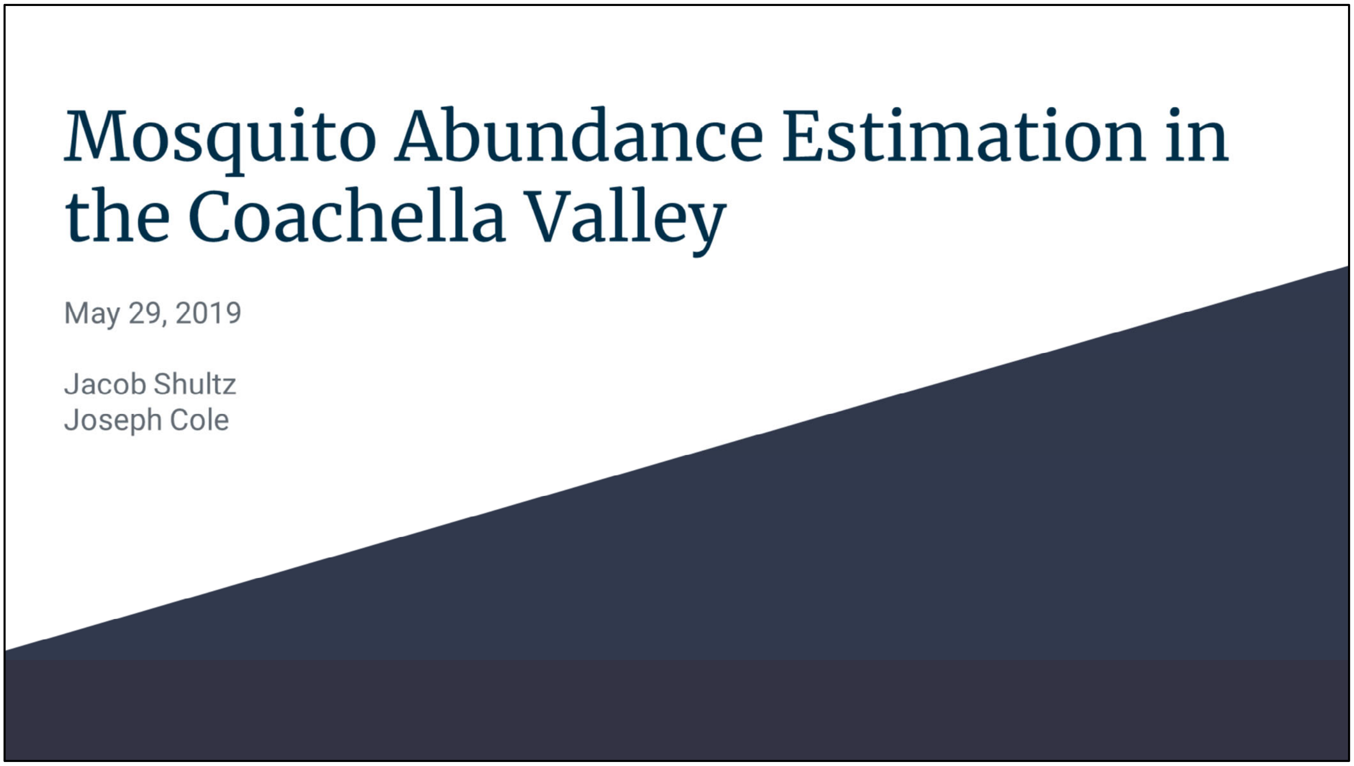


Mosquito Abundance Estimation in the Coachella Valley

May 29, 2019

Jacob Shultz
Joseph Cole

Problem Summary and Project Objectives

Given:

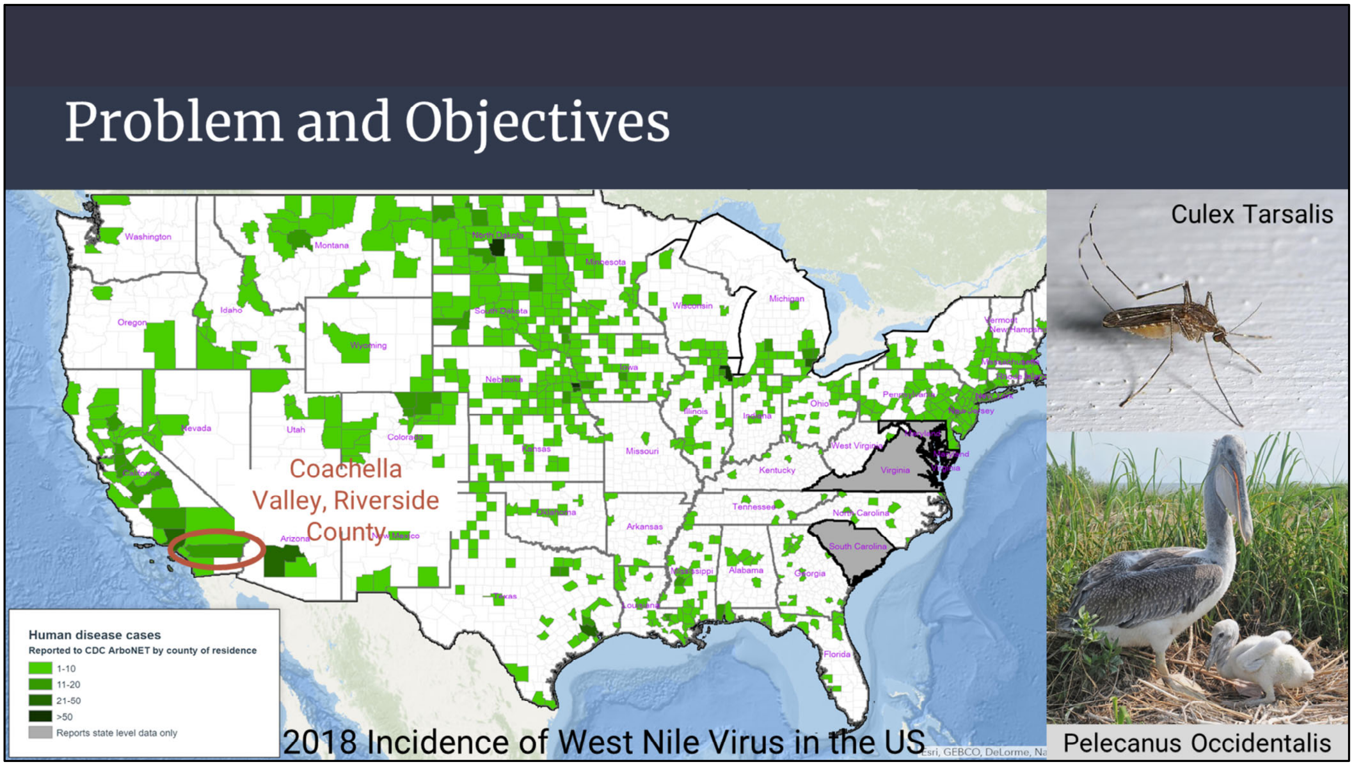
- Data set from [Reisen and Lothrop. J. Am. Mosquito Control Assoc. 15(2):105-114. 1999.]
  - counts from 63 traps set in the Coachella Valley (CA).
  - Collected from Apr 1994 to Nov 1995
  - Focus on *Culex tarsalis*
- Various predictor data
  - Daily temperature data for the Coachella Valley area over the entire study period
  - Latitude and longitude of each trap
  - Biome information for each trap

Objectives:

- Read [Reisen and Lothrop. J. Am. Mosquito Control Assoc. 15(2):105-114. 1999.] and reproduce the results therein
  - Mostly summary statistics. No predictive modeling
- Fit an N-Mixture model for open populations, as shown in [Dail and Madsen. Biometrics 67(2):577-587. 2011.]
- Assess fit: Cross Validation

Task:

- Use modern methods to fit a statistical model to this data
  - Estimate overall abundance of *Culex tarsalis*
  - Identify significant predictors
  - Apply Bayesian methods if possible
- Carry out model diagnostics and assess model fit

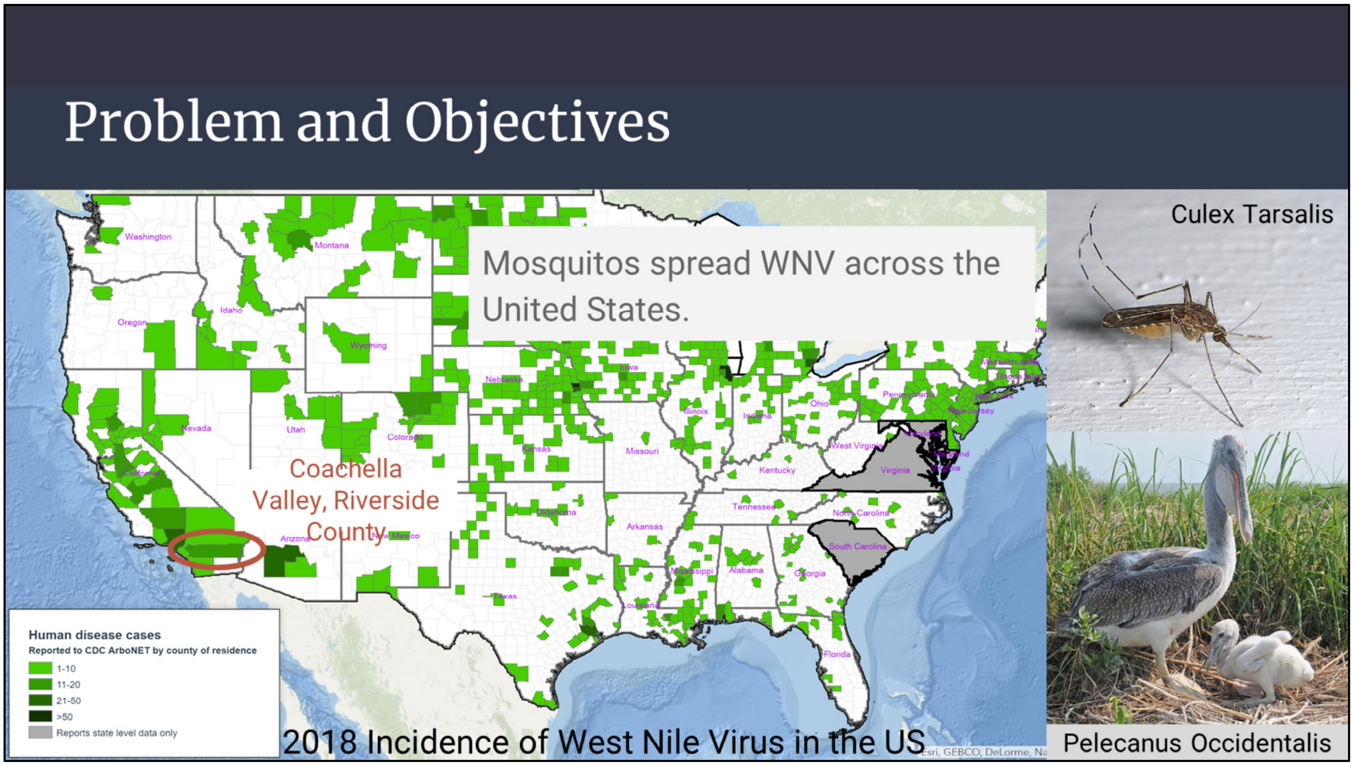


Problem and Objectives

Human disease cases
Reported to CDC ArboNET by county of residence
🟩 1-10
🟩 11-20
🟩 21-50
■ >50
▨ Reports state level data only

2018 Incidence of West Nile Virus in the US

1) West Nile Virus is spread by mosquitos, often with birds as intermediary carriers. In less than 1 percent of infected people, the virus causes a serious neurological infection, including inflammation of the brain (encephalitis) and of the membranes surrounding the brain and spinal cord (meningitis). Symptoms include paralysis and comas. Severe cases can lead to death.
2) Mosquito control is often paid for through county level taxes, so resources are scarce. Accurate mosquito population prediction can help decision makers better target abatement measures.
3) Experts from the Coachella Valley Mosquito and Vector Control District, along with the University of California, Davis, collected an extensive dataset in 1994-1995 to help make trapping efforts more efficient while maintaining a level of accuracy and precision. They used frequentist summary statistics to compare various sampling schemes and showed that stratified random sampling provides an optimal balance between accuracy & precision versus trapping effort. These methods do not allow estimation of total population abundance because the probability of successfully trapping an individual mosquito is unknown and unmodelled.



Problem and Objectives

Mosquitos spread WNV across the United States.

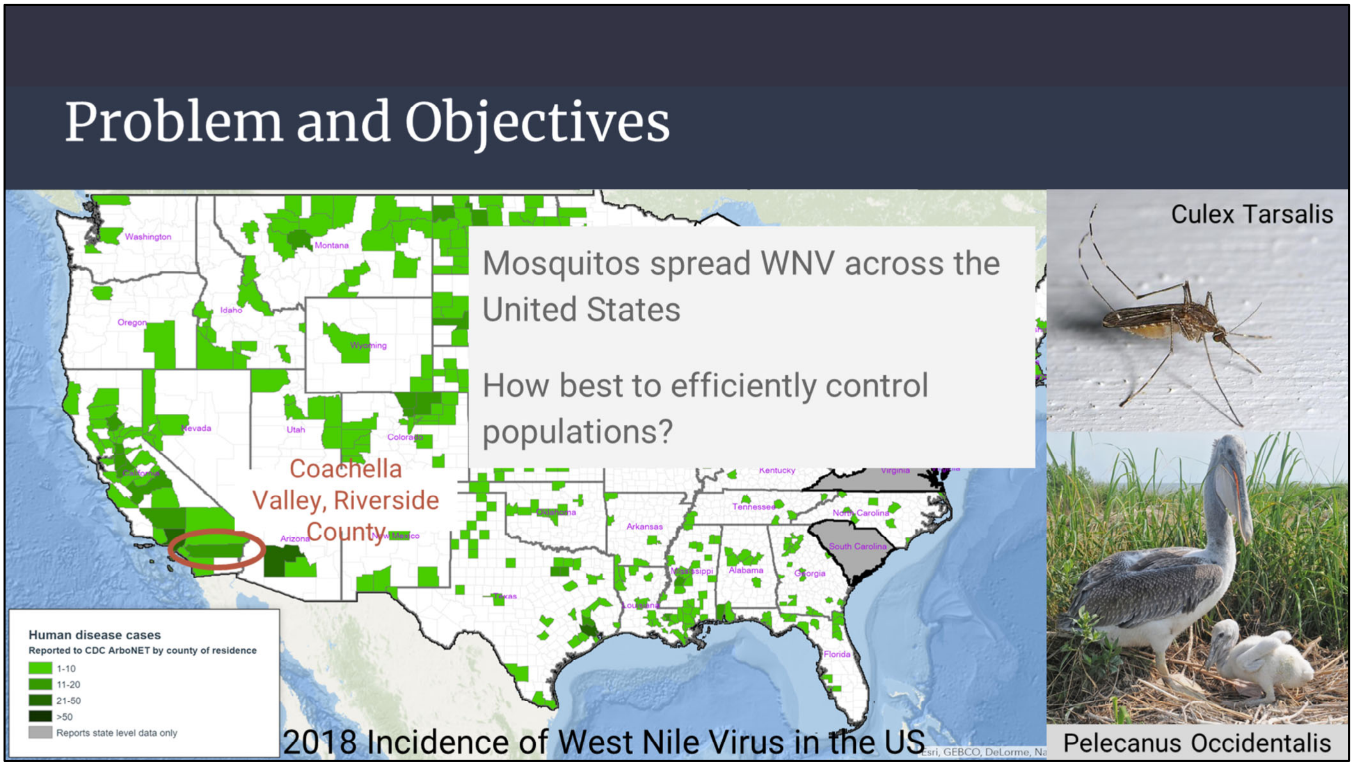
Coachella Valley, Riverside County

2018 Incidence of West Nile Virus in the US

Culex Tarsalis

Pelecanus Occidentalis

1) West Nile Virus is spread by mosquitos, often with birds as intermediary carriers. In less than 1 percent of infected people, the virus causes a serious neurological infection, including inflammation of the brain (encephalitis) and of the membranes surrounding the brain and spinal cord (meningitis). Symptoms include paralysis and comas. Severe cases can lead to death.
2) Mosquito control is often paid for through county level taxes, so resources are scarce. Accurate mosquito population prediction can help decision makers better target abatement measures.
3) Experts from the Coachella Valley Mosquito and Vector Control District, along with the University of California, Davis, collected an extensive dataset in 1994-1995 to help make trapping efforts more efficient while maintaining a level of accuracy and precision. They used frequentist summary statistics to compare various sampling schemes and showed that stratified random sampling provides an optimal balance between accuracy & precision versus trapping effort. These methods do not allow estimation of total population abundance because the probability of successfully trapping an individual mosquito is unknown and unmodelled.



Problem and Objectives

Mosquitos spread WNV across the United States

How best to efficiently control populations?

Coachella Valley, Riverside County

Human disease cases
Reported to CDC ArboNET by county of residence

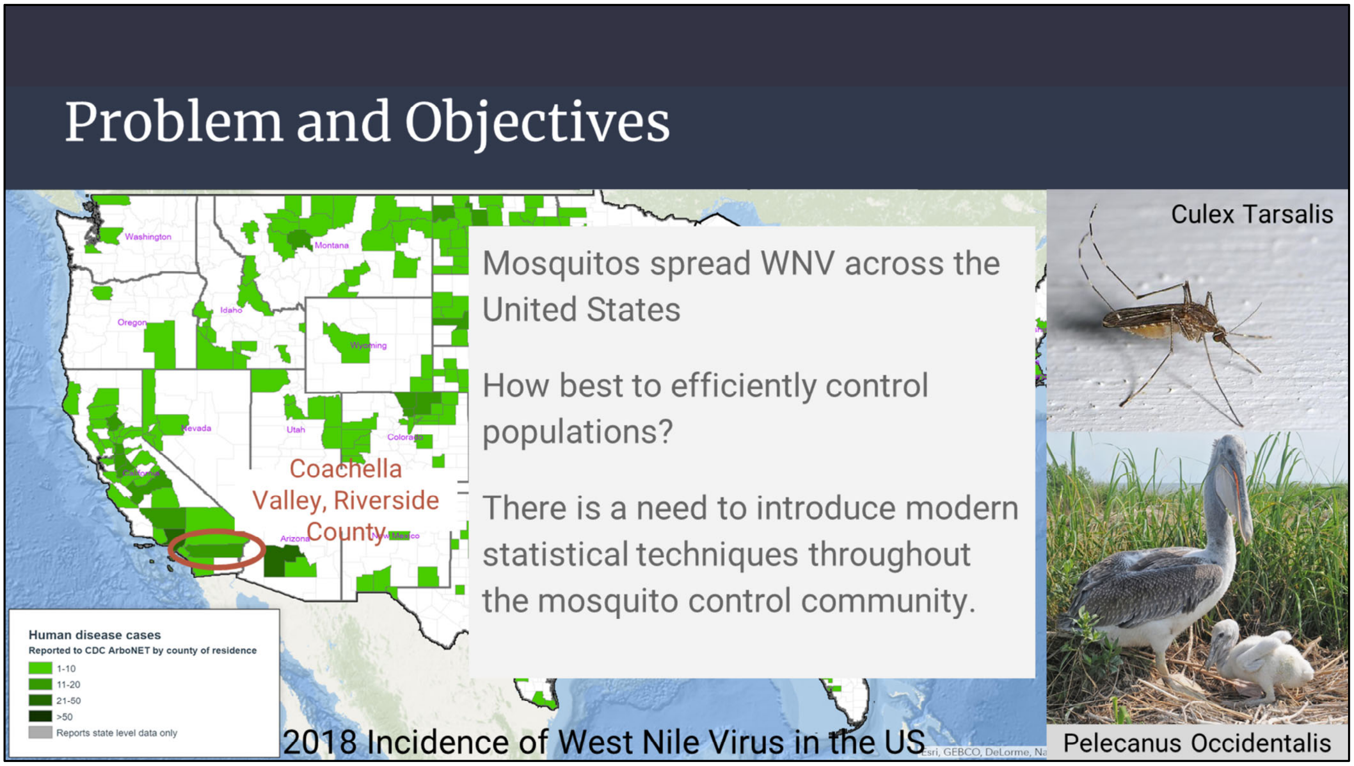
🟩 1-10
🟩 11-20
🟩 21-50
⬛ >50
⬜ Reports state level data only

2018 Incidence of West Nile Virus in the US

Culex Tarsalis

Pelecanus Occidentalis

1) West Nile Virus is spread by mosquitos, often with birds as intermediary carriers. In less than 1 percent of infected people, the virus causes a serious neurological infection, including inflammation of the brain (encephalitis) and of the membranes surrounding the brain and spinal cord (meningitis). Symptoms include paralysis and comas. Severe cases can lead to death.
2) Mosquito control is often paid for through county level taxes, so resources are scarce. Accurate mosquito population prediction can help decision makers better target abatement measures.
3) Experts from the Coachella Valley Mosquito and Vector Control District, along with the University of California, Davis, collected an extensive dataset in 1994-1995 to help make trapping efforts more efficient while maintaining a level of accuracy and precision. They used frequentist summary statistics to compare various sampling schemes and showed that stratified random sampling provides an optimal balance between accuracy & precision versus trapping effort. These methods do not allow estimation of total population abundance because the probability of successfully trapping an individual mosquito is unknown and unmodelled.



Problem and Objectives

Mosquitos spread WNV across the United States

How best to efficiently control populations?

There is a need to introduce modern statistical techniques throughout the mosquito control community.

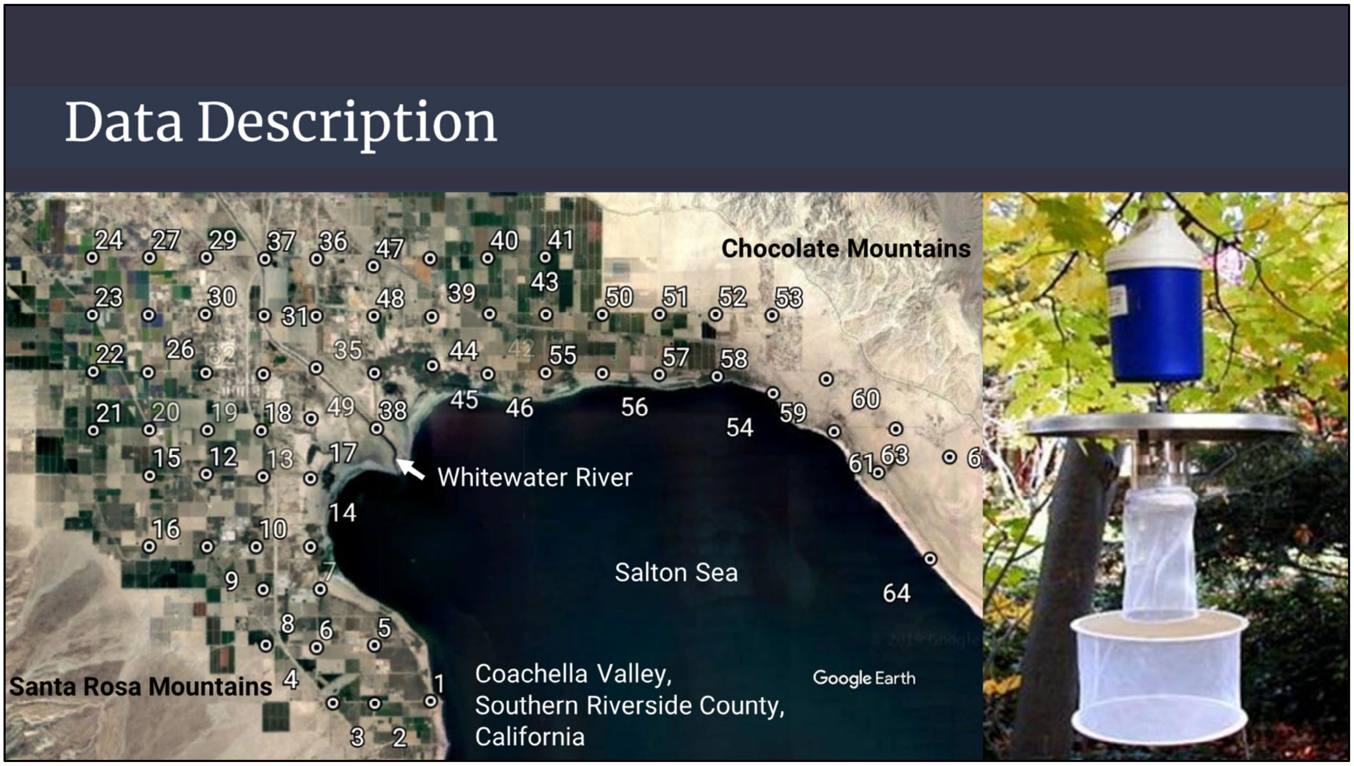
Coachella Valley, Riverside County

2018 Incidence of West Nile Virus in the US

Culex Tarsalis

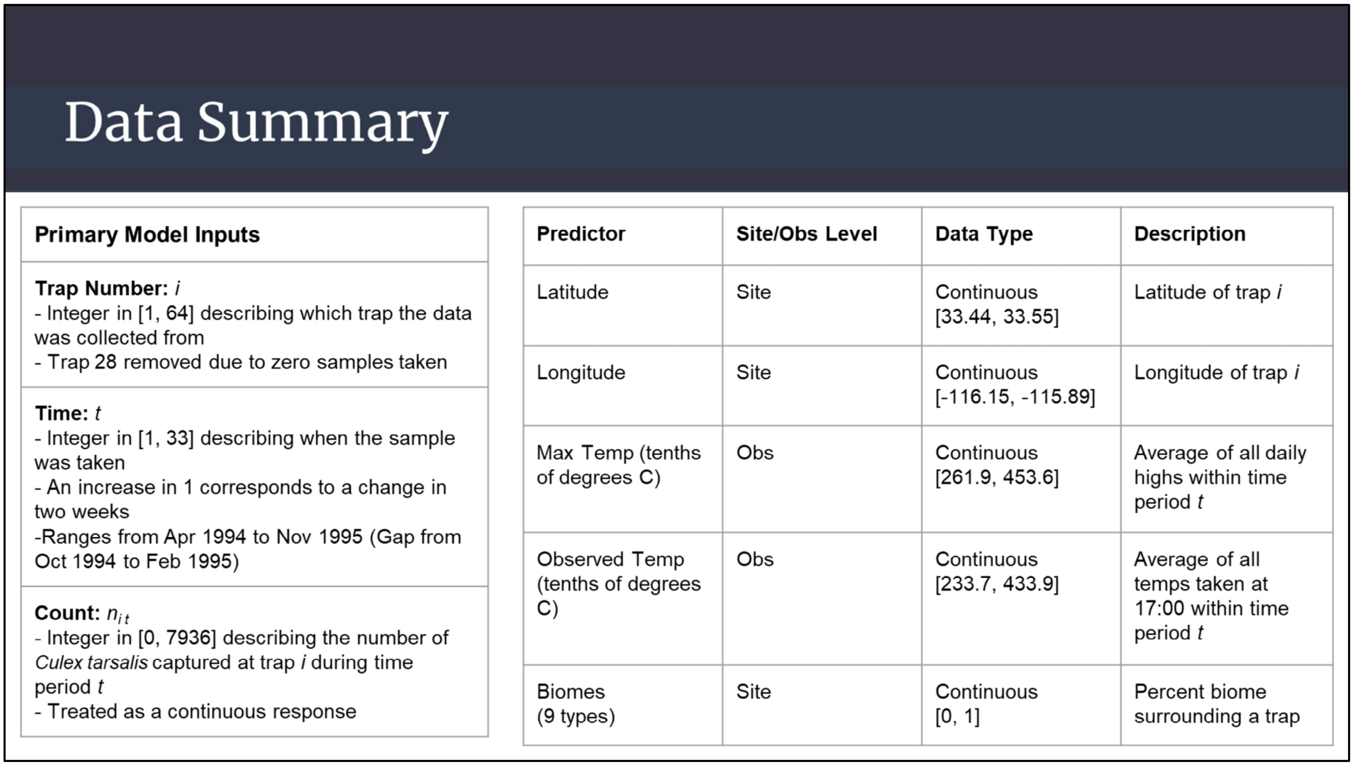
Pelecanus Occidentalis

1) West Nile Virus is spread by mosquitos, often with birds as intermediary carriers. In less than 1 percent of infected people, the virus causes a serious neurological infection, including inflammation of the brain (encephalitis) and of the membranes surrounding the brain and spinal cord (meningitis). Symptoms include paralysis and comas. Severe cases can lead to death.
2) Mosquito control is often paid for through county level taxes, so resources are scarce. Accurate mosquito population prediction can help decision makers better target abatement measures.
3) Experts from the Coachella Valley Mosquito and Vector Control District, along with the University of California, Davis, collected an extensive dataset in 1994-1995 to help make trapping efforts more efficient while maintaining a level of accuracy and precision. They used frequentist summary statistics to compare various sampling schemes and showed that stratified random sampling provides an optimal balance between accuracy & precision versus trapping effort. These methods do not allow estimation of total population abundance because the probability of successfully trapping an individual mosquito is unknown and unmodelled.



Data Description

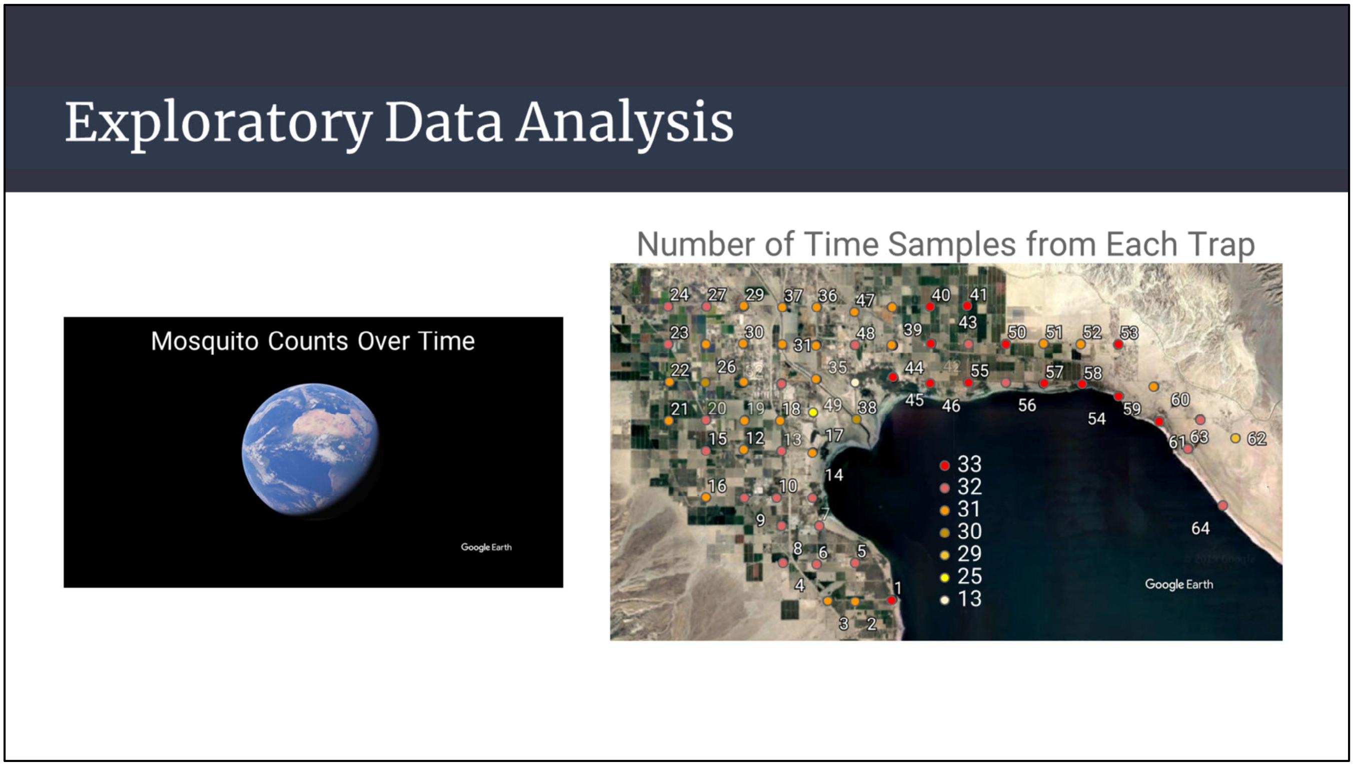
1) 63 dry ice baited traps were set at 1 mile intervals on a grid. Chunks of dry ice produce CO2 at a rate similar to that of a large mammal. A fan sucks insects into the netted area where they are trapped. Traps were left overnight from dusk until dawn once every 2 weeks. Only female mosquitos are attracted to CO2 because they require the protein from a blood meal to lay a clutch of eggs. They can sense gradients of the gas from up to 30 meters away. Collected mosquitos were anesthetized, identified by species, and counted for each trap site.
2) Pelicans and other migratory water fowl nest along the beach of the Salton Sea and throughout duck marshes that are flooded each year for hunting season (late October and November) along the Whitewater River. Chicks of these birds are the preferred source of a blood meal for Culex Tarsalis.
3) Mosquitos typically live for about 3 weeks, so a particular individual will usually be exposed to 2 trap-nights under this collection scheme.
4) Additional predictive covariates provided to us include: habitat classifications around each trap site (desert, salt marsh, duck pond, various types of farms, etc.) and daily temperature collected at the firestation in the town of Mecca for the NOAA (Rexmsrep$Sgiermg$erh$Exqswtlivmg$Ehqmrmwxvexmsr).



Data Summary

Primary Model Inputs
Trap Number: *i* - Integer in [1, 64] describing which trap the data was collected from - Trap 28 removed due to zero samples taken
Time: *t* - Integer in [1, 33] describing when the sample was taken - An increase in 1 corresponds to a change in two weeks -Ranges from Apr 1994 to Nov 1995 (Gap from Oct 1994 to Feb 1995)
Count: $n_{i\,t}$ - Integer in [0, 7936] describing the number of *Culex tarsalis* captured at trap *i* during time period *t* - Treated as a continuous response

Predictor	Site/Obs Level	Data Type	Description
Latitude	Site	Continuous [33.44, 33.55]	Latitude of trap *i*
Longitude	Site	Continuous [-116.15, -115.89]	Longitude of trap *i*
Max Temp (tenths of degrees C)	Obs	Continuous [261.9, 453.6]	Average of all daily highs within time period *t*
Observed Temp (tenths of degrees C)	Obs	Continuous [233.7, 433.9]	Average of all temps taken at 17:00 within time period *t*
Biomes (9 types)	Site	Continuous [0, 1]	Percent biome surrounding a trap

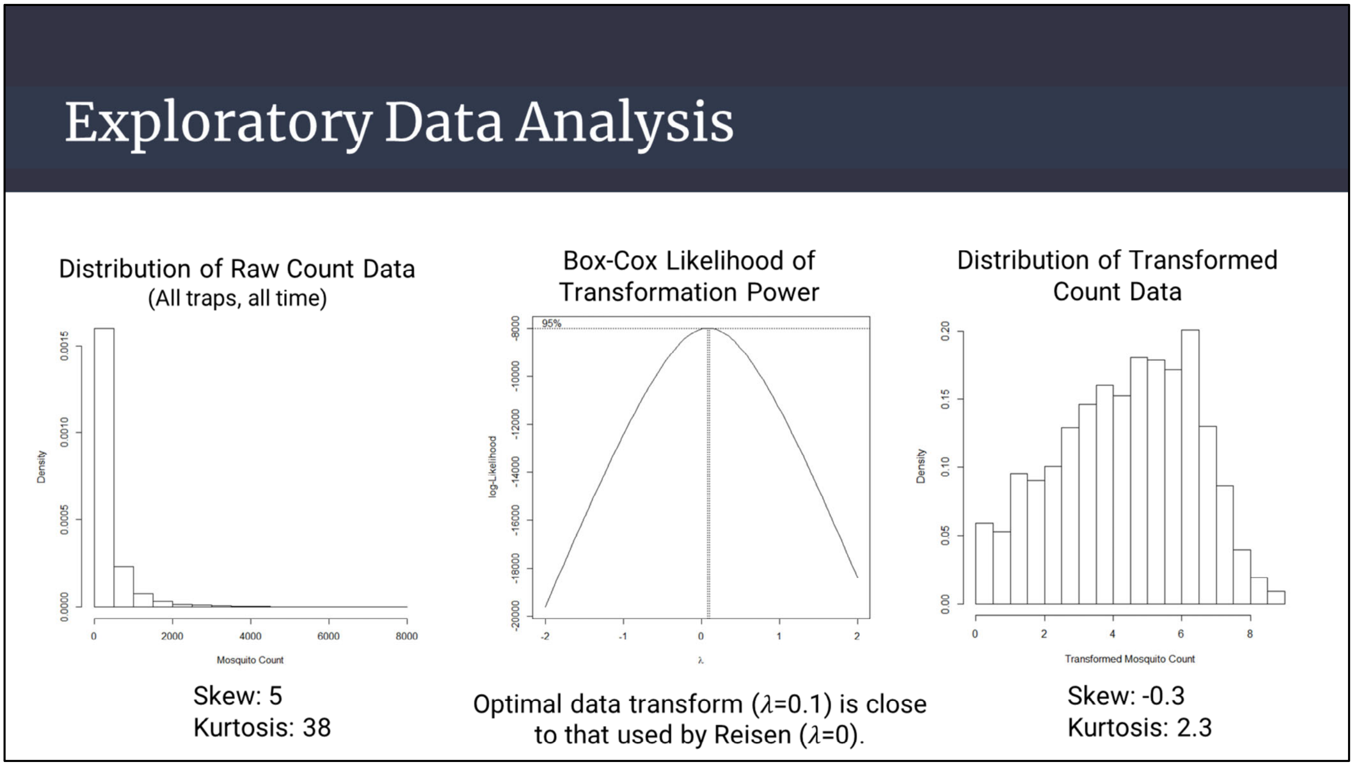


Exploratory Data Analysis

Mosquito Counts Over Time

Number of Time Samples from Each Trap

1) Animating the mosquito counts allowed us to check for indications of mosquito population dynamics in time and space. Selecting the color scale of the overlays was important. We ended up dividing the empirical CDF into 64 equal quantiles and assigning each a color. Due to the long tail of high count data, this choice causes yellow-like colors to signify a wider range of counts than blue-like colors. For example one shade of yellow might indicate 6000-8000 mosquitos, while a single shade of blue might indicate 50-60 mosquitos.
2) We noted that the spatial distribution of time samples was fairly uniform with the exception of the area near the mouth of the Whitewater river. In particular traps 17, 38, and 49 had relatively few observations. We suspect the accessibility of these trap sites may have been problematic. We also noted that during one of the time periods (T=32) data was only collected in the eastern half of the study area.



Exploratory Data Analysis

Distribution of Raw Count Data
(All traps, all time)

Skew: 5
Kurtosis: 38

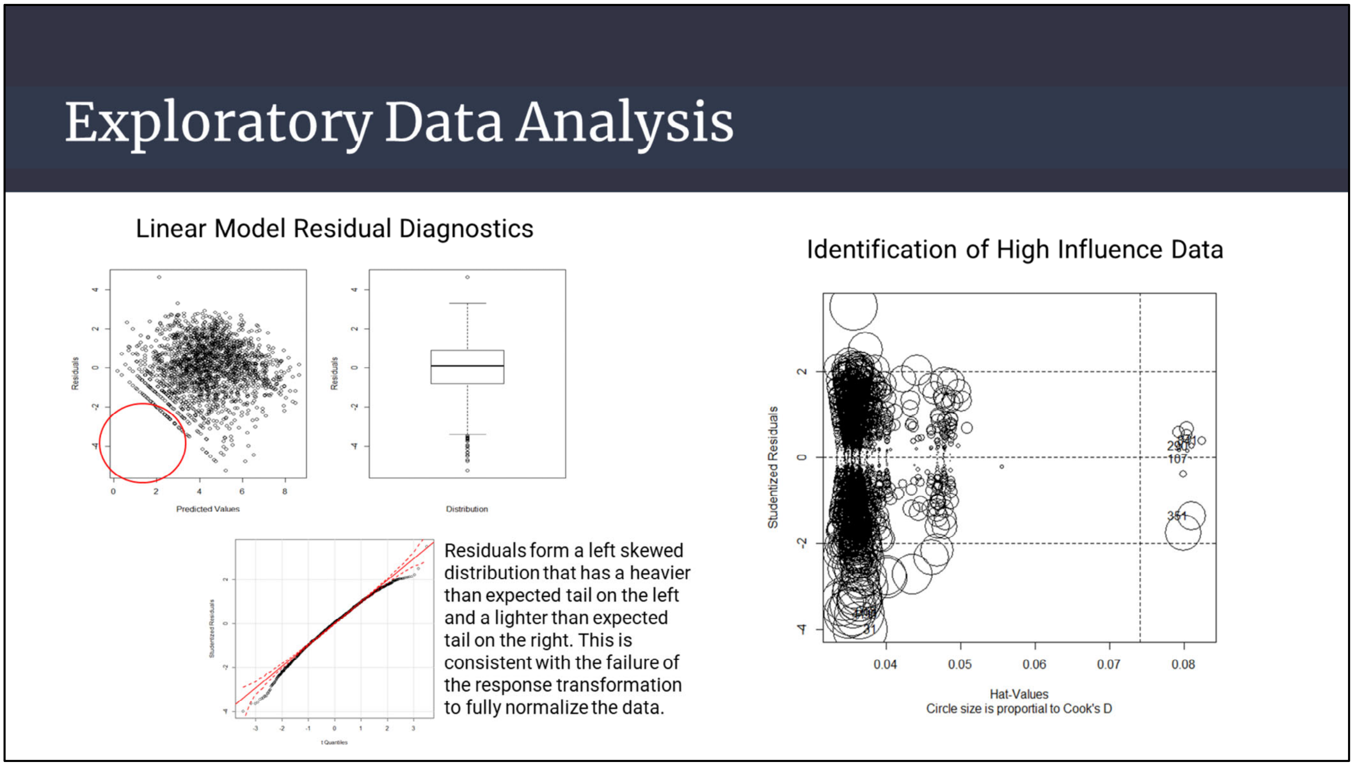
Box-Cox Likelihood of Transformation Power

Optimal data transform (λ=0.1) is close to that used by Reisen (λ=0).

Distribution of Transformed Count Data

Skew: -0.3
Kurtosis: 2.3

The raw mosquito count data was highly right skewed, making Reisen’s original regression analysis in 1999 problematic. He made an adequate attempt to address the issue by applying a log transform to the response variable, and this choice was about the best that could be done when limited to models that assume normality. However, there appears to be plenty of room for improvement!



Exploratory Data Analysis

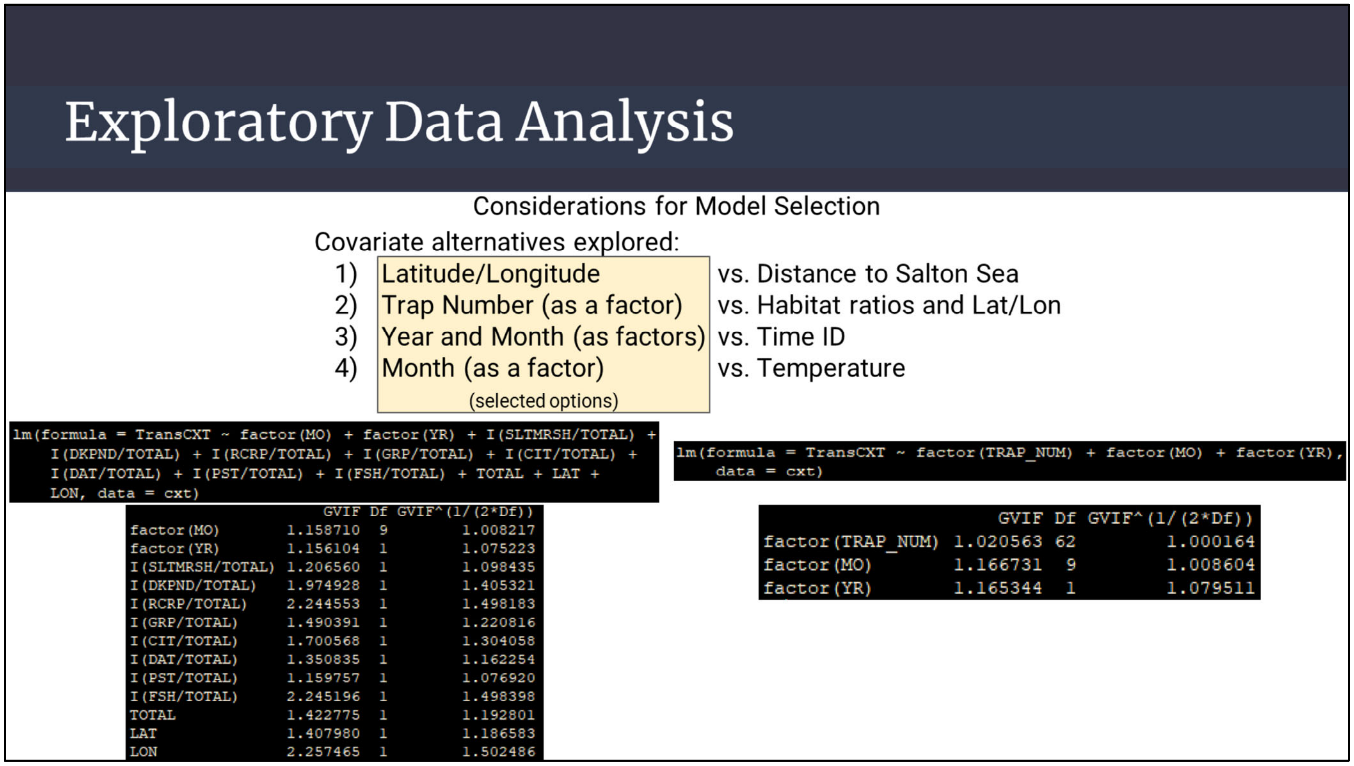
Linear Model Residual Diagnostics

Identification of High Influence Data

Residuals form a left skewed distribution that has a heavier than expected tail on the left and a lighter than expected tail on the right. This is consistent with the failure of the response transformation to fully normalize the data.

Data divides into 3 levels of leverage. This phenomenon is still under investigation.

We discovered the following data points exert the greatest influence on the regression:
31, 95, 107, 290, 351, 411, 613, and 841
Manual inspection of these data points didn’t reveal anything obviously unordinary except they all occur in the year 1994.



Exploratory Data Analysis

Considerations for Model Selection

Covariate alternatives explored:

1) Latitude/Longitude vs. Distance to Salton Sea
2) Trap Number (as a factor) vs. Habitat ratios and Lat/Lon
3) Year and Month (as factors) vs. Time ID
4) Month (as a factor) vs. Temperature

(selected options)

```
lm(formula = TransCXT ~ factor(MO) + factor(YR) + I(SLTMRSH/TOTAL) +
    I(DKPND/TOTAL) + I(RCRP/TOTAL) + I(GRP/TOTAL) + I(CIT/TOTAL) +
    I(DAT/TOTAL) + I(PST/TOTAL) + I(FSH/TOTAL) + TOTAL + LAT +
    LON, data = cxt)
```

```
                     GVIF Df GVIF^(1/(2*Df))
factor(MO)       1.158710  9        1.008217
factor(YR)       1.156104  1        1.075223
I(SLTMRSH/TOTAL) 1.206560  1        1.098435
I(DKPND/TOTAL)   1.974928  1        1.405321
I(RCRP/TOTAL)    2.244553  1        1.498183
I(GRP/TOTAL)     1.490391  1        1.220816
I(CIT/TOTAL)     1.700568  1        1.304058
I(DAT/TOTAL)     1.350835  1        1.162254
I(PST/TOTAL)     1.159757  1        1.076920
I(FSH/TOTAL)     2.245196  1        1.498398
TOTAL            1.422775  1        1.192801
LAT              1.407980  1        1.186583
LON              2.257465  1        1.502486
```

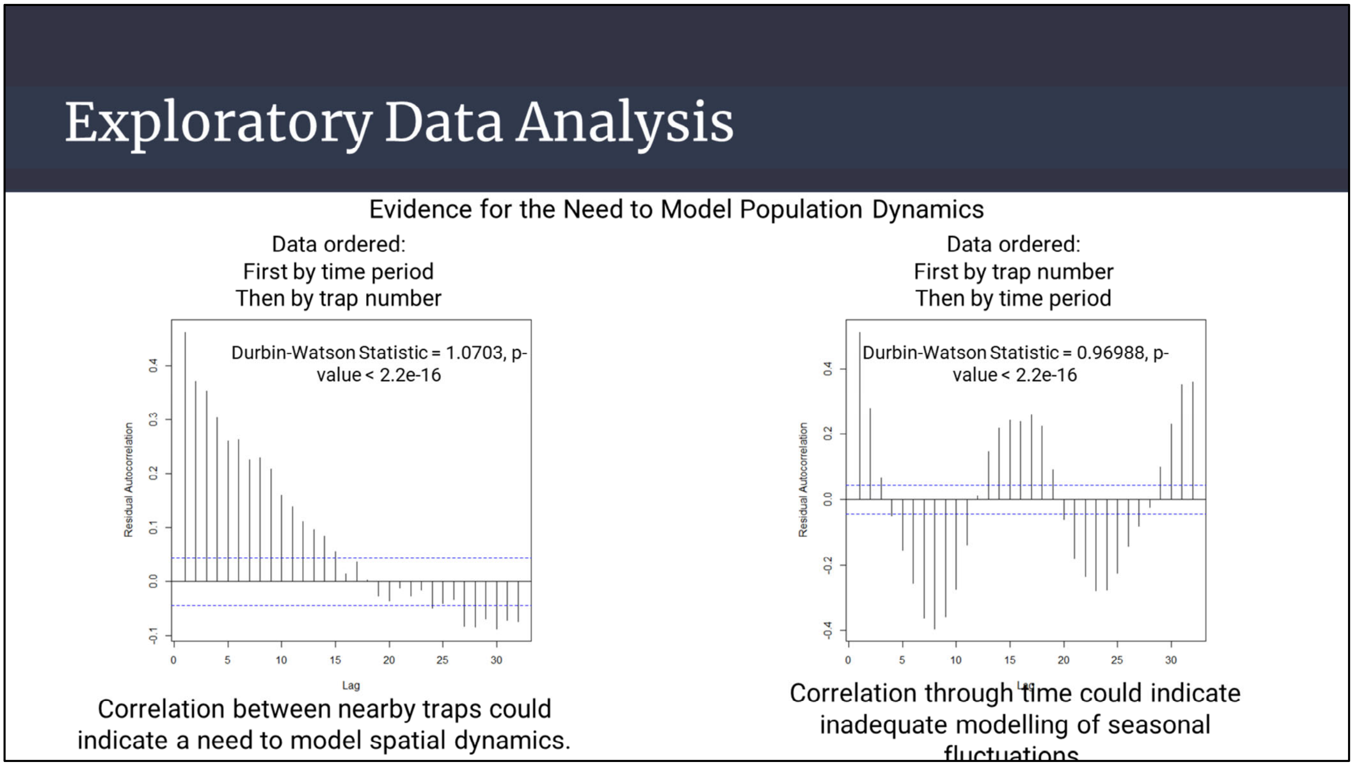
```
lm(formula = TransCXT ~ factor(TRAP_NUM) + factor(MO) + factor(YR),
    data = cxt)
```

```
                     GVIF Df GVIF^(1/(2*Df))
factor(TRAP_NUM) 1.020563 62        1.000164
factor(MO)       1.166731  9        1.008604
factor(YR)       1.165344  1        1.079511
```

Notes on multicollinearity:

1) temperature (Observed and/or max) conflicts with factor(MO)
2) factor(TRAP_NUM) conflicts with LAT, LON, and the habitat covariates
3) including all the habitat types as raw covariates causes problems, best to leave out DESERT

Rule of thumb is that variance inflation factors greater than 5 could indicate model problems due to multicollinearity.



Exploratory Data Analysis

Evidence for the Need to Model Population Dynamics

Data ordered:
First by time period
Then by trap number

Durbin-Watson Statistic = 1.0703, p-value < 2.2e-16

Correlation between nearby traps could indicate a need to model spatial dynamics.

Data ordered:
First by trap number
Then by time period

Durbin-Watson Statistic = 0.96988, p-value < 2.2e-16

Correlation through time could indicate inadequate modelling of seasonal fluctuations

Linear models require the assumption of independent errors. Therefore, the ideal correlation between residuals should be 0, and a rule of thumb is that correlations greater than 2 are problematic. There should be no discernable pattern in the residual correlations. In our case the residual correlations are not particularly strong, but there are obvious patterns. 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).

Methodology: N-Mixture Model for Closed Populations

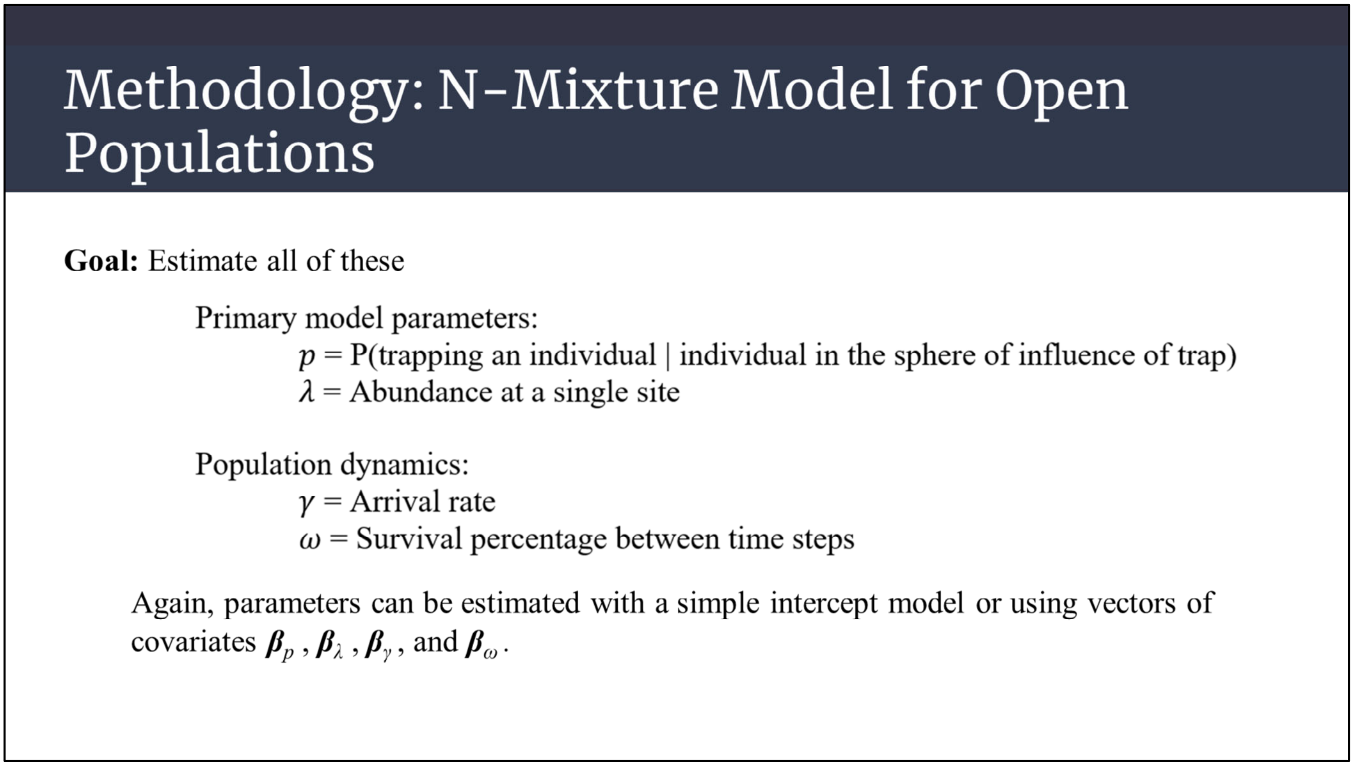
Goal: Estimate these primary model parameters

p = P(trapping an individual | individual in the sphere of influence of trap)
λ = Abundance at a single site

Both parameters can be estimated with a simple intercept model or using vectors of covariates $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_\lambda$.

Open/Closed Populations:

- Closed populations have a constant site-level populations over time
- Open populations can have "additions or deletions" in site-level populations (Dail and Madson, 2011)
    - Require modeling of population dynamics...



Methodology: N-Mixture Model for Open Populations

Goal: Estimate all of these

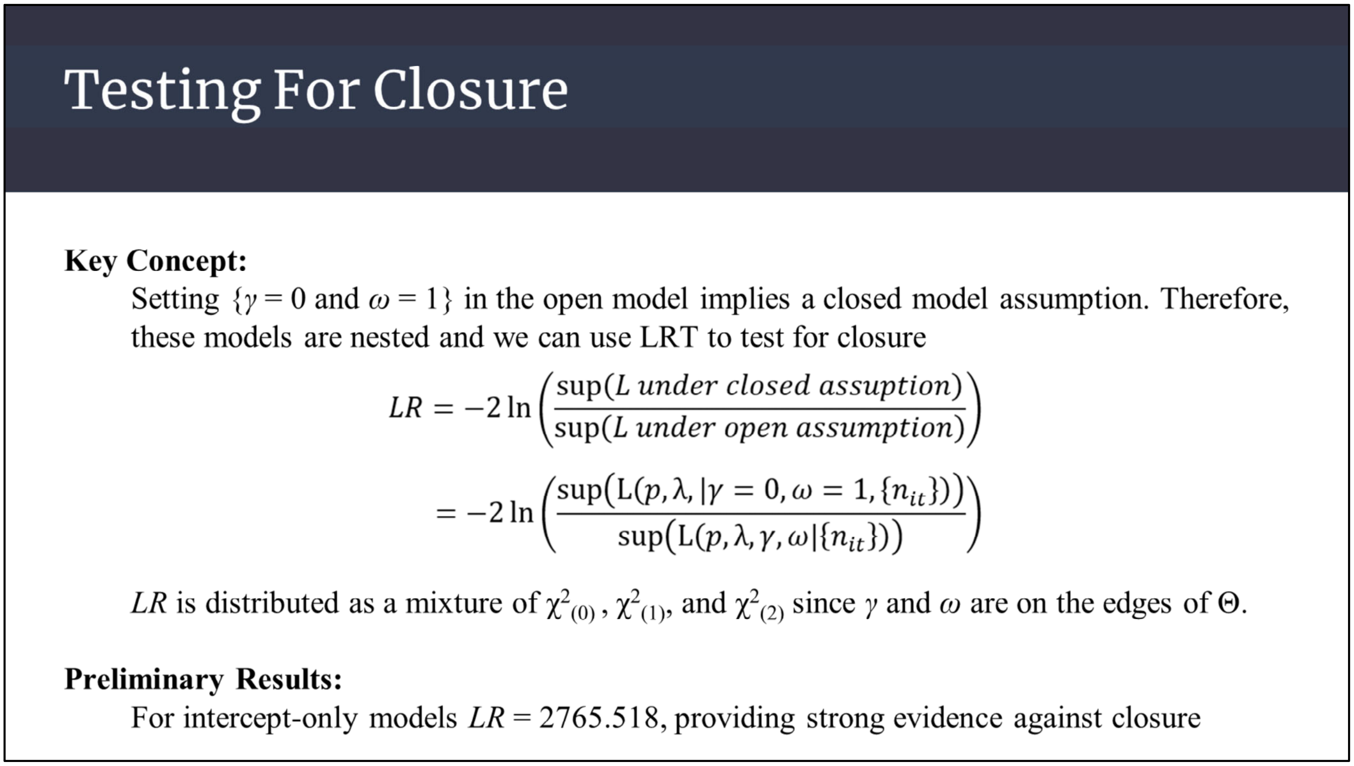
Primary model parameters:

p = P(trapping an individual | individual in the sphere of influence of trap)
λ = Abundance at a single site

Population dynamics:

γ = Arrival rate
ω = Survival percentage between time steps

Again, parameters can be estimated with a simple intercept model or using vectors of covariates $\boldsymbol{\beta}_p$, $\boldsymbol{\beta}_\lambda$, $\boldsymbol{\beta}_\gamma$, and $\boldsymbol{\beta}_\omega$.



Testing For Closure

Key Concept:

Setting {$\gamma = 0$ and $\omega = 1$} in the open model implies a closed model assumption. Therefore, these models are nested and we can use LRT to test for closure

$$LR = -2\ln\left(\frac{\sup(L\ under\ closed\ assuption)}{\sup(L\ under\ open\ assumption)}\right)$$

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

LR is distributed as a mixture of $\chi^2_{(0)}$, $\chi^2_{(1)}$, and $\chi^2_{(2)}$ since γ and ω are on the edges of Θ.

Preliminary Results:

For intercept-only models $LR = 2765.518$, providing strong evidence against closure

Preliminary Results: Parameter Estimates

All estimates were fit under intercept models (no covariates)

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

Open population assumption seems more reasonable

- Fitted abundance is more consistent with count summary statistics
- Smaller AIC in open population model

Data Format

Trap number	Longitude	Latitude	Max Temp1…	…Max Temp 33	Obs Temp 1…	…Obs Temp 33	Time 1…	…Time 33	Count 1….	…Count 33
1	Long @ Trap 1	Lat @ Trap 1	First maxTemp @ Trap 1	33rd maxTemp @ Trap 1 (or NA)	First obsTemp @ Trap 1	33rd obsTemp @ Trap 1 (or NA)	Time period *t* of first observation @ Trap 1	Time period *t* of 33rd observation @ Trap 1 (or NA)	# *Culex tarsalis* in trap 1 during time period 1	# *Culex tarsalis* in trap 1 during time period 33 (or NA)
…	…	…	…	…	…	…	…	…	…	…
64	Long @ Trap 64	Lat @ Trap 64	First maxTemp @ Trap 64	33rd maxTemp @ Trap 64 (or NA)	First obsTemp @ Trap 64	33rd obsTemp @ Trap 64 (or NA)	Time period *t* of first observation @ Trap 64	Time period *t* of 33rd observation @ Trap 64 (or NA)	# *Culex tarsalis* in trap 64 during time period 1	# *Culex tarsalis* in trap 64 during time period 33 (or NA)