# Group 1 GC 1-2

Yizhou Gu Matthew McAnear Sam Rosenberg Victor Verma

format: beamer: linkcolor: blue link-citations: true citation-color: purple bibliography: references.bib execute: echo: FALSE fig-margin: TRUE

# Introduction

- Our goal is to investigate the ranges of bat populations located along the shores of the Great Lakes. Specifically, we are interested in the following research questions:
  - Which species have the largest ranges
  - How population density varies with distance to the shore
  - How stable different species' ranges have been, i.e., how much their ranges have changed over time
- We downloaded a point-referenced dataset called the USFWS Great Lakes and Upper Midwest Acoustic Bat Dataset ("USFWS Great Lakes and Upper Midwest Acoustic Bat Dataset" 2019).
  - The downloaded dataset contains one row for each pair of a site and a night on which data was recorded at that site. For each of several species, there are columns giving the number of detected bats belonging to the species.

# Downloaded Dataset Structure

Some of the columns in the first few rows of the downloaded dataset are shown below. In total, there are 33168 observations from 276 sites.

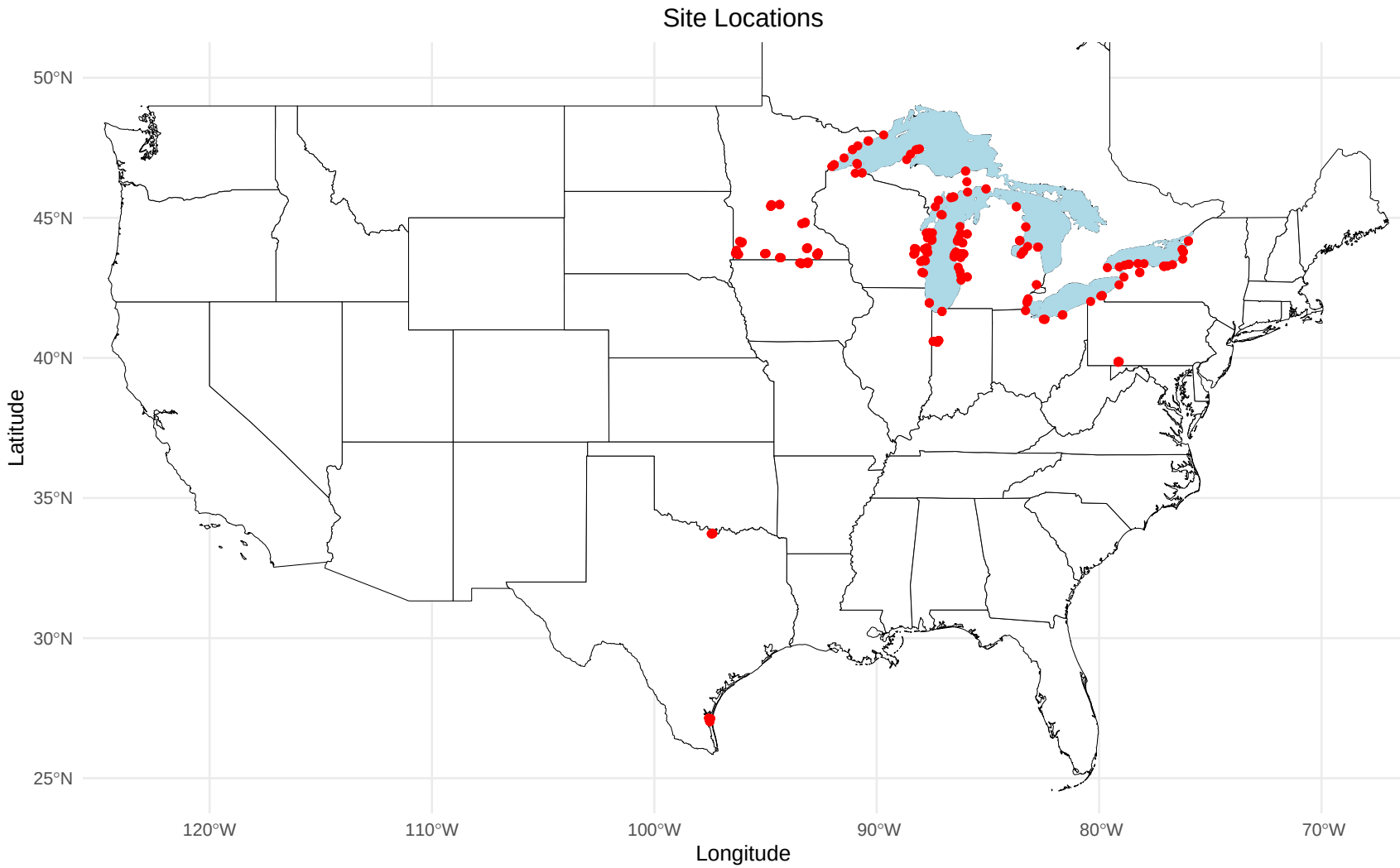
```
# A tibble: 6 x 4
  AcousticSite NightOf              EPTFUS LASBOR
  <chr>        <dttm>                <dbl>  <dbl>
1 B1           2010-08-01 00:00:00       0      0
2 B2           2010-08-01 00:00:00       0      0
3 B3           2010-08-01 00:00:00       0      0
4 BM1500       2010-08-01 00:00:00       0      0
5 BM5K         2010-08-01 00:00:00       0      0
6 BMR          2010-08-01 00:00:00       0      0
```

# Downloaded Dataset Structure

- We have data on 14 species. Some examples are the Big Brown Bat (Eptesicus fuscus) (referred to as EPTFUS in the data), Eastern Red Bat(Lasiurus borealis) (LASBOR), and Hoary Bat (Lasiurus cinereus) (LASCIN).
- Hence, the dataset we downloaded contains 33168 observations, each of which contains a site, a date, and counts of detections for 14 different species.

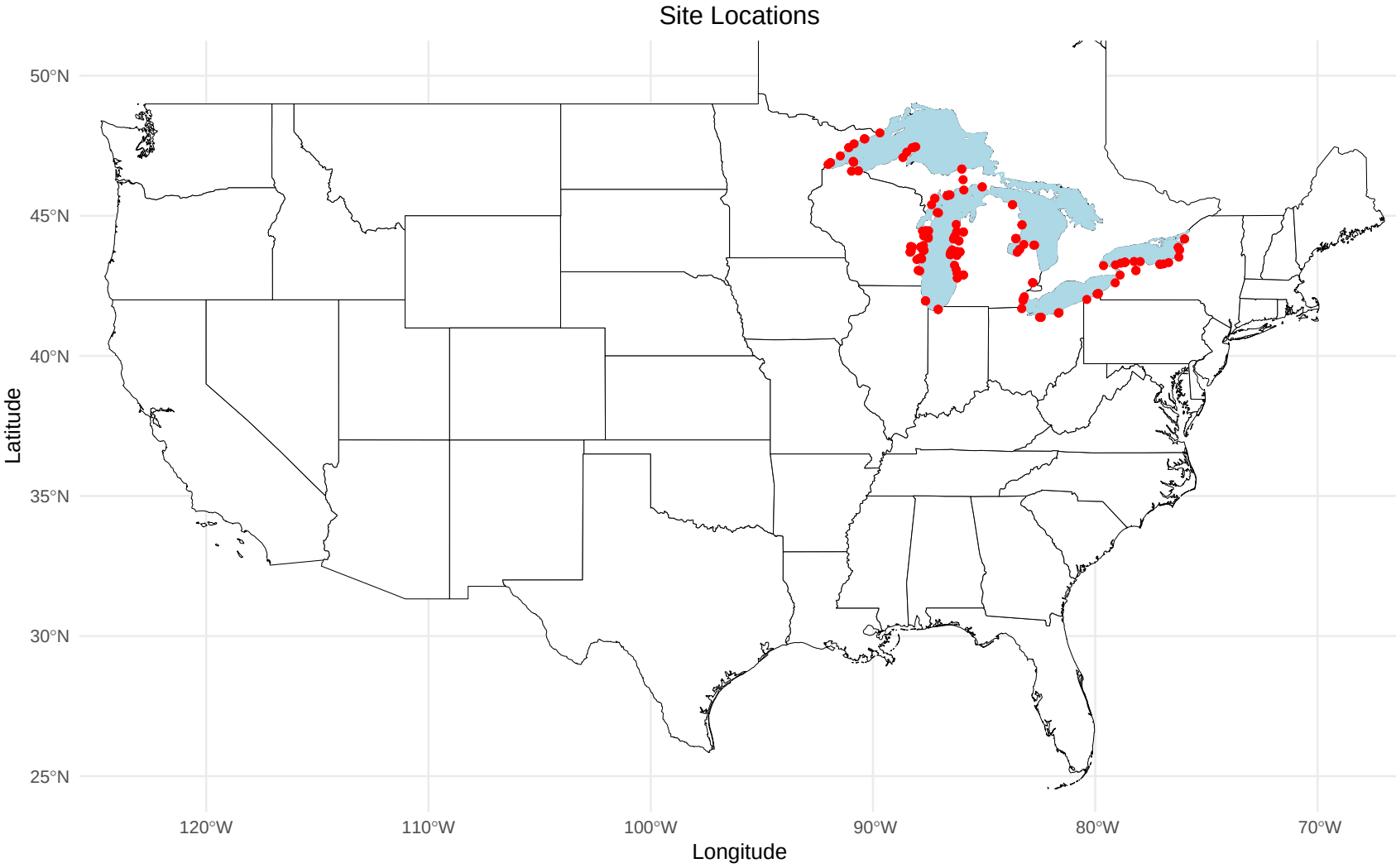
# Sites

The map below reveals that a number of sites are far away from the Great Lakes. Those sites are in southern Minnesota, northern Iowa, central Indiana, southern Pennsylvania, and Texas.



# Sites

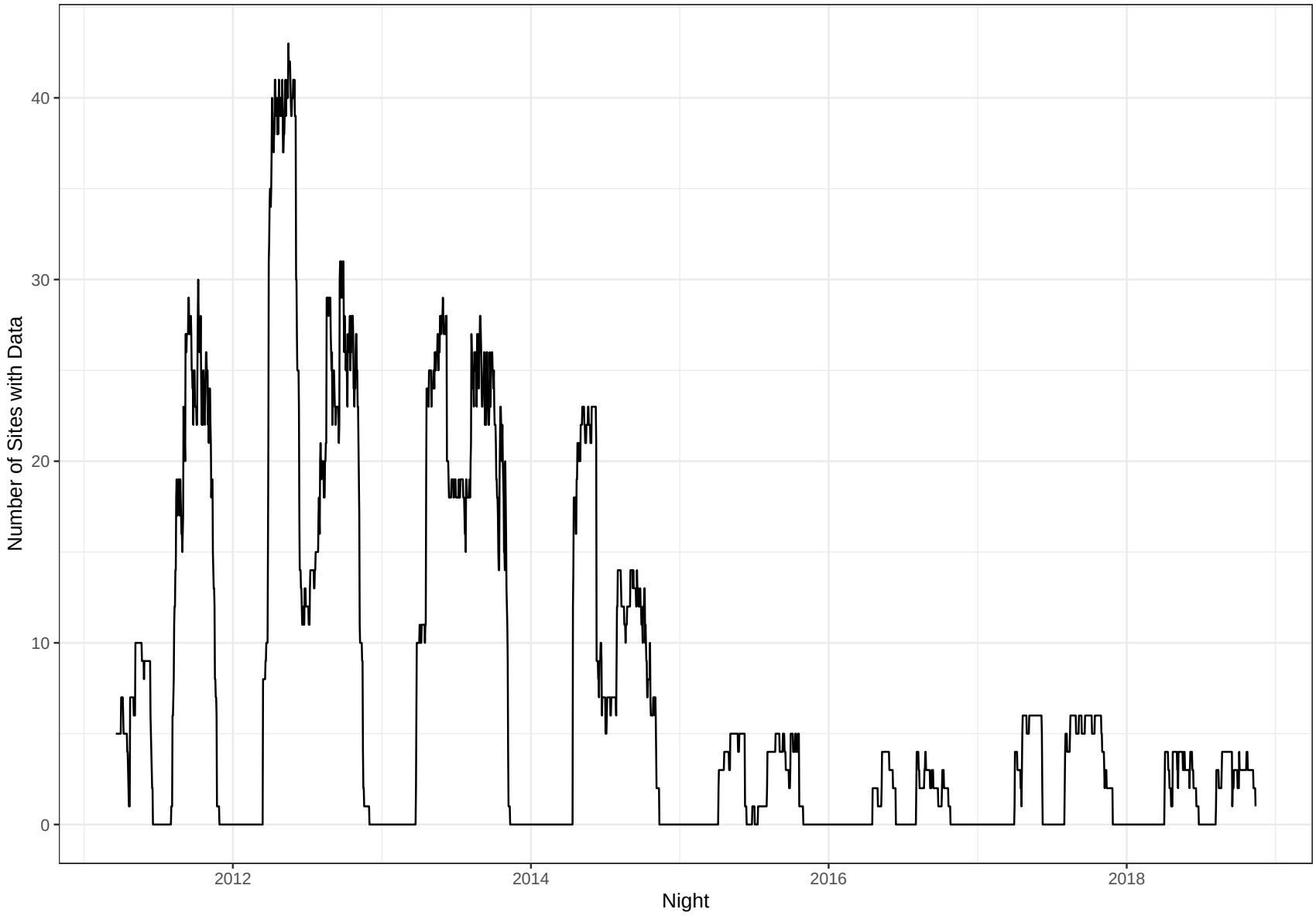
We eliminated data from sites far away from the Great Lakes. We are left with 18869 observations from 176 sites.



# Missingness

- There are no *explicit* missing values. However, there are quite a few *implicit* missing values, as the sensors are only active for certain nights of the year.
- Also, we're interested in bats generally, but the sensors are only equipped to detect certain species of bats.
- The line plot on the next slide shows how many sites have data by night.
  - There are stretches of time where no sites were producing data; they start toward the end of one year and end early the next year.
  - There are strange dips that occur in the middle of the year.
  - Many more sites produced data in the early 2010s than later in the decade.

# Missingness



# Data Aggregation

We aggregated the data by site. For each site, we computed the number of nights with data, and for each species, we computed the number of detections per night. A portion of the aggregated dataset is shown below.

```
# A tibble: 6 x 4
  AcousticSite num_nights EPTFUS LASBOR
  <chr>             <int>  <dbl>  <dbl>
1 2MILECREEKCA        156  3.77   2.41
2 50POINTCA           130  1.47   8.62
3 AUDO                 79  1.13  12.5
4 BAILEY              434  1.33  28.5
5 BETSIE              154  1.58  11.8
6 BLUE1               247  0.101  0.267
```

# Spatial Aspects of the Data

- The spatial domain of the data consists of the shores of the Great Lakes region.
- The dimension of the spatial domain is two, with the spatial locations of observations being indexed by latitude and longitude.
- It seems reasonable to model covariance as decaying with distance, as sites near each other should be in the ranges of roughly the same bat populations, so their detection figures should be roughly the same. In contrast, sites far away from each other should be in the ranges of different bat populations and should thus have dissimilar detection figures.

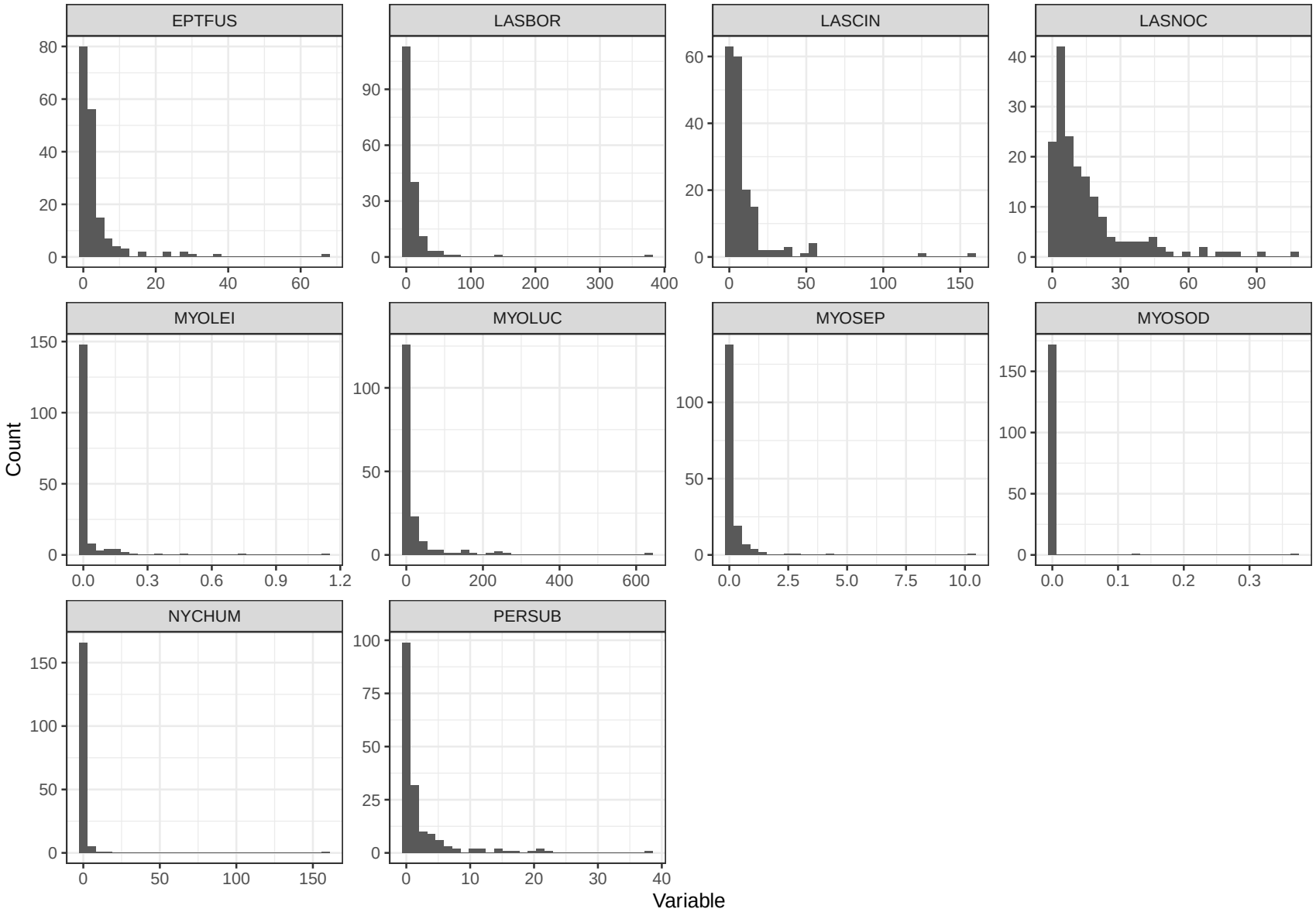
# Spatial Aspects of the Data

- We imagine that spatial dependence will play a role between the variables in the dataset because there should be a large spatial dependence in the counts of bats for nearby detectors. We expect that because certain species each have limited ranges that are spatially determined by habitat factors, then the individual sensors are likely to pick up only certain bats, and therefore nearby sensors should see similar counts of bats.

# Histograms

On the next slide are histograms that show the distributions of the detections/night variables for the various species. Each variable is right-skewed; for most sites, the ratios are close to zero, and for a few, the ratios are large.

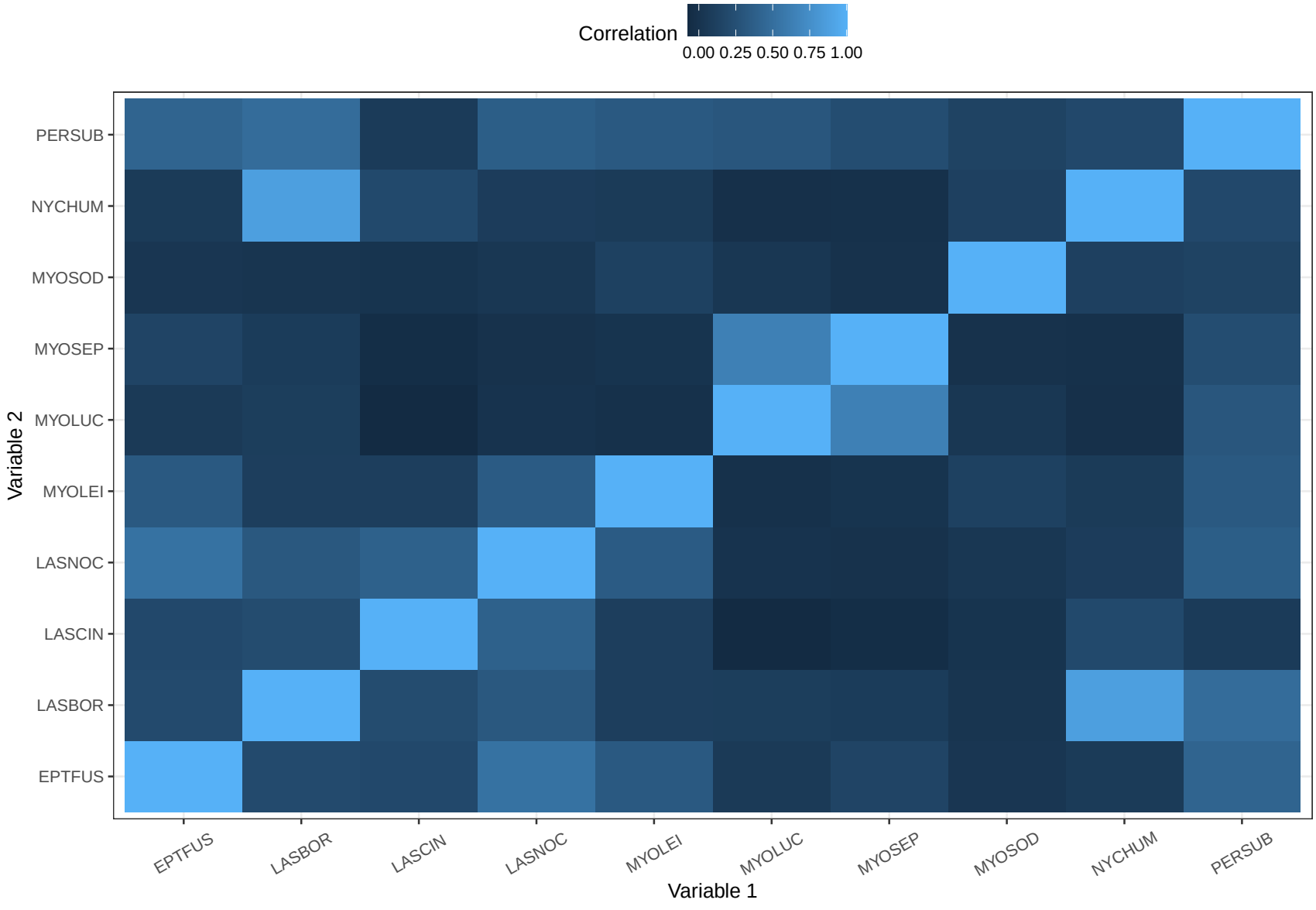
# Histograms


EPTFUS
LASBOR
LASCIN
LASNOC
MYOLEI
MYOLUC
MYOSEP
MYOSOD
NYCHUM
PERSUB
Count
Variable

# Correlations

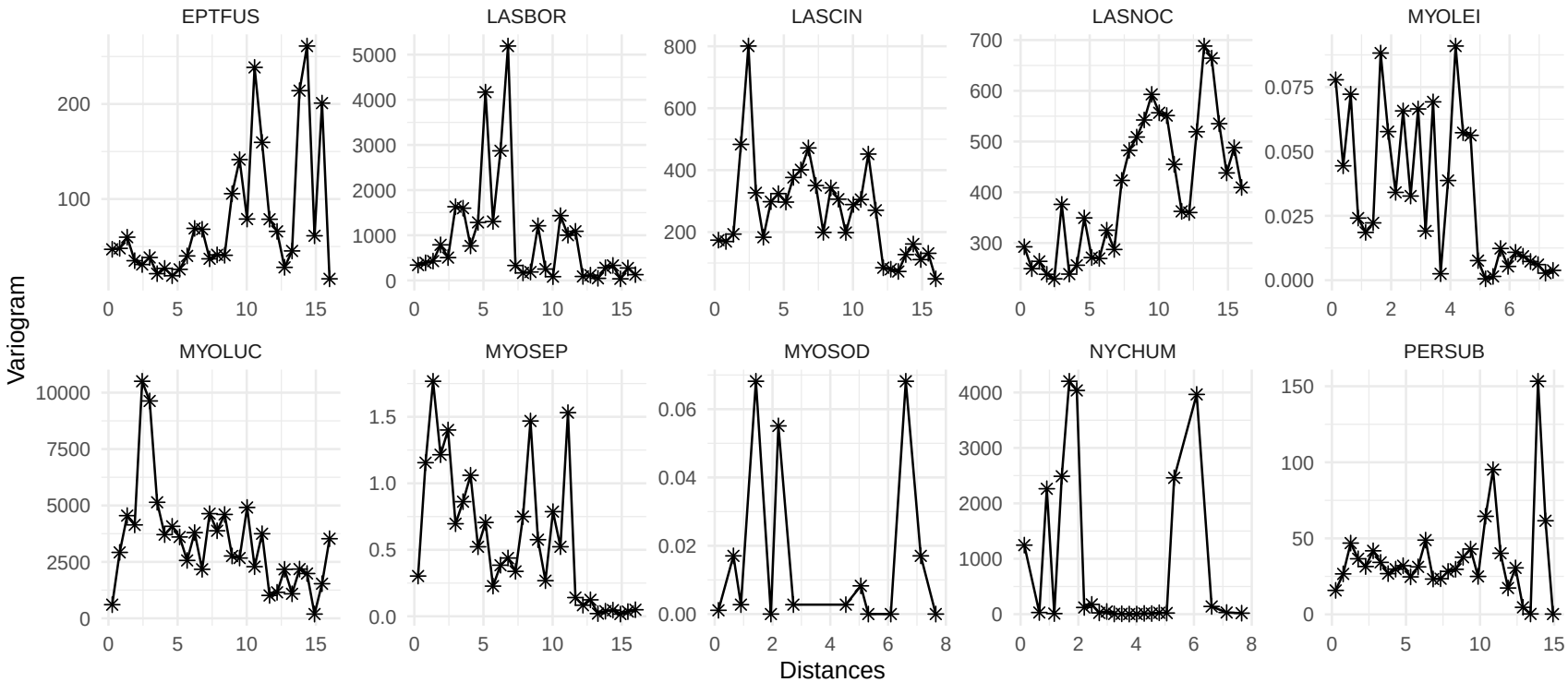
On the next slide is a correlogram that shows the correlations between the detections/night variables for the various species. The only strong correlation between different species is between LASBOR (Eastern Red Bat(Lasiurus borealis) ) and NYCHUM (Evening Bat (Nycticeius humeralis)); it equals 0.87.

# Correlations



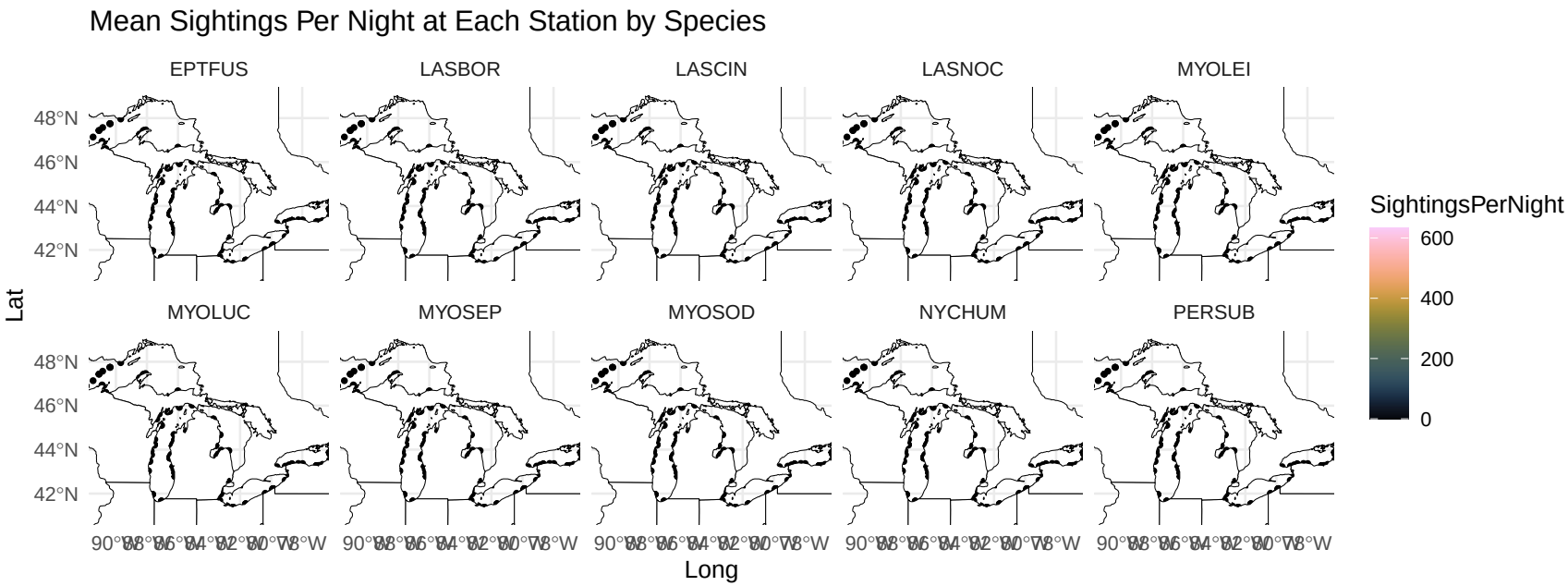
# Empirical Covariograms

The covariograms computed from the aggregated data suggest wildly differing spatial dependences across bat species when ignoring temporal variation, including both positive and negative spatial dependence.



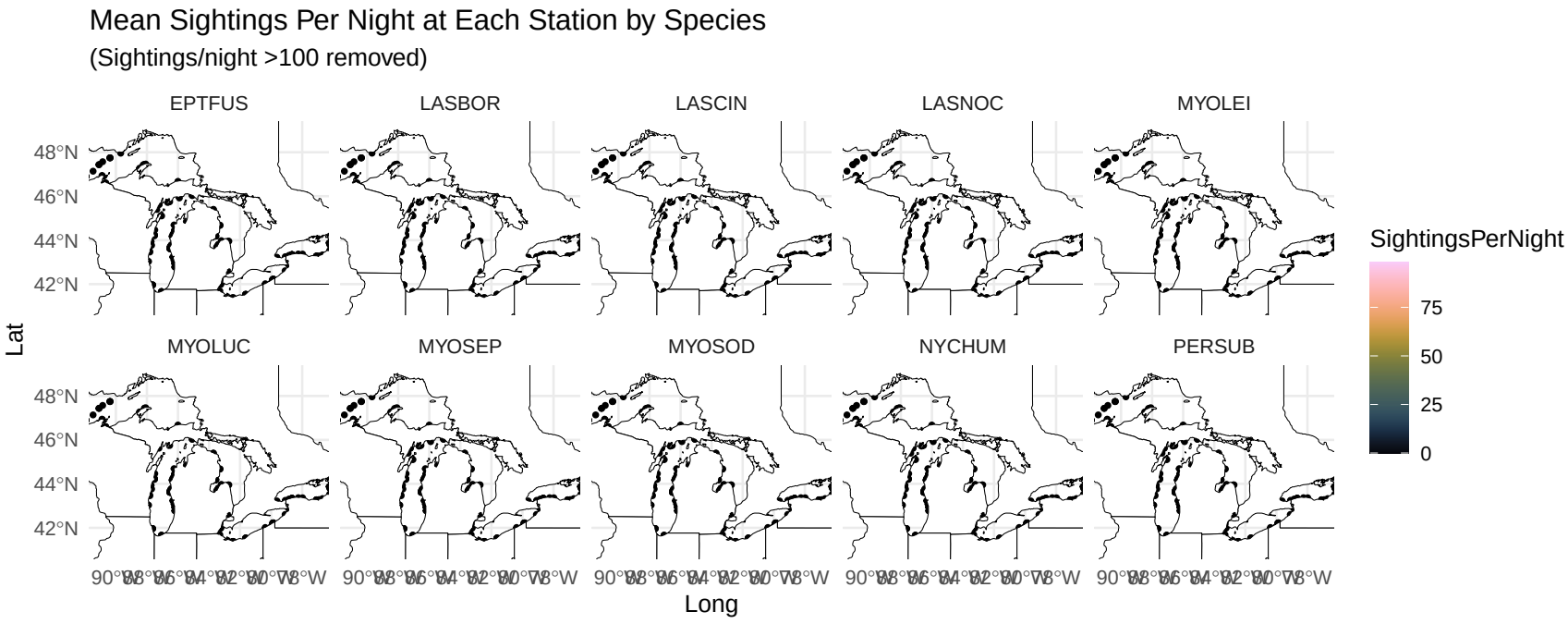
# Spatial Visualizations

It is difficult to discern any patterns in the spatial variation of the mean number of sightings per night by species. The lack of a clear spatial pattern continues even when we filter out seeming outliers in the number of mean number of mean sightings per night.



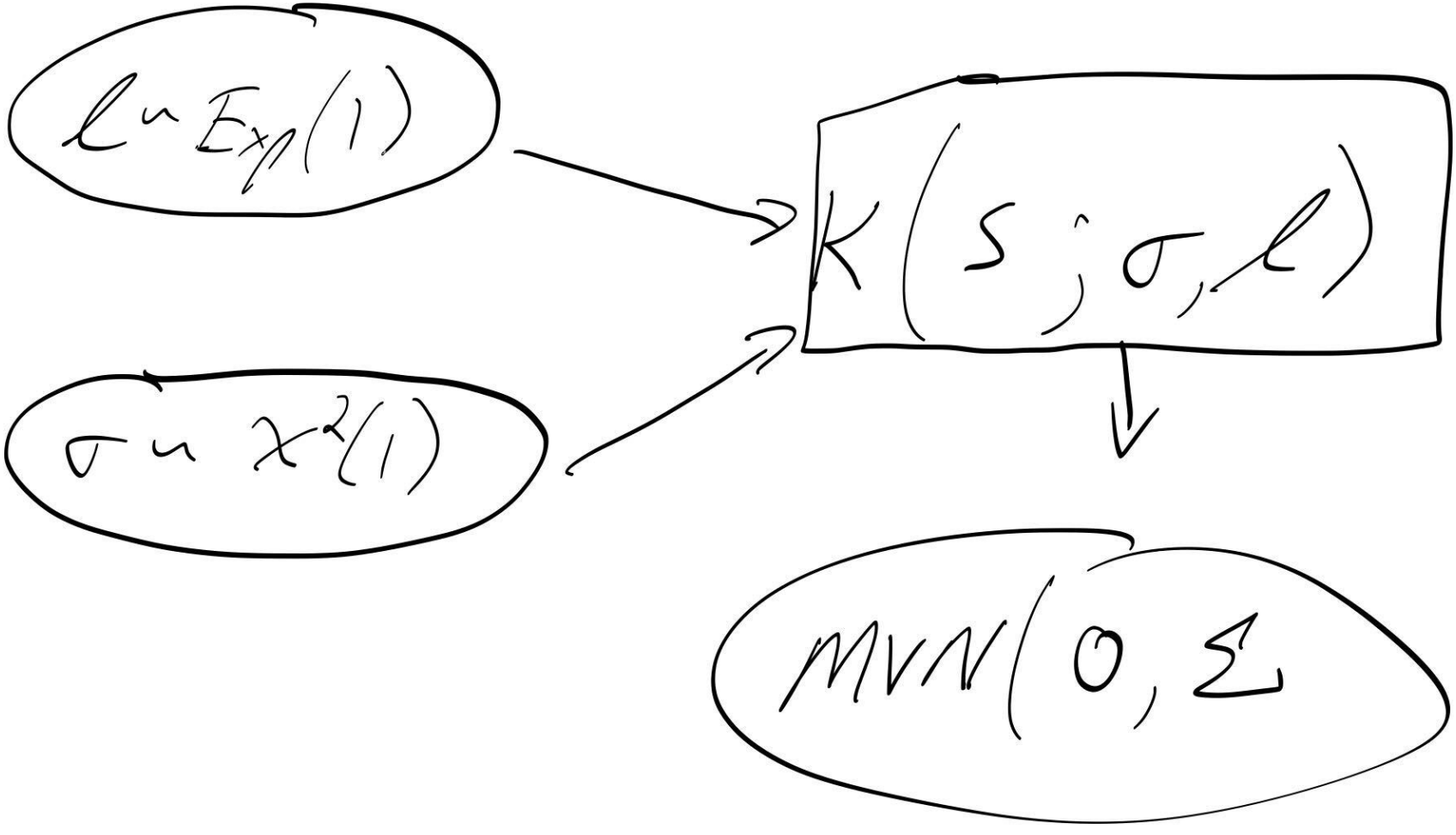
# Spatial Visualizations

The lack of a clear spatial pattern continues even when we filter out seeming outliers in the number of mean number of mean sightings per night.



# A Potential DAG

Below is a potential DAG for our problem.



# A Potential DAG

In the DAG, MVN$(0, \Sigma)$ represents a multivariate Gaussian random vector consisting of detections per night for various sites. We assume a mean zero Gaussian process, where the covariance is modeled using a kernel function (e.g., a Matern kernel) and all hyperparameters are independent. The covariance matrix $\Sigma$ is based on a kernel represented by $\kappa$; $\ell$ and $\sigma$ represent the length and variance scales, respectively.

For our DAG, we only have observations of spatial variables and currently have no additional covariates.

# Anticipated Results, Potential Problems, and Additional Data

- We anticipate that bat populations will drop off as one moves inland, as bats require bodies of water such as rivers in their habitats. Likewise, we anticipate that spatially adjacent sites will have similar populations of bats.
- Our model is limited in that most data is collected along the Great Lakes shores, so extrapolation may be difficult.
- Additional inland bat population observations could be helpful, as well as covariates related to distance from the nearest water way and wind farm.

# Preparation of the Data for Model-Building

There's one record for each pair of `AcousticSite` and `Year`.

```
# A tibble: 0 x 3
# i 3 variables: AcousticSite <chr>, Year <dbl>, n <int>
```

Most sites only have data for a single year.

```
# A tibble: 3 x 2
  num_years num_sites
      <int>     <int>
1         1       197
2         2        59
3         3        17
```

Below is another summary that shows that most sites only have data for one year.

```
# A tibble: 16 x 2
   years           num_sites
   <chr>               <int>
 1 2012                   52
```

# Linear Models

## Non-Spatial Linear Model

To begin with, we fit a simple linear model of counts across all species. Under the model, at any location $s$,

$$\log(1 + Y(s)) \sim N(\beta_0 + X(s)^\top \beta, \sigma^2), \tag{1}$$

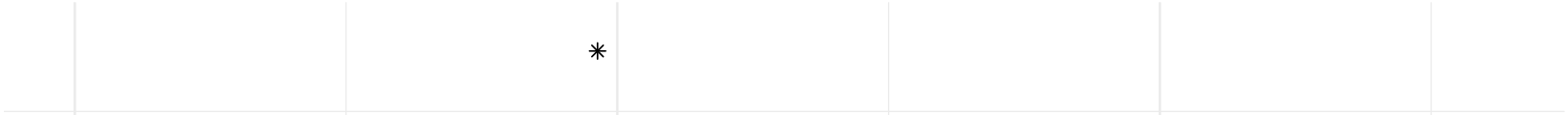
where $X(s)$ is the vector whose components are the distance from $s$ to the shore, the number of operational nights, and the number of detectable species.

The residual variance from the nonspatial linear model is ~2.4, which corresponds to about ~1.3 standard deviations on the logged counts scale.

```
[1] "Residual variance: 2.40022224103286"
```

The residuals show strong spatial correlation (with longitude and latitude as spatial coordinates), as demonstrated in the empirical semivariogram.

```
variog: co-locatted data found, adding one bin at the origi
```



# References

"USFWS Great Lakes and Upper Midwest Acoustic Bat Dataset." 2019. U.S. Fish; Wildlife Service. https://catalog.data.gov/dataset/usfws-great-lakes-and-upper-midwest-acoustic-bat-dataset.