# New Orleans Tire Survey

## Mark Myer

### Progress Report 1/15/2020

The last report (12/17/19) detailed the use of exploratory statistics and graphing to choose candidate models for the larval data and a Bayesian variable selection algorithm to choose 5 candidate predictors for each mosquito species.

Since then, I have conducted further exploratory work to determine whether each mosquito species was sufficiently represented in the data to create a reliable model, and plot the observations in space and created variogram plots to evaluate spatial dependency. I then used `R-INLA`, a Bayesian statistics package in R, to fit **zero-altered (or hurdle) models** that considered the zero-inflation of the data. For each species I fit a base model, a model with random intercepts by site to account for pseudoreplication, a model with spatial effects, a model with temporal effects, and a model with spatial and temporal effects (spatiotemporal model).

Models with spatial and temporal random effects are notorious for overfitting, so I evaluated each model for each species based on a combination of quantitative criteria (WAIC and log-likelihood for model fit, hyperparameter estimates for random effects) and qualitative examination of residual plots, observed vs. fit plots, and plots of spatial/temporal effects.

The effect of each selected variable for each species with sufficient representation to create a model is discussed. In brief, the models for *Ae. albopictus* and *Cx. quinquefasciatus* were the only ones with enough non-zero responses to fit a reliable model **(Table 1).**



**Table 1. Percent of tires surveyed that contained zero larvae by mosquito species.**

Of note is the general guideline that ~30 observations are required to fit a reliable mixed effects regression model. While a “zero” response is still technically an observation, in my experience with ecological modeling, at least 30 positive responses are required for a reliable model. Only *Ae. albopictus* and *Cx. quinquefasciatus* reach that threshold in this dataset. **I performed the entire analysis start-to-finish for every species regardless of this limitation**, and I am happy to discuss the results of modeling for those species, but in my opinion only the ones for those two well-represented species are what I would consider useful, and I will discuss those here. The methods are the same for each species.

#### Variable Selection

The last report detailed the variable selection process, so I will be extremely brief. Approximately 70 candidate variables were considered including tire detritus, co-existing species including mosquito larval predators, NLCD-derived LULC, educational attainment metrics, and residential occupancy statistics. Of these, before algorithmic variable selection was applied, several were discarded as unlikely to affect the outcome based on ecological reasoning. Detritus types were consolidated into two variables (organic/inorganic), the NLCD land cover types were converted into fractional cover within each 2km buffer area.

From these variables, correlation analysis was conducted to remove highly-correlated variables. Any variable pair with a Pearson’s correlation coefficient above 0.5 was noted, and one of the two variables was removed based on its mean absolute correlation with the other variables in the dataset. The variable that was the least correlated with other variables in the dataset was retained.

Next, variables that were “0” or “NA” for more than 95% of observations were eliminated as being too rare to include in a reliable model. These variables, when present, are basically outliers in the dataset that may by sheer coincidence appear to have a highly influential effect on mosquito larvae presence. Considering the relatively low number of tires sampled in this study, a variable that appears in only 5% of observations was observed three or fewer times in the entire study.

Finally, Bayesian variable selection was used to sample from the set of all possible combinations of variables and rank the remaining candidates based on their suitability for modeling **(Table 2).**

|  |  |
| --- | --- |
| **Ae. albopictus** | **Cx. quinquefasciatus** |
| count of **mites** in tire | percentage of buffer covered by **woody wetlands** |
| **number of tires** at site | count of **chironomids** in tire |
| percentage of buffer covered by **woody wetlands** | count of **miscellaneous invertebrates** in tire |
| percentage of **canopy cover** | **presence of** **vegetation** growing within center of tire |
| amount of **organic debris** in tire | amount of **organic debris** in tire |

**Table 2. Selected variables for mosquito larvae modeling**.

#### Model Evaluation

For each species, I fit a **binomial logistic model** for Occurrence, whether a species was found or not, and a **negative-binomial count model** for Abundance, how many larvae were observed when present. The reason for using a hurdle model was that the data is extremely zero-inflated **(Figure 1a).** Count models are not equipped to deal with such a large number of zeroes. I initially considered a Poisson model for the Abundance, but found that the data was right-skewed even after eliminating the zeroes **(Figure 1b).** The Negative Binomial model accounts for the right-skewness of the response distribution, which is sometimes referred to as overdispersion.



**Fig. 1. A. Distribution of *Aedes albopictus* larvae counts. B. Distribution with zeroes excluded.**

For each model-species combination (Occurrence and Abundance), I fit a GLM with no random effects **(Nonspatial)**, a GLMM with random intercepts by Site to mitigate pseudoreplication **(Randint)**, a model with spatially correlated random effects **(Spatial)**, a model with a temporally correlated random effect by Week, with a first-order autoregression **(Temporal)**, and a model with both spatial and temoral effects **(Spatiotemporal)**. To evaluate models, I first used the CPO (conditional predictive ordinate) score, which is a leave-one-out cross validation technique to assess model fit. In brief, points that are not well-fit by the model have a higher CPO score than those that fit well. Models with a high CPO score have many problematic points. I then compared models by WAIC (Watanabe-Akaike Information Criterion) and log-likelihood **(Table 3)**. Sometimes models will show an improvement in WAIC but not log-likelihood, and in those cases I think they’re not actually better models but instead may be cases of overfitting. Similarly, it is important to check whether better fit is actually a result of overfitting. This can be done by checking diagnostic plots and parameter values in the model results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **species** | **type** | **model** | **cpo** | **waic** | **loglik** |
| A. albo | Occurrence | Nonspatial | 0 | 129.57236 | -84.38634 |
| A. albo | Occurrence | Randint | 0.2 | 127.44637 | -84.36917 |
| A. albo | Occurrence | Spatial | 0.32 | 103.26182 | -74.23442 |
| A. albo | Occurrence | Temporal | 0.01 | 113.38131 | -76.71915 |
| **A. albo** | **Occurrence** | **Spatiotemporal** | **0.3** | **101.47878** | **-75.38092** |
| A. albo | Abundance | Nonspatial | 2.49 | 670.92931 | -421.29564 |
| A. albo | Abundance | Randint | 6.85 | 615.85619 | -374.55047 |
| A. albo | Abundance | Spatial | 2.53 | 610.13437 | -343.73195 |
| A. albo | Abundance | Temporal | 3.03 | 613.87154 | -342.69302 |
| **A. albo** | **Abundance** | **Spatiotemporal** | **2.35** | **601.53539** | **-340.9152** |
| **Cx. quinq** | **Occurrence** | **Nonspatial** | **0** | **118.60504** | **-74.2854** |
| Cx. quinq | Occurrence | Randint | 0 | 118.61316 | -74.28691 |
| Cx. quinq | Occurrence | Spatial | 0 | 119.13828 | -74.56491 |
| Cx. quinq | Occurrence | Temporal | 0 | 118.99503 | -74.76468 |
| Cx. quinq | Occurrence | Spatiotemporal | 0 | 120.22492 | -75.74023 |
| Cx. quinq | Abundance | Nonspatial | 1.32 | 421.65962 | -273.16172 |
| **Cx. quinq** | **Abundance** | **Randint** | **4.47** | **361.89433** | **-239.03084** |
| Cx. quinq | Abundance | Spatial | 4.09 | 380.30472 | -229.12912 |
| Cx. quinq | Abundance | Temporal | 3.36 | 380.54278 | -228.37284 |
| Cx. quinq | Abundance | Spatiotemporal | 4.14 | 378.33879 | -227.61351 |

**Table 3. Model validation statistics for Ae. albopictus and Cx. quinquefasciatus models. The best model identified for each species-type combination is highlighted in Bold face.**

As can be seen from Table 3, for *Ae. albopictus* the best fit was the **spatiotemporal model**, while for *Cx. quinquefasciatus* the occurrence model was best fit with the basic, no random effects **nonspatial model** and the abundance model was best fit with the **random intercepts** by site.

For each of the models where the spatiotemporal model seemed to be the best fit, we looked at diagnostic plots of the spatial covariance of the residuals. These are known as **variograms** **(Figure 2).** A variogram that looks like a J-curve is indicative of true spatial correlation, because observations that are closer together are more related than those that are apart.

****

**Figure 2. Variogram for *Ae. albopictus*. Distance between observations in kilometers is on the X axis, while semivariance (a measure of relatedness) is on the Y axis. Lower semivariance means that observations are more related. The blue line is a loess curve added to aid in visual interpretation.**

This variogram is **not especially suggestive of spatial correlation**, because with the exception of sites approximately 5km apart, semivariance is about the same for all distances. The increase in semivariance at 5km may simply be noise, or random chance. If true spatial correlation was present, we would expect to see relatedness decrease in a predictable fashion with distance.

Because of the lack of spatial correlation observed in the variogram, I concluded that the increase in model fit seen with the spatial and spatiotemporal models in *Ae. albopictus* was due to overfitting rather than a true effect. This characterization is further supported by the variable coefficients in the spatial/temporal/spatiotemporal models, which showed that despite the nominally better fit, few variables were significantly associated with the outcome. My conclusion is that the spatial models are overfitting as a result of a **missing, but important, variable that varies in space.** Because of this, I have recommended that we add some variables to the analysis for *Ae. albopictus* including NDVI and human population.

#### Model Results

###### *Cx. quinquefasciatus*, occurrence model

|  |  |  |
| --- | --- | --- |
| Variable | Mean (scaled logit) | 95% Credible Interval |
| Woody Wetlands | 0.5 | -0.064 : 1.155 |
| **Chironomids** | **0.771** | **0.137 : 1.498** |
| Invertebrates | 0.278 | -0.682 : 1.779 |
| **Vegetation in Tire** | **0.587** | **0.101 : 1.160** |
| Organic Debris | 0.011 | -0.522 : 0.481 |

##### *Cx. quinquefasciatus*, abundance model

|  |  |  |
| --- | --- | --- |
| Variable | Mean | 95% Credible Interval |
| **Woody Wetlands** | **0.872** | **0.514 : 1.311** |
| **Chironomids** | **0.739** | **0.397 : 1.123** |
| **Invertebrates** | **-0.477** | **-0.789 : -0.085** |
| **Vegetation in Tire** | **0.421** | **0.120 : 0.779** |
| Organic Debris | 0.195 | -0.092 : 0.672 |

##### *Ae. albopictus*, occurrence model

|  |  |  |
| --- | --- | --- |
| Variable | Mean (scaled log) | 95% Credible Interval |
| **Mites** | **1.399** | **0.205 : 2.985** |
| Number Tires at Site | -0.052 | -0.758 : 0.625 |
| Woody Wetlands | 0.052 | -0.761 : 0.974 |
| Canopy Cover | -0.273 | -0.863 : 0.293 |
| Organic Debris | -0.061 | -0.637 : 0.622 |

##### *Ae. albopictus*, abundance model

|  |  |  |
| --- | --- | --- |
| Variable | Mean | 95% Credible Interval |
| Mites | 0.266 | -0.059 : 0.627 |
| Number Tires at Site | 0.428 | -0.020 : 0.872 |
| Woody Wetlands | 0.223 | -0.202 : 0.774 |
| Canopy Cover | -0.063 | -0.428 : 0.323 |
| Organic Debris | 0.228 | -0.062 : 0.581 |

Table 4. Predictor variable coefficients with 95% credible intervals for mosquito larvae models. **“Significant” variables (those whose 95% CI do not encompass zero) are bold and highlighted.** The coefficients are scaled such that a one-standard deviation increase in the predictor from the mean has the indicated effect upon the logit-odds of larval presence in the case of the occurrence models, and the log-expected number of larvae found in the case of the abundance model.

In the case of *Cx. quinquefasciatus*, both the **presence of chironomids** and **vegetation growing in the center of the tir**e had a positive effect on the odds of larval occurrence, while for larval abundance **woody wetlands, chironomids, and vegetation in the tire** had positive associations with the number of larvae, while the count of **miscellaneous invertebrates** had a negative association.

For *Ae. albopictus*, the presence of **spider mites** was positively associated with larval occurrence, and no variables were significant for the abundance model. Yet, these spatiotemporal models were the “best fit” according to WAIC and log-likelihood. The spatial effect **(Figure 3)** and temporal effect **(Figure 4)** showed distinctive patterns, although these may be artifacts of sampling time and choice of sample locations. Note that the spatial effect is almost zero at anywhere other than the sampling locations, which limits the model’s applicability to predicting mosquito numbers at out-of-sample locations. The temporal effect indicates that Ae. albopictus should slightly increase in both occurrence and abundance up until approximately Epi Week 30, which is a trend that is slightly borne-out by the data **(Figure 5)**.



**Figure 3.** Spatial effect from spatiotemporal models for *Ae. albopictus* occurrence model (left) and abundance model (right). Areas with hotter colors are more likely to harbor *Ae. albopictus* larvae.



**Figure 4.** Temporal trends from spatiotemporal model for *Ae. albopictus* occurrence (left) and abundance model (right). Both indicate a small increase in mosquito prevalence up until approximately Epi Week 30, when the trend begins to reverse. Note the Y-axis, the value of the temporal coefficient, upon which the change is very small throughout the study period.

****

**Figure 5.** Observed Ae. albopictus larvae by week. The trend does not quite match that predicted by the temporal portion of the spatiotemporal model, which may be the result of explanatory variables that change with time throughout the study period.