Optimizing Trap Placement to Predict West Nile Virus Cases

Anwesha Chakravarti, Bo Li, Dan Bartlett, Patrick Irwin, Rebecca Smith

Abstract

The rapid spread of West Nile Virus (WNV), the primary cause of mosquito-borne illness in the continental United States, is a growing concern. The lack of vaccines or medications to treat WNV makes prevention through mosquito control the only solution for controlling the spread of the infection. Mosquito traps that are used to test for the presence of the virus in mosquito populations play a crucial role in monitoring WNV and informing response. But how do you decide where to place a mosquito trap for WNV surveillance? And what makes a good trap location, anyway?

We present a statistical approach to determine the ability of a mosquito trap to predict human WNV cases in the next two weeks within a 1500m radius of the trap. We then use that value to understand what landscape, demographic and socioeconomic factors cause a mosquito trap to have the ability to predict correctly when human cases will occur, and when they will not, in the Chicago metropolitan area and its suburbs. This approach enables resource-limited mosquito control programs to identify better locations for their trap-based surveillance to increase trap efficiency while reducing the number of traps needed.

1 Introduction

Since its emergence in the New York Metropolitan area in the fall of 1999, the West Nile Virus (WNV) has been spreading rapidly. With over 52,000 cases nationwide to date, among which around 2,700 are from Illinois, WNV is the leading cause of mosquito-borne disease in the continental United States. Thus, taking adequate measures for the control of WNV has become vital. The primary mode of transmission for this virus is through host-seeking female *Culex* spp. mosquitoes. Cases in Illinois are mainly seen from late summer to early fall, as this is the peak of both mosquito season and WNV infection within mosquitos and the primary avian hosts. Most people infected by WNV show no symptoms, which results in under-reporting of WNV infection and limited information about the true extent of the spread of the infection in humans. The absence of vaccines or medications to treat the virus makes prevention through mosquito control the only solution for controlling the spread of the infection.

The use of mosquito traps is vital for monitoring the spread of the virus and providing information for response efforts. To monitor mosquito populations for WNV, Mosquito Abatement Districts (MADs) and Public Health Departments across different spatial extents deploy mosquito traps throughout their jurisdictions. These traps capture a variety of mosquito species, from which the *Culex* spp. are identified and collected in pools. These pools are then tested for the presence of the virus. The resulting data from trap testing is used to predict the risk of a human case in the vicinity of the location of the trap, which is then used to inform mosquito abatement practices.

Currently, mosquito traps are placed using a grid-based approach and the use of each of these traps can be expensive in terms of labor. In areas where trap deployment is minimal due to lack of resources, this grid approach results in poor surveillance accuracy and may limit predictive performance for WNV cases. Locating the best areas for trap placement can improve the effectiveness of each individual trap. Thus, identifying optimal locations for trap deployment can help health departments optimize their surveillance efforts by reducing the number of traps needed. In this work, we aim to identify a statistical approach to determine optimal trap placement. This line of research is quite novel in the field of West Nile Virus mosquito modelling and little research has been done in this direction.

[BL: The above is the scientific motivation why we study this problem. After that, should do literature review of related work on trap studies, and then briefly describe what you would propose to do, then literature review of related work to the method you would use.]

2 Dataset

Our dataset spans from 2004 to 2018 and comprises data collected from traps placed in Chicago and its surrounding suburbs (Figure 1). For our analysis, we consider a total of 1062 traps distributed across 275 distinct zip codes, with Cook and DuPage counties having the highest trap density. Pools of mosquitoes are typically collected from these traps every week, with some traps having multiple pools of mosquitoes collected on the same day, aligning with the WNV testing guidelines that recommend pools of 50 or fewer mosquitoes. Subsequently, each mosquito pool undergoes testing for the presence of WNV. If any individual mosquito within a pool is identified with the WNV virus, the entire pool is classified as positive. Note that this implies that not every mosquito in a pool undergoes testing. **[BL: What is the measure for testing results? may introduce MLE? we may need to exlpain a little why thsi MLE is not our MLE as this can be confused with our MLE.]** The variables in our dataset can be broadly divided into four categories:

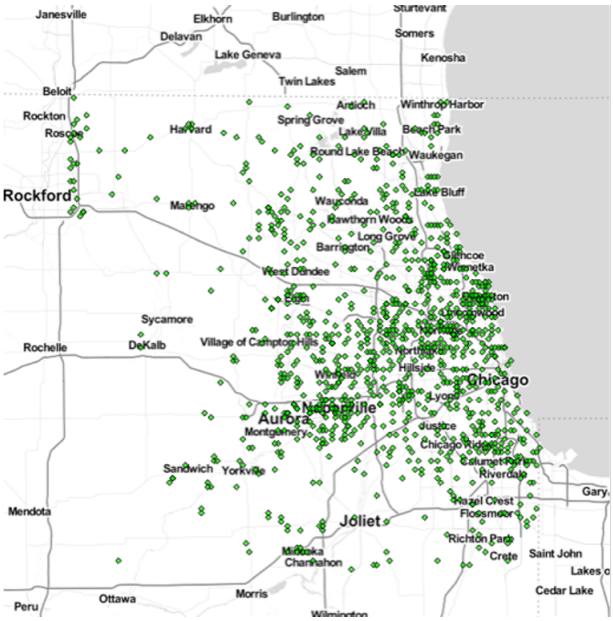


Figure 1: Location of the traps

- **Mosquito related data** contains information on the mosquito pools, the test results for each of these pools and whether a human case was detected in the vicinity of the traps.
- **Landcover data** contains data on the landcover characteristics of the trap sites. It includes variables like canopy cover, percentage of impervious land, and indicators for whether the traps were placed in highly developed areas, near open water bodies, etc.
- **Demographic data** contains information on the total population, population by race, age and so on, in a 1500m buffer area around the trap.
- **Socio-economic data** contains data on poverty levels, percentage occupied housing and education levels of the population in a 1500m buffer area around the trap.

The mosquito-related data was provided to us by the Illinois Department of Public Health. The landcover data has been collected from the National Land Cover Dataset (NLCD)[1] (Dewitz, 2019). The Census block-level population data has been taken from the U.S. Census Bureau (2021)[2]. The Census data for educational attainment and poverty has been collected from the U.S. Census Bureau (2020)[3].

[1]Data available at https://www.mrlc.gov/data/nlcd-2019-land-cover-conus.

[2]Data available at https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html.

[3]Data available at https://www.census.gov/newsroom/press-kits/2020/acs-5-year.html.

3 Modelling Methodology

A major challenge in determining the best trap locations is quantifying the "goodness" of a trap's placement. Our proposed method addresses this problem by assigning a score to each trap based on the trap's historical performance in predicting human cases of WNV, followed by utilizing this score to determine the locations that correspond to traps with a higher score. This method enables us to identify trap locations that contain traps that perform well or poorly at predicting human cases. Data associated with these locations can then be used to determine the characteristics of a good trap location. To achieve this, we present a three-phase approach

- **Phase 1:** Model historic performance of the traps for predicting human cases.
- **Phase 2:** Define score for individual traps.
- **Phase 3:** Find characteristics of locations which correspond to higher trap scores.

3.1 Phase 1: Modelling the historic performance of the traps

To determine the historical performance of the traps, data on the mosquito pools collected from the traps is used to model the incidence of human cases of WNV in the proximity of the traps. The observations in this model correspond to a particular mosquito pool collected from a trap. Thus, a trap has multiple observations corresponding to it, each of them denoting a particular mosquito pool. The response variable in this model indicates whether a human case was detected within a 1.5 km radius around the trap within two weeks of the pool's collection. For instance, if a pool was collected during week 31, the response indicates the incidence of a human case during weeks 32 and 33. We use this leading response variable due to the biology of mosquito-borne spreading of WNV, and due to the latency period of human WNV infections. The variables used in the model are as follows:

- **Pool Size** - Number of mosquitoes in the pool.
- **Test Indicator** - Indicator variable denoting whether the pool tested positive for the presence of WNV.
- **Risk** - The risk associated with a pool. Since not every mosquito undergoes testing, and the prevalence of the virus in the tested pool remains unknown, we require a measure to estimate the risk associated with a given pool. To do so, we first estimate the prevalence of infected mosquitoes in the pool. This is done using the Maximum Likelihood Estimation (MLE) method. The MLE method is the standard method in WNV surveillance to

estimate the prevalence of the virus in a pool. Details regarding the MLE method and its advantages can be found in Gu et al. (2008). Given the MLE value, the risk (for a pool of a trap) is defined using the Vector Index equation in mosquito surveillance:

$$Risk = \frac{Daily\ Abundance * MLE}{1000}$$

where

$$Daily\ Abundance = Avg\ Abundance \times\ Number\ of\ pools\ in\ that\ trap\ on\ that\ day;$$

$$Avg\ Abundance = \frac{Number\ of\ total\ mosquitoes\ caught\ by\ that\ trap\ in\ that\ week}{Number\ of\ pools\ in\ that\ trap\ in\ that\ week}$$

- **Week and Year** - Week in the year, and the year when the pool was collected.
- **Latitude and Longitude** - Location of the trap.

3.1.1 Model description and evaluation

We model the data using a Spatial Generalized Linear Mixed-effects Model (Spatial GLMM) with a binomial family. In simpler terms, we use a logistic regression model with a spatial component. Given the proximity of the traps, we expect to observe spatial correlation in our data, as closely situated traps are likely to exhibit similar behaviour. To handle these spatial correlations, we use a GLMM with spatially autocorrelated random effects. The spatial random effects are taken over the location of the traps and are modelled using a Matern kernel. The rest of the variables described above are incorporated into our model as fixed effects. Thus, our statistical model is as follows:

$$Y_{pool} \sim Poolsize + TestInd + Risk + Week + Year + Matern(1|long + lat),$$

where Y_{pool} is the indicator response variable for whether a human case was detected in a 1.5 km radius of the trap within a two-week timeframe following the collection of the pool. The fixed effect and the random effects parameters, which include the scale parameter ρ and the smoothness parameter ν of the Matern kernel, are estimated via likelihood-based estimators using the R package "spaMM" (Rousset and Ferdy, 2014).

The performance of the traps is measured using the sensitivity and specificity values as metrics. These values are calculated through a 5-fold cross-validation. Since we want to evaluate the historical performance of the traps, the data for a particular year is divided into training and testing sets in a 4:1 ratio. The model is applied to the training data and the predictions

for the testing data are evaluated to calculate the sensitivity and specificity values for a particular trap. This is repeated across the 5-fold cross-validation, and the average sensitivity and specificity scores for each of the traps are computed. Mathematically, this can be described as follows:

$$AvgSens_t = \frac{1}{5}\sum_{cv=1}^{5}\left(\frac{TP_t}{P_t}\right)_{cv} \quad and \quad AvgSpec_t = \frac{1}{5}\sum_{cv=1}^{5}\left(\frac{TN_t}{N_t}\right)_{cv},$$

where TP_t = number of positive cases corresponding to trap t, that are predicted as positive, TN_t = number of negative cases corresponding to trap t, that are predicted as negative, P_t = total number of positive cases corresponding to trap t, and N_t = total number of negative cases corresponding to trap t. **[BL: Write out your statistical model]**

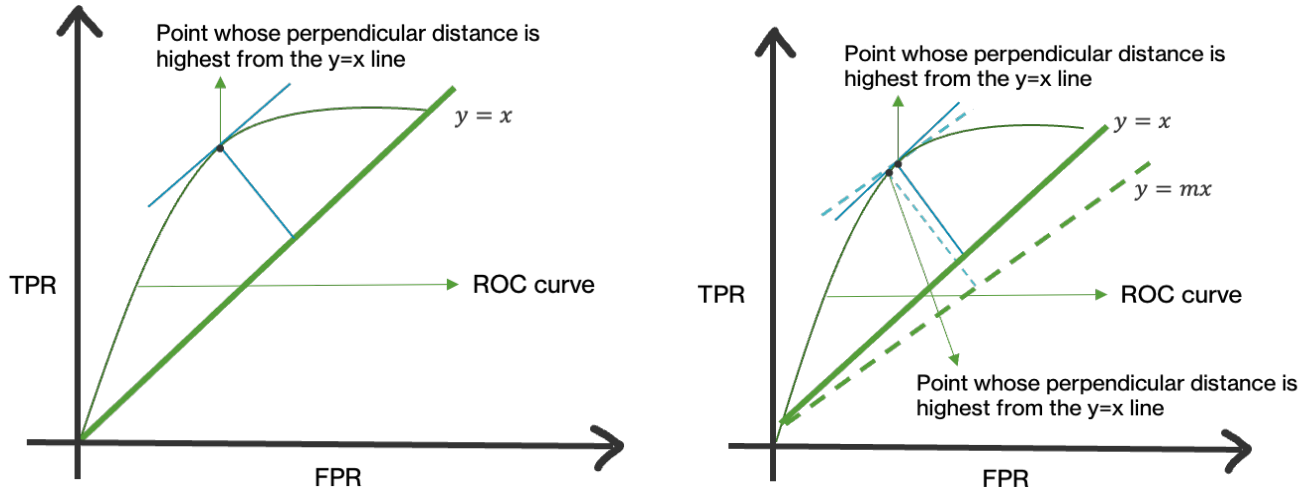
3.1.2 Problem of class imbalance and thresholding

Since the occurrence of human cases of WNV is low, the model suffers from the problem of class imbalance. In particular, we have 184101 instances of $y = 0$ compared to only 2951 instances of $y = 1$. Consequently, the output probabilities predicted by the logistic regression model tend to be higher for the event $y = 0$. An effective way to deal with this is to set an appropriate threshold (instead of the default of 0.5) on the predicted probabilities to obtain the predicted class labels of 0 or 1 (Goorbergh et al., 2022). The best threshold can be computed using the ROC (Receiver operator characteristics) curve (Figure 2). The ROC curve plots the False Positive Rate (FPR) on the X–axis vs the True Positive Rate (TPR) on the Y-axis at different thresholds. A random classifier model would have the y = x line as its ROC curve. Thus, to get the best threshold, we choose the point which has the highest perpendicular distance from the $y = x$ line (Figure 2a).

Due to the nature of our data, our priority is to accurately predict the occurrence of human cases, even if it may come at the expense of predicting some false negatives. In other words, we want to increase our sensitivity (True Positive Rate) at the expense of reduced specificity (True Negative Rate). Thus, we compute the best threshold using the point with the highest perpendicular distance from the $y = mx$ line where $m < 1$, since the $y = mx$ line, is the ROC curve for a random classifier with specificity = $m \times 100\%$ as important as sensitivity (Figure 2b). In particular, we compute the best threshold to be the point

$$argmax(TPR - m * FPR) \quad \text{where } m < 1$$

[BL: should this be perpendicular distance?] The parameter m makes our methodology flexible to the needs of the mosquito control organizations. The value of m can be chosen through a



(a) Sensitivity and specificity having equal importance

(b) Specificity set as m times as important as sensitivity with $m < 1$.

Figure 2: Choosing the threshold using ROC curve. **[BL: both distances drawn here are not perpendicular. In plot (b), the label y=x and y=mx should be switched?]**

cost analysis of losses incurred due to false positives as opposed to losses incurred due to false negatives. The final prediction labels are then assigned as: predicted $P(Y = 1) > threshold$ then $Y_{pred} = 1$, else $Y_{pred} = 0$.

3.2 Phase 2: Defining a score for each of the traps

Considering the model developed in Phase 1 as a baseline, we use the average sensitivity and specificity values computed for each of the traps (as defined in Section 3.1.1) to assign a score to the individual traps:

$$score_t = \frac{m * AvgSpec_t + AvgSens_t}{m + 1}, \tag{1}$$

where $m < 1$ is as defined in Section 3.1.2. The defined score is a weighted mean of the sensitivity and specificity values of the trap, with the weight m controlling the importance we give to false negatives as compared to the false positives. We choose a weighted mean for the score after analysing the results of Phase 1, further details regarding which are outlined in Section 4.1.

3.3 Phase 3: Finding locations corresponding to higher trap scores

We now return to our primary objective of finding the best locations to place the traps. To do so, we consider the demographic, socio-economic and landcover variables associated with

each trap, and identify the variables that may cause the score of a trap, as calculated in Phase 2, to be higher or lower. Specifically, we analyze the causal effect of the variables on the score and not just an association between them. Let us denote the covariates as $T = (T_1, T_2, \ldots, T_p)$ and let t_i and t_i' ($t_i \neq t_i'$) be two possible values that the covariate T_i takes. Then the causal effect of moving covariate T_i from t_i to t_i' is $Y(t_i) - Y(t_i')$. Since our covariates are continuous variables, it is not feasible to estimate the causal effect for any such pair of values. Thus we estimate the average causal effect for moving the covariate T_i from t_i to t_i' using

$$\mathbf{E}\left(Y(t_i)\right) - \mathbf{E}\left(Y(t_i')\right),$$

where $Y(t_i)$ is the potential outcome when covariate T_i takes the value t_i. A practical approach to computing the average causal effect for moving between any pair of values of a covariate is to calculate the average dose-response function (also known as the exposure-response function). The average dose-response function (ADRF) is defined as

$$\mu_{T_i} = \mathbf{E}\left(Y(t_i)\right).$$

It contains information on all comparisons for the average causal effect. A vital part of any causal inference is identifying the confounder variables which could affect the estimate of the causal effect of a covariate on the score. Thus, we draw a directed acyclic graph (DAG) (shown in figure 3) to graphically represent and visualize the causal relationships between the variables in our model. We use the R package "causaldrf" (Schafer, 2022; Galagate, 2016) to estimate the ADRF. In particular, we use a doubly robust method for estimating the ADRF using a generalized additive model (GAM) estimator. Further details on this method of estimation can be found in Galagate (2016); Hirano and Imbens (2004) and Flores et al. (2012). The ADRF helps us identify the variables which have an impact on the score and thus points us towards a good trap location.

4 Results

4.1 Sensitivity and specificity values from Phase 1 and scores of the traps from Phase 2

We analyze the average sensitivity and specificity values that we compute from Phase 1 of our method (Section 3.1) to understand our data and to present the intuition behind our choice of score. In our data, there exists many traps which have not had any human cases of WNV in

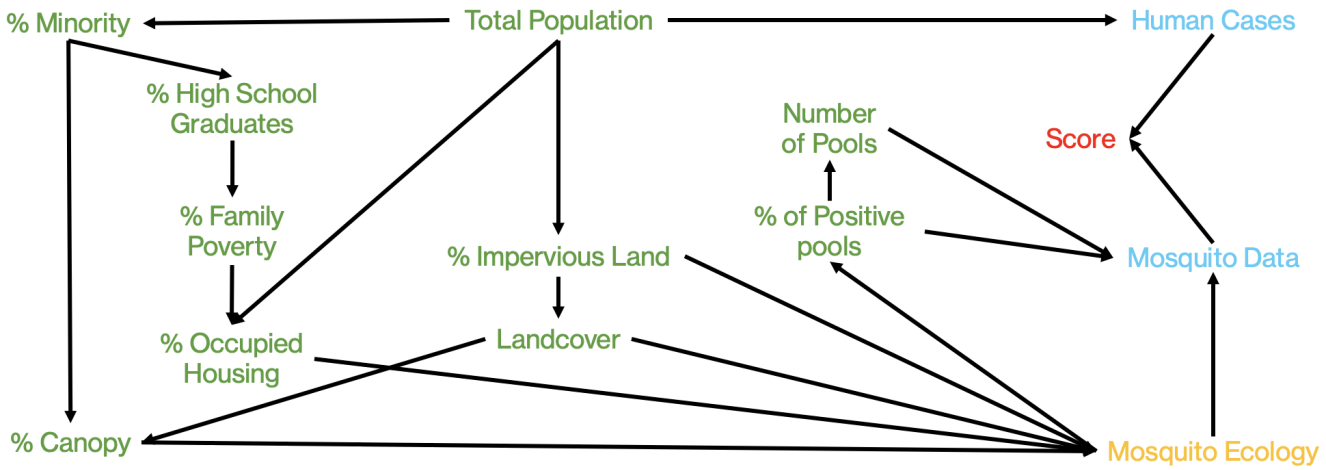
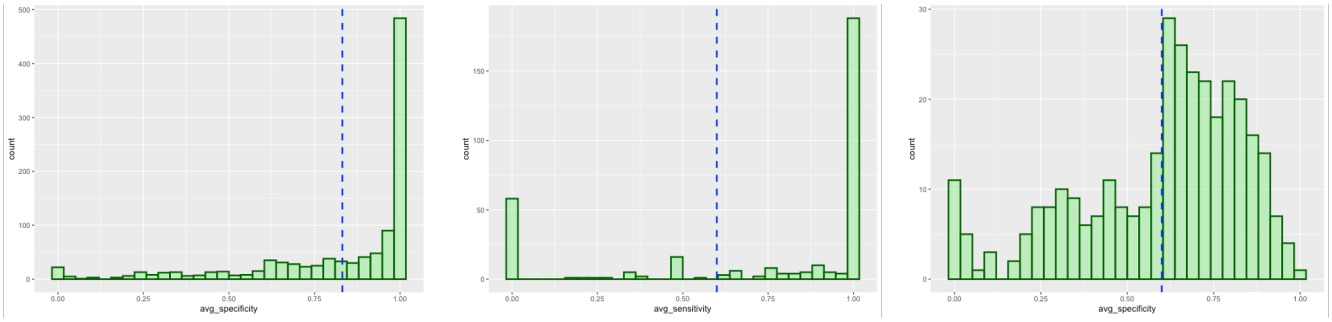


Figure 3: The DAG which represents the causal relationships between our variables. The names in green denote the covariates considered in our model, the names in blue denote the hidden variables which directly affect the score, the names in yellow denote the other hidden variables and the name in red denotes our response.

their vicinity (in a 1.5 km radius around them). That is there exist traps which do not have any observations with a $Y = 1$ associated to them. Thus we cannot compute the sensitivity values for these traps due to the non-existence of any true positives corresponding to them. There are 326 traps (see figure) out of the total 1062 traps, for which the sensitivity values can be computed. For simplicity of notation, we denote the set of all traps as T and the subset of traps which have had at least one human case in their vicinity as T^*. Thus, while we use the data from all the traps in T to build our model in Phase 1, we compute the scores in Phase 2 for only the 326 traps in T^*.

Figure 4 depicts the sensitivity and specificity scores for the traps. The first histogram (Figure 4a) displays the specificity scores for all traps (T). We can see that there is a prominent peak at 1, indicating that the majority of traps effectively predict the absence of human cases ($Y = 0$). The second (figure 4b) and third (figure 4c) histograms correspond to the traps in T^*. In the second histogram, representing the sensitivity values of the traps, a peak at 1 suggests that most traps in T^* effectively predict the occurrence of human cases ($Y = 1$). Additionally, a smaller peak at 0 implies the existence of some traps that underperform in this regard. However, examining the specificity values for traps in T^* in Figure 4c, we observe that the histogram is more spread compared to the distribution in Figure 4a. This indicates that traps are better at predicting $Y = 0$ outcomes when they have not encountered a human case ($Y = 1$), a notion that aligns with intuition. This observation motivates the definition of the score in Phase 2 as the weighted mean of the specificity and sensitivity values of the traps.

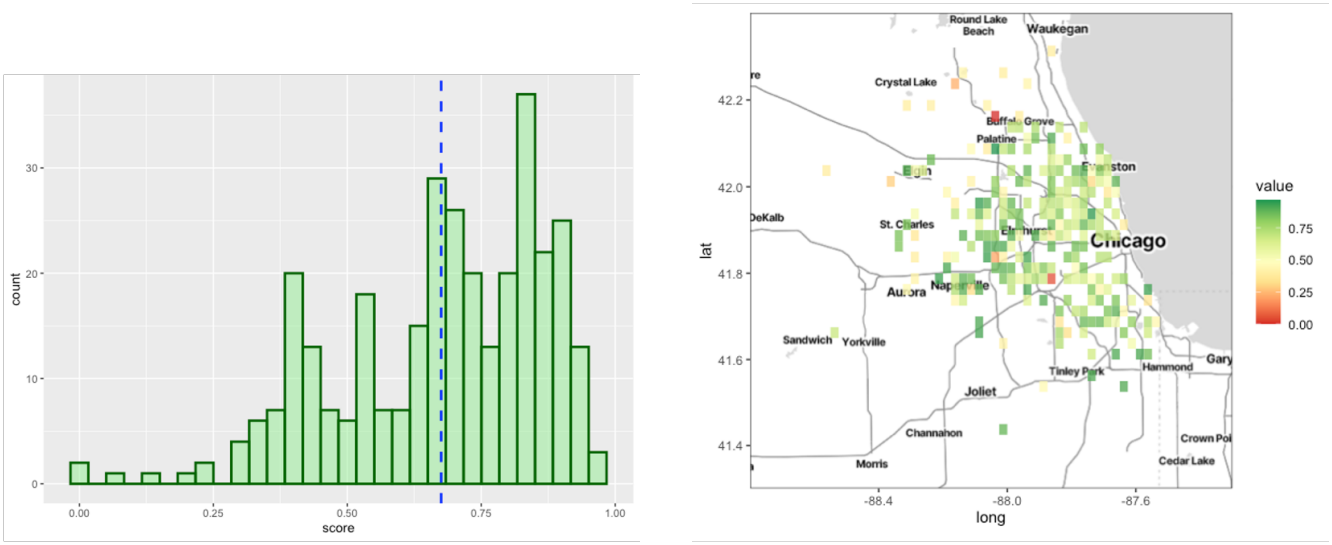


(a) Specificity values for the traps $\in T$ (1062 traps)

(b) Sensitivity values for the traps $\in T^*$ (326 traps)

(c) Specificity values for the traps $\in T^*$ (326 traps)

Figure 4: Specificity and Sensitivity values computed from the model in Phase 1.

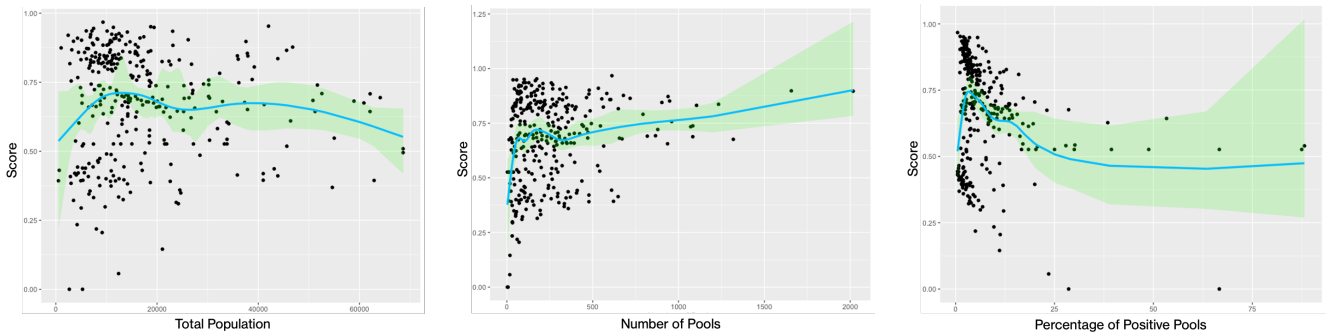


(a) Histogram of trap scores.

(b) Geospatial distribution of trap scores.

Figure 5: Distribution of the scores computed in Phase 2 of traps in T^* (326 traps).

The scores plotted in figure 5 are computed with $m = 0.9$ in equation 1. Write about why m = 0.9 was chosen. We see from figure 5a that the distribution of scores is left skewed with a mean score of approximately 0.72 (depicted by the blue line). This suggests that while most traps perform well, there are some traps that exhibit poor performance. Analysing figure 5b, where each marker represents a trap location, distinct clusters of regions with similar scores become apparent. This motivates us to explore the features which could result in a region having a higher or lower score and leads us to Phase 3 of our method.



(a) ADRF for total population in a buffer region around the trap.

(b) ADRF for the number of pools collected from the trap.

(c) ADRF for the number of positive pools collected from the trap.

Figure 6: Estimate of the ADRF for a variable's causal effect on the score with 95% pointwise standard errors (denoted in green). The standard errors are estimated by bootstrapping the entire estimation process.

4.2 Finding efficient trap locations using Phase 3

Figure 6 contains plots representing the estimates for the average dose-response functions (ADRFs) calculated from Phase 3 of our method (as described in Section 3.3). Interpretation of these ADRFs is straightforward: a straight line suggests no causal effect, while an ascending or descending line indicates a positive or negative causal effect, respectively. In figure 6, we only include the ADRFs for the three variables that show a significant causal effect. The first of these three variables is the total population in a 1500m buffer region around the traps. Upon examining the ADRF (figure 6a), we find a positive causal effect of the variable on the score until approximately a population of 10,000, after which the curve levels off. This implies that traps in locations with higher populations contribute to higher scores up to a certain population threshold. The second variable demonstrating a significant causal effect is the number of pools collected from a trap. This variable is highly correlated to the mosquito population in the vicinity of the trap, as a larger number of mosquitoes implies the collection of more pools. Figure 6b shows a positive causal effect of the number of pools collected from the trap on the score. This finding aligns with our intuition, as a higher mosquito population in an area implies more data, thereby making a trap more effective in yielding results. The last figure (figure 6c) depicts the causal effect of the percentage of pools tested as positive by the traps. The positive result corresponds to the testing of mosquitoes for the presence of the WNV. The figure illustrates an initial positive effect, succeeded by a diminishing impact that eventually levels off. Given that the score is a weighted mean of sensitivity and specificity, a low or high percentage of positive pools can both adversely influence the score. Thus, a location receiving a balanced mix of positive and negative test results is expected to be the most effective, which is confirmed by the

last ADRF plot. All the other variables considered in the analysis showed no causal effect on the score.

References

Dewitz, J. (2019). National land cover database (nlcd) 2016 products. *US Geological Survey data release*, 10:P96HHBIE.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *Review of Economics and Statistics*, 94(1):153–171.

Galagate, D. (2016). *Causal inference with a continuous treatment and outcome: Alternative estimators for parametric dose-response functions with applications*. PhD thesis, University of Maryland, College Park.

Goorbergh, R. v. d., van Smeden, M., Timmerman, D., and Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *arXiv preprint arXiv:2202.09101*.

Gu, W., Unnasch, T. R., Katholi, C. R., Lampman, R., and Novak, R. J. (2008). Fundamental issues in mosquito surveillance for arboviral transmission. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 102(8):817–822.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84.

Rousset, F. and Ferdy, J.-B. (2014). Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, 37(8):781–790.

Schafer, J. (2022). *causaldrf: Estimating Causal Dose Response Functions*. R package version 0.4.2.

U.S. Census Bureau (2020). *2015-2019 American Community Survey 5-Year Estimates*.

U.S. Census Bureau (2021). *2020 Census State Redistricting Data (Public Law 94-171)*. Summary File Prepared by the U.S. Census Bureau.