5: Performance of Different Predict Models on National Model Using Subsampling

Model	MSE	Q^2
Linear Regression	314.90	0.08
Random Forest *	325.12	0.05
SVM	309	0.09

* The MSE and Q^2 for RF is average results of 100 runs

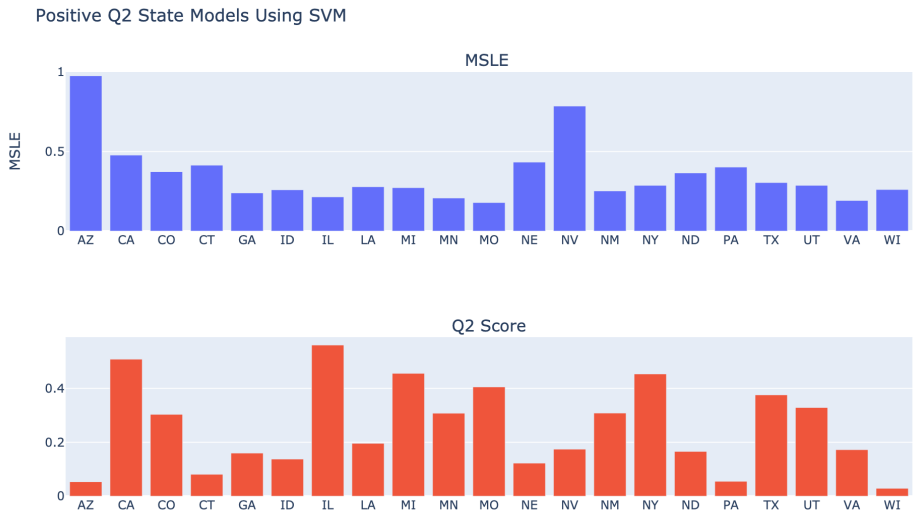


Figure 12: Comparative Analysis of SVM Models for WNND Across States Using Subsampling

3.3. Data Granularity

To examine the impact of environmental variable granularity on model accuracy, we utilized a newly compiled dataset containing six years of weekly West Nile Neuroinvasive Disease (WNND) reports from all counties in California. By aggregating this weekly data into monthly reports, we constructed a more detailed dataset. The first five years served as the training set, with the final year reserved for testing. Figure 13 illustrates the enhanced correlation between monthly county-level WNND counts and a range of variables—including environmental factors, *Cx. Tarsalis* SNP data, and demographic information. This refined granularity reveals a stronger and more discernible relationship between environmental conditions and disease incidence.

Figure 14 presents the predictive results from employing four different modeling approaches: linear regression, random forest, ensemble modeling, and SVM, aimed at forecasting monthly WNND counts for California counties. The random forest model outperforms its counterparts, achieving the lowest Mean Squared Error (MSE) and the highest average Q2 score, signifying its superior predictive capability within the context of enhanced data resolution

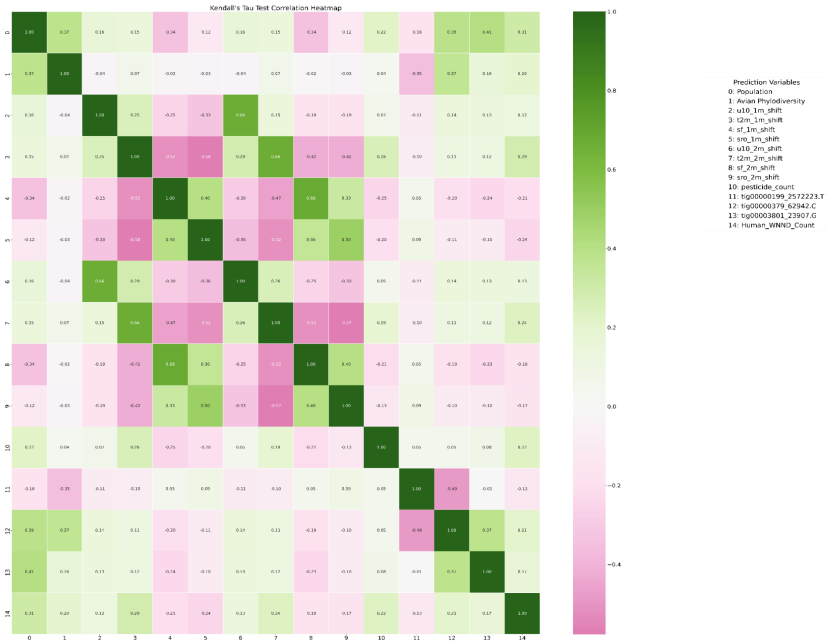


Figure 13: Kendall's Tau Correlation Heatmap for California Public Health Dataset: Monthly Human Neuroinvasive WNV Cases vs. Environmental and Demographic Factors

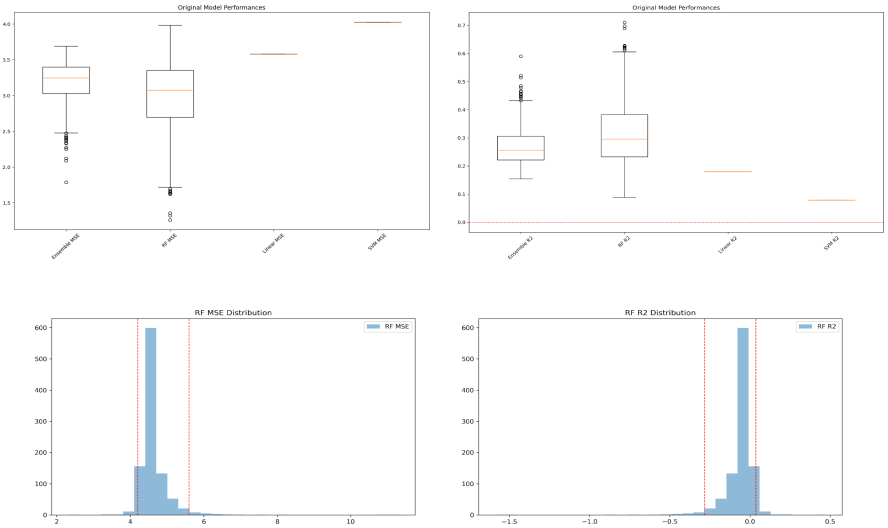
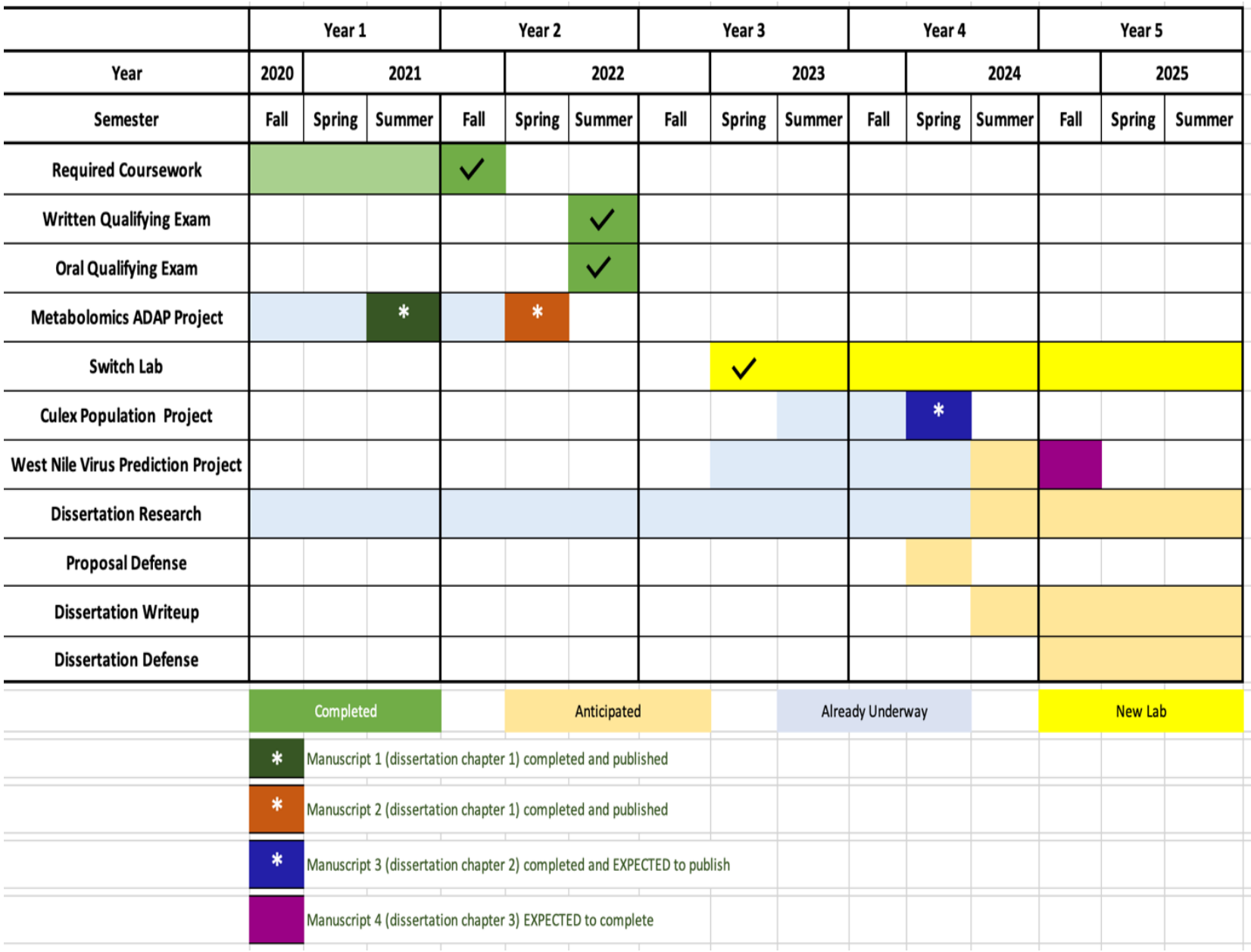


Figure14: Model Performance of Predicting Monthly County Level Data using California Data. (A) and (B) are MSE and Q^2 of running Ensemble, RF, Linear Regression and SVM 1000 times. (C) and (D) are MSE and Q^2 Permutation Test on RF with 1000 iterations, the dotted red line is 0.025 and 0.975 Percentile of Q^2.

The preliminary data analysis for West Nile Virus (WNV) modeling revealed challenges in inconsistent data collection and significant class imbalance, which skewed predictive models towards zero-case outcomes. To address these issues, I applied subsampling techniques and shifted our approach to binary classification, particularly for states with robust reporting systems like New York, California, and Illinois. This change in strategy yielded models with improved performance, as binary prediction proved to be more effective than predicting case numbers. Enhanced data granularity can also refine our models, as seen in detailed monthly analyses from California, leading to stronger correlations and better model accuracy as demonstrated by the random forest's performance in both MSE and average Q2 scores.

VII. Project Timeline



VIII. References

1. Sejvar JJ. Clinical manifestations and outcomes of West Nile virus infection. Viruses. 2014;6: 606–623.
2. Ozdenerol E, Taff GN, Akkus C. Exploring the spatio-temporal dynamics of reservoir hosts, vectors, and human hosts of West Nile virus: a review of the recent literature. Int J Environ Res Public Health. 2013;10: 5399–5432.
3. Bauer AM, Guralnick RP, Whitehead SA, Barve N, Allen JM, Campbell LP. Land use predicts proportion of West Nile virus vector-competent mosquitoes. Ecosphere. 2024;15. doi:10.1002/ecs2.4771
4. Wilke ABB, Chase C, Vasquez C, Carvajal A, Medina J, Petrie WD, et al. Urbanization creates diverse aquatic habitats for immature mosquitoes in urban areas. Sci Rep. 2019;9: 15335.
5. Kolimenakis A, Heinz S, Wilson ML, Winkler V, Yakob L, Michaelakis A, et al. The role of urbanisation in the spread of Aedes mosquitoes and the diseases they transmit—A systematic review. PLoS Negl Trop Dis. 2021;15: e0009631.
6. Turell MJ, Dohm DJ, Sardelis MR, Oguinn ML, Andreadis TG, Blow JA. An update on the potential of north American mosquitoes (Diptera: Culicidae) to transmit West Nile Virus. J Med Entomol. 2005;42: 57–62.
7. Goddard LB, Roth AE, Reisen WK, Scott TW. Vector competence of California mosquitoes for West Nile virus. Emerg Infect Dis. 2002;8: 1385–1391.
8. Holcomb KM, Staples JE, Nett RJ, Beard CB, Petersen LR, Benjamin SG, et al. Multi-Model Prediction of West Nile Virus Neuroinvasive Disease With Machine Learning for Identification of Important Regional Climatic Drivers. Geohealth. 2023;7: e2023GH000906.
9. García-Carrasco J-M, Muñoz A-R, Olivero J, Segura M, Real R. Predicting the spatio-temporal spread of West Nile virus in Europe. PLoS Negl Trop Dis. 2021;15: e0009022.
10. Humphreys JM, Pelzel-McCluskey AM, Cohnstaedt LW, McGregor BL, Hanley KA, Hudson AR, et al. Integrating Spatiotemporal Epidemiology, Eco-Phylogenetics, and Distributional Ecology to Assess West Nile Disease Risk in Horses. Viruses. 2021;13. doi:10.3390/v13091811
11. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 2016;44: D463–70.
12. Yurekten O, Payne T, Tejera N, Amaladoss FX, Martin C, Williams M, et al. MetaboLights: open data repository for metabolomics. Nucleic Acids Res. 2024;52: D640–D646.
13. Smirnov A, Liao Y, Fahy E, Subramaniam S, Du X. ADAP-KDB: A Spectral Knowledgebase for Tracking and Prioritizing Unknown GC-MS Spectra in the NIH's Metabolomics Data Repository. Anal Chem. 2021;93: 12213–12220.
14. Smirnov A, Liao Y, Du X. Memory-Efficient Searching of Gas-Chromatography Mass Spectra Accelerated by Prescreening. Metabolites. 2022;12. doi:10.3390/metabo12060491
15. adap-big.github.io. Github; Available: https://github.com/ADAP-BIG/adap-big.github.io
16. Borrego SL, Fahrmann J, Datta R, Stringari C, Grapov D, Zeller M, et al. Metabolic changes associated with methionine stress sensitivity in MDA-MB-468 breast cancer cells. Cancer Metab. 2016;4: 9.
17. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19: 1655–1664.
18. Slatkin M. ISOLATION BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS. Evolution. 1993;47: 264–279.
19. Wang IJ, Bradburd GS. Isolation by environment. Mol Ecol. 2014;23: 5649–5662.
20. Jiang S, Luo M-X, Gao R-H, Zhang W, Yang Y-Z, Li Y-J, et al. Isolation-by-environment as a driver of genetic differentiation among populations of the only broad-leaved evergreen shrub Ammopiptanthus mongolicus in Asian temperate deserts. Sci Rep. 2019;9: 12008.

21. Weir BS, Cockerham CC. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. Evolution . 1984;38: 1358–1370.
22. Meuti ME, Stone M, Ikeno T, Denlinger DL. Functional circadian clock genes are essential for the overwintering diapause of the Northern house mosquito, Culex pipiens. J Exp Biol. 2015;218: 412–422.
23. Chang V, Meuti ME. Circadian transcription factors differentially regulate features of the adult overwintering diapause in the Northern house mosquito, Culex pipiens. Insect Biochem Mol Biol. 2020;121: 103365.
24. Shetty V, Meyers JI, Zhang Y, Merlin C, Slotman MA. Impact of disabled circadian clock on yellow fever mosquito Aedes aegypti fitness and behaviors. Sci Rep. 2022;12: 6899.
25. Surveillance Resources. 4 Apr 2023 [cited 28 Mar 2024]. Available: https://www.cdc.gov/westnile/resourcepages/survResources.html
26. Westnile.ca.gov. [cited 28 Mar 2024]. Available: https://westnile.ca.gov/resources_reports?report_category_id=6
27. Copernicus Climate Change Service. ERA5-Land monthly averaged data from 2001 to present. ECMWF; 2019. doi:10.24381/CDS.68D2BB30
28. EarthEnv. [cited 22 Feb 2024]. Available: https://www.earthenv.org/
29. Tuanmu M-N, Jetz W. A global 1-km consensus land-cover product for biodiversity and ecosystem modelling. Glob Ecol Biogeogr. 2014;23: 1031–1045.
30. Sullivan BL, Wood CL, Iliff MJ, Bonney RE, Fink D, Kelling S. eBird: A citizen-based bird observation network in the biological sciences. Biol Conserv. 2009;142: 2282–2292.
31. Hoyer S, Hamman J. xarray: N-D labeled arrays and datasets in Python. Journal of open research software. 2017;5. doi:10.5334/JORS.148
32. Bradski G, Kaehler A, Others. OpenCV. Dr Dobb's journal of software tools. 2000;3. Available: http://roswiki.autolabor.com.cn/attachments/Events(2f)ICRA2010Tutorial/ICRA_2010_OpenCV_Tutorial.pdf
33. Kendall MG. A NEW MEASURE OF RANK CORRELATION. Biometrika. 1938;30: 81–93.
34. Semenza JC, Tran A, Espinosa L, Sudre B, Domanovic D, Paz S. Climate change projections of West Nile virus infections in Europe: implications for blood safety practices. Environ Health. 2016;15 Suppl 1: 28.
35. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20: 273–297.
36. Ho TK. Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition. IEEE Comput. Soc. Press; 2002. doi:10.1109/icdar.1995.598994