Unveiling Mosquito Patterns in Chicago (2007-2024): A Data Analytics and Machine Learning Study

Ilyas Dr Potamitis

Unveiling Mosquito Patterns in Chicago (2007-2024): A Data Analytics and Machine Learning Study.

Ilyas Potamitis

Hellenic Mediterranean University, Heraklion Crete, Greece

Abstract

We apply data analytics to the publicly available and recently updated Chicago 2007-2024 Mosquito Database. In this database, 195 traps have been deployed in Chicago, Illinois, USA, from 2007 to 2024. Every year, from late May to early October, public health workers in Chicago set up mosquito traps scattered across the city. These traps collect mosquitoes, which are then partitioned into batches of fifty specimens. Each batch has been assessed using Polymerase Chain Reaction (PCR) for the presence of West Nile virus before the end of each week. The database records include the number of mosquitoes, the mosquito species, geographical information, and whether West Nile virus is present in each cohort.

In its first part, this work explores the application of mosquito data analytics to the manually collected data, focusing on the potential to identify trends, find the outbreaks, and localize hotspots to support vector control strategies. In its second part, we investigate at what extent a virus-positive batch can be predicted using the rest of the variables recorded in the database, showing that an AUC score of approximately 81% can be achieved on a 2-year held out subset without including weather data.

Finally, we discuss our findings in the context of integrating automated insect counting traps (e-traps) with mosquito data analytics. We argue that optical counters with environmental sensors embedded in traps can provide the supplementary information used in the Chicago database to predict the probability of an infected cohort without the use of PCR analysis. The probability of an infested cohort is less accurate than PCR but comes at no extra cost and is delivered almost real-time contributing to public awareness and resource allocation to intervention activities.

Introduction

Mosquito-borne diseases continue to pose a significant threat to public health globally, with over one million people dying from these diseases each year (World Health Organization, 2023) [1]. The ability to effectively monitor and predict the spread of mosquito populations is therefore critical in mitigating the risks associated with diseases such as West Nile Virus (WNV) [2-4], malaria, Zika and dengue among others [5-7]. This has led to the emergence of mosquito monitoring programs as an essential tool in public health planning, vector control, and disease prevention efforts [8-15].

The application of data analytics to mosquito surveillance data allows for the identification of spatial and temporal trends in mosquito activity, providing insights into the drivers [8] and dynamics [9] of disease transmission. By integrating data from multiple sources, including mosquito traps, weather stations, and population health records, predictive models can be developed to estimate disease risk in different areas. For instance, predictive modeling has been used successfully in [8] to estimate WNV occurrences based on mosquito population data and environmental variables such as temperature and precipitation.

Recent advances in machine learning and spatial analysis have further enhanced the capabilities of mosquito data analytics (see [9]-[10] and [11] for statistical challenges). Spatial analysis tools such as geographic information systems (GIS) have enabled researchers to map high-risk areas and understand how environmental factors contribute to mosquito breeding and disease spread [12-13].

The use of statistics is particularly important to understand mosquito surveillance Data in Arizona [14] and elsewhere, where recent studies have utilized detailed mosquito trapping and WNV occurrence data to identify outbreaks and guide targeted interventions [15]. Data analytics and machine learning has been applied on mosquito-related dataset in various contexts [16-19]. Research in [20] models the distribution of invasive mosquito species using several machine learning techniques on tabular datasets. In [21], the authors use climate data to predict malaria incidence, which is linked to mosquito populations. In [22], researchers explore the use of tabular data to forecast mosquito vector abundance. In [23], machine learning models are applied to mosquito occurrence data, analyzing mosquito habitat based on regional climate data. In [24-25], data mining and machine learning techniques are used to understand relationships among vectors, hosts, and pathogens. The established procedure for identifying mosquitoes with a virus load is to subject them to PCR testing.

PCR is a molecular biology technique used to detect the presence of specific pathogens or viruses in mosquito samples. It works by amplifying small segments of DNA or RNA, allowing researchers to identify and confirm the presence of disease-causing agents, such as West Nile Virus, Dengue Virus, or Malaria Plasmodium in mosquitoes. While PCR is highly effective for detecting pathogens, it has several practical disadvantages: (a) PCR requires specialized reagents (such as enzymes, primers, and nucleotides) and consumables (e.g., tubes and plates). The cost per test can add up significantly, making it expensive for large-scale mosquito surveillance programs. (b) PCR requires skilled personnel and sophisticated laboratory equipment, such as thermal cyclers, which are costly to purchase and maintain, especially in under-resourced regions. (c) PCR is not a real-time monitoring tool; the process involves collection, transportation to a lab, sample preparation, and testing, which introduces delays. It may take days or weeks to process and analyze samples from the field, leading to a lag between data collection and actionable results. Near-infrared (NIR) spectrometry has been suggested as an alternative approach for virus detection in mosquitoes. Although it relaxes some of the strict requirements of PCR, such as reagent use, it still requires specialized personnel and costly equipment [26-30]. NIR spectrometry is faster than PCR but not instantaneous, and it requires careful placement of the sensing probe on a mosquito specimen, making it unsuitable for automated analysis of large numbers of mosquitoes.

Although is a strong statement, we argue that traditional mosquito surveillance practices are time-consuming, expensive, and lack scalability [31]. In this work, we are mainly interested in investigating whether we can predict the probability of an infected WNV batch in mosquito traps based on other variables such as the date, location, number of batches per trap, and number of mosquitoes per batch given historical data with manually verified virus presence. Machine learning models have been employed to predict mosquito populations and disease outbreaks with high accuracy, often outperforming traditional statistical approaches. This work seeks to explore the growing role of mosquito data analytics and machine learning on the publicly available, tabular dataset of Chicago Mosquito records (2007-2024) in addressing public health challenges posed by mosquito-borne diseases.

While our findings indicate that the probability attributed to each batch of being infected is not as accurate as PCR, it is a cost-effective and instantaneous approach. We then discuss the technical challenges that must be overcome so that automated optical

counters embedded in mosquito traps [32-38], which can extract the variables used in this study, could be adapted to report an informed probability of a WNV-positive cohort. Additionally, in terms of resource allocation, which is always an issue in practice, we suggest that it is more effective to allocate PCR analysis to the cohorts flagged as positive by automated traps.
We open-source the code used to analyze the public data and classify the Chicago mosquito (2007-2024) database, making it applicable to any mosquito database with a similar structure (see Appendix).

Materials & Methods

The Chicago Database (2007-2024)

The Chicago West Nile Virus (WNV) Mosquito Database last updated October 4, 2024, is a publicly available dataset focused on mosquito surveillance in the city of Chicago, Illinois, for monitoring the spread of the WNV [39]. The dataset is primarily used by public health agencies, researchers, and data scientists to study mosquito population trends, virus prevalence, and the effectiveness of vector control strategies. The database reflects the city's mosquito control and disease surveillance efforts. The data is collected weekly during mosquito season, typically between late-May to early-October when mosquitoes are most active. The mosquitos are grouped in batches of up to fifty specimens and each batch is tested for the presence of WNV before the end of the week. The test results include the number of mosquitos in the batch, the mosquito's species, and whether WNV is present in the cohort.

The location of deployed traps

The database is centered around the Chicago area and includes community areas, trap addresses, and environmental factors like latitude and longitude coordinates for the mosquito traps. The location of the traps is described by the block number and street name.

The Trap Types

In mosquito surveillance and control, various types of traps are employed to monitor mosquito populations, detect disease presence, and assist in vector management. Each trap type targets mosquitoes at different stages or conditions, using specific attractants or designs [40]. In the Chicago database, four mosquito trap types are mentioned:
1. GRAVID Traps: Gravid traps are designed to attract and capture female mosquitoes that are ready to lay eggs (gravid mosquitoes). These traps typically use organic matter-infused water, mimicking the stagnant water sites where females prefer to lay their eggs. Gravid traps are particularly effective for collecting mosquitoes from the Culex genus, known vectors of the WNV. By targeting gravid mosquitoes, which have already fed on blood and are potentially infectious, these traps are critical for disease surveillance.
2. CDC Light Traps: The CDC (Centers for Disease Control and Prevention) light traps are among the most widely used tools for mosquito surveillance. These traps utilize light as an attractant, usually a small incandescent or LED bulb, combined with a fan to capture flying mosquitoes. In many cases, CO_2 is also used as an additional lure to mimic the presence of a warm-blooded host. CDC traps are effective in capturing a wide variety of mosquito species, including Anopheles, Aedes, and Culex, making them versatile in mosquito population monitoring.

3. OVI Traps (Oviposition Traps): Oviposition traps, or OVI traps, are designed to attract female mosquitoes looking for a site to lay eggs. These traps often consist of dark containers filled with water and a rough surface for mosquitoes to deposit their eggs. Oviposition traps are useful for detecting mosquito species like Aedes aegypti and Aedes albopictus, which are known carriers of diseases such as dengue, Zika, and chikungunya. By collecting eggs rather than adults, these traps provide early indications of mosquito activity and help in monitoring invasive species.
4. SENTINEL Traps: Sentinel traps are used primarily for long-term mosquito monitoring and disease surveillance. These traps are often baited with animal hosts (e.g., live birds) or attractants such as CO_2 or pheromones. The primary function of sentinel traps is to capture mosquitoes that are actively seeking blood meals. They are instrumental in tracking potential disease outbreaks, particularly in areas with a high risk of vector-borne diseases. Their design allows for continuous operation, making them valuable in both research and public health monitoring programs.
Each of these traps serves a specific purpose in mosquito surveillance, with different characteristics tailored to the behavioral ecology of the target mosquito species.
These four trap types are mentioned in the Chicago database though OVI is practically not employed. Another trap-type called 'Magnetic' is mentioned but with no valid measurements.

The database's fields

The database is tabular, and it is important to note that each row corresponds to a batch of mosquitoes and several rows can belong to the same trap visit as the catches are partitioned in groups of fifty specimens. The database is highly unbalanced as less than 10% of the batches have a WNV positive label. The mosquito occurrences dataset contains the following columns:
SEASON YEAR: The year of data collection.
WEEK: The week of the year that has been assessed with PCR.
TEST ID: Unique identifier.
BLOCK: General location of the mosquito trap.
TRAP: Trap ID.
TRAP_TYPE: Type of trap used (GRAVID, CDC, OVI, SENTINEL).
TEST DATE: Date and time the test was performed.
NUMBER OF MOSQUITOES: Number of mosquitoes collected.
RESULT: Outcome of the test (positive or negative for the presence of WNV). A positive case means that the batch has been subjected to PCR and has been found positive due to an unknown number of infected mosquitoes in the batch.
SPECIES: Mosquito species found.
COMMUNITY AREA NUMBER: Number identifying the community area.
COMMUNITY AREA NAME: Name of the community area.
LATITUDE: Latitude of the trap location.
LONGITUDE: Longitude of the trap location.

Automatic Mosquito counters

In the Chicago database, specialized personnel maintain and manually annotate (count mosquitos, recognize species composition from catch bags and perform data entry). In this work we suggest that mosquito monitoring can be automatized to a certain extent. The cooperation of data analytics/AI with automated mosquito traps is feasible at the server level and commercial automated mosquito traps already produce most of the variables in the Chicago database. We show in this work that given historical data; these variables can predict if a batch is infected without resorting to PCR. Therefore, we include

a basic introduction to the principles of this technology to inform the interested reader about the potential of this technology and the possibility to predict WNV infection in batches without employing PCR. The integration of automatic counters in mosquito traps, such as optical sensors and automated imaging systems, offers significant advantages for monitoring mosquito populations. The primary benefit is the wireless transmission of mosquito catches on daily basis that reduce manpower and budget constraints, allowing these systems to be deployed on a large spatial scale. Additionally, automatic counters facilitate the delivery of data from remote, hard-to-reach, and often hostile environments in near real-time (see [32-38] for such approaches).
In the Chicago Database, GPS coordinates, timestamps and mosquito counts per species are provided manually and historical data of these variables can be used to predict the probability of an infected batch. We argue that the same variables can also be provided by automated traps, allowing the inference in real-time, but this needs to be verified in practice.

Basic principles

Optical counters typically use an array of light-emitting diodes (LEDs) as an emitter and photodiodes as receivers. The light can be modulated in two ways. In the first approach, a thin flow of infrared light is interrupted as a mosquito is drawn in by the trap's fan, casting a shadow onto the photodiodes. This shadow results in a voltage fluctuation, and the root mean square (RMS) value of light intensity at the receiver is used to generate a binary signal indicating the presence or absence of an insect. When the emitter shapes a field of view (FOV) with small width, the optoelectronic counter can sense the presence of an insect but nothing more as the suction imposes high speed movement and the insect has no time to beat its wings inside the FOV. Devices described in [36] are of this type and return a binary value (presence versus absence). In an enhanced setup, the emitter-receiver pair can be arranged in a 2D array to expand the (FOV), so that the incoming insect can have the time to complete some full wingbeat cycles thus enabling the registration of short recordings of wingbeat, as demonstrated in [33-36].
The second approach involves placing both the emitter and receiver on the same side to record the backscattered light from the mosquito's main body and wingbeat as it enters the trap (see [41-43] for a related technology applied in a different context). The backscattered-light setup captures more detailed information by allowing for the detection of additional wingbeat harmonics. The apparatus described in [34-35] and [36] exemplifies this type of device. Overall, capturing wingbeat frequencies from a sufficiently large FOV enables the analysis of species, sex, or genus of mosquitoes by comparing the incoming signal to pre-existing wingbeat prototypes. These prototypes are derived in laboratory settings using mosquito colonies contained in net cages with enclosed e-traps. Although there are thousands of mosquito species globally, only a limited number coexist in designated locations. For example, in the Chicago dataset, three species are mainly responsible for nearly all WNV positive cases. Automatic counters provide an efficient method for identifying these key species, contributing to more effective surveillance and targeted vector control strategies.

Technical challenges

Automatic mosquito traps are not currently without limitations and in an automated monitoring setting connected to data analytics modules their errors will affect the quality of predictions. There are several technical challenges that must be addressed to ensure the effective operation of these systems.
One major issue is that automatic counters often struggle to accurately differentiate between mosquito species and other non-targeted species. The ability to discern insect size from the shadow or backscattered signal is inherently limited due to variability in

insect orientation and flight patterns as they pass through the detection area. Consequently, size-based differentiation has significant constraints, and only broad classifications can be made. This leads to non-target insects being erroneously counted, especially in environments where mosquito traps are exposed to high populations of other insects. In an evaluation of this technology in [37] it has been reported that in areas where the relative mosquito abundance is especially low, the optical counter has been ineffective and unable to provide data that are reflective of the actual number of mosquitoes suffering many false alarms.

In another independent, field-evaluation [38], automatic counters have been evaluated, yielding mixed accuracy levels that can be very low in some settings and highlighting certain technical challenges. If the technical obstacles identified in [38], are valid, then they can be addressed using the current maturity status in technology. For example, in [38] is stated that 15% of deployed devices did not connect and failed to deliver data. If a device cannot connect to the mobile network, then it does not get a success confirmation from the server and can proceed into storing that data internally until the next successful programmed connection and therefore, loss of data is not inevitable. Moreover, in a quite recent advancement, new affordable communication modems and global SIM cards now support satellite communications, and the modem can switch to this option when terrestrial communications fail. In the same evaluation it is also reported that the counters can transmit all data to an online server every 15 min, probably because they need to extract activity pattern of mosquitoes. However, frequent transmissions must be avoided if they are not of absolute necessity because during transmission the optical counter is deactivated to avoid the effects of electromagnetic interference from the emitting antenna. Additionally, GPS communication must be set once during installation and not in every transmission as traps are not moving platforms and GPS communication affects a delay and increased power consumption. Therefore, during transmission of data any mosquitoes sucked-in are not counted by the e-trap. Finally, commercial mosquito traps use a suction to trap and retain mosquitoes in an embedded net/bucket. In long deployments the number of mosquitoes can be so large that it affects air circulation which, in turn, affects the imposed pressure from the suction fan that is reduced and may lead to escaping mosquitoes as there is not enough pressure to contain them in the catch net. These failures will end up to false alarms in the counting process and lead to errors that day by day pile up.

Finally, if the sex, species and genus is to be discerned then temperature dependent corrections/normalizations need to apply as temperature variations affect the wingbeat of insects and its attribution to species classes [44].

Current optoelectronic counters in mosquito traps can wirelessly transmit environmental variables (typically temperature, humidity and ambient light intensity), GPS coordinates, mosquito counts, timestamps of captures and battery status as main variables. Experimental approaches can also provide sex, genus, species categorization based on its wingbeat at various accuracies, after being adapted to the species composition of the operating location [34-36]. We suggest that there is a need for another round of technical improvements in optical counters embedded in mosquito traps and then they need to be assessed in the field by independent organizations/institutes at medium to large scale deployment. This is imperative for advanced devices that beyond counts they report sex, genus and species composition of their mosquito catches.

Evaluation using ROC curve and AUC

In the Results section we evaluate classification results using the Receiver Operating Characteristic (ROC) metric and the area under this curve (AUC). ROC curves are widely used in classification problems to evaluate the performance of a binary classifier. ROC curves plot the True Positive Rate (TPR) (sensitivity) against the False Positive Rate (FPR)

at various classification thresholds, providing a comprehensive view of how a model's performance changes across different thresholds. Unlike metrics like accuracy, precision, or recall, which depend on a specific threshold, the ROC curve provides an aggregate measure of performance across all possible thresholds. This is particularly important in applications where there is no natural or predefined threshold. ROC curves are less affected by class imbalance compared to metrics like accuracy. In a highly imbalanced dataset, accuracy can be misleading, as the classifier might simply predict the majority class. The Chicago database is imbalanced because the WNV-positive cases are rare compared to the negative cases (<10% of the batches). ROC curves provide a way to visualize the trade-off between correctly identifying positives and mistakenly classifying negatives as positives. The Area Under the ROC Curve (AUC) is often used as a summary statistic for model performance. A perfect classifier has an AUC of 1.0, while a random classifier has an AUC of 0.5. Higher AUC values indicate better performance, capturing how well the model discriminates between positive and negative classes over all thresholds. In practical applications, choosing an appropriate threshold—also called an operational point—depends on the specific requirements of the use case and the cost of erroneous decisions. A ROC curve figure helps to select the best operational point by visualizing the trade-off between True Positive Rate and False Positive Rate.

Results

Data Preprocessing

Traps and trap-types

Traps in the Trap column of the database named T240, T240B, T143 have missing data. For T240, T240B we have tracked the address: 24 Lincoln Park with Latitude: 41.9187, Longitude: -87.6715. The T143 address is Norwood Park. The approximate GPS coordinates for this area are 41.995 latitude and -87.799 longitude.

Some traps in the Chicago database are "satellite traps". These are traps that are set up near (usually within 6 blocks) an established trap to enhance surveillance efforts. Satellite traps are postfixed with letters. For example, T220A is a satellite trap to T220 [46]. This dataset is organized in such a way that when the number of mosquitos found in the catch bug/backet exceed fifty, they are split into another record (another row in the dataset), such that the number of mosquitos is capped at fifty. Therefore, the maximum number of mosquitoes per batch is fifty with only 2 exceptions in the records in 2014 and 2022 with 77 and 61 mosquitoes respectively (probably outliers).

Regarding trap types, in this dataset the OVI trap exists only in a single valid record in 2007 and, therefore, has no influence on the statistics. There are other entries as well, but the crucial parameter of the infection status is missing and for that reason these records are dropped.

There are 195 unique traps but the traps in Table 1 add up to 210. The discrepancy arises because some traps appear under multiple TRAP_TYPE categories (e.g. Trap 009). When one sums these counts, it treats each instance of a trap across different TRAP_TYPE categories as unique, leading to an inflated total. This is either a data entry mistake or TRAP ids are reserved for a location, but the type of trap can change during the monitoring period.

Table 1. There are 195 trap ids in the dataset. 169 are GRAVID traps, 28 CDC, 12 SENTINEL and 1 OVI. These numbers add up to 210 because some IDs appear with two different trap types.

TRAP TYPE	Multirow counts	Traps
GRAVID	34573	169
CDC	1256	28
SENTINEL	319	12
OVI	1	1

Missing values

After we impute traps: T240, T240B, T143, we drop all records that do not have data entry (NaN value) in the column of RESULTS (infection status) as this is the most crucial variable, it is rare, and we refrain from imputing it. We end up with a database of 36233 rows and we keep 14 relevant columns (variables): 'SEASON YEAR', 'WEEK', 'TEST ID', 'BLOCK', 'TRAP', 'TRAP_TYPE', 'TEST DATE', 'NUMBER OF MOSQUITOES', 'RESULT', 'SPECIES', 'COMMUNITY AREA NUMBER', 'COMMUNITY AREA NAME', 'LATITUDE', 'LONGITUDE'.

Data Analytics

In this section we proceed to pose useful questions of practical value. These are the questions that would affect policy decisions, would be used to improve public awareness and to evaluate intervention strategies. The code is provided in the appendix and would be applicable to any other infection and mosquito database with corresponding structure.

What is the distribution of WNV positive cases by year?

In Figure 1 we meant to visualize the distribution of the West Nile Virus presence by year to see if there is a potential trend in this pattern. Each bar represents the number of occurrences where the virus was detected in that specific year. By doing so, the histogram helps to identify which years had a higher or lower incidence of West Nile Virus presence and the trend. We see that the number of incidents in Chicago, based on this particular database, has been relatively stable over the years. This picture can be used to assess the impact of an intervention policy. We are not aware of the intervention policies currently applied but there is no steady decline of the phenomenon in Figure 1.

The mosquito species distribution of the whole database is gathered in Table 2. This Table shows all the species that are included in the database. The *Culex pipiens/restuans* categorization is the most prevalent, followed by *Culex restuans* and *Culex pipiens*. The term *Culex pipiens*/restuans is sometimes used when the differentiation between the two species is not clear, especially in mixed pools of collected mosquitoes. Because of their similarities in appearance and overlapping habitats, many mosquito surveillance programs use the combined term *Culex pipiens*/restuans when distinguishing between the two is difficult, especially without genetic testing. Therefore, *Culex pipiens* and *Culex restuans* are distinct species but are often grouped together due to their similarity. This vagueness in class attribution imposes an additional difficulty in the classification experiments. What this data definitely suggests is that the majority of the captured mosquitoes belong to the Culex genus, known for their role in transmitting diseases like West Nile Virus.

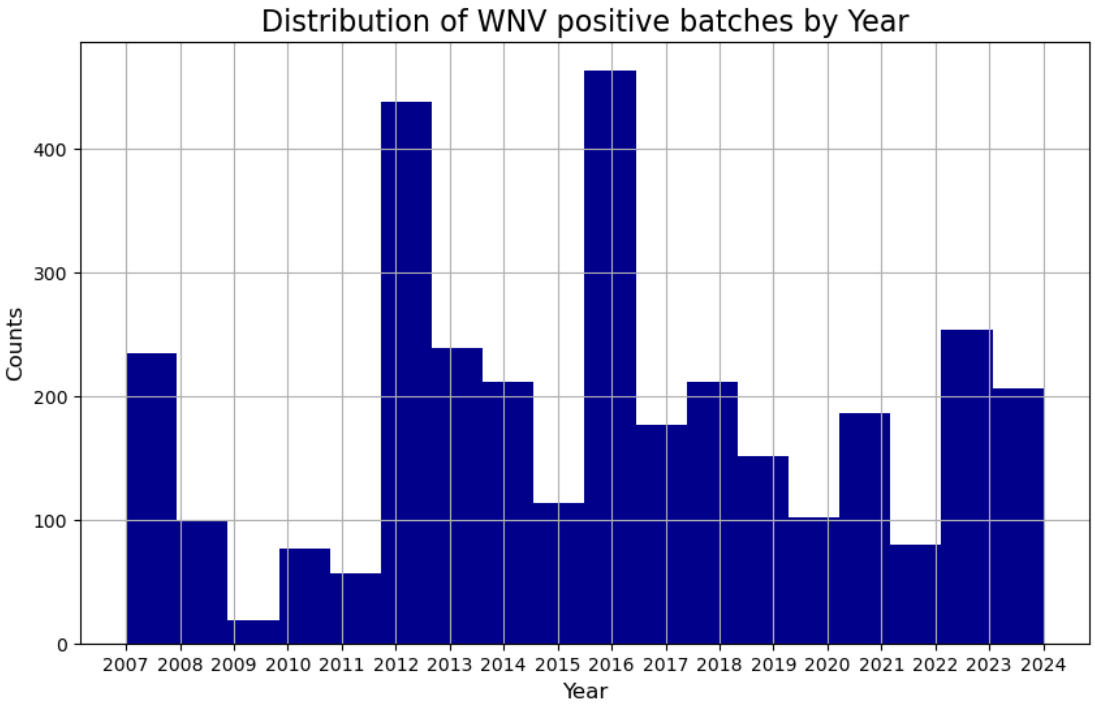

Figure 1. WNV positive cases out of all traps with respect to the year. The y-axis holds the number of batches that have been found positive for WNV.

Table 2. The Table holds the Species composition, the total # of mosquitoes and the positive batches that correspond to them. *Culex pipiens and Culex restuans* have been the main carriers of WNV virus in the Chicago database.

Species	**# of mosquitoes**	**# batches with WNV present**
CULEX PIPIENS/RESTUANS	280858	2038
CULEX RESTUANS	115613	780
CULEX PIPIENS	68122	489
CULEX TERRITANS	1967	4
CULEX SALINARIUS	492	3
CULEX TARSALIS	97	0
UNSPECIFIED CULEX	52	0
CULEX ERRATICUS	46	0

How does the species composition in catches of mosquito traps evolve over time?

Figure 2 provides an overview of mosquito trends in Chicago, focusing on variations in mosquito species and the prevalence of WNV over time. The analysis reveals the evolving population of different species, which may indicate changes in environmental factors, mosquito control measures, or virus prevalence.

The *Culex pipiens* (orange line in Figure 2) shows an initial peak in 2007, reaching the highest count among all species at that time. After 2007, the population rapidly declines in 2008, and stays consistently low from 2009 onwards, with only minor fluctuations in 2013 and 2014. The *Culex pipiens/restuans* (green line in Figure 2) is the dominant case throughout most of the time period (mind though that this is not a species but a collective characterization). Peaks can be observed in 2007, 2012, 2015, 2021, and 2023. The

population shows a cyclical pattern, with significant rises and falls. Notably, there is a sharp decline in 2024 for the attribution to the mixed class *Culex pipiens/restuans* indicating either potential classification errors or some advancement in discerning these species.

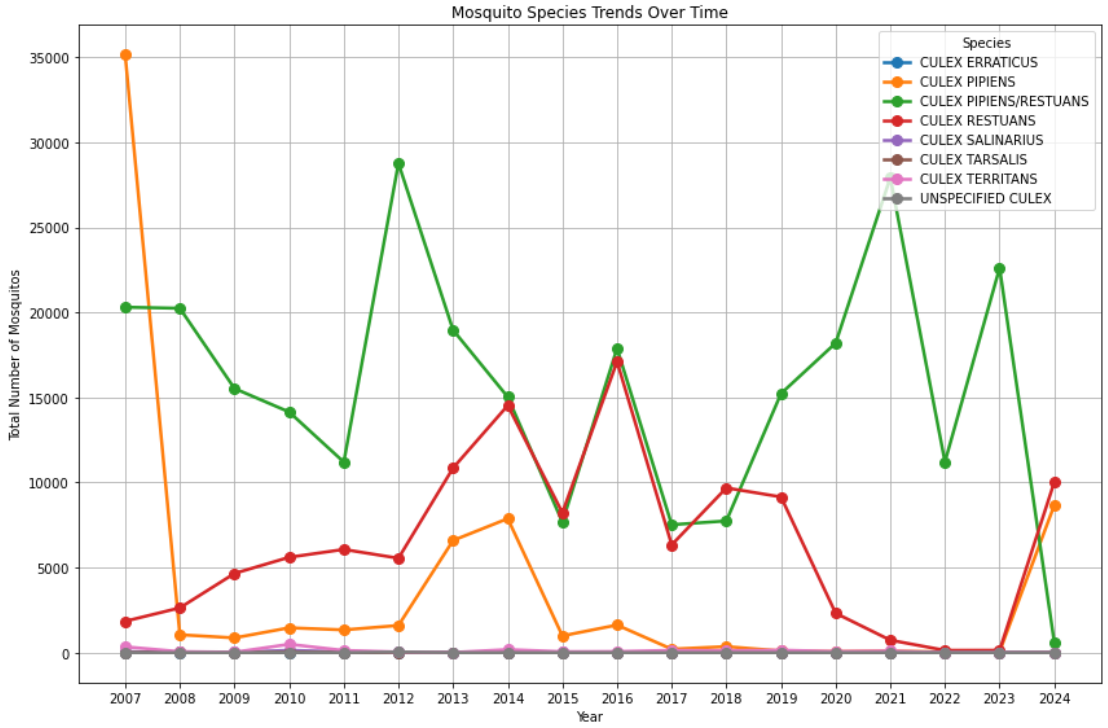

Figure 2. Mosquito species composition trend over the years in the Chicago WNV database. The y-axis holds the counted number of mosquitoes in trap catches.

The *Culex restuans* (red line) initially has a low population but begins a steady increase from around 2013 to 2015. There are some year-to-year fluctuations but generally stay moderate from 2015 onwards. Notable peaks occur in 2015 and 2023, with a general trend of maintaining a steady presence.
Culex erraticus (blue line) demonstrates very low numbers throughout the entire period. This species shows no significant spikes, suggesting either low prevalence or limited environmental suitability in the study area.
Culex salinarius, Culex tarsalis, Culex territans, Unspecified Culex (purple, yellow, grey lines) consistently have nearly zero to low populations throughout the time period.
This suggests that these species are either not as prevalent in the area or may be more challenging to trap using the specific trap types.

When is it most probable to detect WNV infection in batches of traps' catches?

Figure 3 is, in our view, the most significant figure in this work as it highlights the peak and distribution of WNV occurrences over time, providing insight into the seasonal pattern of outbreaks. It illustrates WNV-positive batches in relation to the weeks and months in which the virus was detected, across all years and species. This visualization is especially valuable for identifying potential seasonal trends, such as spikes in virus presence during specific months. It allows us to pinpoint periods of heightened activity, which can inform vector control strategies and public health responses. Notably, peak activity is observed between the last week of August and the first week of September.

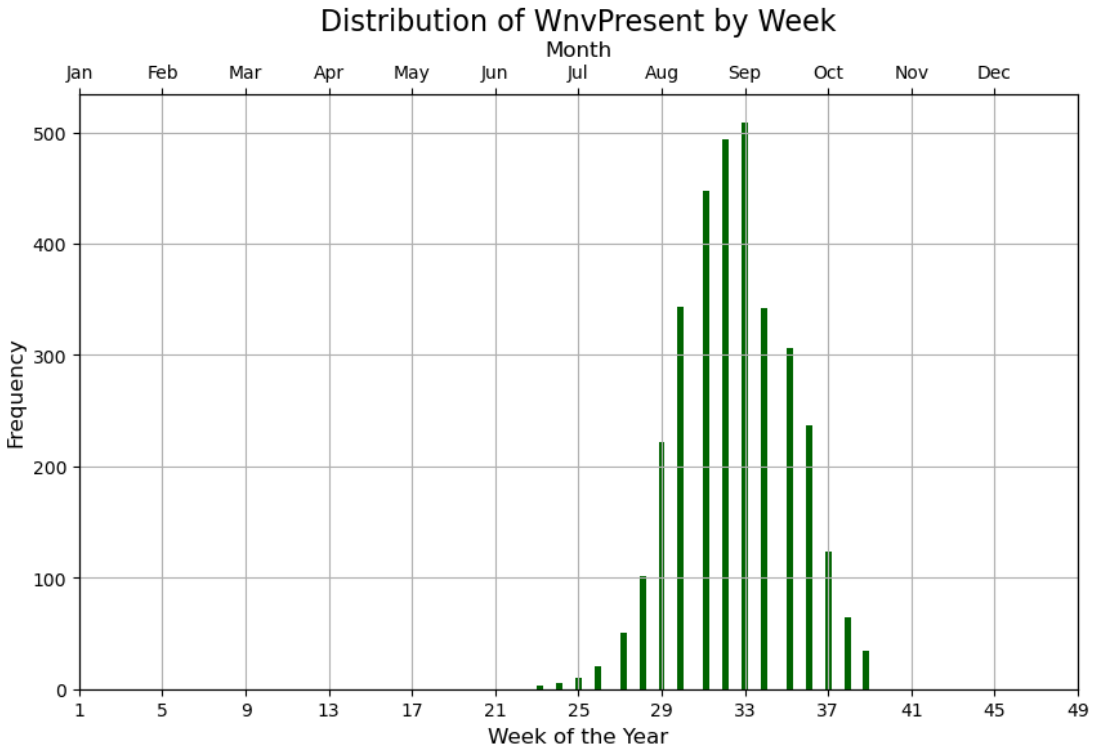

Figure 3. WNV positive batches (y-axis) with respect to the week and month they have been identified. The main x-axis shows the week numbers (from 1 to 52), while a secondary x-axis on-top displays the months. Between the last week of August and the first week of September we have peak activity from data pooled from 2007 to 2024.

Which trap types are most effective in catching most mosquitos and have the most WNV-infected batches? What about species composition?

To compare the effectiveness of each trap type fairly, we need to account for the unequal distribution of traps among trap types (see again Table 1). Since each TRAP_TYPE has a different number of traps, directly comparing the total mosquito counts would be biased. Normalizing by the number of traps within each TRAP_TYPE allows us to account for this imbalance, providing a fairer comparison of each trap type's effectiveness. Figure 4 visualizes the effectiveness of different trap types in catching mosquitoes, broken down by species. The data is grouped by the TRAP_TYPE and SPECIES variables of the database, aggregating the normalized number of mosquitoes caught (variable NUMBER OF MOSQUITOES) for each species by each trap type. Since each bar in the bar-plot of Fig. 4 represents a specific trap type and species combination, we can see which traps are more successful at capturing certain species. This insight can guide the deployment of different trap types to target specific mosquito populations more effectively, focusing on species that are major vectors of diseases like WNV. It is also helpful in optimizing trapping strategies by selecting the most effective trap types based on the target species in an area. The outcome of this analysis is that the GRAVID trap type is found to be the most effective trap in Culex catches followed by CDC. Note the difference in species caught by each trap type. The SENTINEL trap does not perform very well with Culex.

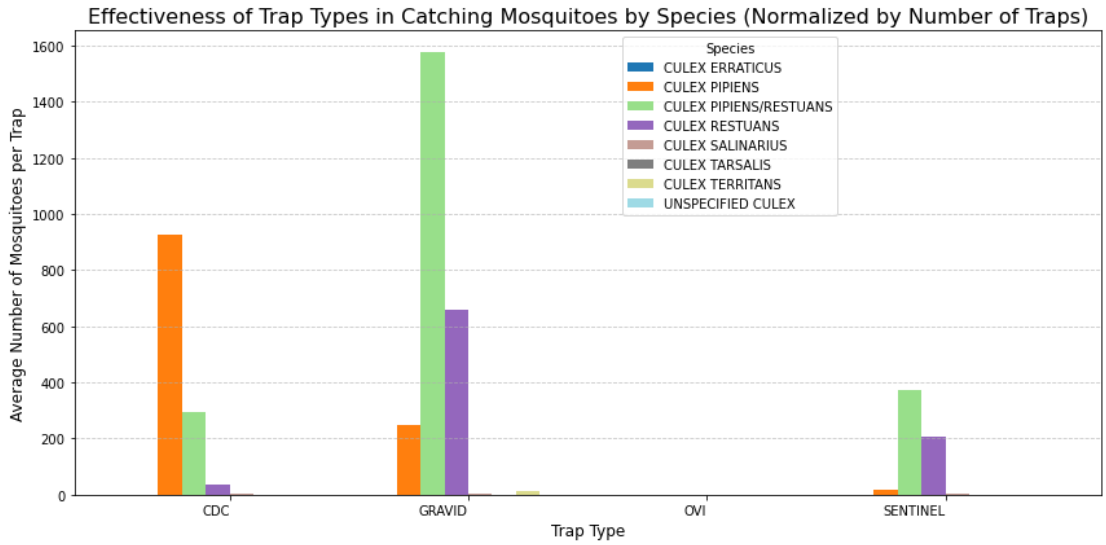

Figure 4. The visualization shows how effective different trap types are in catching various mosquito species. Y-axis is normalized. It answers the question: Which trap types are best suited for capturing high numbers of mosquitoes overall.

Figure 5 is related to Figure 4, but the y-axis now holds the WNV positive cases normalized by the number of traps in each trap type. Therefore, it depicts the correlation between the trap type, the species and only the WVN positive batches. Gravid traps are the trap-types found with most cases of WNV-positive mosquito batches, particularly *Culex pipiens/restuans* and *Culex restuans*. Note again that there are two species *Culex pipiens* and *Culex restuans* and the combined class is introduced when distinguishing between them is not possible, often due to mixed mosquito collections or limitations in identification methods.
Other traps like CDC and sentinel have limited success, highlighting the need for a strategic approach in mosquito surveillance, focusing on trap types that maximize the likelihood of capturing disease-carrying species. Again, the GRAVID type is associated with the larger number of catches (normalized) followed by the SENTINEL trap type this time followed by CDC.

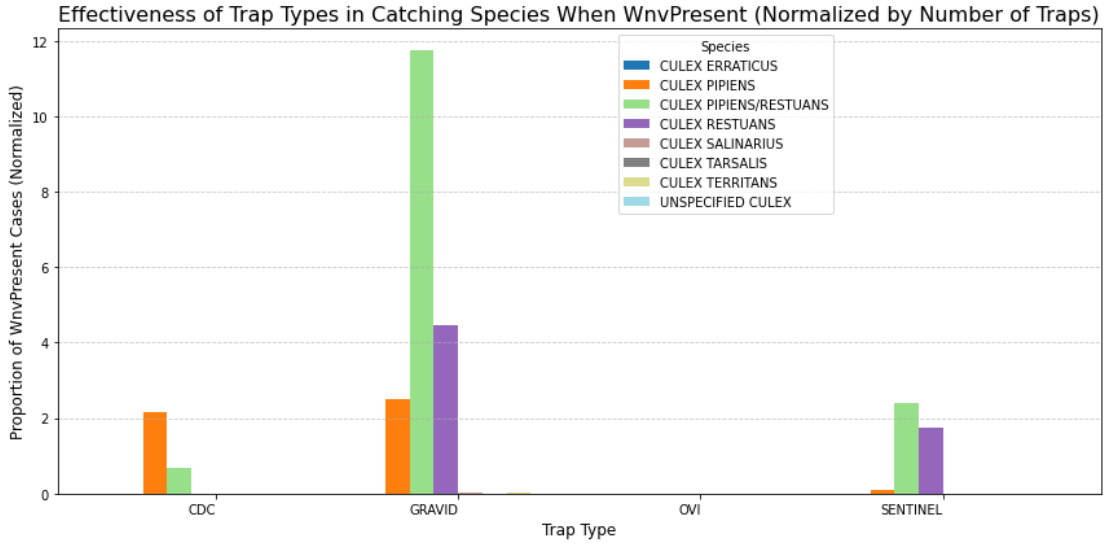

Figure 5. Visualizes which trap type have been found with the most WNV positive cases normalized by the number of traps in each trap type. Gravid type trap again has most of the captures for the cases of *Culex pipiens*, and *Culex restuans*.

Which trap IDs were responsible for more mosquito catches and WNV infected batches with respect to species? Where are the corresponding addresses?

There are 195 unique mosquito Trap Numbers that the public health workers in Chicago set up and scattered across the city. In Figure 6, we identify the top-performing mosquito traps based on two key metrics: West Nile Virus presence (variable RESULT in the left y-axis) and the total number of mosquitoes caught (in the right y-axis). If a trap has high mosquito counts but low WNV detections, it may indicate that the mosquitoes caught are not the primary carriers of the virus, suggesting a lower risk. Conversely, a high number of WNV detections, even with a moderate number of mosquitoes, points to a high concentration of infected mosquitoes, indicating that the location has a heightened risk of virus transmission.

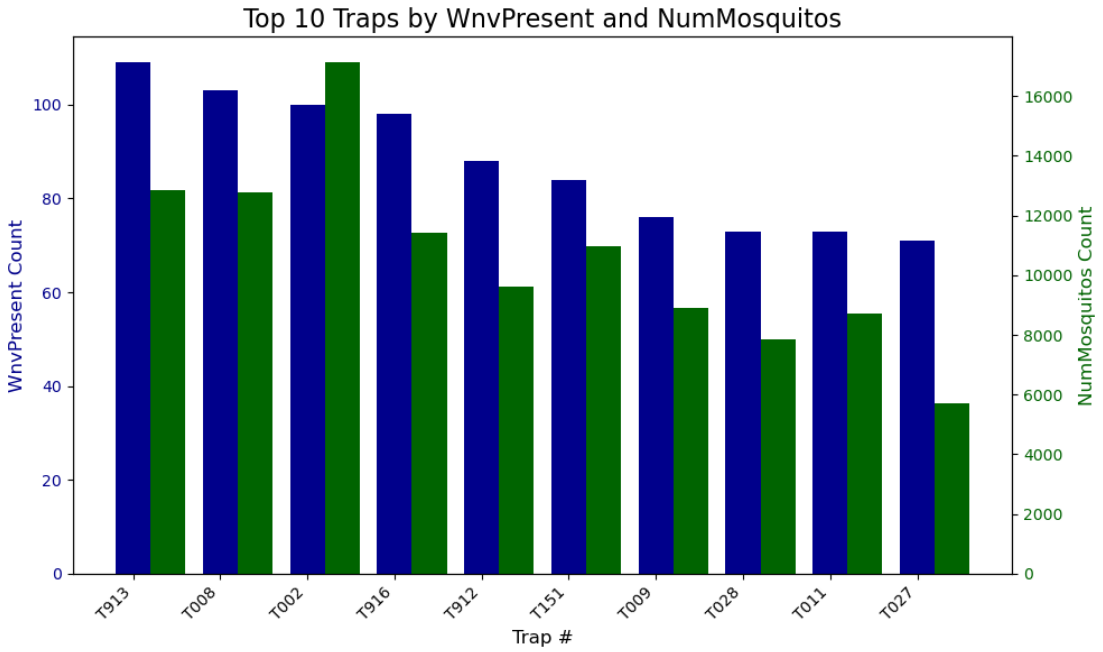

Figure 6. The top 10 traps ranked based on WNV-positive batches and the number of mosquitoes per batch.

Once we have the best performing trap names, we proceed into composing a table of their Address in Table 3.

Table 3. Addresses and Trap type of the best performing trap IDs. Note that many traps with different trap-types can be installed in the same location.

Trap ID	Location	Trap type
T913	100XX W OHARE AIRPORT	GRAVID
T008	70XX N MOSELLE AVE	GRAVID
T002	41XX N OAK PARK AVE	GRAVID
T916	100XX W OHARE AIRPORT	GRAVID
T912	100XX W OHARE AIRPORT	GRAVID
T151	70XX W ARMITAGE AVE	GRAVID
T009	91XX W HIGGINS RD	CDC
T009	91XX W HIGGINS RD	GRAVID
T028	58XX N WESTERN AVE	GRAVID
T011	36XX N PITTSBURGH AVE	GRAVID

Trap T009 is located at 91XX W HIGGINS RD and appears with two different trap types—CDC and GRAVID. This is either a data entry mistake or they have the same physical trap location reused over time, but the trap type is changed during different trapping periods.

Which locations in the city constitute a hotspot for WNV batches?

We proceed to identify the geographic location of hotspots. The first approach is to find the community areas (variable COMMUNITY AREA NAME) associated with virus-positive cases and sort them by value. Then we derive heatmaps of the trap locations with the highest numbers of WNV-positive cases. The histogram in Fig. 7 visualizes the distribution of WNV detections across different areas of the town. Fig 7 helps in identifying high-risk locations, to guide public health efforts for targeted vector control and preventing the spread of WNV. This information can help in prioritizing vector control efforts, such as targeting these high-risk areas for increased spraying, public awareness campaigns, or other preventive measures. Understanding which traps consistently detect the virus can help in allocating resources efficiently. Health authorities can use this information to optimize monitoring locations, ensuring that the most significant risk areas are continuously observed to prevent outbreaks. This allows you to see which geographic blocks have higher instances of WNV presence, indicating potential hotspot areas like O'Hare airport.

Figure 8a and 8b depict two types of geospatial visualizations that can be used to analyze the spatial distribution of WNV presence in the region covered by the dataset. The heatmap displays the intensity of WNV occurrences geographically. Each point on the map represents a location with the attributes of latitude and longitude, with the color intensity indicating the presence of the virus. Note that the points contribute to traps' locations and not the actual distribution of WNV in the field over the area.

The heatmap helps in identifying hotspot regions where the density of infected mosquitoes is highest. Areas with darker colors indicate higher virus activity, suggesting areas of greater risk. Health officials can use this information to focus vector control efforts like pesticide spraying or mosquito breeding habitat elimination in the most affected regions.

The convex hull can be used to define the boundary of the region that needs to be monitored or controlled for WNV. It gives an idea of the geographical limits of areas where traps have detected the virus. By looking at how the traps are distributed within the convex hull, authorities can assess the spatial spread and identify areas where traps may be missing (i.e. identifying gaps in monitoring). Regions within the hull but with fewer traps could need additional monitoring.

Both figures are useful for effective resource allocation, monitoring coverage, public health interventions, and communicating risk to stakeholders and the public. They can be used in public health campaigns to inform communities of areas with a high risk of WNV transmission and encourage protective behaviors, such as avoiding outdoor activities at peak mosquito times or using insect repellent.

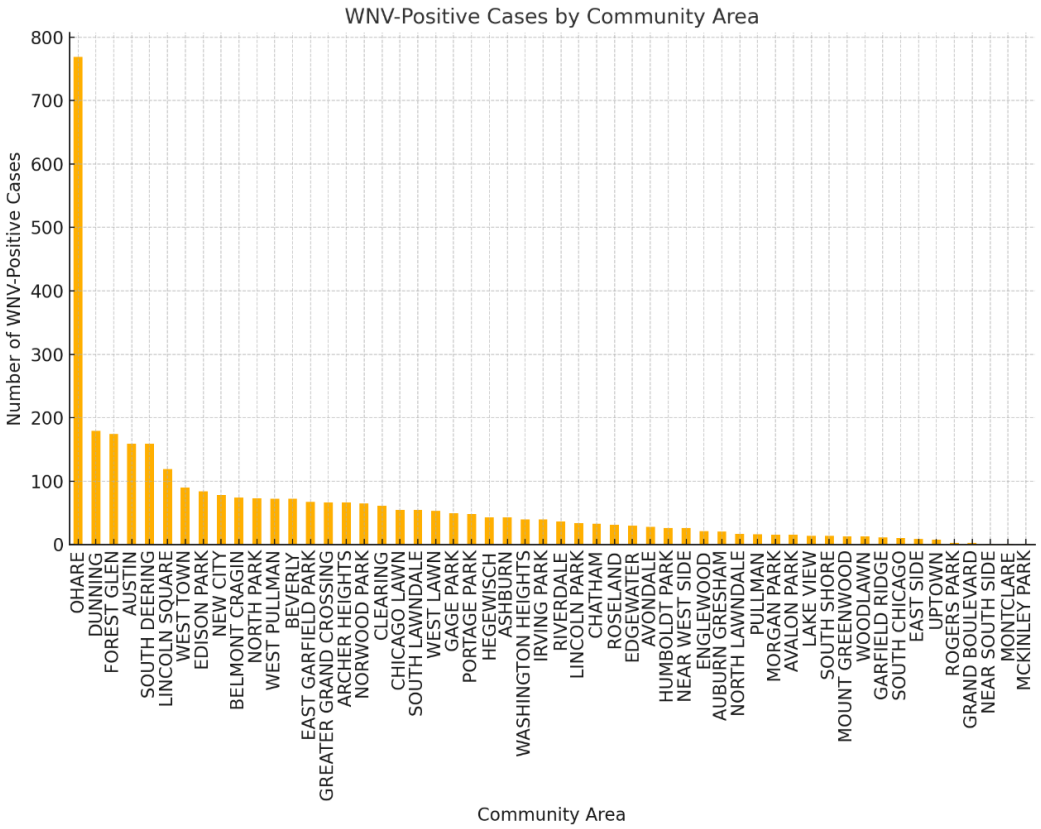

Figure 7. We mark the Addresss where most incidents of WNV positive occurred. The address that stands out corresponds to the station at the Community area Name O' Hare International Airport.

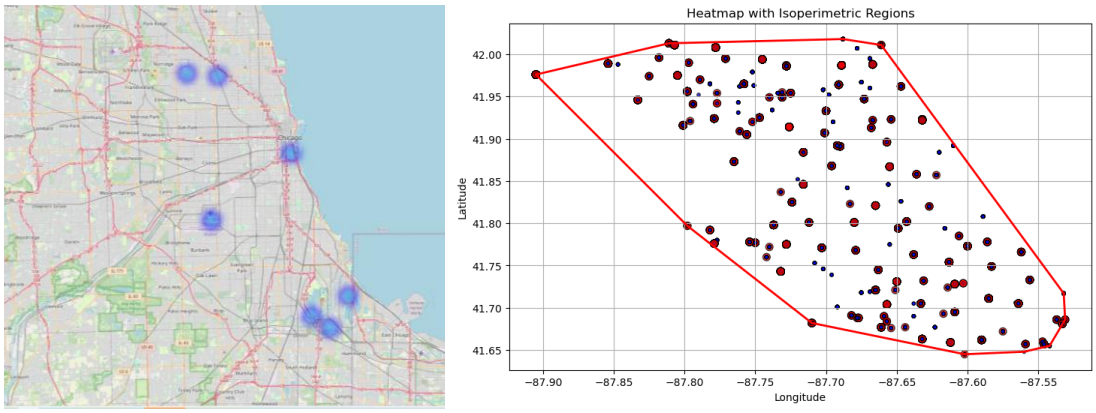

Figure 8. a) Heatmap of WNV positive traps in Chicago b) The convex hull is a mathematical boundary that encapsulates all the points representing trap locations (i.e., latitude and longitude of each trap). Larger size of spots for WNV, default size otherwise.

Identify outbreaks

To identify outbreaks of West Nile Virus (WNV), we can look for clusters of positive cases within a certain time period and/or geographic area. Outbreaks can be defined by: A high number of positive cases in a short time frame (e.g., several days to weeks). Geographic clustering in the context of mosquito outbreaks refers to multiple WNV-positive traps located close to each other. Identifying outbreaks of WNV over the years has practical value for public health planning, resource allocation, and risk mitigation. By identifying periods of outbreaks, public health authorities can plan and execute targeted interventions such as mosquito control, spraying campaigns, and public awareness

initiatives. Knowing the precise periods when outbreaks tend to occur helps in taking proactive measures rather than reactive responses, thereby reducing the spread of WNV. By analyzing the timing of outbreaks over multiple years, authorities can understand whether they follow a predictable seasonal pattern or are influenced by certain environmental or climatic conditions. This information can be used to forecast future outbreaks, thereby allowing for preparedness and mitigation planning can evaluate the effectiveness of previous public health interventions and mosquito control efforts. If the frequency or intensity of outbreaks decreases over time, it may indicate that current strategies are effective.

Figure 9 identifies and plots WNV outbreaks, defined as periods with at least 3 consecutive weeks of WNV-positive cases (i.e. identifying the temporal pattern). Outbreaks are highlighted in red on the weekly WNV occurrence plot.

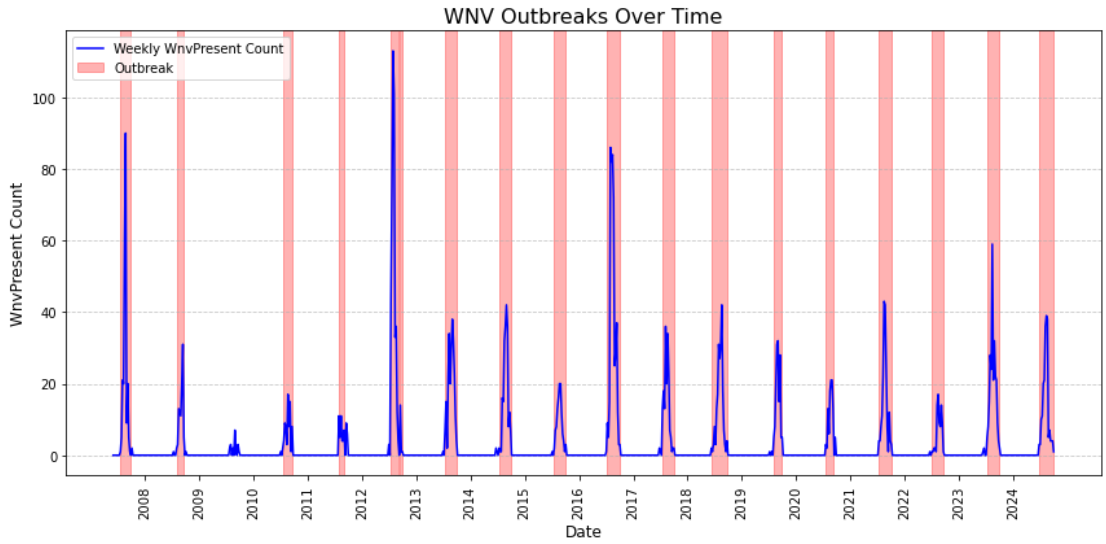

Figure 9. Identifying West Nile Virus Outbreaks Over Time: Highlighted periods indicate consecutive weeks of heightened WNV activity. We consider an outbreak as a period with 3 or more consecutive weeks of number of WNV incidences in any trap >= 1

What is the distribution of WNV positive cases over batch size?

In the Chicago database, 9.15% of the mosquito batches are classified as infected with WNV, meaning some mosquitoes in those batches tested positive for the virus. In Figure 10, we examine the batch sizes when they were found to be WNV-positive. The histogram displays the distribution of the number of mosquitoes in each batch where the virus was detected. This visualization helps reveal the relationship between batch size and WNV presence, offering insights into the data distribution. As expected, larger batches of fifty mosquitoes were more likely to test positive, but positive cases were also observed in smaller batches.

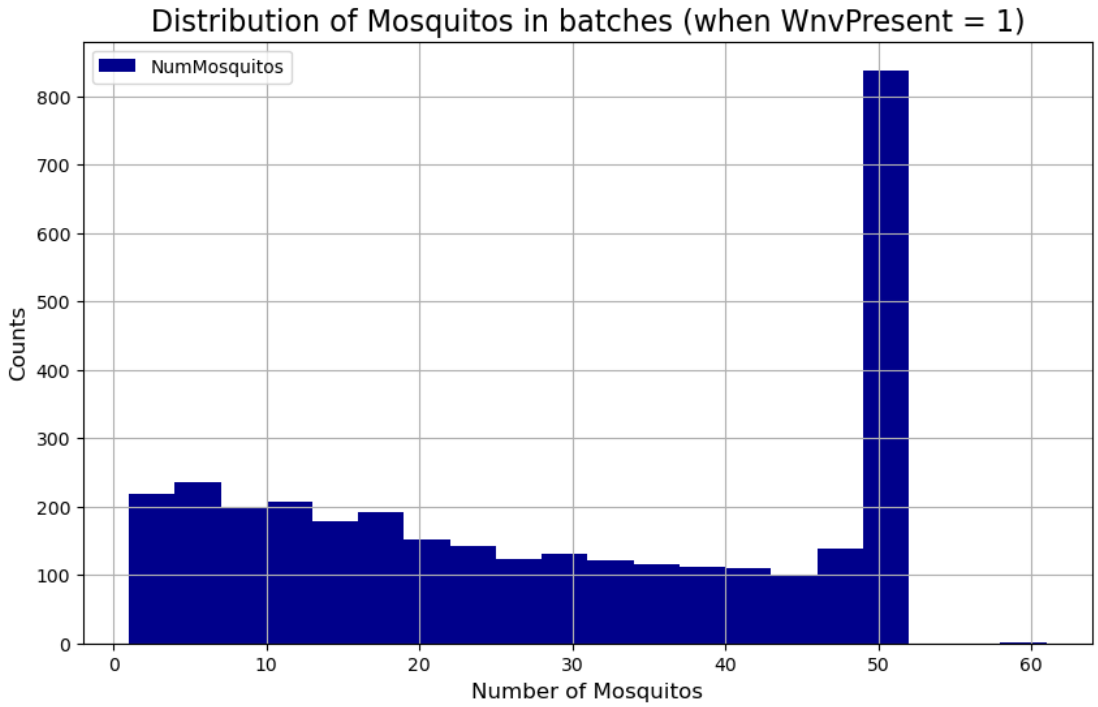

Figure 10. WNV positive cases with respect to the size of the batch in mosquito captures (1-50 specimens).

PRREDICTION

A prediction model in the context of this work would provide a forecast of mosquito catches, but most importantly, infer which batches are going to be found infested based on the rest of the variables. Note that such an approach relies on reliable, historical data. This information can guide public health officials on when to implement interventions such as pesticide spraying, public awareness campaigns including alerting, and vector control measures. By understanding the probability of WNV occurrence over a time span, resources can be optimized, instead of evenly distributing resources in time and locations. This helps in efficient allocation of resources—for example, more frequent mosquito trapping and testing during peak times, reducing resource use during periods with low risk. For instance, if the peak occurrence falls around mid-August, health authorities can plan proactive measures just before this peak, focusing and geographic locations (hotspots) to minimize mosquito populations and, consequently, the transmission of WNV.

In this work we are interested only in the accuracy of a single model implementing a core idea, and we do not examine approaches like stacking or voting of a group of classifiers. We also focus only on the data of the Chicago database, and we do not integrate environmental factors such as spraying records, temperature, precipitation, and humidity, which are not part of this database, but it is known to greatly affect mosquito activity each year.

The Kaggle Competition

A portion of this dataset, spanning from 2007 to 2014, was used in a Kaggle competition where participants were tasked with predicting which mosquito batches were infected with WNV in the years (2008, 2010, 2012, 2014) given the data in years (2007, 2009, 2011, 2013) [46]. In the Kaggle competition, submissions were evaluated using a hidden portion of the dataset, with performance scores displayed publicly on a leaderboard. The final

ranking was based on a separate hidden portion of the data, which was evaluated after the competition deadline (the private part of the dataset).
In that competition, winning participants often used the publicly available leaderboard results to fine-tune their algorithms. If the data distribution of the private portion closely matches that of the public portion, this fine-tuning can be beneficial; otherwise, it may have a detrimental effect on the accuracy of the predictor. In this particular competition, fine-tuning on the public subset proved advantageous and many competing teams used it. However, this approach represents a form of data leakage, where information from the test set influences the shaping of the predictor —a scenario that is unrealistic in real operational settings.
Kaggle competitions are valuable sources of knowledge but several competing practices such as averaging of many models (stacking) and test-set leakage through fitting the leaderboard are not met in real situations. In this work we use the last two years as a test set, and this is a tougher prediction task as the training-test set is not partitioned into adjacent years. Our contributions are the following: a) We examine several ideas presented in the Kaggle winning solution which was sophisticated in view of becoming 'features' in different modeling approaches. In the 2023-2024 test-set they are not that helpful, but we explain the statistics behind the coding as they are complex and not commented on in the literature. b) We introduce a new approach based on a bivariate Normal fitting with trap significance assessment (see Appendix for code and mathematical derivation). c) We refactor the Kaggle approach that now executes at 10% of the original time and is suitable for large databases, and its application to the 2023-2024 test-set after removing the data leakage practices. d) We upload the dataset and the associated code so that different approaches can be tested for different splits of training and testing years of the Chicago database.

The Gaussian as a basic predictor

The Gaussian distribution (or normal distribution) is a common statistical tool used to model natural phenomena. The approach in [46] starts with fitting a Gaussian distribution on the histogram of WNV occurrences to a time span centered around August 1 (see Fig. 11). The 1st of August was selected because it is in the middle of the monitoring period and close to the peak. The value of such a figure lies in its ability to model and estimate the probability of WNV occurrences simply based on time. This base estimator is further refined significantly with subsequent steps that give detail to the Gaussian fit. As an example, the probability of a specific batch (a row in the Chicago database) being infected must be updated on the fact that this batch may be in the middle of an outbreak or near a hotspot or this specific trap is less/more reliable in shaping a probability. That is, the probability of a specific row is affected by the information in other rows. This is analyzed in the Multirow and TrapBias section.
The monitoring period extends from late May to the first week of October. By fitting a Gaussian distribution to data pooled across all years with respect to each day in the monitoring period, one creates a baseline probabilistic model that estimates the likelihood of observing WNV on specific dates, relying solely on the date. This approach has the advantage of simplicity, requiring only two parameters: the mean and standard deviation, which makes it less susceptible to overfitting compared to other classification methods with numerous parameters.
In essence, this model suggests that when a dataset adheres to a straightforward probabilistic structure, using a maximum likelihood approach with a Gaussian model can be more effective. This framework is preferable to complex models, which may easily overfit, especially when dealing with limited data.

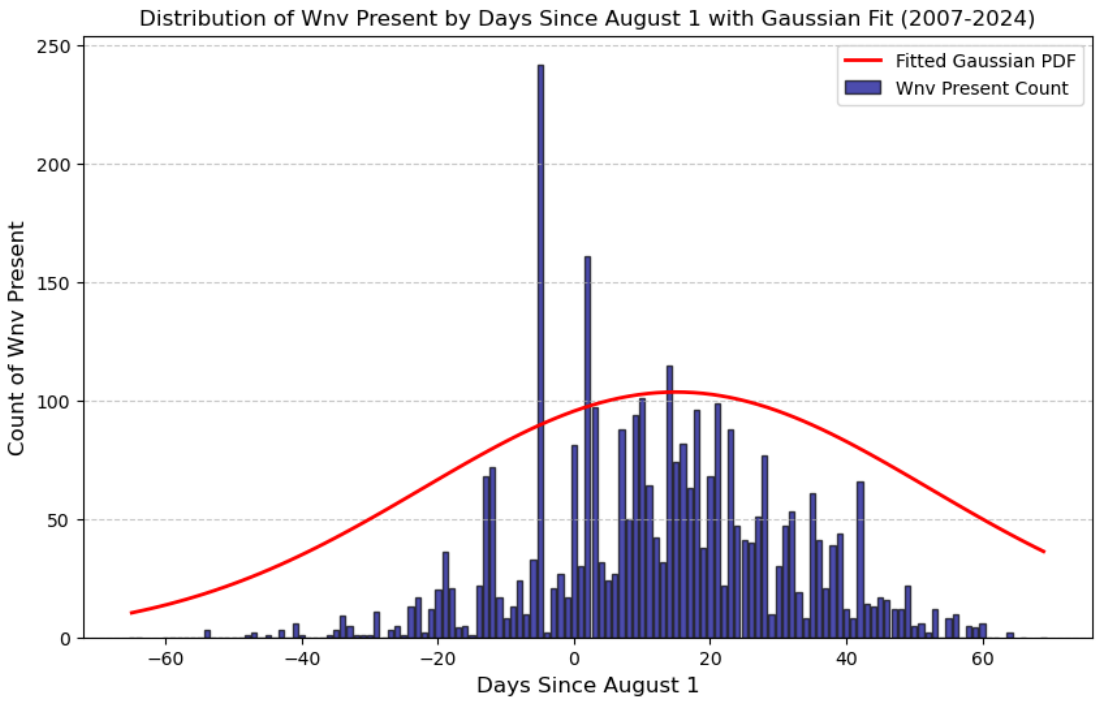

Figure 11. The histogram of the positive West Nile Virus has been re-indexed with respect to 1st of August (day 0) which is approximately in the middle of the monitoring period. A Gaussian is fitted on the positive WNV cases with maximum likelihood approximation of the mean and standard deviation.

We have evolved the base predictor by fitting a bivariate Normal jointly on the variables of the train dataset: 'Dates' re-indexed from 1st of August and 'Number of mosquitoes' (the log of it) for the WNV positive class and WNV negative class. This can be seen in Fig. 12 and Fig. 13 (see also Appendix for code and mathematical derivation).
Using basic Bayesian statistics, we can derive the probability of an infested batch given the test date and test 'NumMosquitos' variable that we have access to assuming that the same bivariate fit holds for the test set. The 'NumMosquitos' variable was not available at the Kaggle's competition test-set, but our view is towards connecting data analytics with automated mosquito traps that report this variable. The suggested approach alone will give an AUC of almost 80% without using any other variable or processing on the database (vs 78.38% for the approach in Fig 11).
This is a result of interest to our point of view as it returns an accuracy very close to more complex classification methods but still is embeddable in microprocessors with few lines of code (see Appendix).
In Fig. 12 we see that the bivariate Gaussian fits on Days and log(Number of Mosquitoes) for WNV positive and negative classes of the train set are partly disjoint. This will allow to extract some information on the probability of a batch being infected and this probability can only be better than the approach in the winning solution (see Fig. 11) as the extra dimension allows the histograms of virus-negative mosquito catches and positive cases to be better separated. Note also in Figs 12 that the peak of WNV positive cases comes a bit after the peak in mosquito catches and is more concentrated in this 2-D feature space.
In Fig. 13 we can easier see gross decision boundaries: a) before August, it is unlikely to have WNV positive batches especially in batches with small number of catches, b) between the last week of August and the first week of September, in batches with high number of mosquitoes, the probability of an infested batch is in its peak.

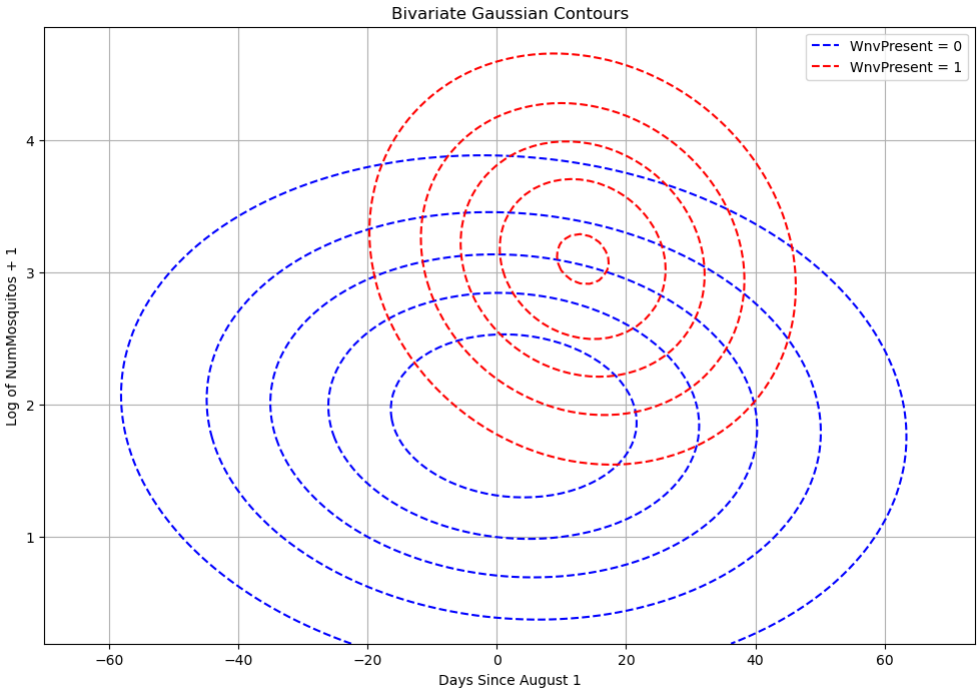

Figure 12. A bivariate Gaussian fits on Days and log(Number of Mosquitoes) for WNV positive and negative classes of the train set. The pdfs are partly disjoint. The bivariate fit is applied on the training set, and it is assumed to characterize also the test. Using the dates and the corresponding number of mosquitoes of the test set we can predict the probability of the infested batches of the test set. Note also that the peak of positive cases comes after about two weeks after the peak in mosquito catches.

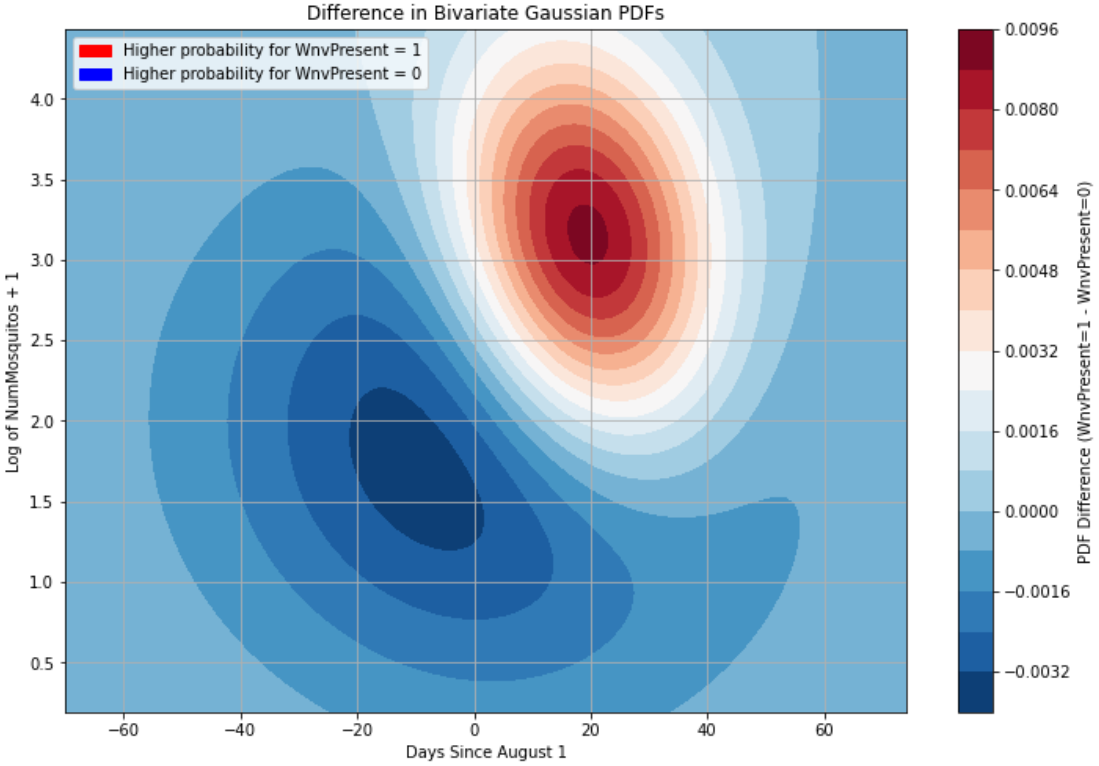

Figure 13. A maximum likelihood plot of the difference between the pdf of WNV positive and the pdf of negative classes. We can identify regions where one decision is more likely than the other (i.e. identify "decision boundaries," which help distinguish between two classes).

The multirow counts

In the Chicago Database, each week the mosquito catches of the traps are grouped in batches of fifty specimens and are subsequently processed manually to identify species and WNV infection status through PCR. Therefore, although we have a single event (i.e. the opening of the catch bug at a specific date and trap), if the number of mosquitoes traps is large, this is catalogued in the database as many rows having the same Species, Trap number, Address, and Date but different numbers of mosquitoes and WnvPresence for each batch. This grouping constitutes a multirow record.
The data is grouped (i.e. a multi-index is created) using columns (Species, Trap, Address, Date). This uniquely identifies different multirows in the dataset. Then, the number of occurrences of each unique multirow in the dataset is computed. This essentially tells us how many records we have for each unique combination of (Species, Trap, Address, Date). In the Kaggle competition, multirow counts have been used as proxy for the total number of mosquitoes as this variable was removed from the test set. However, our work moves towards the direction of connecting data analytics with automated traps and automated traps do transmit the number of captured mosquitoes. Nevertheless, multirow counts constitute a feature of interest and we present it, as it is not encountered in literature. In Table 4 we show the distribution of multirows in the whole database (i.e. grouping together 'Species', 'Trap number', 'Address', and 'Date').

Table 4. Multirow counts in the Chicago 2007-2024 Mosquito Database. These are rows with the same Species, Trap, Address and Date. We show only cases up to 13 rows. Most cases are single entry, but a notable percentage is multirow (i.e. the mosquito bag in a single visit of a specific trap is partitioned in many 50-specimen batches due to its large size).

Multirow Count	# Mosquitos	# Mean cases
1	30614	8.4
2	2596	30.6
3	1026	36.8
4	464	43.2
5	255	44.1
6	234	45.6
7	168	46.0
8	104	46.7
9	108	46.1
10	80	47.8
11	22	48.1
12	48	48.0
13	39	48.1

The trap biases

Another interesting idea in [46] is to calculate Trap-biases. One first calculates the observed WNV incidence rate for each trap relative to the overall incidence rate. Then one computes a p-value using the hypergeometric distribution to assess the statistical significance of the observed WNV cases in each trap. Then adjusts the trap bias by considering both the incidence rate ratio and the p-value, to account for traps with few observations. In the hold out test-set based on the years 2023-2024 they do not help much but again we decide to present them as an interesting feature.

We define in (1) the global ratio of WNV (ratio_WNV_global) that is calculated from the training data and represents the baseline occurrence of WNV across all traps:

$$ratio\_WNV\_global = \frac{num\_WNV}{num\_Total} \quad (1)$$

And the ratio per trap (ratio_WNV_trap):

$$ratio\_WNV\_trap = \frac{num\_WVN\_trap}{num\_Total\_global} \quad (2)$$

We then define the ratio of (1) and (2), weighted with a factor a. This is described in (3)

$$bias = \left(\frac{ratio\_WNV\_trap}{ratio\_WNV\_global}\right)^{a} \quad (3)$$

This ratio indicates whether a trap has a higher or lower mean incidence of WNV compared to the overall mean. If a trap has very few observations, a high or low incidence rate might not be statistically significant. A hypergeometric statistical test assessed the significance of each trap's performance in comparison to the overall data, which adjusts the calculated ratio to avoid overfitting or bias due to small sample sizes. By incorporating statistical significance, this feature aims to provide a reliable and interpretable measure of each trap's effectiveness while mitigating the risk of overfitting to small number of mosquito catches.
The p-value assesses the probability of observing k or more (or k or fewer) WNV cases in the trap under the null hypothesis that WNV cases are randomly distributed among traps. This bias is adjusted through 'a 'by statistical significance.

$$logit(p) = \log\frac{p}{1-p}$$

α = logit(1-prob)/bias_factor. The bias_factor is set to 45. The prob is calculated using the hypergeometric distribution forming essentially a two-tail test (see code in Appendix). In Figure 14, the trap bias of all traps is sorted by value. In the probability estimation task, the Trap Bias of each trap is calculated only for the training set and is applied to the same Traps in the test. If a trap in the test set does not exist in the training set it receives a bias of 1.

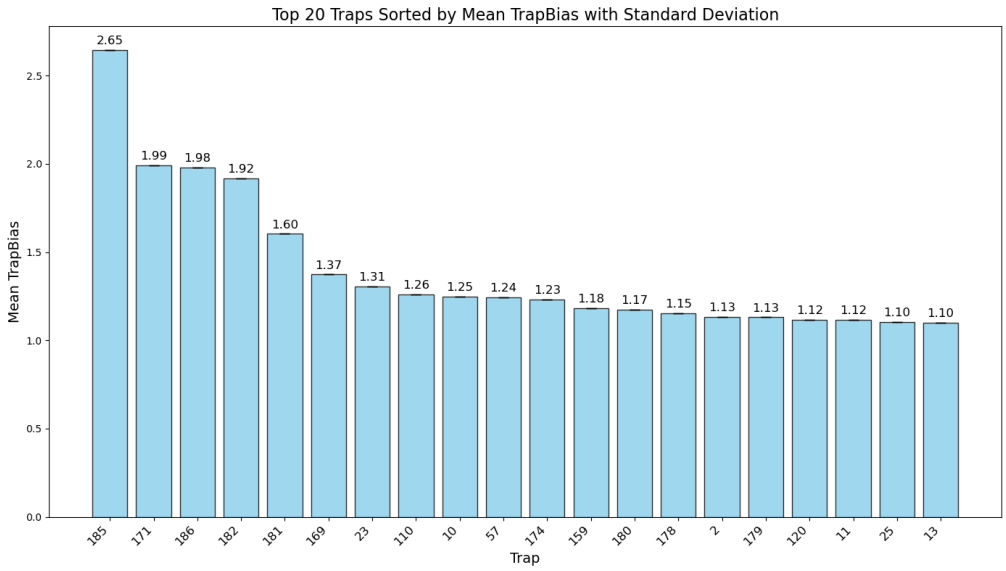

Figure 14. Trap bias of the first 20 traps sorted by value.

Kernel weighted regression by applying distances in time and space

In [46] the code uses a kernel-weighted regression approach to estimate the expected mosquito count for each observation, considering neighboring data points in time and space. Kernel-weighted regression is a type of 'local regression' where data points closer to a given input point are given more weight in estimating the probability of that batch being infected. The code estimates the number of mosquitoes per row based on the count of mosquitoes from similar rows. The "similarity" is defined by temporal proximity, spatial proximity, and optionally species or trap type. The final estimate is computed as a weighted average, where weights are determined by a custom distance function that considers both temporal and spatial distances. The statistical significance of each nearby row's contribution is adjusted by the count of mosquitoes. The estimates are influenced by distance-based weighting (giving more weight to closer rows in time and geographic location). The specific characteristics of the row, including species and trap, determine which nearby observations are relevant. This approach helps in providing a more robust estimate of mosquito counts, particularly when the original data may be sparse or inconsistent across different spatial and temporal dimensions. All these ideas and their influence on the public and private part of the test set are gathered in Table 4 (see also Appendix).

Classification results on the Chicago Kaggle (2007-2014) and 2007-2024 data

In Table 4 we gather the results for the Kaggle dataset which is based on the (2007-2014) Chicago database. In the Appendix, we include the submission files to the Kaggle competition that recreate Table 4.

Table 4. Winning solution on the Kaggle (2007-2014) Chicago database. Given the data in years (2007, 2009, 2011, 2013) predict infected batches in (2008, 2010, 2012, 2014). We show how the different biases affect the AUC score of the base predictor.

Original Cardal	**PUBLIC**	**PRIVATE**	**COMMENT**
1_Normal distribution as base predictor	0.62906	0.62689	Apply a Gaussian on WNV-positive vs date as base predictor
2_Applying yearly biases	0.74882	0.75322	Yearly biases derived from multirow counts
3_Applying species biases	0.79820	0.79360	Species biases derived from multirow counts
4_Applying date_location_trap - based outbreak	0.86284	0.84246	Outbreak biases (leaderboard feedback)
5_Applying trap bias	0.86352	0.84363	Trap biases
6_Applying combined multirow counts probabilities	0.88315	0.85992	Multirow counts probabilities

We partitioned the data into a training set containing all years from 2007 to 2022 and held out the years 2023 and 2024 as a test set. We removed leaderboard fitting practices and applied the refactored code from the Kaggle competition to a different data split. This presents a more challenging scenario, as we need to predict two consecutive years based on training data that includes years from the distant past.

Table 5. Refactored winning solution of the Kaggle competition as applied to the (2007-2024) Chicago database without data-leakage practices. The training set includes the years 2007-2022 and the test set the last two remaining years, 2023 and 2024. AUC score.

Refactored Cardal	**PRIVATE**	**COMMENT**
1_Normal distribution as base predictor	0.7845	Apply a Gaussian on WNV-positive vs date as base predictor
2_Applying yearly biases	0.7843	Yearly biases derived from multirow counts
3_Applying species biases	0.7922	In 2024 there is a fundamental change in Species tagging
4_Applying geo_location outbreak	0.7970	Outbreak biases
5_Applying trap bias	0.7973	Trap biases
6_Applying combined multirow counts probabilities	0.8056	Multirow counts probabilities

Other classifiers

The Chicago Database is a tabular one, and this kind of data structure is typically treated with tree-based classifiers. Tree-based classifiers are particularly effective for structured/tabular data due to their ability to compare features and handle mixed data types, manage non-linear relationships, and deal with missing values. We used the following:

GradientBoostingClassifier is an ensemble machine learning technique that builds a sequence of weak learners (typically decision trees), each correcting the errors made by the previous one. By combining these weak learners, Gradient Boosting can produce a much better predictive model, often achieving high accuracy for both regression and classification tasks. This method iteratively minimizes a loss function (AUC score in our case), making it effective at capturing complex relationships in the data.

XGBClassifier (XGBoost) is a specific implementation of the Gradient Boosting approach. It introduces features like regularization, which helps reduce overfitting. XGBoost is widely applied for its performance in data science competitions due to its accuracy and flexibility in handling a wide range of data types and problems, including those with high dimensionality and class imbalance (like the Chicago database).

ExtraTreesClassifier (Extremely Randomized Trees) is an ensemble method that builds multiple decision trees using random splits of the dataset and random feature selections. Unlike Random Forests, Extra Trees make splits using random thresholds, which introduces more randomness. This often helps improve generalization and reduces overfitting. ExtraTreesClassifier is particularly effective in reducing variance and improving prediction performance, especially in datasets with a high number of features.

HistGradientBoostingClassifier is a variant of Gradient Boosting that employs histogram-based techniques to optimize decision trees. Instead of processing each data point individually, it bins continuous features into discrete intervals, significantly improving training speed, especially on large datasets. All tree-based classifiers have been adjusted for class imbalance.

The following variables have been converted to categorical: 'TRAP_TYPE', 'Species', 'Trap', 'Address', 'COMMUNITY AREA NAME'])

The columns used are ['Block', 'Species', 'TRAP_TYPE', 'Trap', 'Latitude', 'Longitude', 'month', 'week', 'NumMosquitos', 'Address', 'COMMUNITY AREA NAME']

Table 6. Tree-based classifiers as applied to the (2007-2024) Chicago database. The training set includes the years 2007-2022 and the test set the last two remaining years, 2023 and 2024. AUC score.

Model	**AUC score**
GradientBoostingClassifier	0.80
XGBClassifier	0.81
ExtraTreesClassifier	0.80
HistGradientBoostingClassifier	0.81

The ROC curve rises above the diagonal (gray dashed line), indicating that the model is better than random chance in distinguishing between positive and negative classes (see Fig. 13).
The area under the curve (AUC) is 0.81, which suggests that the model has a reasonably good ability to discriminate between the two classes. An AUC closer to 1 would indicate a very strong classifier, while an AUC of 0.5 would represent a classifier that performs no better than random guessing. The True Positive Rate (TPR), also known as sensitivity, is around 0.78 at the operational point. This means that the model correctly identifies approximately 78% of the actual positive cases and suffering 28% of false positives.

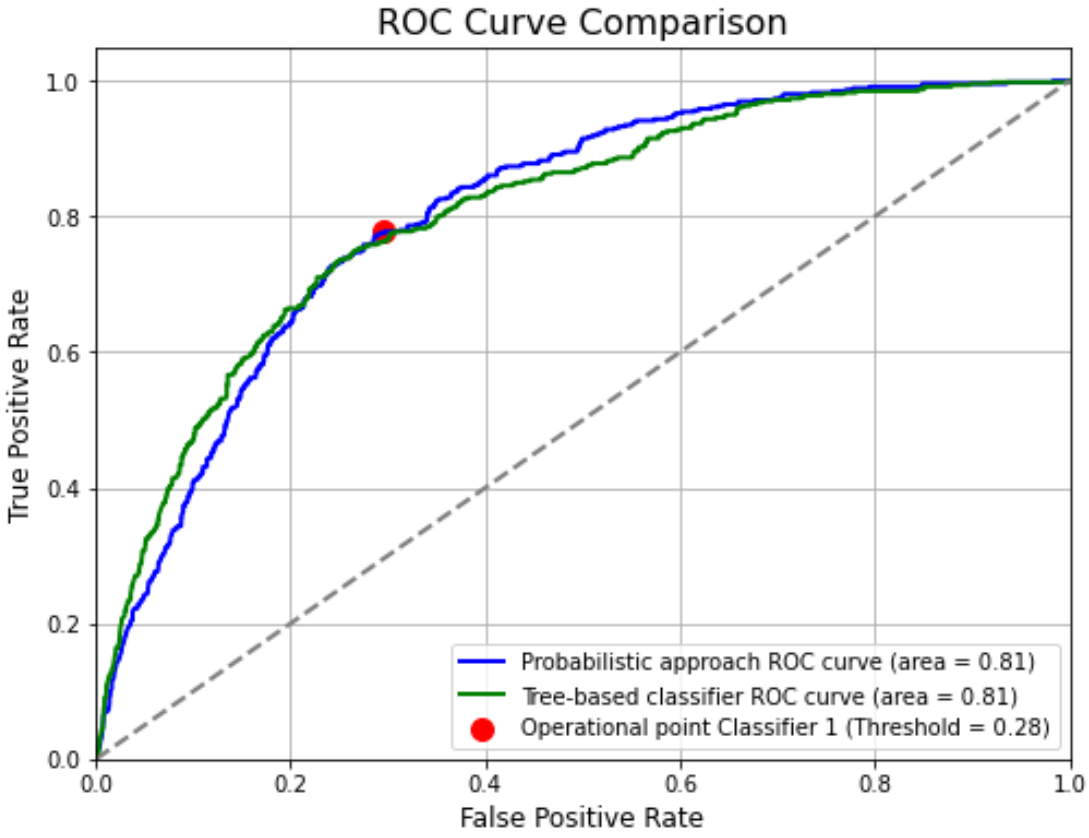

Figure 13. ROC curves of classification approaches. The probabilistic approach and tree-based classification return equal results.

Additional Parameters

Spraying records and weather conditions can significantly influence the probability of a mosquito batch testing positive for WNV. Spraying—a common mosquito control

measure—can directly reduce mosquito populations, particularly those carrying WNV, thereby lowering the likelihood of WNV-positive batches. Weather factors such as temperature, humidity, and rainfall also play a crucial role; warmer temperatures and increased rainfall create favorable conditions for mosquito breeding, potentially raising WNV transmission risk. Therefore, integrating spraying records and weather data into predictive models could enhance accuracy by accounting for these critical environmental factors.

Discussion

The key findings of this paper reveal that maintaining detailed historical data from a widespread network of mosquito traps, including counts of WNV incidents, enables the prediction of future WNV-infected batches. We show that a probabilistic approach is quite effective but is on par with other data-driven, tree-based techniques. This predictive capability is rooted in the analysis of hotspots, species composition, geographic location, and spatiotemporal correlations of outbreaks. While predictions based solely on variables like location, species composition, and mosquito counts do not match the accuracy of PCR-based methods, they significantly outperform random guessing. The accuracy can increase with the incorporation of weather and spraying data.

Another important finding is that detailed records of mosquito catches from a distributed network of traps allow for the rapid identification of hotspots, community names, and specific trap locations responsible for the highest number of infected batches. This information enables swift identification of problematic areas. Rising trends may signal the need for more aggressive control measures, while declining trends could indicate the effectiveness of current practices, providing a reference point for future efforts.

In this study, we analyzed manually entered data, but we propose that commercial, automated mosquito traps could report all key variables from the Chicago database, including GPS coordinates, precise timestamps of mosquito catches, trap addresses, and mosquito counts per hour. Note that the time stamping capabilities of automatic traps are far superior to manual practices as each mosquito is time-stamped at its entrance time allowing to reveal the circadian rhythm of the species, an information that is lost in the manual weekly counting. In principle, species composition could be determined by classifying the wingbeat patterns of incoming mosquitoes. Note that in the Chicago database the main collective Species group is *Culex pipiens/restuans* has not been sorted out denoting the difficulty of this task if carried out manually. Accurately predicting WNV-infected batches requires a long-term network of traps in fixed locations, supplemented by PCR analysis over several years to correlate variables with WNV presence. Current commercial mosquito counting devices need upgrades to address criticisms regarding their varying accuracy, and advanced traps capable of reporting sex, species, and genus must progress from TRL-7 to TRL-9, with independent evaluation on a medium scale at least.

Conclusion

We envision a connected world where extensive, permanently installed networks of mosquito traps are seamlessly integrated with the internet via terrestrial and satellite communication, enabling real-time monitoring of parameters such as mosquito counts, timestamps of captures, and the sex, species, and genus of trapped mosquitoes. Leveraging current data, historical trends, and data analytics, these traps could send alerts directly to mobile phones of individuals near hotspot areas in urban/suburban areas. Based on mosquito species and their known circadian rhythms, these warnings could inform vulnerable visitors near hotspots, such as the elderly, when to avoid outdoor exposure during peak activity times. Additionally, digital signage could display alerts only

during critical periods to avoid overloading the public with frequent alerts, based on the probability of infected batches in traps, reminding people to wear protective clothing and avoid high-risk areas. A public agency that merges reports from automated traps with anonymized hospitalization records of WNV-positive cases would allow for deeper cross-correlation and more comprehensive public health responses.

The large-scale deployment of automated mosquito counters is expected to grow with the advent of cost-effective modems and SIM cards utilizing satellite technology. The integration of advanced data analytics with reliable automated mosquito traps offers accurate predictions of infection likelihood in captured batches, eliminating the need for PCR testing and at no extra cost. Beyond issuing early warnings, these predictions could help public health agencies to efficiently allocate resources for prevention and intervention efforts and secure necessary funding for mosquito control, which remains a top concern for policymakers.

However, significant technical challenges persist in achieving effective automated mosquito monitoring, including maintaining data quality, integrating information from diverse sources, and ensuring timely action on predictions. Addressing these challenges and the social and ethical considerations that come with them will require a multidisciplinary collaboration, combining public health expertise, data science, and ecological knowledge. Comprehensive surveillance data, like that from the Chicago database, is crucial for identifying high-risk areas, understanding mosquito species involved, and evaluating the effectiveness of current monitoring programs.

Appendix

Code

The Chicago database can be found [39] and in our github account [47] we include a slightly preprocessed version.

The original code for the winning solution of the Kaggle competition is in [45]. In [46] one is still able to make late submissions after logging in. A refactored version of ours needing 10% of the original execution time can be found in [47]. In the same link, we include the files to reproduce the findings in Table 4 concerning the Kaggle West Nile Virus prediction competition. We also provide python code for reproducing all figures, calculating the bivariate predictions and running the classification tasks.

Bivariate fitting on WNV counts and Mosquito Counts

We aim to predict the probability of West Nile Virus (WNV) presence (WNV=1) in the test using two features of the database:

x_1: DaysSinceAug1: Number of days since August 1, and

x_2: NumMosquitos: Number of mosquitoes captured.

For the two classes c, $c \in \{0, 1\}$ (WNV=0, WNV=1)

Our approach involves:

a) Modeling the joint distribution of features (x_1, x_2) for each class (WNV = 0 and WNV = 1) independently using a bivariate Gaussian distribution fitted on (x_1, x_2) for each class separately.
b) Computing the likelihoods of the observed data under each class's distribution.
c) Applying Bayes' Theorem to compute the posterior probability that WnvPresent = 1 given the observed features.

Steps (a-c) in mathematical terms are presented below.

We assume that the features (x_1, x_2) follow a bivariate Normal pdf within each class c.

The bivariate Gaussian distribution models the joint probability of two continuous random variables (x_1, x_2). The probability density function (PDF) is:

$$f(\mathbf{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu)\right) \quad (1)$$

where the joint feature vector is $x = (x_1, x_2)$. The mean of **x** in (1) is $\mu = (\mu_1, \mu_2)$
The covariance matrix in (2) is:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (2)$$

The determinant of the covariance matrix is $|\Sigma| = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$

The means of each class *c* are calculated from the features (x_1, x_2) of the training set as in (3):

$$\mu_c = \frac{1}{N_c}\sum_{i=1}^{N_c} x_i \quad (3)$$

The covariance of each class is calculated from the features of the training set as in (4):

$$\Sigma_c = \frac{1}{N_c - 1}\sum_{i=1}^{N_c}(x_i - \mu_c)(x_i - \mu_c)^{\top} \quad (4)$$

The likelihood of each class *c* (i.e. WNV=0 vs WNV=1) is in (5)

$$L_c = P(\, x \mid \text{WnvPresent} = c\,) = f_c(x) \quad (5)$$

The priors of each class *c* are derived from the training set as in (6):

$$P(\text{WnvPresent} = c) = \frac{N_c}{N} \quad (6)$$

The evidence ***P(x)*** is calculated in (7) and requires the priors in (6):

$$P(x) = L_0 \cdot P(\text{WnvPresent} = 0) + L_1 \cdot P(\text{WnvPresent} = 1) \quad (7)$$

The likelihood of each class needed in (7) is calculated in (8):

$$L_c = \frac{1}{2\pi\sqrt{|\Sigma_c|}} \exp\left(-\frac{1}{2}(x-\mu_c)^{\top}\Sigma_c^{-1}(x-\mu_c)\right) \quad (8)$$

The exponent can be calculated in terms of the so called Mahalanobis distance in (9):

$$D_c^2 = (x-\mu_c)^{\top}\Sigma_c^{-1}(x-\mu_c) \quad (9)$$

So (8) is rewritten as in (10) for each class *c* and plugged into (7).

$$L_c = \frac{1}{2\pi\sqrt{|\Sigma_c|}} \exp\left(-\frac{1}{2}D_c^2\right) \quad (10)$$

Finally using Bayes theorem, we calculate the probability of an infested batch as in (11), using (10), and (6):

$$P(\,\text{WnvPresent} = 1 \mid x\,) = \frac{L_1 \cdot P(\text{WnvPresent}=1)}{L_0 \cdot P(\text{WnvPresent}=0) + L_1 \cdot P(\text{WnvPresent}=1)} \quad (11)$$

References

1. World Health Organization (2023). Mosquito-borne diseases. WHO. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases
2. Uelmen, J.A., Lamcyzk, B., Irwin, P. et al. Human biting mosquitoes and implications for West Nile virus transmission. Parasites Vectors 16, 2 (2023). https://doi.org/10.1186/s13071-022-05603-1
3. DeFelice, N., Little, E., Campbell, S. et al. Ensemble forecast of human West Nile virus cases and mosquito infection rates. Nat Commun 8, 14592 (2017). https://doi.org/10.1038/ncomms14592
4. DeFelice NB, Birger R, DeFelice N, et al. Modeling and Surveillance of Reporting Delays of Mosquitoes and Humans Infected with West Nile Virus and Associations With Accuracy of West Nile Virus Forecasts. JAMA Netw Open. 2019;2(4):e193175. doi:10.1001/jamanetworkopen.2019.3175
5. Villena, O.C., McClure, K.M., Camp, R.J. et al. Environmental and geographical factors influence the occurrence and abundance of the southern house mosquito, Culex quinquefasciatus, in Hawai'i. Sci Rep 14, 604 (2024). https://doi.org/ 10.1038/s41598-023-49793-9
6. Petruff, T.A., McMillan, J.R., Shepard, J.J. et al. Increased mosquito abundance and species richness in Connecticut, United States 2001–2019. Sci Rep 10, 19287 (2020). https://doi.org/10.1038/s41598-020-76231-x
7. Haddawy P, Wettayakorn P, Nonthaleerak B, Su Yin M, Wiratsudakul A, Schöning J, et al. (2019) Large scale detailed mapping of dengue vector breeding sites using street view images. PLoS Negl Trop Dis 13(7): e0007555. https://doi.org/10.1371/journal.pntd.0007555.
8. Karki S, Brown WM, Uelmen J, Ruiz MO, Smith RL (2020) The drivers of West Nile virus human illness in the Chicago, Illinois, USA area: Fine scale dynamic effects of weather, mosquito infection, social, and biological conditions. PLoS ONE 15(5): e0227160. https://doi.org/10.1371/journal.pone.0227160
9. Smith DL, Battle KE, Hay SI, Barker CM, Scott TW, McKenzie FE (2012) Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens. PLoS Pathog 8(4): e1002588. https://doi.org/10.1371/journal.ppat.1002588
10. Monaghan AJ, et al,. On the Seasonal Occurrence and Abundance of the Zika Virus Vector Mosquito Aedes Aegypti in the Contiguous United States. PLoS Curr. 2016 Mar 16;8: doi: 10.1371/currents.outbreaks .50dfc7f46798675fc63e7d7da563da76.
11. Bowman LR, Runge-Ranzinger S, McCall PJ (2014) Assessing the Relationship between Vector Indices and Dengue Transmission: A Systematic Review of the Evidence. PLoS Negl Trop Dis 8(5): e2848. https://doi.org/10.1371/journal.pntd.0002848
12. Dale PE, Ritchie SA, Territo BM, Morris CD, Muhar A, Kay BH. An overview of remote sensing and GIS for surveillance of mosquito vector habitats and risk assessment. J Vector Ecol. 1998 Jun;23(1):54-61. PMID: 9673930.
13. Xu J., and Wang X.Y., and Zhou Y.L., 2024, Applications of geographic information systems in mosquito monitoring, Journal of Mosquito Research, 14(3): 161-171 (doi: 10.5376/jmr.2024.14.0016)
14. Brown HE, Sedda L, Sumner C, Stefanakos E, Ruberto I, Roach M. Understanding Mosquito Surveillance Data for Analytic Efforts: A Case Study. J Med Entomol. 2021 Jul 16;58(4):1619-1625. doi: 10.1093/jme/tjab018. PMID: 33615382; PMCID: PMC8285009.

15. Moutinho, S.; Rocha, J.; Gomes, A.; Gomes, B.; Ribeiro, A.I. Spatial Analysis of Mosquito-Borne Diseases in Europe: A Scoping Review. Sustainability 2022, 14, 8975. https://doi.org/10.3390/su14158975
16. Sheard Julie Koch, et al., 2024. Emerging technologies in citizen science and potential for insect monitoring. Phil. Trans. R. Soc. B37920230106, https://doi.org/10.1098/rstb.2023.0106
17. Ananya Joshi, Clayton Miller, Review of machine learning techniques for mosquito control in urban environments, Ecological Informatics, Volume 61, 2021, 101241, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2021.101241.
18. Lee, DS., Lee, DY. & Park, YS. Interpretable machine learning approach to analyze the effects of landscape and meteorological factors on mosquito occurrences in Seoul, South Korea. *Environ Sci Pollut Res* **30**, 532–546 (2023). https://doi.org/10.1007/s11356-022-22099-5
19. Chevalier V, Tran A, Durand B. Predictive modeling of West Nile virus transmission risk in the Mediterranean Basin: how far from landing? Int J Environ Res Public Health. 2013 Dec 20;11(1):67-90. doi: 10.3390/ijerph110100067.
20. Linus Früh, Helge Kampen, Antje Kerkow, Günter A. Schaub, Doreen Walther, Ralf Wieland, Modelling the potential distribution of an invasive mosquito species: comparative evaluation of four machine learning methods and their combinations, Ecological Modelling, Volume 388, 2018, Pages 136-144, ISSN 0304-3800, https://doi.org/10.1016/j.ecolmodel.2018.08.011.
21. Odu Nkiruka, Rajesh Prasad, Onime Clement, Prediction of malaria incidence using climate variability and machine learning, Informatics in Medicine Unlocked, Volume 22, 2021, 100508, ISSN 2352-9148, https://doi.org/10.1016/j.imu.2020.100508.
22. Ana Ceia-Hasse, Carla A. Sousa, Bruna R. Gouveia, César Capinha, Forecasting the abundance of disease vectors with deep learning, Ecological Informatics, Volume 78, 2023, 102272, ISSN 1574-9541, https://doi.org/10.1016/j.ecoinf.2023.102272.
23. Ralf Wieland, Katrin Kuhls, Hartmut H.K. Lentz, Franz Conraths, Helge Kampen, Doreen Werner, Combined climate and regional mosquito habitat model based on machine learning, Ecological Modelling, Volume 452, 2021, 109594, ISSN 0304-3800, https://doi.org/10.1016/j.ecolmodel.2021.109594.
24. Md. Siddikur Rahman, Chamsai Pientong, Sumaira Zafar, Tipaya Ekalaksananan, Richard E. Paul, Ubydul Haque, Joacim Rocklöv, Hans J. Overgaard, Mapping the spatial distribution of the dengue vector Aedes aegypti and predicting its abundance in northeastern Thailand using machine-learning approach, One Health, Volume 13, 2021, 100358, ISSN 2352-7714, https://doi.org/10.1016/j.onehlt.2021.100358.
25. Diing D.M. Agany, Jose E. Pietri, Etienne Z. Gnimpieba, Assessment of vector-host-pathogen relationships using data mining and machine learning, Computational and Structural Biotechnology Journal, Volume 18, 2020, Pages 1704-1721, ISSN 2001-0370, https://doi.org/10.1016/j.csbj.2020.06.031.
26. Santos LMB, et al., High throughput estimates of Wolbachia, Zika and chikungunya infection in Aedes aegypti by near-infrared spectroscopy to improve arbovirus surveillance. Commun Biol. 2021 Jan 15;4(1):67. doi: 10.1038/s42003-020-01601-0.
27. Sikulu-Lord MT, Milali MP, Henry M, Wirtz RA, Hugo LE, Dowell FE, et al. (2016) Near-Infrared Spectroscopy, a Rapid Method for Predicting the Age of Male and Female Wild-Type and *Wolbachia* Infected *Aedes aegypti*. PLoS Negl Trop Dis 10(10): e0005040. https://doi.org/10.1371/journal.pntd.0005040
28. Goh B, Ching K, Soares Magalhães RJ, Ciocchetta S, Edstein MD, Maciel-de-Freitas R, et al. (2021) The application of spectroscopy techniques for diagnosis of malaria parasites and arboviruses and surveillance of mosquito vectors: A systematic review

and critical appraisal of evidence. PLoS Negl Trop Dis 15(4): e0009218. https://doi.org/10.1371/journal.pntd.0009218
29. Fernandes, J. N. et al. Rapid, noninvasive detection of Zika virus in Aedes aegypti mosquitoes by near-infrared spectroscopy. Sci. Adv. 4, eaat0496 (2018).
30. Sikulu-Lord, M. T. et al. Rapid and non-destructive detection and identification of two strains of Wolbachia in Aedes aegypti by near-infrared spectroscopy. PLoS Negl. Trop. Dis. 10, e0004759 (2016).
31. Uelmen, J.A., Clark, A., Palmer, J. et al. Global mosquito observations dashboard (GMOD): creating a user-friendly web interface fueled by citizen science to monitor invasive and vector mosquitoes. Int J Health Geogr 22, 28 (2023). https://doi.org/10.1186/s12942-023-00350-7
32. Lai Z, Wu J, Xiao X, Xie L, Liu T, Zhou J, et al. (2022) Development and evaluation of an efficient and real-time monitoring system for the vector mosquitoes, *Aedes albopictus* and *Culex quinquefasciatus*. PLoS Negl Trop Dis 16(9): e0010701. https://doi.org/10.1371/journal.pntd.0010701
33. Kim, D., DeBriere, T.J., Cherukumalli, S. *et al.,* Infrared light sensors permit rapid recording of wingbeat frequency and bioacoustic species identification of mosquitoes. *Sci Rep* **11**, 10042 (2021). https://doi.org/10.1038/s41598-021-89644-z
34. González-Pérez, M.I., Faulhaber, B., Williams, M. *et al.,* A novel optical sensor system for the automatic classification of mosquitoes by genus and sex with high levels of accuracy. *Parasites Vectors* **15**, 190 (2022). https://doi.org/10.1186/s13071-022-05324-5
35. González-Pérez, M.I., Faulhaber, B., Aranda, C. *et al.* Field evaluation of an automated mosquito surveillance system which classifies *Aedes* and *Culex* mosquitoes by genus and sex. *Parasites Vectors* **17**, 97.2024. https://doi.org/10.1186/s13071-024-06177-w
36. Johnson, B.J., Weber, M., Al-Amin, H.M. *et al.,* Automated differentiation of mixed populations of free-flying female mosquitoes under semi-field conditions. *Sci Rep* **14**, 3494 (2024). https://doi.org/10.1038/s41598-024-54233-3
37. Corey A. Day, Stephanie L. Richards, Michael H. Reiskind, Michael S. Doyle, Brian D. Byrd, Context-Dependent Accuracy of the BG-Counter Remote Mosquito Surveillance Device in North Carolina, J Am Mosq Control Assoc (2020) 36 (2): 74–80. https://doi.org/10.2987/19-6903.1
38. Rauhöft, L., Șuleșco, T., Martins Afonso, S.M. *et al.* Large-scale performance assessment of the BG-Counter 2 used with two different mosquito traps. *Parasites Vectors* **17**, 273 (2024). https://doi.org/10.1186/s13071-024-06338-x
39. https://data.cityofchicago.org/Health-Human-Services/West-Nile-Virus-WNV-Mosquito-Test-Results/jqe8-8r6s/about_data (accessed at 26/10/2024)
40. Gorsich, E.E., Beechler, B.R., van Bodegom, P.M. *et al.* A comparative assessment of adult mosquito trapping methods to estimate spatial patterns of abundance and community composition in southern Africa. *Parasites Vectors* **12**, 462 (2019). https://doi.org/10.1186/s13071-019-3733-z
41. I. Potamitis, I. Rigakis, N. Vidakis, M. Petousis, M. Weber, Affordable bimodal optical sensors to spread the use of automated insect monitoring, J. Sens., 2018 (2018), pp. 1-25, 10.1155/2018/3949415
42. Rigakis, I.; Potamitis, I.; Tatlas, N.-A.; Livadaras, I.; Ntalampiras, S. A Multispectral Backscattered Light Recorder of Insects' Wingbeats. *Electronics* **2019**, *8*, 277. https://doi.org/10.3390/electronics8030277
43. Rydhmer, K., Bick, E., Still, L. *et al.* Automating insect monitoring using unsupervised near-infrared sensors. *Sci Rep* **12**, 2603 (2022). https://doi.org/10.1038/s41598-022-06439-6

44. Saha, T.; Genoud, A.P.; Park, J.H.; Thomas, B.P. Temperature Dependency of Insect's Wingbeat Frequencies: An Empirical Approach to Temperature Correction. Insects 2024, 15, 342. https://doi.org/10.3390/insects15050342
45. https://github.com/Cardal/Kaggle_WestNileVirus (accessed at 26/10/2024)
46. https://www.kaggle.com/c/predict-west-nile-virus (accessed at 26/10/2024)
47. https://github.com/potamitis123/Chicago-Mosquito-Database (accessed at 26/10/2024)