Mosquito Population Abundance the Coachella Valley Region of Riverside County, California

Prepared for George Peck

By: Jacob Schultz and Joseph Cole

Executive Summary

Mosquito population modelling can benefit from fitting modern ecological models like the generalized N-mixture model. Furthermore, it is important to account for factors such as detection probability and population dynamics. We estimate the detection probability (64% per trap-night), survival rate (65% over a two week interval), recruitment/birth rate (207% over a two week interval), and initial average population intensity throughout the study area (953 individuals within the sphere of influence of a single trap). Further improvements are possible by introducing spatially and time varying covariates to the model.

1 Introduction

Ecological modeling has grown more sophisticated over the past two decades as computing power and improved optimization techniques continually push forward the state-of-the-art. However, the mosquito control community is sometimes slow to adopt new modeling techniques because they may be operating in resource constrained environments with stretched budgets. Our goal is to explore advanced statistical models that may increase the prediction accuracy and precision of mosquito population models, leading to more optimized abatement decision making [6]. As a starting point we were given mosquito count data collected in the Coachella Valley area of southern California. The data were collected by Reisen and Lothrop [8], and they created an ordinary least squares (OLS) regression model for abundance predictions. We will provide a mathematically rigorous assessment of Reisen's model and show that this data set violates standard assumptions necessary for the application of an OLS model. Then, we will introduce the generalized N-mixture model from Dail and Madsen [2] and show that this model provides a promising avenue to explore for improving mosquito modelling. We provide strong statistical evidence that it is important to model mosquitoes as an open population to achieve optimal predictions. The remainder of the report documents the status of our efforts at finding the most informative group of covariates to use in fitting the Madsen model.

2 Background

Predicting mosquito abundance in the wild is important primarily because mosquitos are a vector for a variety of serious, and sometimes fatal, diseases. In the United States, West Nile virus is the leading cause of disease that results from mosquito bites [1]. Figure 1 shows the incidence of West Nile cases through the United States in 2018. Most people exposed to West Nile virus never feel sick, but severe symptoms emerge in about 1 of 150 infections. West Nile affects the central nervous system and can result in encephalitis (inflammation of the brain) or meningitis (inflammation of the membranes surrounding the brain and spinal cord). Outward signs include high fever, convulsions, paralysis, and occasionally death [1]. Resources dedicated to control mosquito populations are typically scarce, as control often falls within the jurisdiction of county level agencies with small budgets. As a result, accurate and precise predictions of mosquito abundance can help to optimize the allotment of limited abatement resources.

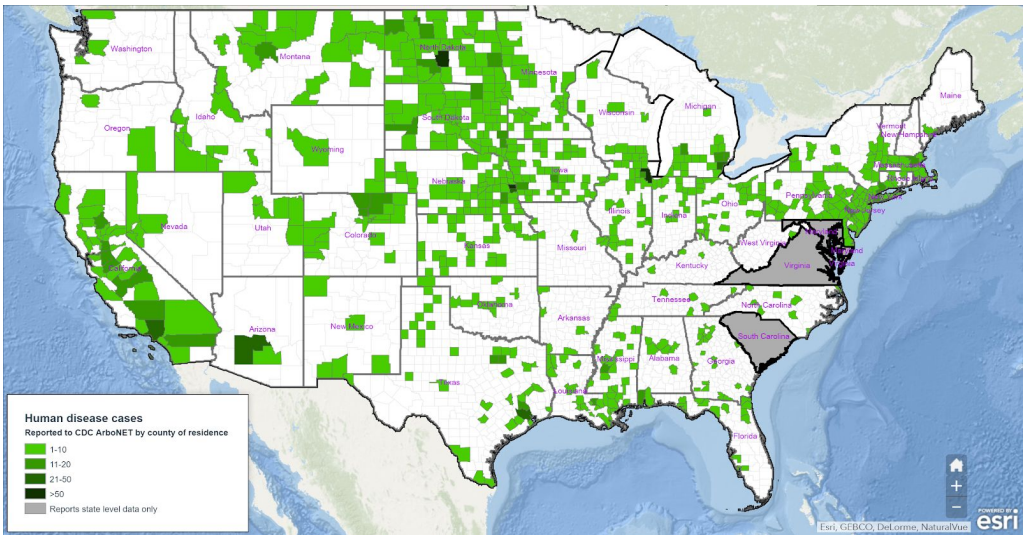

Figure 1. Incidence of West Nile Virus in the US, 2018

Given that reducing incidence of West Nile infections is a key underpinning the need for mosquito control, *Culex tarsalis* becomes the most logical target (figure 2). *Culex tarsalis* is a particularly efficient vector of encephalitis diseases among species present in the United States [13]. Females prefer birds as a source of a blood meal needed prior to laying eggs, but they will not hesitate to bite humans as well. Since birds provide a reservoir for West Nile virus, *Culex tarsalis* ends up spreading the disease from the bird population to humans. It is sometimes known as the Western Encephalitis Mosquito due to this behavior.

Figure 2. *Culex tarsalis*

Figure 3. *Pelecanus occidentalis*

Management of habitat is one way to naturally control mosquito populations before resorting to more direct methods. Like all species in the genus *Culex, tarsalis* is a type of “standing-water”

mosquito because it lays eggs directly on the water surface. Water is likewise required as an initial environment for the larva. Eggs typically hatch within a few days, and the entire lifecycle of the insect is about three weeks. Thus, some relatively permanent source of standing water must be present to seed a population of this species in a given region. Furthermore, *Culex tarsalis* requires specific habitat traits in order to survive winter. They need a place that is both warm enough to avoid a hard freeze, but cool enough to reduce metabolism in order to conserve fat reserves. Likely places to find them mid-winter include mine shafts, storm sewers, culverts, sheds, basements, and garages [13]. These factors are worthwhile to remember in building an abundance model.

Although the Coachella Valley overall is arid desert, the area north of the Salton Sea is an important agricultural region. Depending on the type of crop the usual method of irrigation may be to flood the fields, leading to ideal *Culex tarsalis* habitat. There are many birds in the area because the coastline of the Salton Sea is a favored habitat for pelicans to nest and hatch young (figure 3). Furthermore, there are duck marshes that are flooded annually in preparation for duck hunting season. Lastly, temperatures are often mild with few lows below freezing. These influences conflate to make the present study area a particular concern for West Nile infection.

3 Data

3.1 Description of the Samples Collected

Reisen and Lothrop [8] collected counts of mosquitos throughout the Coachella Valley starting in April 1994 through November 1995. This was accomplished by setting 63 traps overnight from dusk until dawn once every two weeks. Each of these samples are known as a trap-night. The traps were sited on an approximate one mile grid as shown in figure 4. Data were not collected during winter because *Culex tarsalis*, the targeted species, undergoes a reproductive diapause during those months that causes population counts to plummet. This resulted in a maximum of 33 time samples from each trap site. We checked that the data was collected uniformly throughout both time and space (figure 4). While some sites are missing data, there are no overly concerning gaps. However, we noted that data are missing for the entire western half of the study area during the second two week interval in June 1995. Accuracy and precision of abundance estimates during that time frame could be negatively impacted. Also, relatively few samples were collected from trap sites 17, 38, and 49 near the mouth of the Whitewater River. The reason is unknown, but we speculate access to those sites may have been problematic. Overall the dataset is impressively high quality.

Figure 4. Coachella Valley study area

There are a variety of trap designs for mosquitoes with corresponding efficacies depending on the targeted species [7]. These data were collected using standard CDC style traps [8], baited with dry ice and without light (figure 5). Female mosquitoes are attracted to the CO_2 emanating from the traps from as far away as 30 meters. Once a mosquito approaches, a fan draws it into the netted area. In the morning, the traps were collected and transported to a laboratory where captured mosquitoes were anesthetized and tallied by an expert according to species. Thirteen different species were identified, with *Culex tarsalis* the most common by a large margin. They made up 75% of the total sample. Among the 1972 individual data samples, counts of *Culex tarsalis* range from 0 to 7936, with the maximum occurring in March 1995 at trap site 2.

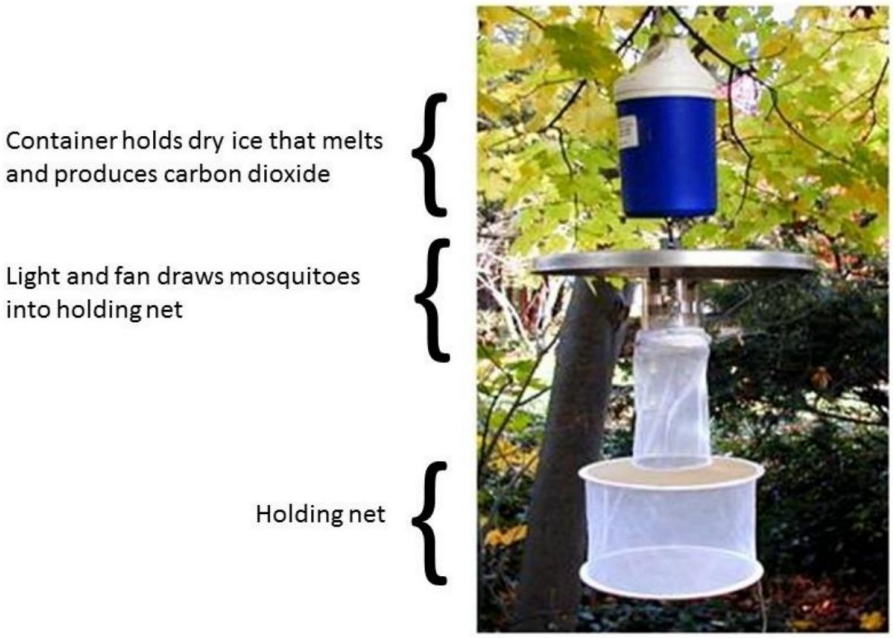

Figure 5. CDC style mosquito trap

3.2 Description of Available Predictive Variables

In addition to the count data, we considered a number of additional covariates for incorporation in the model. As a simplifying assumption based on the data available to us, we divided the covariates into two categories: those that were constant in time but depended on position (site level covariates), and those that were constant in space but varied in time (observation level covariates). We did not have enough information to create any predictors that varied in both time and space.

As a site level covariate, Lothrop documented the habitat types surrounding each trap. Using GIS software, he drew polygons around each of nine habitat categories (desert, saltmarsh, duck pond, row crop, grapes, citrus farm, date farm, pasture, or fish farm) on aerial overlays of the study area. He then calculated the intersection of the habitat polygons with a 1 km radius circle around each trap site to find the number of hectares of each habitat category surrounding the site. In most cases these values sum to about 312.3 hectares, but traps bordering the Salton Sea have lower totals because area over the sea was not counted. [4] Mosquitoes are known to stay within a short distance of the shore in any case.

The remaining site level covariates describe the geographic location of the trap sites. While the longitude and latitude coordinates of each trap site were not available to us, we were able to estimate the locations knowing that traps were typically placed along transects and at intersections of major thoroughfares. We matched the latitude and longitude of those points to the locations indicated by Reisen [8]. These untransformed values were treated as continuous covariates. A third location covariate was created post-hoc describing the distance from the trap sites to the Salton Sea. In order to estimate the distance, we placed 67 closely spaced markers along the shoreline in Google Earth. Then we calculated the minimum great circle distance between each trap and the set of markers. The marker locations selected and the corresponding code are available on GitHub [12]. This covariate was created because having access to both standing water and coastal bird blood meals is important to the reproducing female *Culex tarsalis*. These effects might reasonably be modeled by the distance from the Salton Sea.

The observation level covariates consisted of two different measures of temperature as well as time itself. The time covariate varies from 1 to 42 and describes the time period when the count in question was measured. Since data was collected in two-week intervals, an increase of 1 in this time covariate corresponds to a two-week change in time.

The temperature data came from a publically available NOAA database and was collected daily at the Mecca fire station. This station is located in the Coachella Valley about 5.7 km from the coast of the Salton Sea. There were three different measurements available for each day: maximum temperature, minimum temperature, and a daily reading taken at 17:00. The minimum temperature data was littered with obvious sensor errors and was therefore deemed unreliable. The other two measures, which we will refer to as “max temperature” and “observed temperature” showed very few missing values and errors. To account for the two-week timescale granularity, daily max and observed temperatures were averaged over every two week interval, yielding two-week averages of the max and observed temperatures. Therefore each of the 42 time periods had a distinct average max temperature and average observed temperature. Some of these average temperatures were

calculated using less than 14 values due to missing data and the removal of erroneous measurements.

3.3 Exploratory Analysis

The best way to gain intuition about the dynamics of the mosquito population throughout space is likely to watch an animation of the raw collected counts [12]. We created overlays for each of the 33 time periods with available data by linearly interpolating the trap counts on a 1000x1000 grid over the study area. In order to avoid interpolation over the sea, we added the 67 shoreline markers to the interpolation as “virtual traps” with zero counts. These were animated using Google Earth Studio and Matlab. While it is difficult to make concrete conclusions based on the animation alone, it is easy to imagine individual mosquitoes migrating between trap sites. Observe frames from early and late July 1994 (figure 6). It appears that some of the population from trap 45 may have dispersed to the vicinity of trap 44 to supplement the population there within the two week sampling gap. This observation is consistent with dispersal rates estimated using mark and recapture methods [9]. We also note dramatic swings in counts that drive us to expect that it will be necessary to model parameters like mosquito birth rate and immigration along with death rate and emigration between sites.

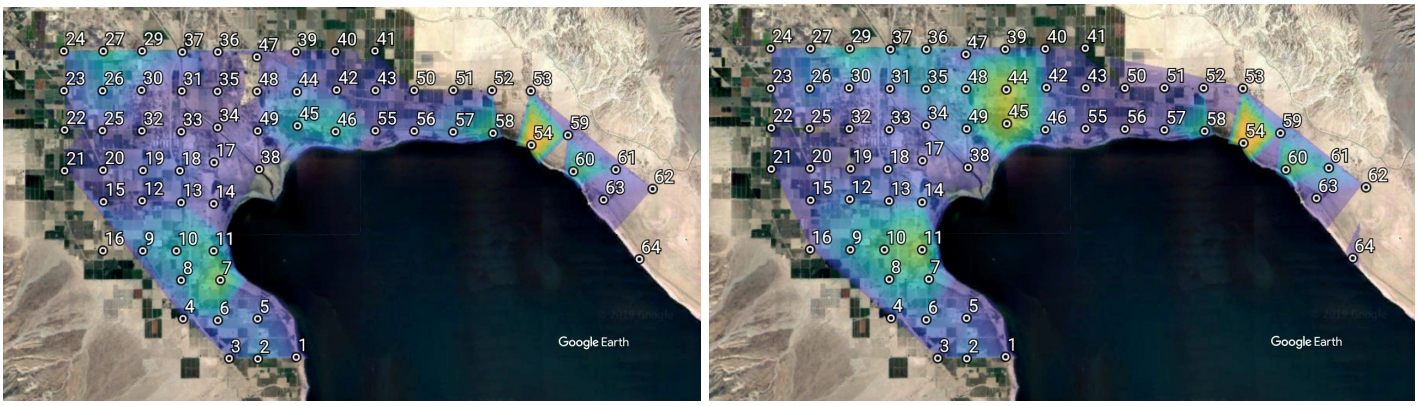

Figure 6. Mosquito counts, early to late July 1994

In order to gain further familiarity with the data, we reproduced much of Reisen’s work [8], as well as checking a number of standard regression diagnostics. Reisen created an ordinary least squares (OLS) regression model of the population abundance. The model implicitly assumes perfect detection (i.e. all mosquitoes that come within the sphere of influence of a trap are successfully captured). Since this assumption leads to highly biased estimates [2,10], Reisen avoided making absolute predictions and limited himself to conclusions that depend only on relative counts. These checks help us assess the value of the Reisen model as a baseline for comparison with more advanced ecological models. See the appendix for a complete description of the diagnostics performed. We concluded that the Reisen model is about the best that can be found for the available covariates when restricted to a model that assumes normality and perfect detection. However, it only achieves an adjusted-R^2 value of 0.55 leaving plenty of room for improvement in the quality-of-fit to the data.

4 Methods

One of the fundamental considerations when modeling trap data is the probability of catching an individual given individual was within the sphere of influence of the trap during the time of trapping. This is referred to as the "detection probability" of trap *i* at time *t* and is denoted p_{it}. It is often unreasonable to assume that $p_{it} = 1 \ \forall (i, t)$, so count data n_{it} should be interpreted as counts under imperfect detection. Modeling ecological data under imperfect detection is a large and growing field, but in this project we focused on one approach: the N-mixture model.

First an N-Mixture model for closed populations was fit to the data, which was developed by Royle in 2004 [10]. An N-mixture model for open populations was then fit, which comes from a paper by Dail and Madsen in 2011 [2]. A likelihood ratio test was then applied to test the open population assumption, as shown in Dail and Madsen's 2011 paper [2]. The fundamental goal of these models is to estimate both abundance of *Culex tarsalis* and the probability of detection p_{it}.

4.1 N-Mixture Model for Closed Populations

The N-mixture model for closed populations models *R* distinct subpopulations that are sampled over *T* sampling occasions. It is assumed that detection probability p_{it} is constant for all animals at site *i* during sampling occasion *t*. One of the primary values of interest is N_{it}, which is the number of individuals at site *i* during sampling occasion *t*. However, the closed population assumption states that the number of individuals at trap site *i* is constant over the course of the study, which implies

$$N_{ij} = N_{ik} = N_i \quad \forall\, j, k \in \{1, \dots, T\}\,.$$

The count data in our possession is denoted n_{it}, which is interpreted as a realization of a random variable distributed $Binom(N_i\ ,\ p_{it})$. The unknown N_i values are treated as realizations from $Poiss(\lambda_i)$ random variables. Therefore estimation of abundance at each site (and overall abundance by extension) depends on obtaining an estimate of the intensity parameters λ_i. This set of definitions and assumptions lead us to the following likelihood[10].

$$L(p_{it}, \lambda_i \mid \{n_{it}\}) \approx \prod_{i=1}^{R} \left(\sum_{N_i = \max_t(n_{it})}^{k} \left(\prod_{t=1}^{T} Bin(n_{it}\ ; N_i, p_{it}) \right) Poiss(N_i\ ; \lambda_i) \right)$$

Where *k* is a large finite integer allowing the sum to be approximated as infinite.

Models are fit by maximizing the above likelihood with respect to λ_i and p_{it} using any number of numerical optimization techniques. Note that both site-level and observational level covariates can be used to predict p_{it} while only site-level covariates may be used to predict λ_i per the closed population

assumption. Of course these two parameters can be modeled without predictors, which implies constant detection probability and/or intensity over time and space. We call the model with no covariates the “Closed Intercept Model” and it plays a significant role in the likelihood ratio test for closure in Section 4.3.

4.2 N-Mixture Model for Open Populations

The N-mixture model for closed populations succeeds in modeling counts for imperfect detection, but the closed population assumption. One way way around this issue is to break the timespan of the study into “primary sampling periods”. The subpopulations are assumed to be closed within each primary period, but may be open between primary periods [11]. Dail and Madson [2] built on this by shrinking the primary periods so that each observation was considered a primary period. They then modeled abundances of the subpopulations N_{it} changing through time according to population dynamics parameters. The resulting models accomplish the goal of modeling abundance under imperfect detection while allowing for open populations, so they are called N-mixture models for open populations.

Most of the notation is retained from Section 4.1 with a few key differences. First, we must refer to abundance as N_{it} instead of N_i to account for subpopulations changing through time. We must add two additional parameters to model population dynamics: ω_{it} and γ_{it}. The rate of arrivals of individuals at site i during time period t is γ_{it} and ω_{it} is the probability of an individual surviving for an additional time step. Note that γ_{it} accounts for both birth and immigration and ω_{it} accounts for both death and emigration. Lastly, we now interpret λ_i as the abundance at $t = 0$ or “initial abundance”.

We then assume that the set of abundances $\{N_{it}\}$ has the markov property so that N_{it+1} depends only on N_{it}. The one-step transition probabilities are calculated using the population dynamics [2]:

$$P_{N_{it-1}\to N_{it}} = \sum_{c=0}^{\min(N_{it-1},N_{it})} Bin(c\,;\, N_{it-1}, \omega_{it-1}) Poiss(N_{it} - c\,; \gamma_{it-1})$$

The site-level initial abundances are assumed to be realizations from a $Poiss(\lambda_i)$ distributions. We can therefore estimate overall abundance at location i during time t by applying fitted parameter values and stepping through the markov chain.

The following likelihood emerges under the open population assumption with population dynamics and the markov property [2]:

$$L(p_{it}, \lambda_i, \gamma_{it}, \omega_{it} \mid \{n_{it}\}) \approx \prod_{i=1}^{R} \left(\sum_{N_{i1}=n_{i1}}^{k} \cdots \sum_{N_{iT}=n_{iT}}^{k} \left(\left(\prod_{t=1}^{T} Bin(n_{it}\,; N_{it}, p_{it}) \right) \frac{e^{-\lambda} \lambda^{N_{i1}}}{N_{i1}!} \prod_{t=2}^{T} P_{N_{it-1}\to N_{it}} \right) \right)$$

Where k is a large finite integer allowing the sum to be approximated as infinite.

Similar to the model in Section 4.1, parameters are estimated by maximizing this likelihood using a numerical optimization technique. Site-level and observational level covariates can be used to predict p_{it}, γ_{it}, and ω_{it} while only site-level covariates may be used to predict λ_i since it is only used for modeling abundance at t =0. We call the model with no covariates the "Open Intercept Model".

4.3 Likelihood Ratio Test for Closure

The approach hinges on the fact that the model under the closed population assumption is a special case of the open population model, leading to a set of nested likelihoods. If we set the arrival rate to zero and the survival percentage to one under the open assumption, the likelihood reduces to the one under the closed population assumption. Therefore, we can easily test the open population assumption using the following likelihood ratio test statistic [2].

$$LR = -2\ln\left(\frac{\sup\left(L_{open}(p,\lambda,|\gamma=0,\omega=1,\{n_{it}\})\right)}{\sup\left(L_{open}(p,\lambda,\gamma,\omega|\{n_{it}\})\right)}\right)$$

$$= -2\ln\left(\frac{\sup\left(L_{closed}(p,\lambda|\{n_{it}\})\right)}{\sup\left(L_{open}(p,\lambda,\gamma,\omega|\{n_{it}\})\right)}\right)$$

Note that two restrictions in the reduced models are located on the boundaries of the parameter space. This means that LR under the null is not distributed χ^2, but follows a particular mixture of χ^2 distributions [2].

$$f_{LR} = [(0.5-\delta)\chi_0^2] - [0.5\chi_1^2] + [\delta\chi_2^2]$$

Where $\chi^2_{(0)}$ is a point-mass at zero. Dail and Madsen mention several approaches for calculating the δ value. However, we suggest an alternative by arguing that the power of a test under a $f_{LR} \sim \chi^2_{(2)}$ null distribution is strictly less than a test employing the mixture null distribution. This is due to the more exaggerated right-skew of the mixture distribution, regardless of the choice of δ**.** Therefore we will use the more convenient $\chi^2_{(2)}$ null and look into calculating δ only if we fail to reject the null hypothesis.

5 Results

R code for fitting both open and closed N-mixture models were provided in an online supplement to Dail and Madsen's paper and these functions are also built into the `unmarked` package in R [3]. We attempted to implement the supplementary code, but issues with the

optimization function let us to the more stable `unmarked` functions.

5.1 Intercept Models

The first step was fitting the intercept models mentioned in Sections 4.1 and 4.2. One issue we immediately encountered was determining the correct *k* for estimating the infinite sums contained in the likelihoods. The chosen *k* must be at least as large as the maximum observed count, which is 7936 for our data. This caused unreasonably long computation time, so we binned the counts every 100, producing a manageable minimum *k* value.

The output of the open and closed intercept models are given below. The closed population model converges in under 30 seconds while the open population model converges after 15-30 minutes depending on the CPU being used. All values shown are back-transformed from their respective log and logit-scale versions and the estimated abundances are multiplied by 100 to account for binned counts.

	AIC	$\widehat{\lambda}$	$\widehat{p}$	$\widehat{\gamma}$	$\widehat{\omega}$
Open Population	12974.97	952.541	0.6439376	2.0719353	0.6509723
Closed Population	15736.49	7069.499	0.05684929	N/A	N/A

We then ran the likelihood ratio test outlined in Chapter 4.3, which produced a test statistic of 408.065, with an associated p-value of 2.45366e-89 under the $\chi^2_{(2)}$ null distribution. Therefore we have strong statistical evidence against the closure assumption. Interpretation of the fitted coefficients supports this conclusion; Royle's analysis shows single-site counts being much closer to 1000 than 7000. The inflated abundance estimation in the closed model is likely due to a gross underestimation of the detection probability. Under the closed population assumption, the vast seasonal differences in counts are modeled as inconsistency in trapping quality, driving down the estimated detection probability.

5.2 Covariate Models

We began fitting models using the `unmarked` package and various sets of covariates, but we quickly encountered some limitations. Through experimentation with both our data and the famous `mallard` data set from `unmarked`, we concluded that only site-level covariates could be used to fit $\widehat{\lambda}$, $\widehat{\gamma}$, and $\widehat{\omega}$ while both site-level and observational-level covariates could be used to fit $\widehat{p}$. According to the formulation of the likelihoods, only $\widehat{\lambda}$ should be limited to site-level covariates. We acknowledge that this could be user error, but a fix has not yet been found.

With this in mind, we decided to model $\widehat{p}$ temporally and model the rest of the parameters using site-level covariates. We attempted fitting intercept models with a temperature covariate on $\widehat{p}$, but the model did not converge due to a singular hessian (identifiability issues). Various starting values were attempted for fitting these models with the same outcome every time. This issue was replicated for both the observed and maximum temperature covariates. We remedied this issue by fitting a categorical “season” covariate instead of the continuous temperature values and the models ran without error. A table showing the model structure and AIC is given below. Note that a “1” in a covariate column should be be read ”intercept only”.

Model Name	$\widehat{\lambda}$ Covariates	$\widehat{p}$ Covariates	$\widehat{\gamma}$ Covariates	$\widehat{\omega}$ Covariates	AIC
Open Intercept	1	1	1	1	12974.97
Closed Intercept	1	1	NA	NA	15736.49
Model 1 Open	d.to.sea	season	d.to.sea	d.to.sea	12048.1
Model 1 Closed	d.to.sea	season	NA	NA	15789.58

5.3 Computational Challenges

Maximizing complicated likelihoods with a large data set such as this is quite challenging computationally, and the problem grows quickly with the number of dimensions at play. For example, the most complicated model we fit (Model 1 Open in section 5.4) took about 12.5 hours to converge. However, its closed-population counterpart (Model 1 Closed) converges in under five minutes. Here we provide a list of the main computational challenges we faced and share how we dealt with them

5.3.1 Setting *k*

Increasing the *k* setting will slow down computation significantly by making the likelihood calculation slower at every time step of the optimization. As mentioned in section 5.1, we want *k* to be large enough to approximate an infinite sum in the likelihood. Our largest binned response was 80 (corresponding to counts on the interval [7001, 8000]) so we chose to set *k* to be 25% larger than this maximum count, leading to *k* = 100.

5.3.2 Modeling Population Dynamics with Covariates

Modeling population dynamics using covariates is much more expensive than using covariates to fit initial abundance or detection probability [2]. To account for this, we were careful to fit models with no more than one predictor on $\widehat{\gamma}$ and $\widehat{\omega}$ so we could attain convergence in a reasonable amount of time.

5.3.3 Modeling Using Trap Number as a Covariate

Modeling one parameter using trap number as a covariate corresponds to 63 additional parameters. Adding this many dimensions to the likelihood slows convergence significantly. We did not fit models using this method, but one could see the appeal. With fitted trap-specific initial abundance and population dynamics estimates, we could estimate each of the 63 subpopulations through time, producing a spatial–temporal areal model for abundance. The trap number covariate could capture random effects that are not accounted for by more general site-level covariates like habitat, but the required computational cost is most likely prohibitive.

6 Conclusions and Recommendations

At this stage of work, we can conclude with confidence that an ordinary least squares regression model of mosquito population abundance violates key assumptions necessary for such a model to be valid. Based on the adjusted-R^2 metric, the Reisen model was the best that could be achieved when limited to this class of models, given the available covariates. However, it only achieved adjusted-R^2 of 0.55. We showed that the Resien model violates the assumption of normality of the errors, the assumption of constant variance of the errors, and the assumption of independence of the errors. From a biological point of view, the obvious gaps in the model are its inability to factor in an imperfect detection probability and important population dynamics that certainly play a role in abundance.

Our experiments in fitting more modern ecological models, like the Madsen model, allow us to conclude these are a promising class of model to apply to mosquito count data. We successfully fit an intercept only, open population model to the data set, which yielded reasonable estimates of detection probability (64% per trap-night), survival rate (65% over a two week interval), recruitment/birth rate (207% over a two week interval), and an estimate of initial average population intensity throughout the study area (953 individuals within the sphere of influence of a single trap). We used a likelihood ratio test to validate the need for an open population model and found extremely strong support for this assumption.

However, an intercept only model cannot provide the site and time specific information needed to help influence mosquito abatement decisions. Work to fit a Madsen model that includes covariates that vary in time and space is ongoing but limited by computation time and the available covariate options exposed in the software package used (`unmarked` package in R [3]). Due to these limitations we recommend that a fully Bayesian version of the Madsen model be implemented in STAN. Such an implementation would be a research project on its own, but it would allow complete flexibility in the specification of covariates and permit the full estimation of posterior distributions for each parameter. That would enable the calculation of credible intervals for the parameter estimates. Once satisfied with the fit of an open population N-mixture model, we recommend performing a cross validation analysis to determine the accuracy and precision of abundance estimates and compare these to the Reisen model as a baseline.

7 References

[1] Center for Disease Control and Prevention. n.d. *West Nile virus.* Accessed 6 1, 2019. https://www.cdc.gov/westnile/index.html.

[2] Dail, D., and L. Madsen. 2011. "Models for Estimating Abundance from Repeated Counts of an Open Metapopulation." *Biometrics* 67 (2): 577-587.

[3] Fiske, Ian, and Richard Chandler. 2011. "unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance." Journal of Statistical Software 43 (10): 1-23. Accessed June 3, 2019. http://www.jstatsoft.org/v43/i10/.

[4] Lothrop, Hugh D. 2014. "Private communication from Hugh Lothrop to George Peck." March 26.

[5] Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. 2012. *Introduction to Linear Regression Analysis.* 5th. Hoboken, New Jersey: John Wiley & Sons, Inc.

[6] Peck, George W. 2015. "Surveillance of Invading Mosquitoes Using Occupancy Estimation and Modeling." *Proceedings and Papers of the Mosquito and Vector Control Association of California* 83: 62-64.

[7] Peck, George W., Fanny Castro-Llanos, Victor M. Lopez-Sifuentes, Gissella M. Vasquez, and Erica Lindroth. 2018. "Comparative Analysis of Mosquito Trap Counts in the Peruvian Amazon: Effect of Trap Type and other Covariates on Counts and Diversity." *Journal of the American Mosquito Control Association* 34 (4): 291-301.

[8] Reisen, William K., and Hugh D. Lothrop. 1999. "Effects of Sampling Design on the Estimation of Adult Mosquito Abundance." *Journal of the American Mosquito Control Association* 15 (2): 105-114.

[9] Reisen, William K., and Hugh D. Lothrop. 1995. "Population Ecology and Dispersal of Culex tarsalis (Diptera: Culicidae) in the Coachella Valley of California." *Journal of Medical Entomology* 32 (4): 490-502.

[10] Royle, J. Andrew. 2004. "N-Mixture Models for Estimating Population Size from Spatially Replicated Counts." *Biometrics* 60 (1): 108-115.

[11] Royle, J. Andrew, and R. M. Dorazio. 2008. *Hierarchical Modeling and Inference in Ecology.* Amsterdam: Academic Press.

[12] Schultz, Jacob, and Joseph R. Cole. 2019. *Mosquito.* June 3. Accessed June 3, 2019. https://github.com/schultz7676/Mosquito/releases/tag/v1.0.

[13] Weissmann, Michael. 2016. *Mosquito of the Month: Culex tarsalis - the Western Encephalitis Mosquito.* 7 28. Accessed 6 1, 2019. http://www.vdci.net/blog/mosquito-of-the-month-culex-tarsalis-western-encephalitis-mosquito.

Appendix

First we evaluated the response transform selected by Reisen: log(y+1). The raw count data is highly right skewed (figure 7), which is likely to invalidate the assumption in OLS regression that errors are normally distributed. We confirm this is the case by checking a q-q plot of a model using the raw counts as the response (figure 8). Furthermore, the variance of the data grouped by trap site

increases with increasing mean (figure 9, a reproduction of figure 4a in [8]), which violates the assumption of constant variance. Power transformation of the response (y^{λ}) is therefore indicated, and Reisen evaluated two options: λ=½ and λ=0. We used the Box-Cox method to find the optimal value for λ and found λ=0.1 (figure 10), which is so close to the value ultimately selected by Reisen (λ=0) as to make no practical difference. We continued the analysis using the same transform as Reisen to maintain consistency.

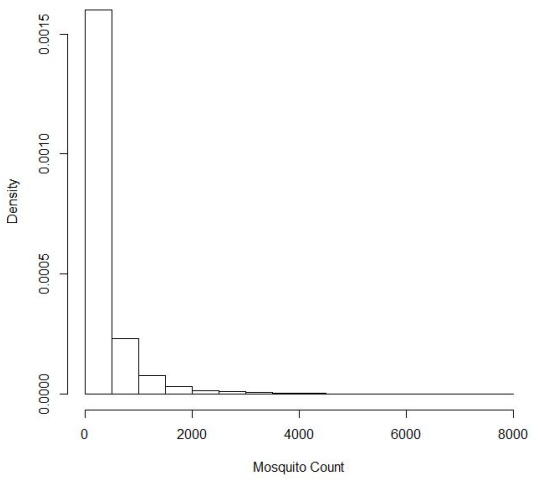

Figure 7. Histogram of raw mosquito count data

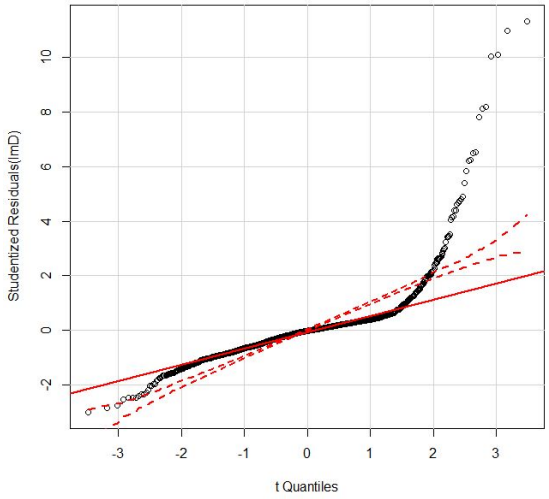

Figure 8. Q-Q plot of model using raw counts

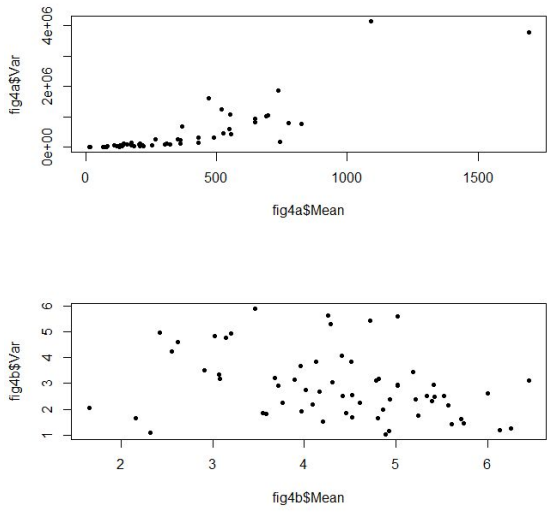

Figure 9. Reproduction of Reisen figure 4 a & b

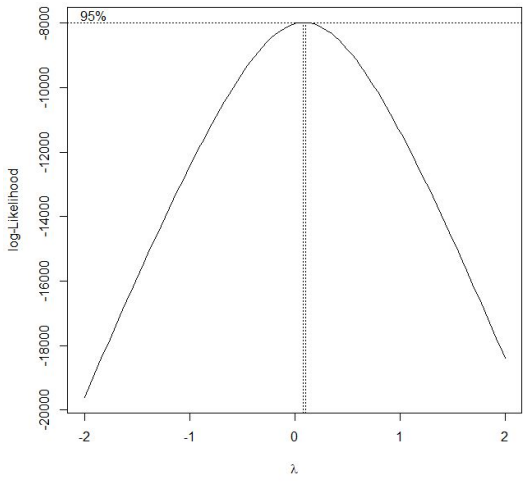

Figure 10. Box-Cox selection of transform power

A histogram of the transformed data is shown in figure 11. While the power transform did help to control the variance (figure 9, a reproduction of figure 4b in [8]), evaluated another way there are still apparent problems. Figure 12 shows a plot of the residuals versus the predicted value of the model. If the assumption of constant variance is valid, the envelope of these points should be constant across the domain of predictions. However, in this case the variance appears to be larger in

the middle of the model's range (predicted values around 4 or 5) than at the extremes (predicted values near 0 or 8). Compounding the violation is the fact that there is a quadrant in the plot mostly devoid of data points. When y=0, the transformed value is log(1)=0, and this is the smallest observation possible. But the (back transformed) model predictions for observations of zero counts range from 0.17 up to 182.84. It consistently overestimates the count causing the residuals to be uniformly negative, not normally distributed as assumed. The fact that we treat discrete count data as a continuous variable also causes the linear patterns seen in the figure. The q-q plot of the model with the transformed response confirms that we did not completely normalize the errors (figure 13). We see that the distribution has a heavier than expected tail on the left and lighter than expected tail on the right.

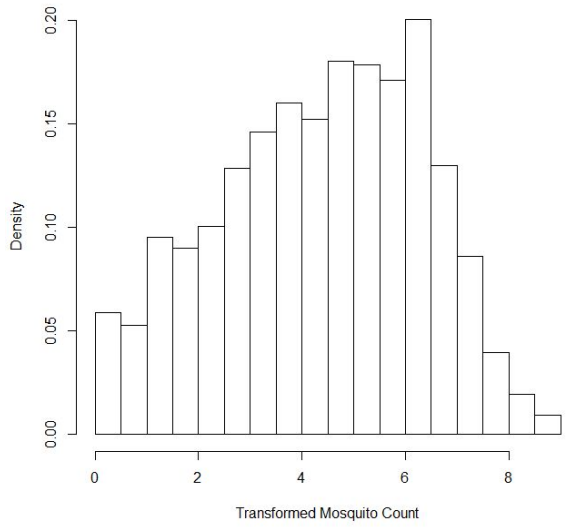

Figure 11. Histogram of back transformed counts

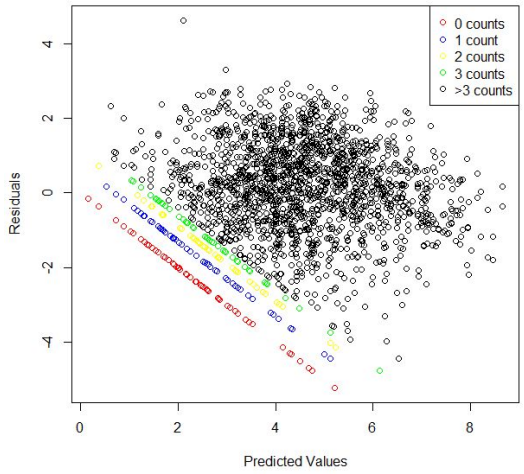

Figure 12. Non-constant error variance

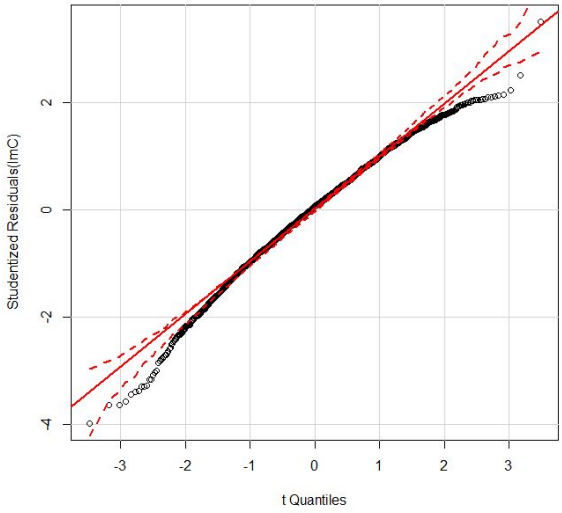

Figure 8. Q-Q plot of model using transformed counts

Next, we checked the autocorrelation of the series of residuals two different ways. Any apparent patterns might demonstrate a violation of the assumption of independent errors. The first series comes from sorting the data first by time period, then by trap number. Therefore a lag in the series roughly indicates correlations in the errors between neighboring traps. For example if the lag is 2, the correlation may be between trap sites 2 and 4 or trap sites 50 and 52, in the same time period. The next series comes from sorting first by trap number, then by time period. So a lag of 2 now indicates, for example, a comparison of early August to early September at a common trap location. The ideal correlation between residuals is 0, and a rule of thumb is that correlations greater than 2 are problematic. Under the Reisen model, the correlations are not particularly strong, but there are clear patterns (figure 14). This indicates relationships within the response that are inadequately captured by the model. The expected value of the Durbin-Watson statistic is 2, so values away from this provide support for a pattern in the correlations (here it is obvious anyway).

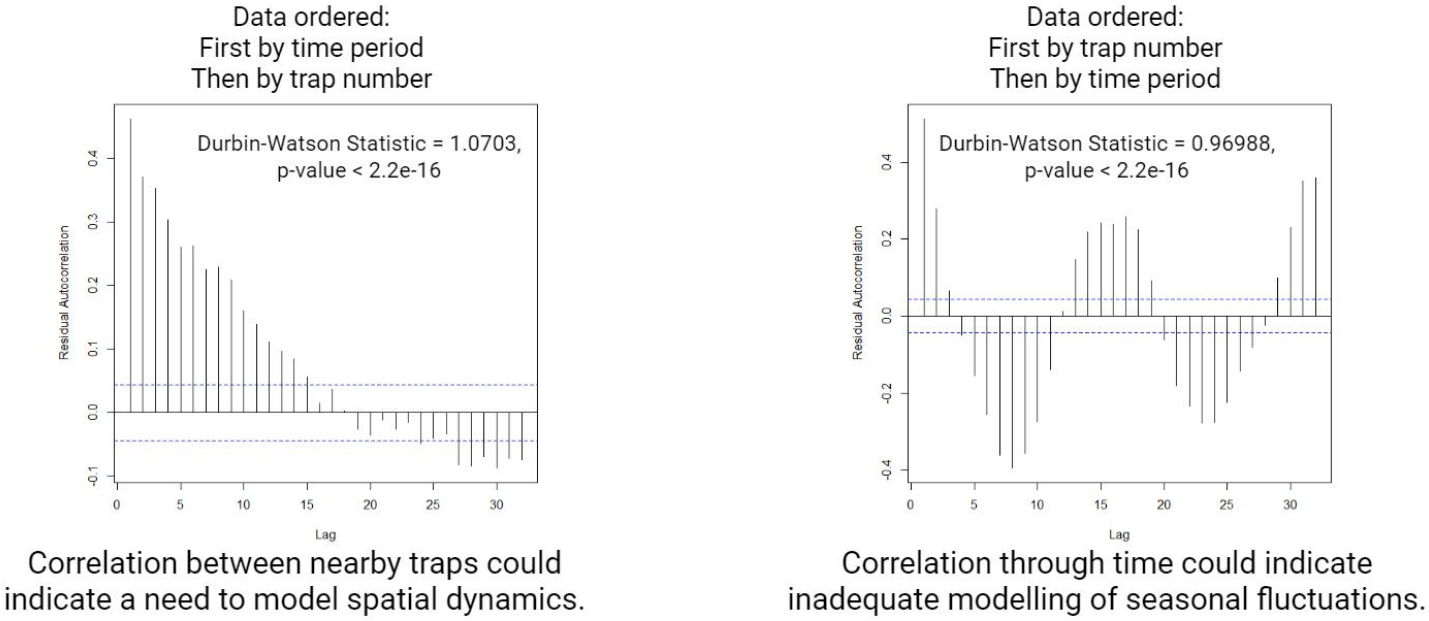

Figure 14. Patterns in residual series autocorrelations

After that, we checked for any data points that exert relatively high influence on the regression model compared to the other data. Influence in regression derives from a combination of “leverage” and conforming response. Leverage is a measure of how close a particular data point sits to the other data in the space spanned by the regressors of a model. Likewise a conforming response is one that matches well with model predictions that would be made if a given data point were removed. When a data point has both high leverage and a nonconforming response, it tends to “pull” the entire regression model to better accommodate itself. See figure 15 for a simple example from Montgomery [5]. Figure 16 is an influence plot of all the data points. The x-axis shows the data’s “hat-value”, which is a measure of leverage, and the y-axis shows the studentized residuals, a measure of conforming response. Data around the perimeter of this plot tends to have high influence. The data points are

further evaluated using Cook's Distance metric (represented by the size of the circles), which is an attempt to combine leverage and response into a single metric.

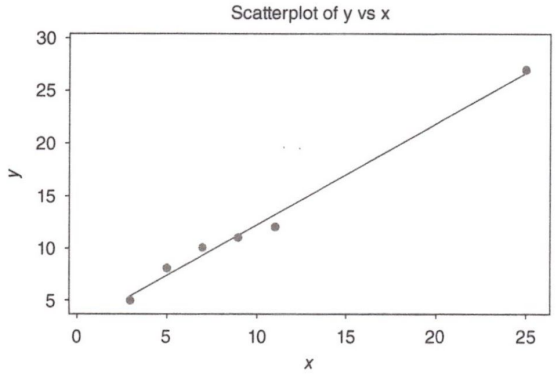

Figure 4.1 Example of a pure leverage point.

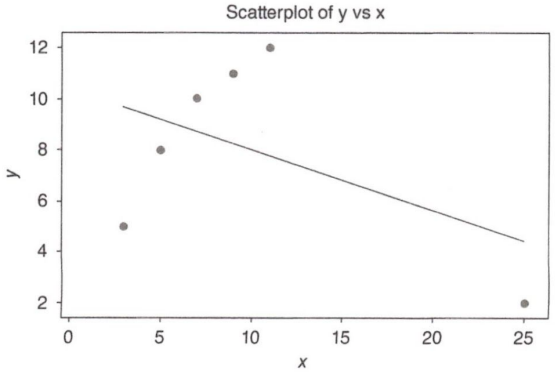

Figure 4.2 Example of an influential point.

Figure 15. Explanation of leverage and influence from Montgomery

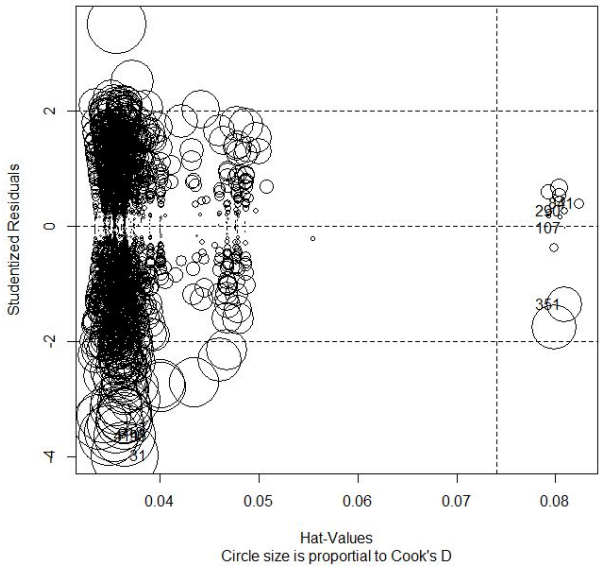

Figure 16. Data point influence plot

With this analysis we discovered the following influential data points: 31, 95, 107, 290, 351, 411, 613, and 841. These can be broken into two groups. Points 107, 290, 351, and 841 have high leverage. In fact, all four come from a set of 13 relatively high leverage points with leverage values near 0.08. This is the set of data collected from trap site 49, and the leverage is high because there were so few samples collected from this trap compared to all the others. The other four points all have low counts and a high count estimate leading to large residuals. They have a corresponding high value of Cook's Distance. These points are another manifestation of the previously noted deficiencies in the model, namely that we are modelling a discrete response using a continuous output and that the power transform used did not completely normalize the errors. We noted that all eight data points were collected in 1994. Other than that, manual inspection of these data points did not reveal anything obviously unordinary, and we continued with the exploratory analysis.

Lastly, we checked for multicollinearity among the regressors available for building a model. Using variance inflation factors as a metric, we found several groups of covariates that are incompatible with each other because they bring overlapping predictive information to the model. There are three groups of conflicting covariates:

1) observed temperature, maximum temperature, and month (as a factor);
2) latitude, longitude, habitats, and trap number (as a factor);
3) all nine habitats together since they are percentages that sum to 1.

Based on these incompatibilities, we only allow one covariate from group 1 and one covariate from group 2 to be present in a model. Also, when we use habitat as a covariate, we satisfy the group 3 conflict by leaving the "desert" percentage out of the model.