Modeling Beetle Infestation Throughout a Forest as a Spatial Stochastic Process

Mentor: Dr. Keisha Cook
Fiona Han, Tasnim Iqbal, Jeffrey Utley • 08.04.2023

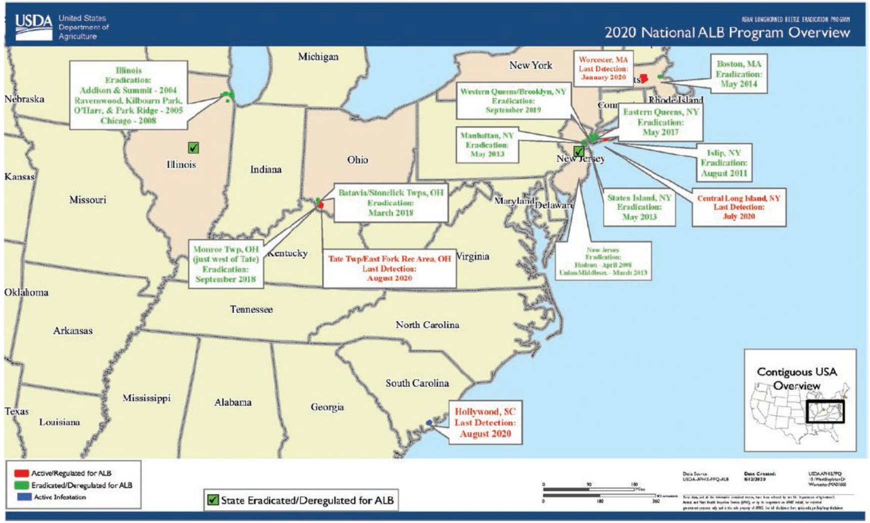
Overview

Problem Statement:

How do we predict the spread of infection caused by Asian Long-horned Beetles in a forest?

Motivation

- Asian Long-horned Beetle (ALB) is...
  - Native to asia
  - Discovered in Massachusetts in 1996
- Causes a disease to trees that **no known cure exists**



Data Analysis per Year

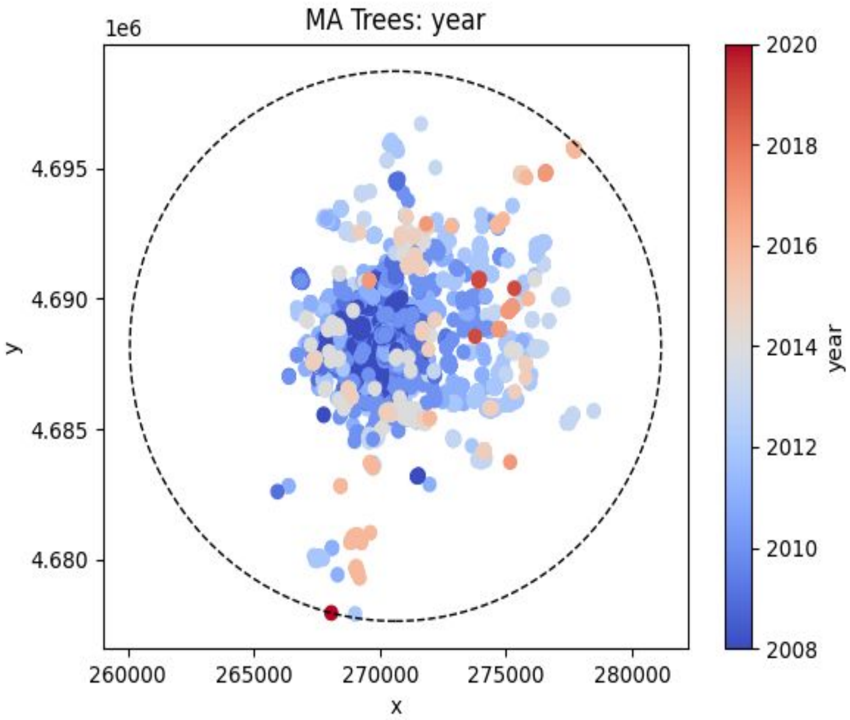
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

- Our data contains 2-D points of locations of infected trees in Massachusetts.
- The data set includes 2 parameters:
    - Year
    - Species



Overview

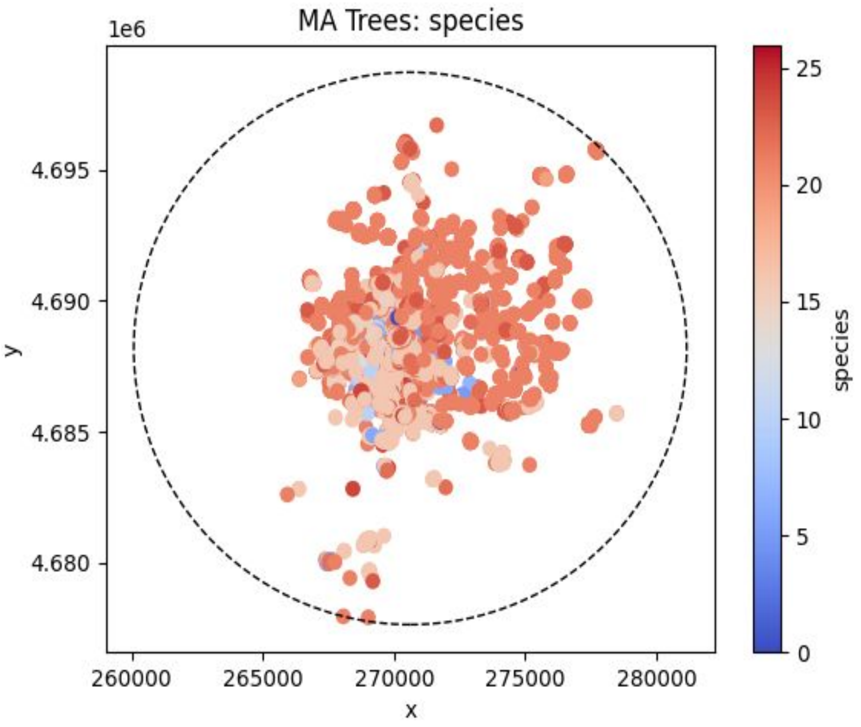
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Data Analysis per Species

- Our data contains 2-D points of locations of infected trees in Massachusetts.
- The data set includes 2 parameters:
  - Year
  - **Species**



Data Analysis per Infestation Level

Overview

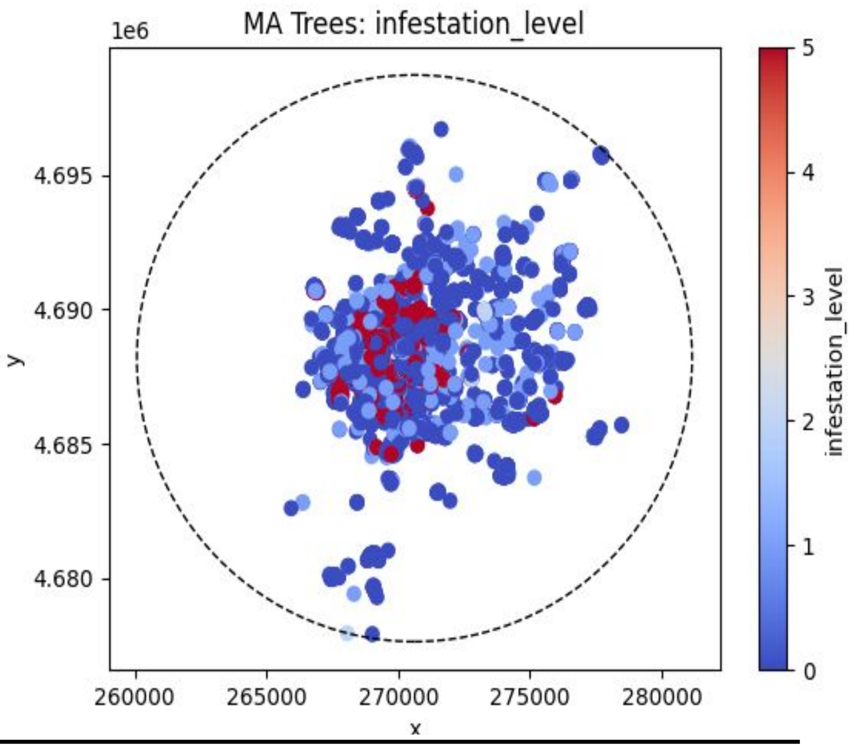
Motivation

Progress

- **Data Analysis**
- Model Construction

Next Steps

- Our data contains 2-D points of locations of infected trees in Massachusetts.
- The data set includes 3 parameters:
  - Year
  - Species
  - Infestation Level



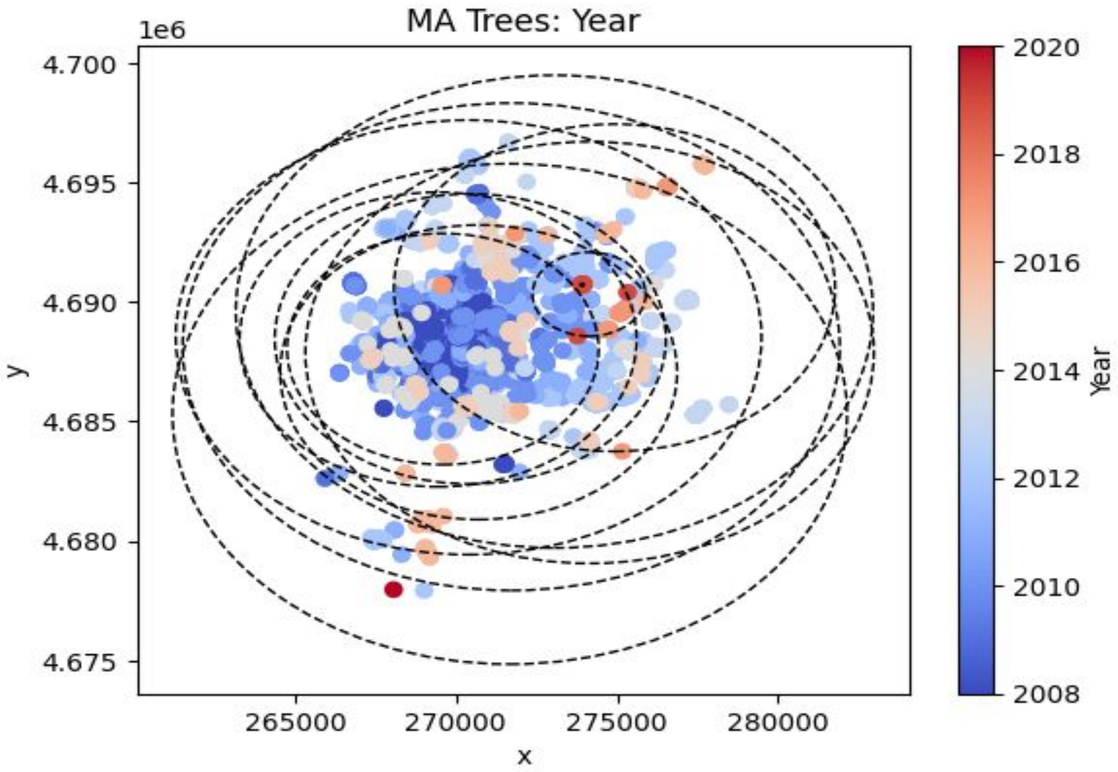
Data Analysis per Year - Grouped

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



Methods Employed

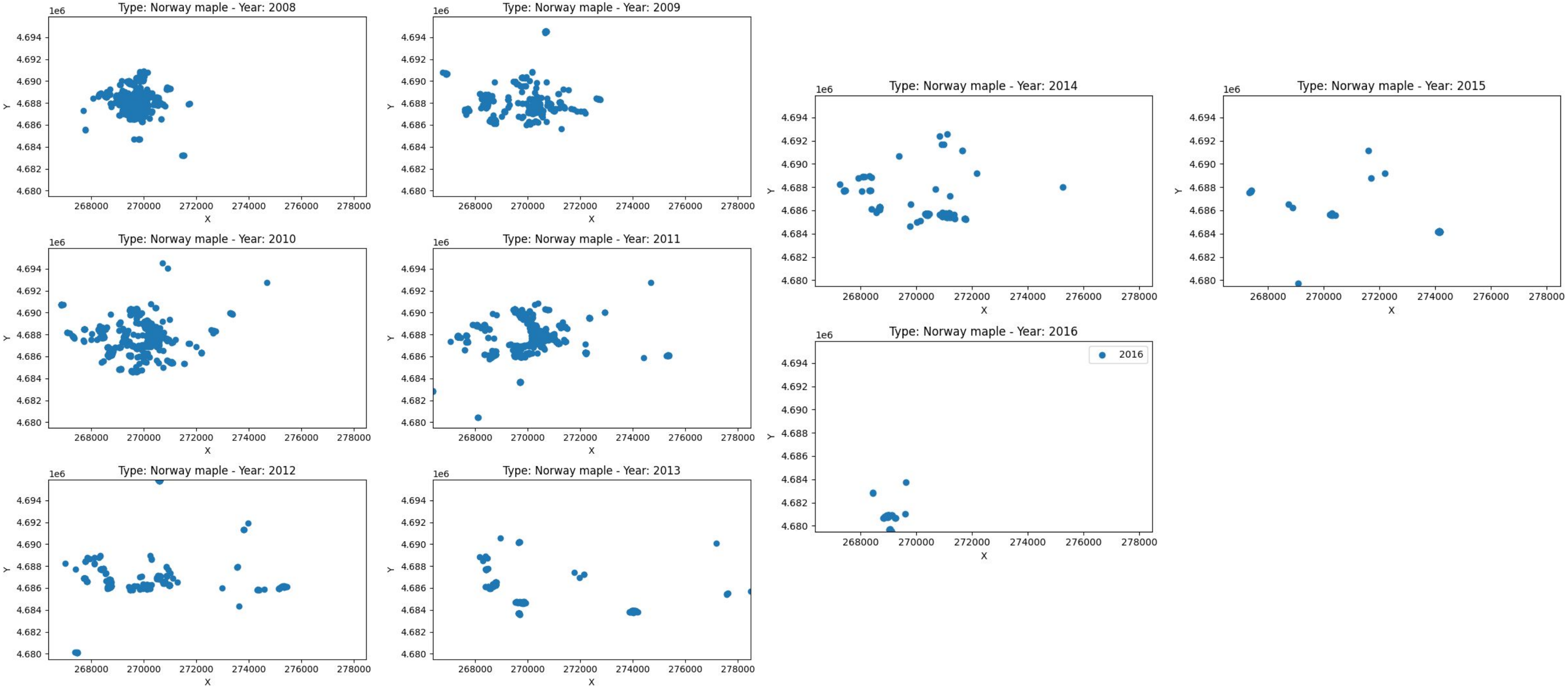
Data analysis methods:

- KDE (Kernel Density Estimation)

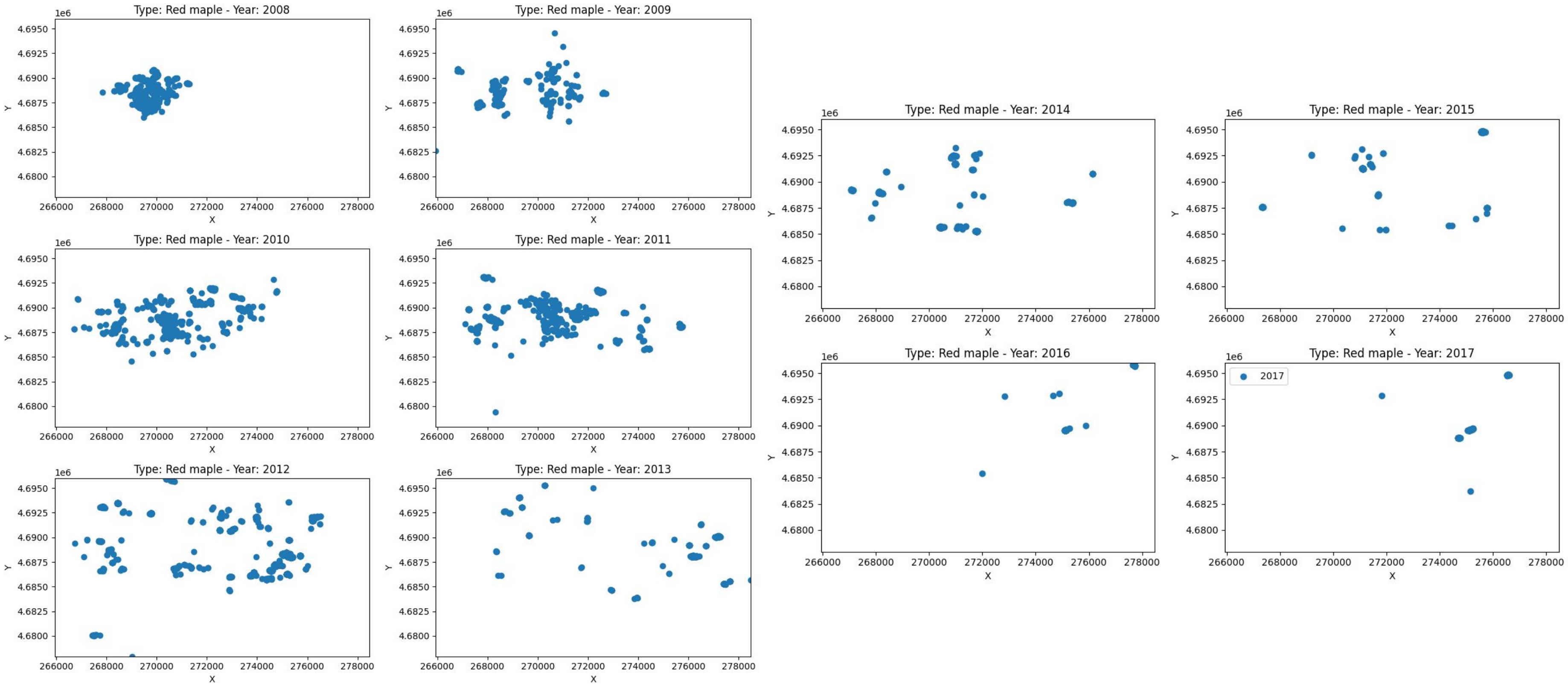
Modeling Methods:

- Bayesian Inference
- Maximum Likelihood Approach (MLE)

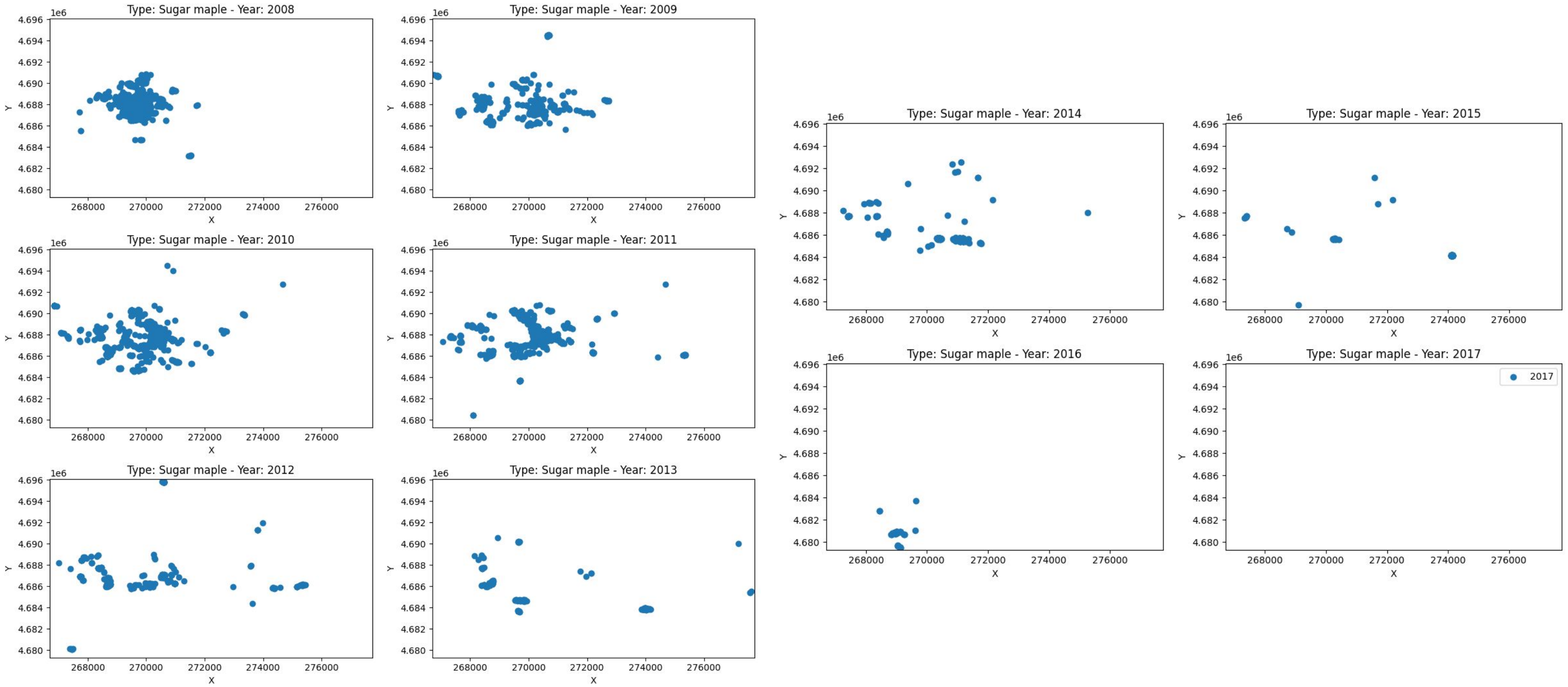
Norway Maple Species by Year



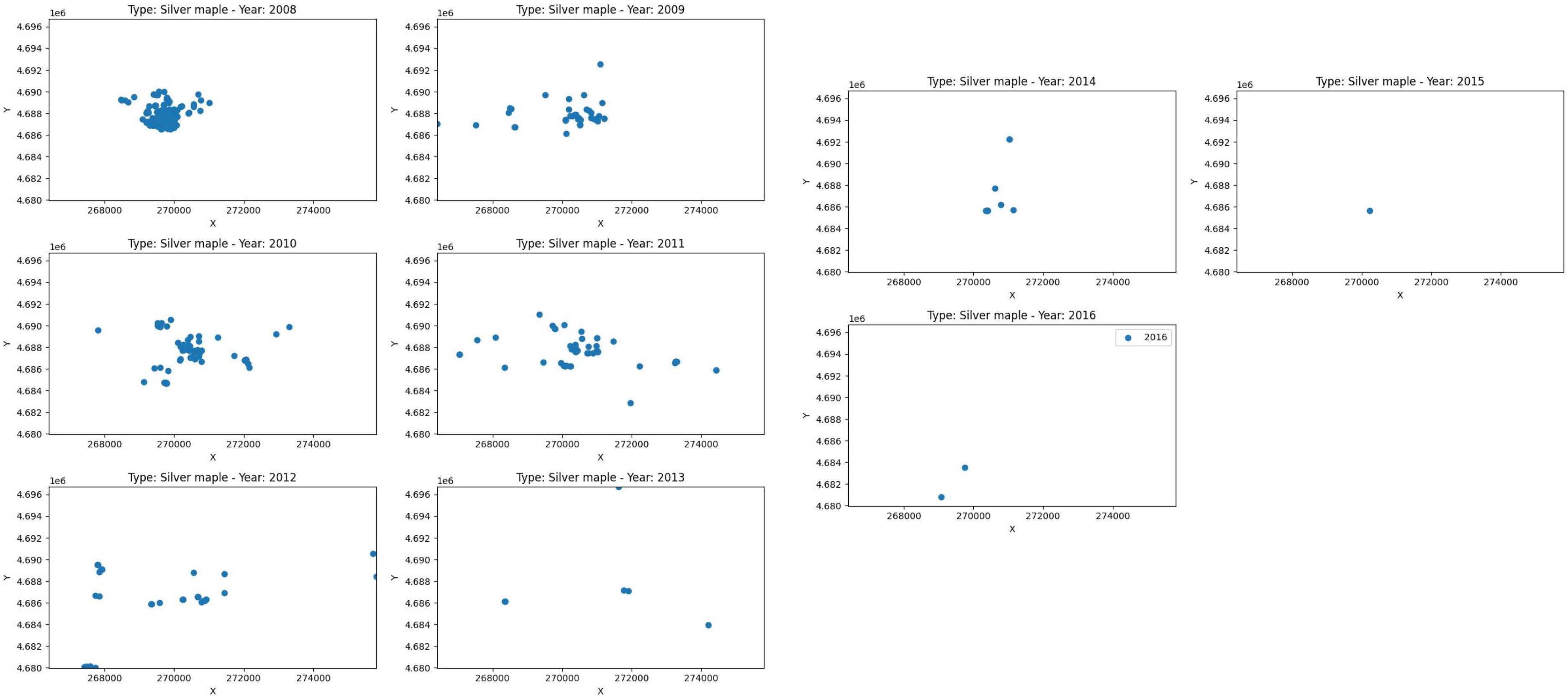
Red Maple Species by Year



Sugar Maple Species by Year



Silver Maple Species by Year



Progress - Data Analysis

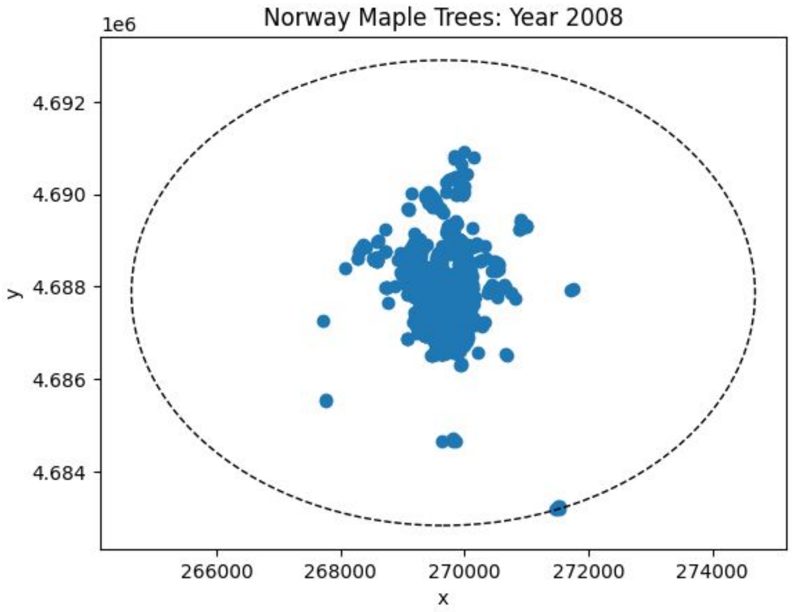
Circle data (Norway maple : year 2008 & 2009)

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



Progress - Data Analysis

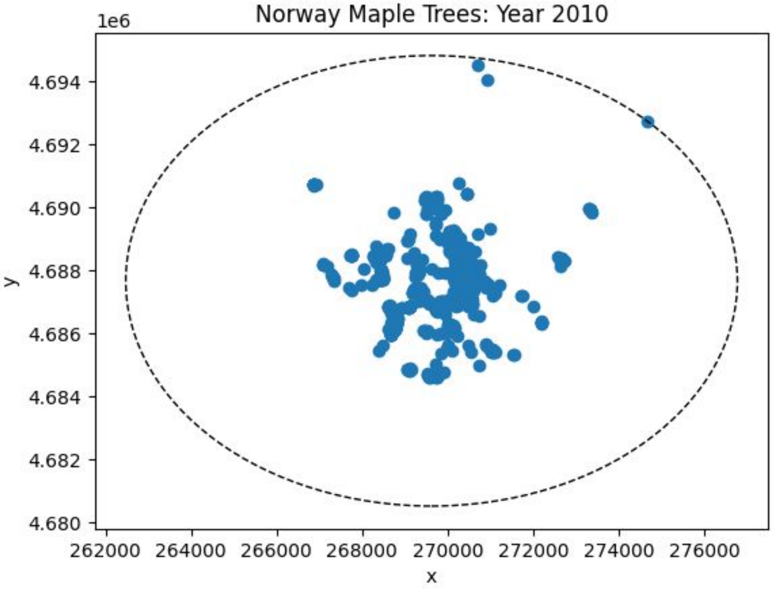
Overview

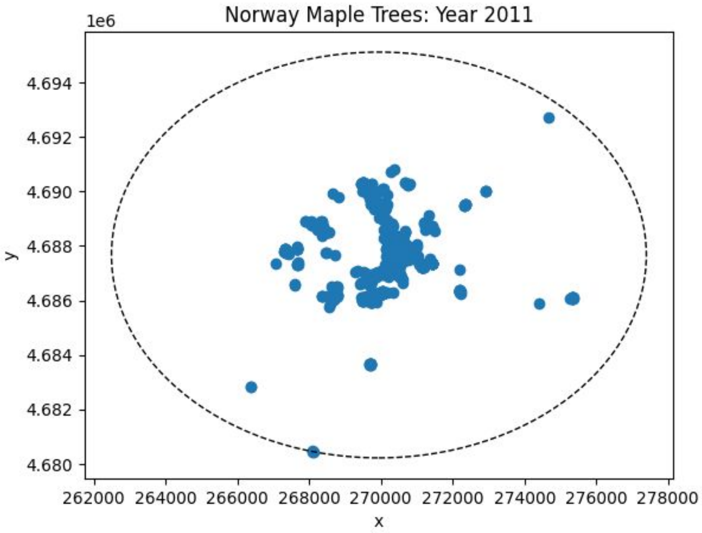
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Circle data (Norway maple: year 2010 & 2011)





Progress - Data Analysis

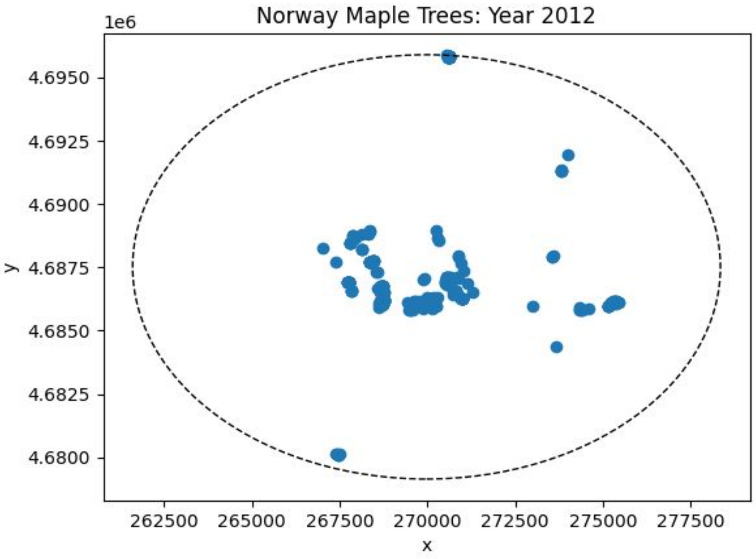
Overview

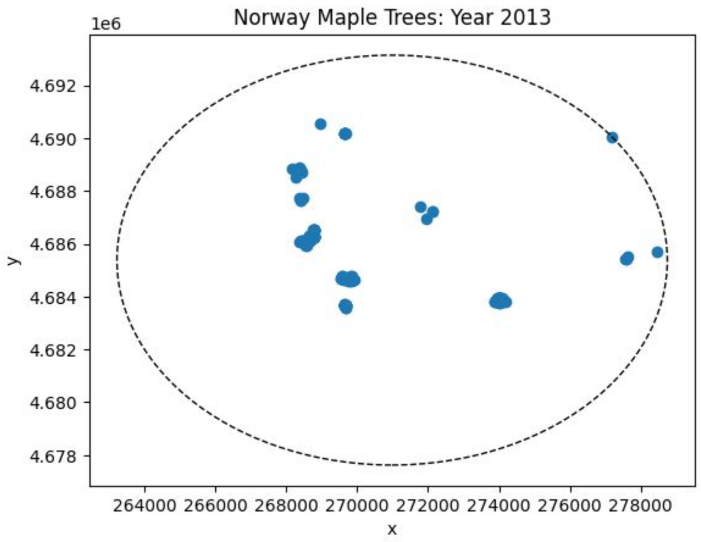
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Circle data(Norway maple: year 2012 & 2013)





Progress - Data Analysis

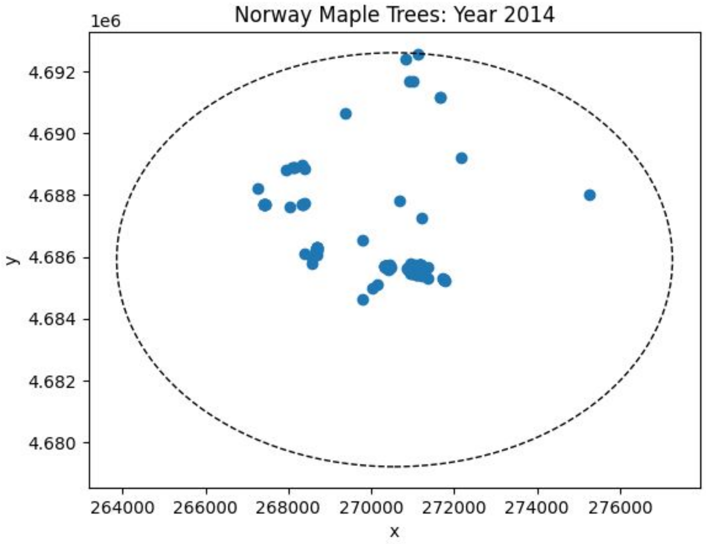
Circle data (Norway maple: year 2014 & 2015)

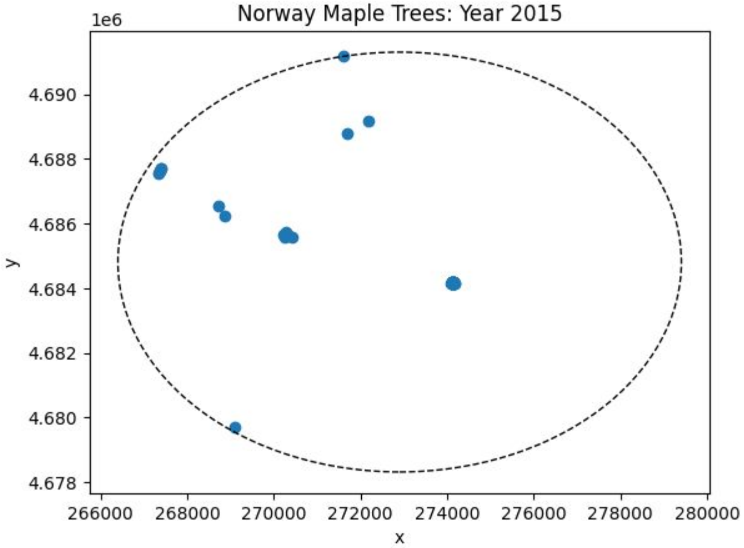
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps





Progress - Data Analysis

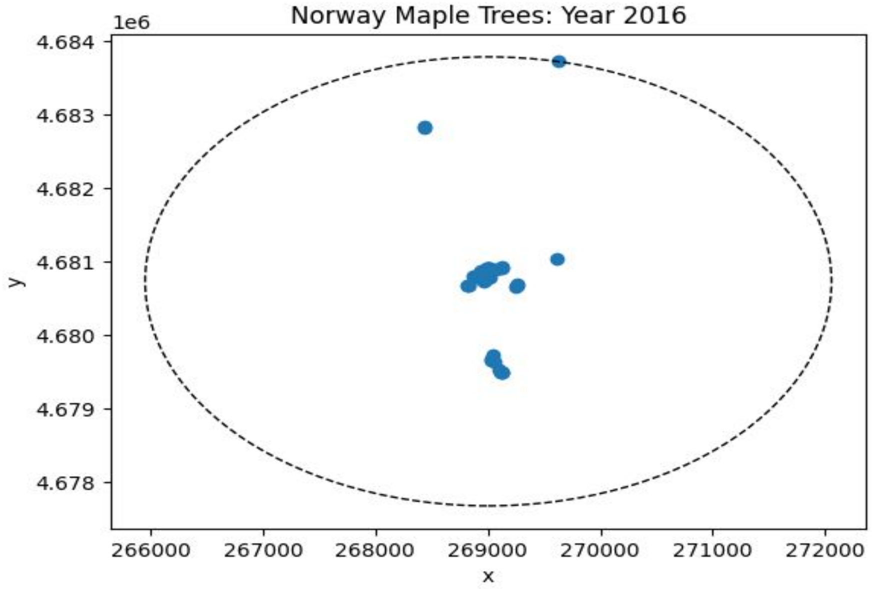
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

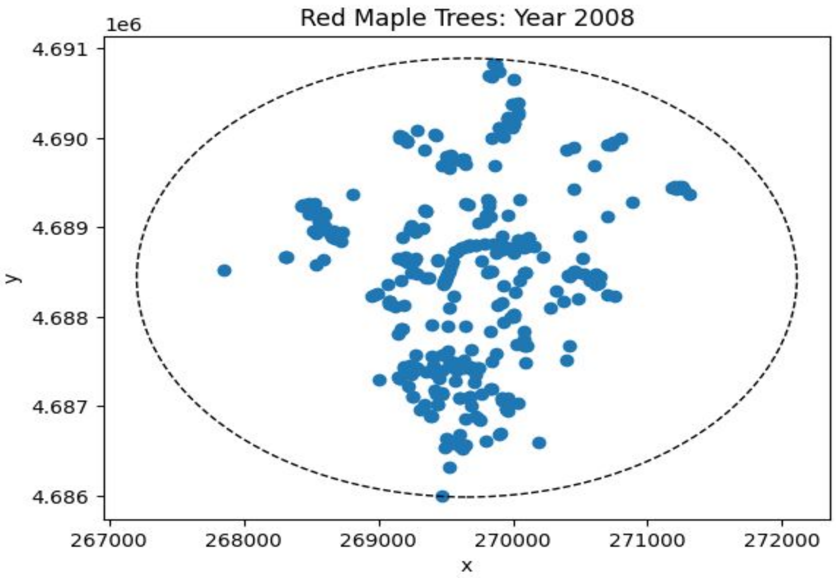
Next Steps

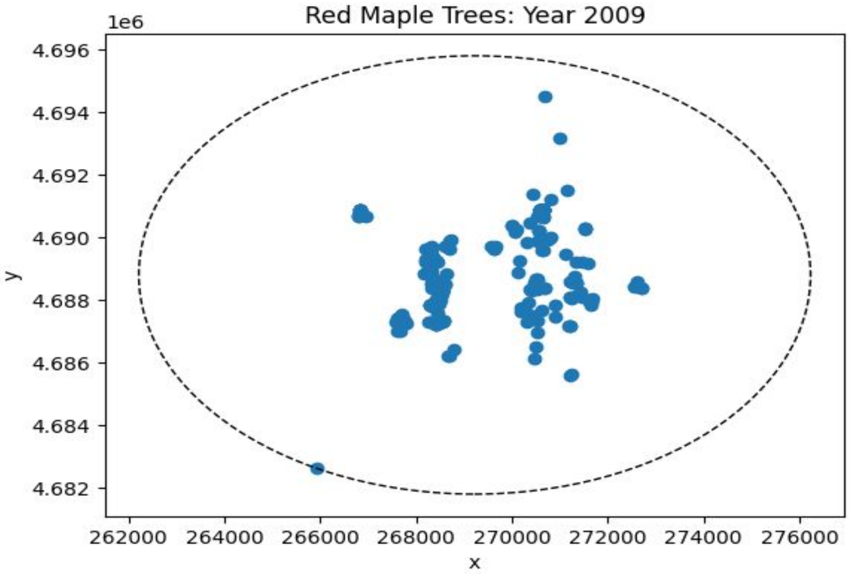
Circle data (Norway maple: year 2016)



Red Maple Species by Year

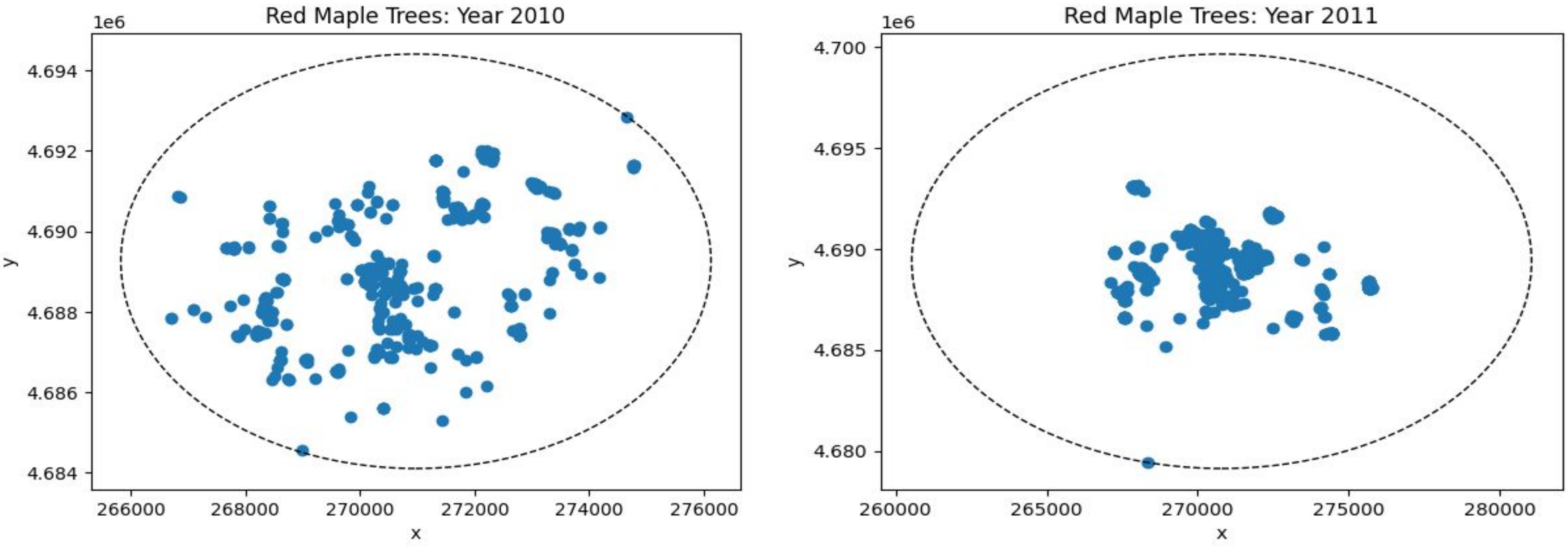
Spatial Poisson point process of infected tree species by year





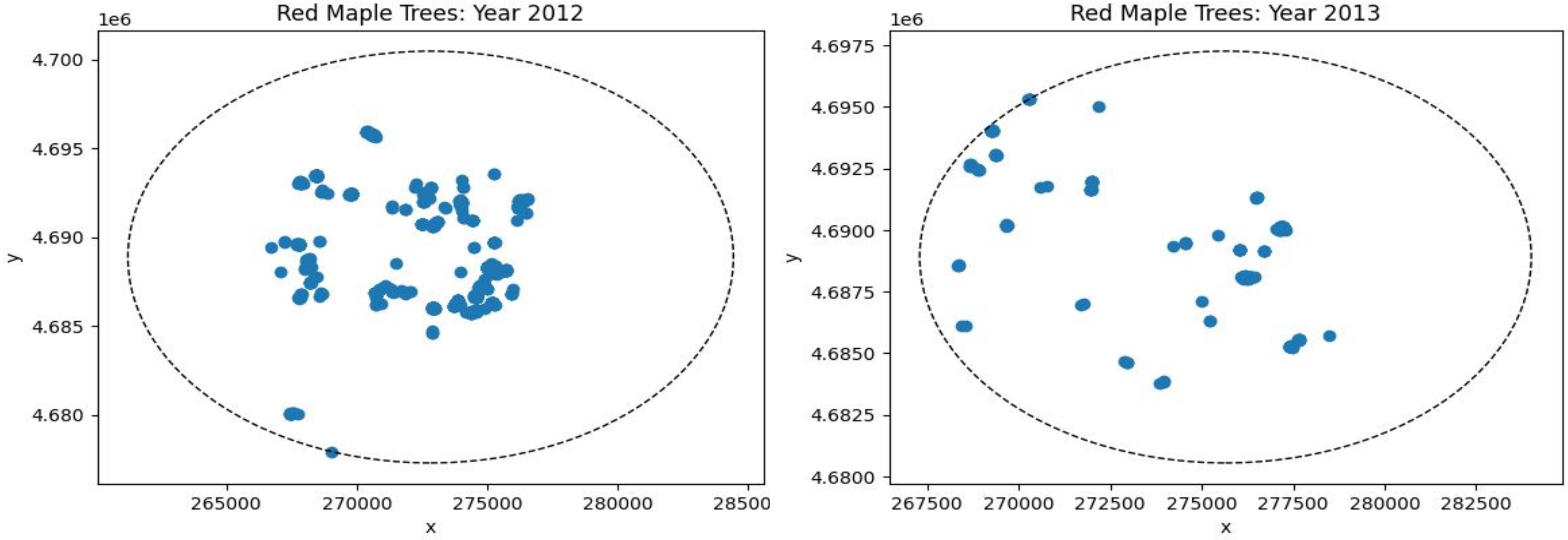
Red Maple Species by Year

Spatial Poisson point process of infected tree species by year



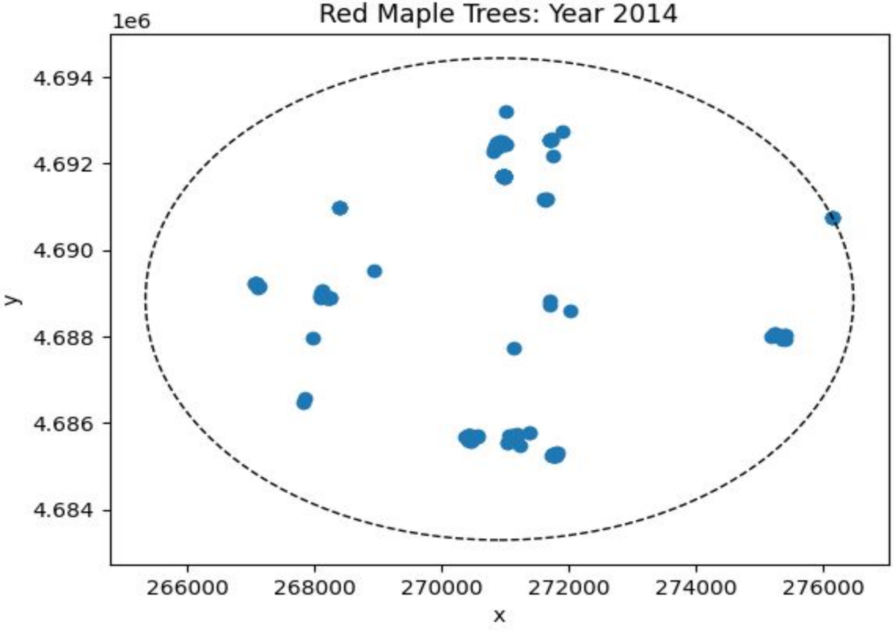
Red Maple Species by Year

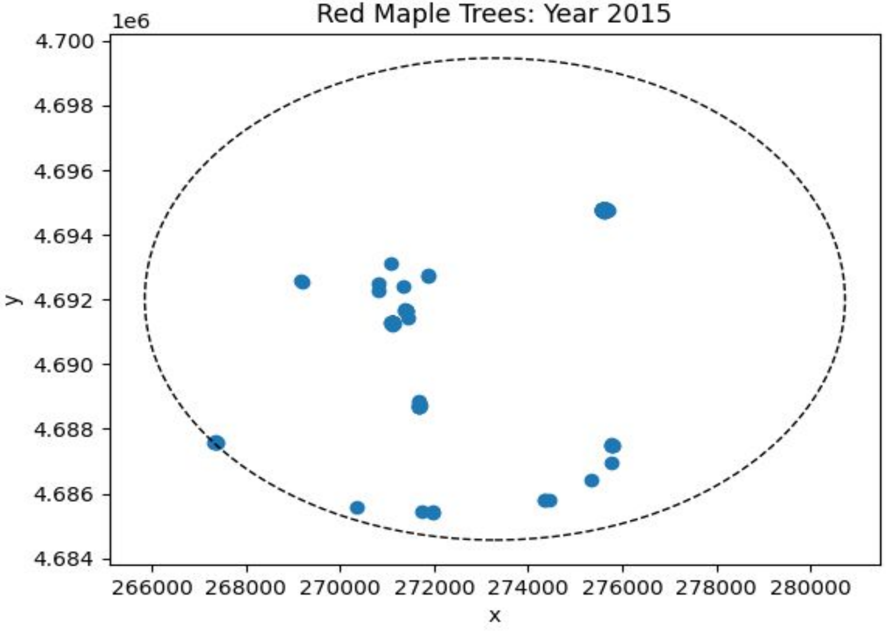
Spatial Poisson point process of infected tree species by year



Red Maple Species by Year

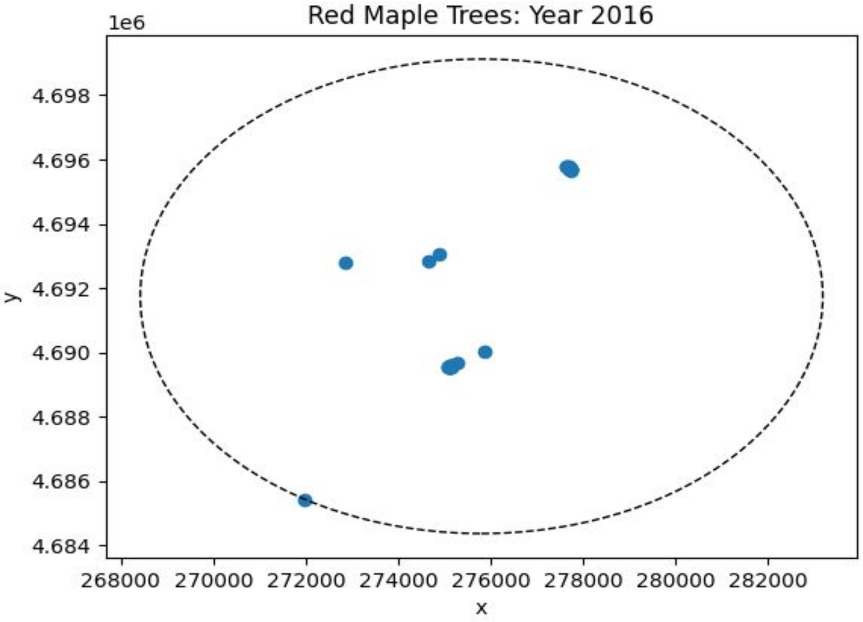
Spatial Poisson point process of infected tree species by year

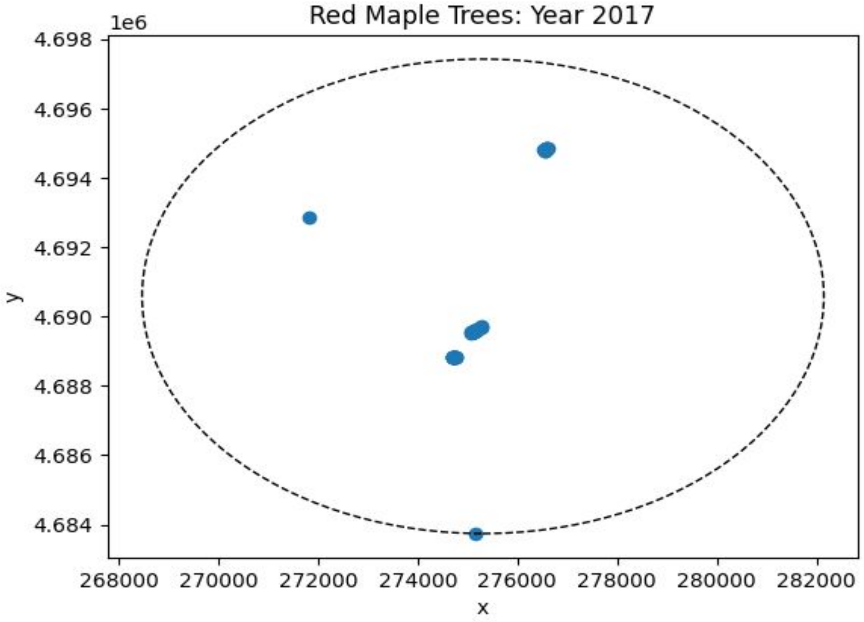




Red Maple Species by Year

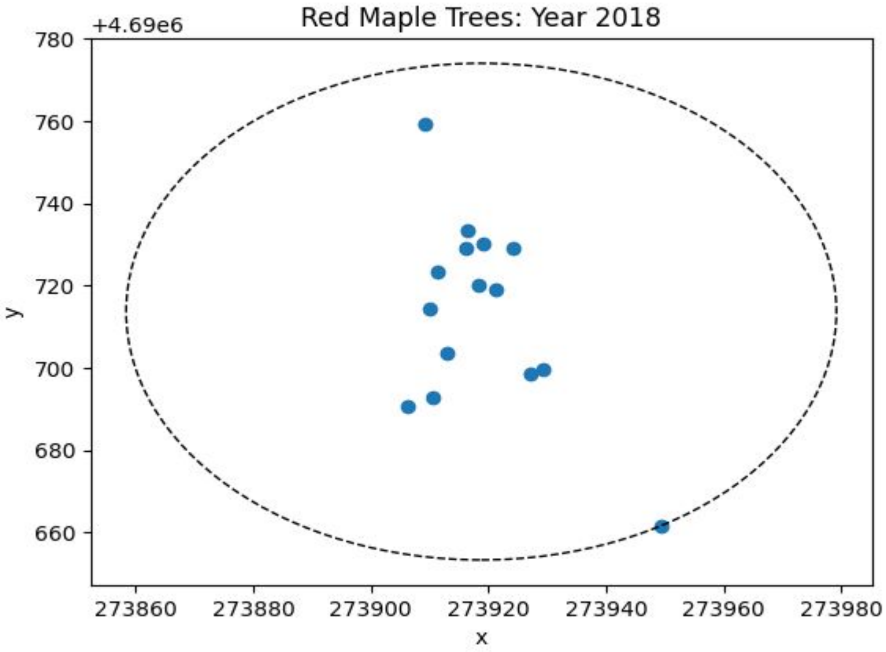
Spatial Poisson point process of infected tree species by year

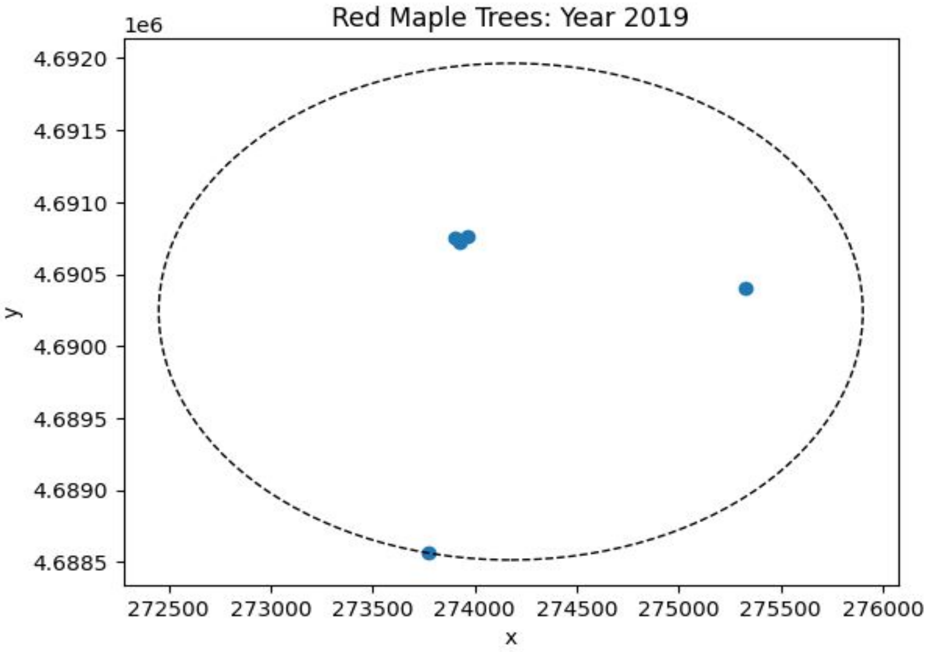




Red Maple Species by Year

Spatial Poisson point process of infected tree species by year





Progress - Data Analysis

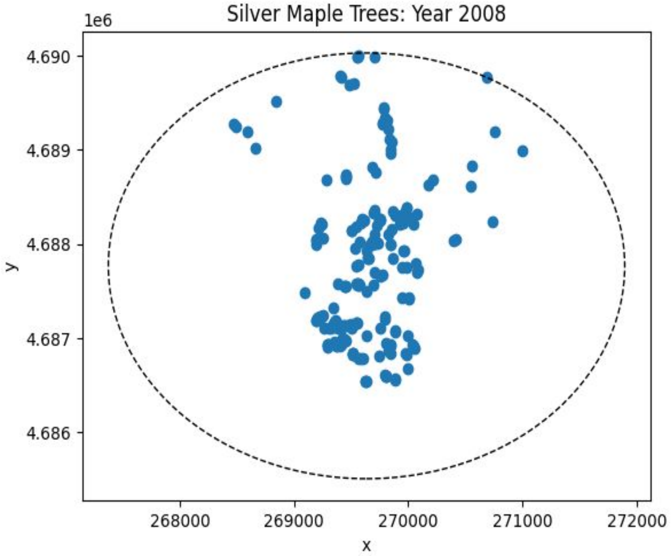
Overview

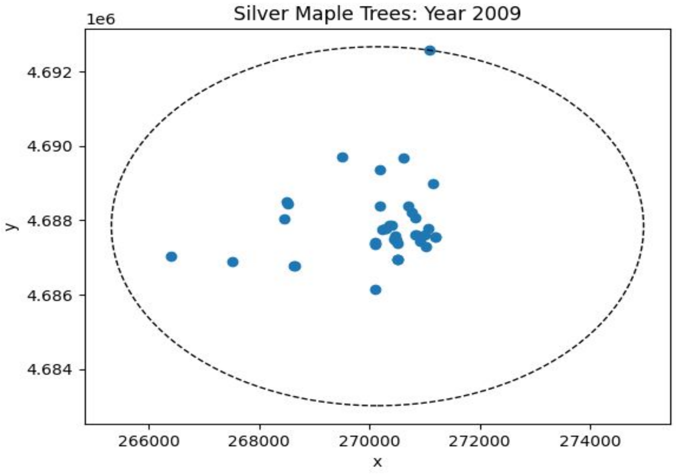
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

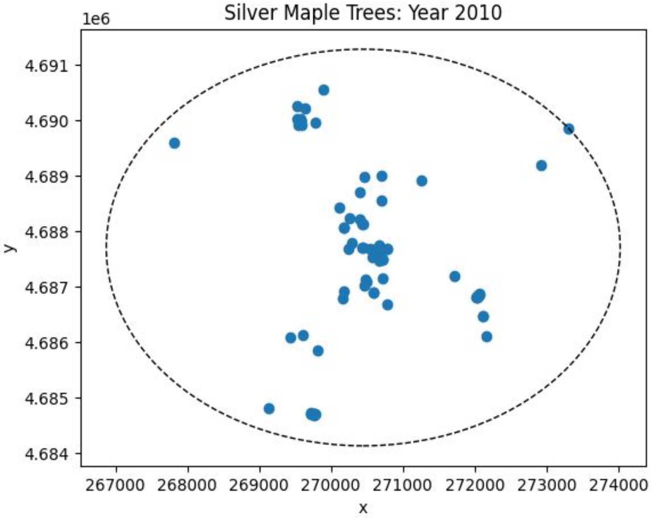
Circle data (Silver maple: year 2008 & 2009)

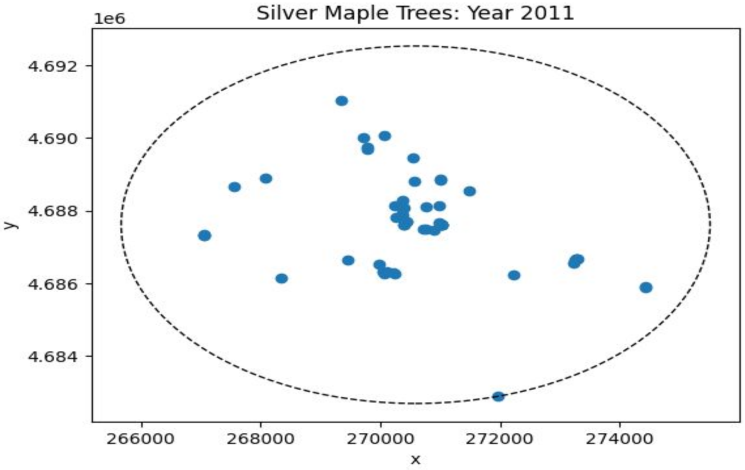




Progress - Data Analysis

Circle data (Silver maple: year 2010 & 2011)





Progress - Data Analysis

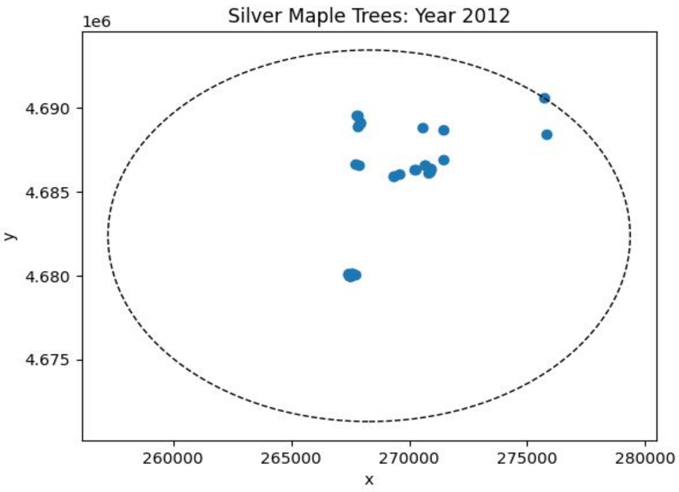
Circle data (Silver maple: year 2012 & 2013)

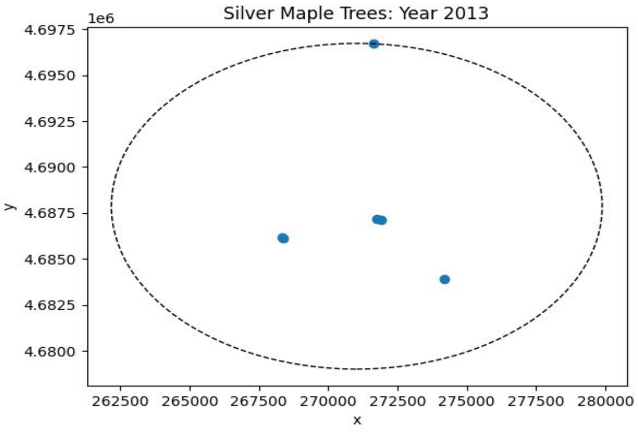
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps





Progress - Data Analysis

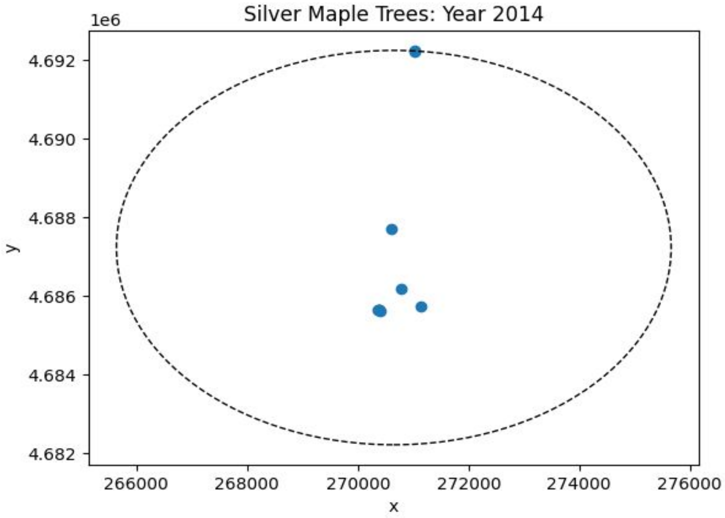
Circle data (Silver maple: year 2014 & 2016)

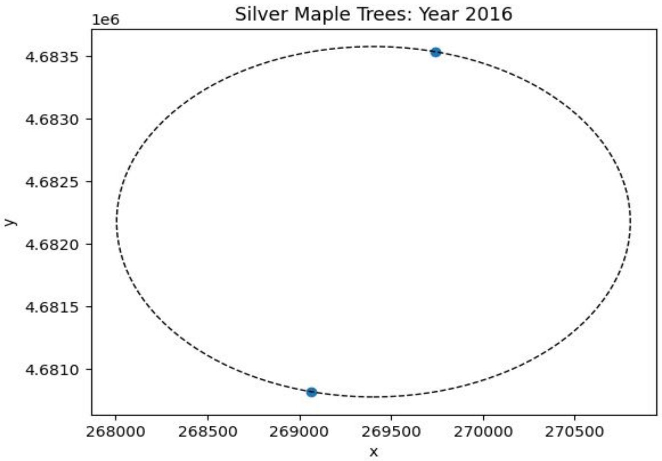
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps





Overview

Motivation

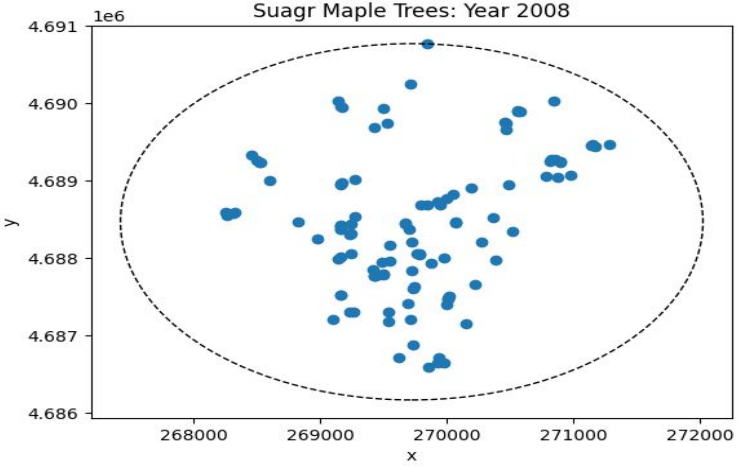
Progress

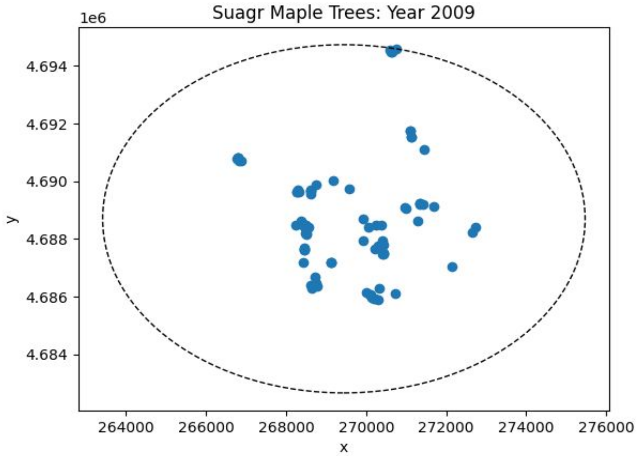
- **Data Analysis**
- Model Construction

Next Steps

Progress - Data Analysis

Circle data (Sugar maple: year 2008 & 2009)





Progress - Data Analysis

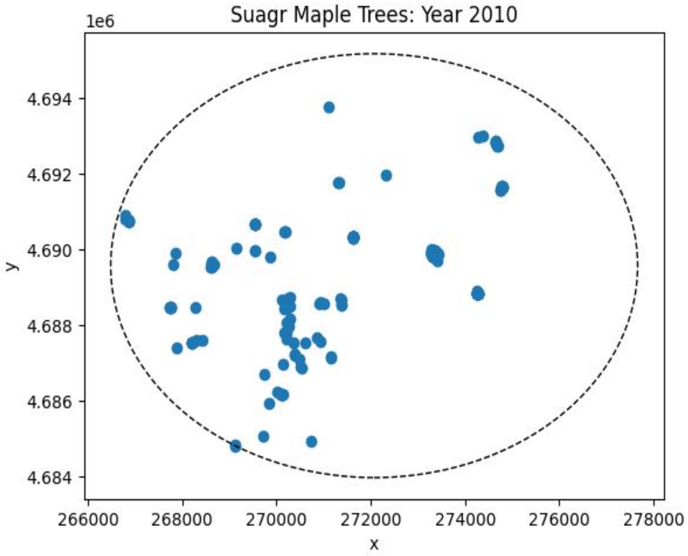
Overview

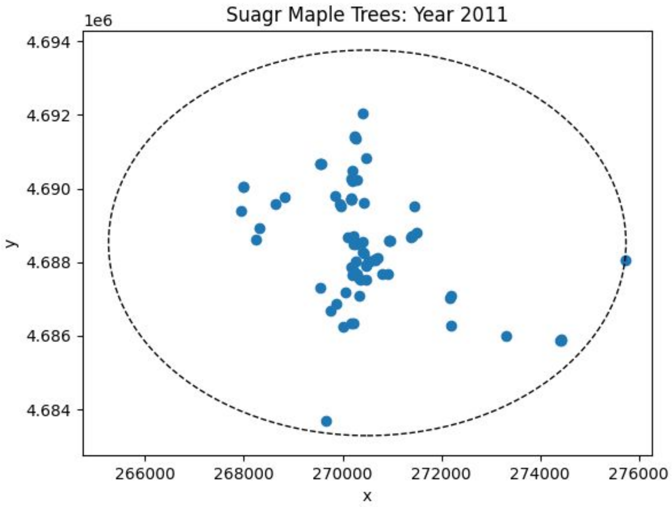
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Circle data (Sugar maple: year 2010 & 2011)





Progress - Data Analysis

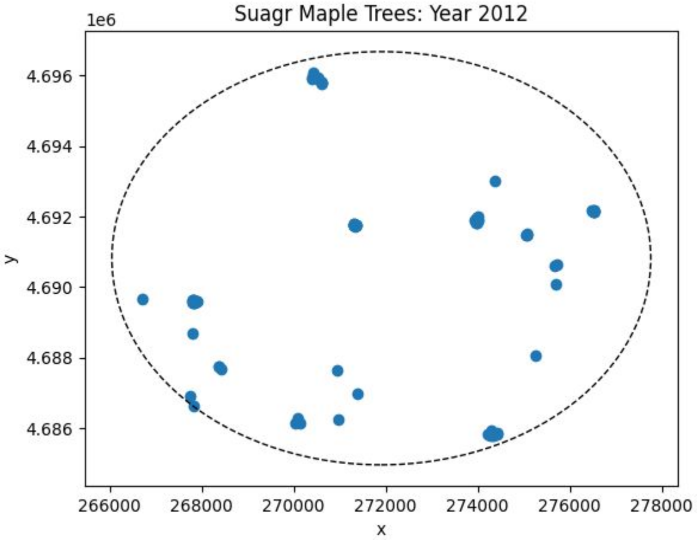
Overview

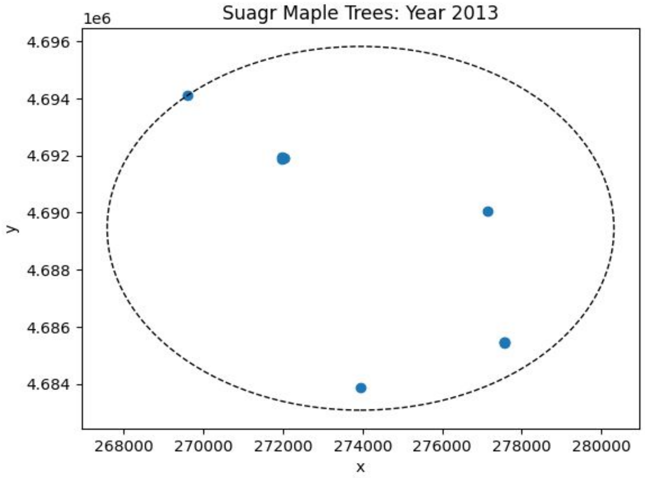
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Circle data (Sugar maple: year 2012 & 2013)





Progress - Data Analysis

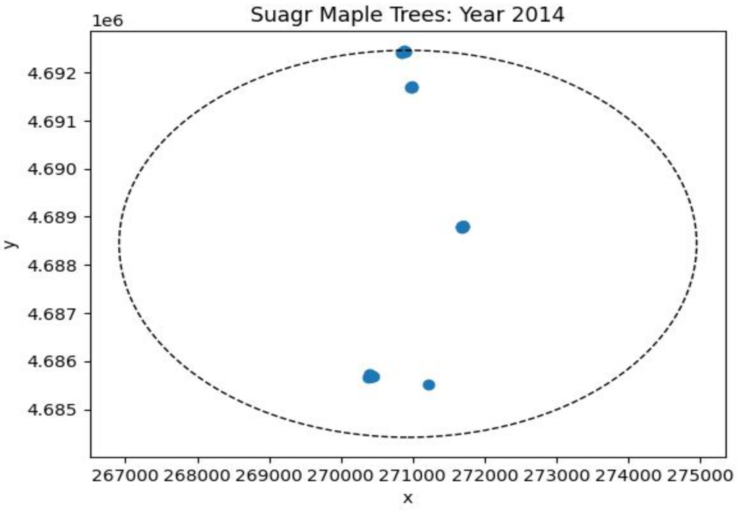
Overview

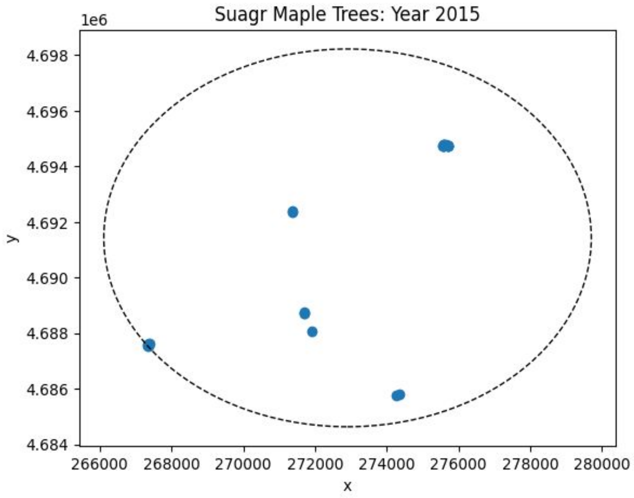
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Circle data (Sugar maple: year 2014 & 2015)





Progress - Data Analysis

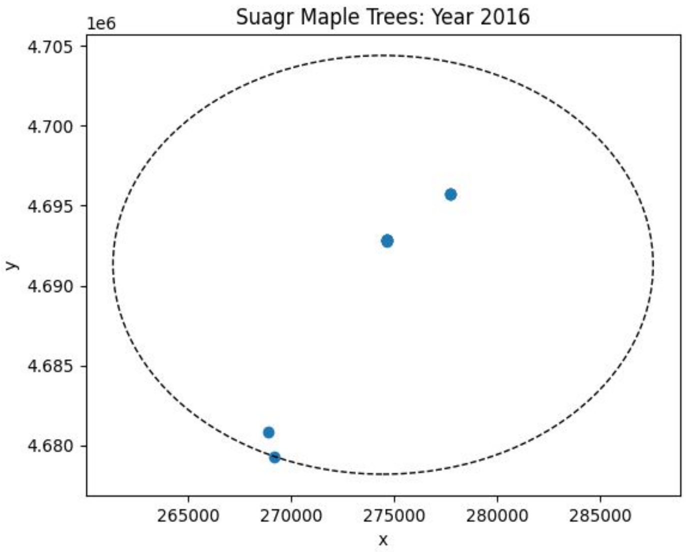
Overview

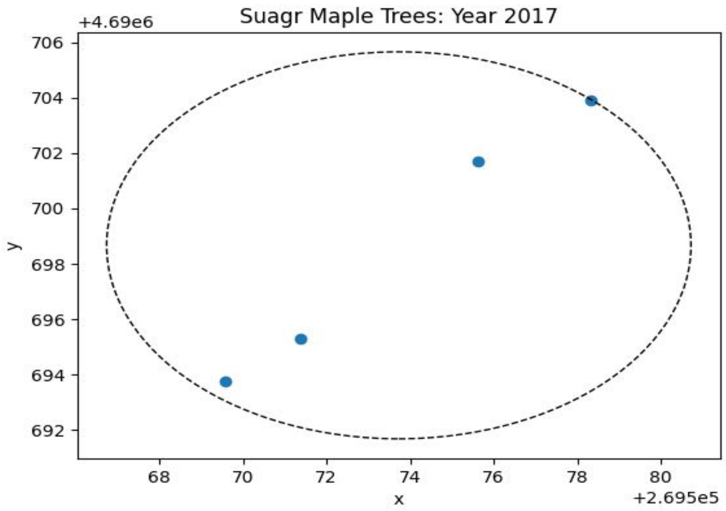
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Circle data (Sugar maple: year 2016 & 2017)





How is the data distributed?

Kernel Density Estimations (KDE)

- We can approximate a 1-D or 2-D distribution as a Gaussian distribution.
- To do this, take a histogram of the data and fit a normal curve to each bin.
- Add all curves together to get our approximate (Gaussian) distribution.

How is the data distributed?

Kernel Density Estimations (KDE)

- If $(x_1, \ldots, x_n)$ are i.i.d. Samples from our distribution, we approximate the probability density function (PDF) via

$$\hat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

- The parameter $h > 0$ is the *bandwidth* of the estimation.
- The function *K()* is the *kernel function*. We use the Gaussian probability density function Normal(0, 1).

Overview

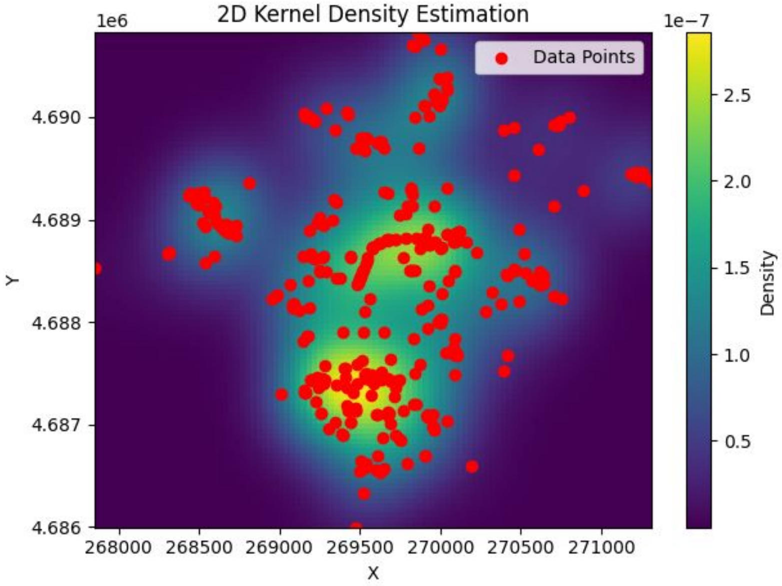
Motivation

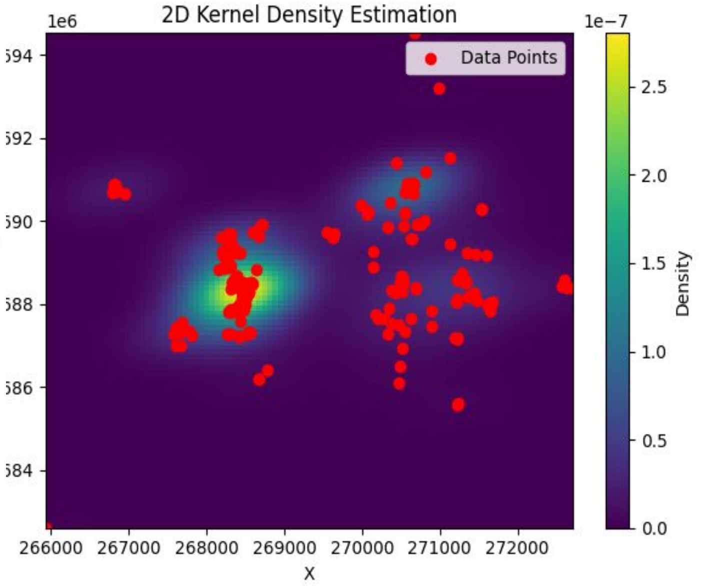
Progress
- **Data Analysis**
- Model Construction

Next Steps

2D Kernel Density Estimation

Red Maple Species: 2008 & 2009





2D Kernel Density Estimation

Overview

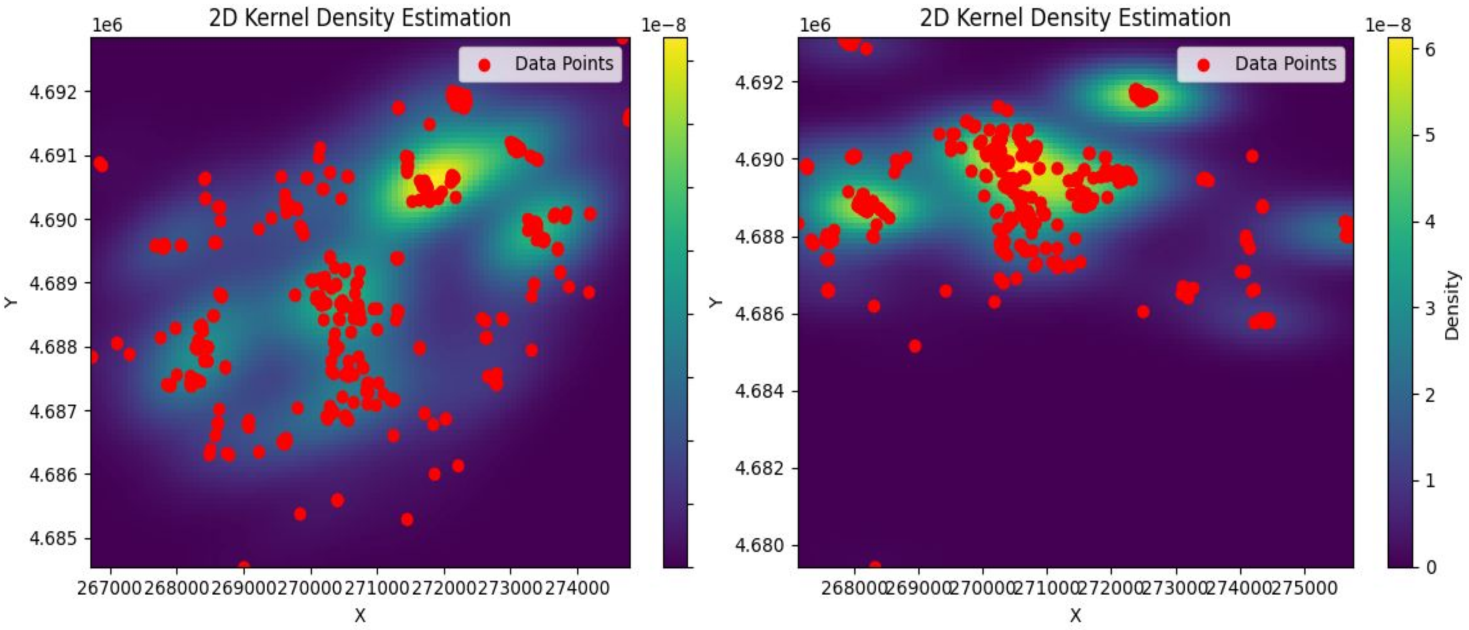
Motivation

Progress

- **Data Analysis**
- Model Construction

Next Steps

Red Maple Species: 2010 & 2011



2D Kernel Density Estimation

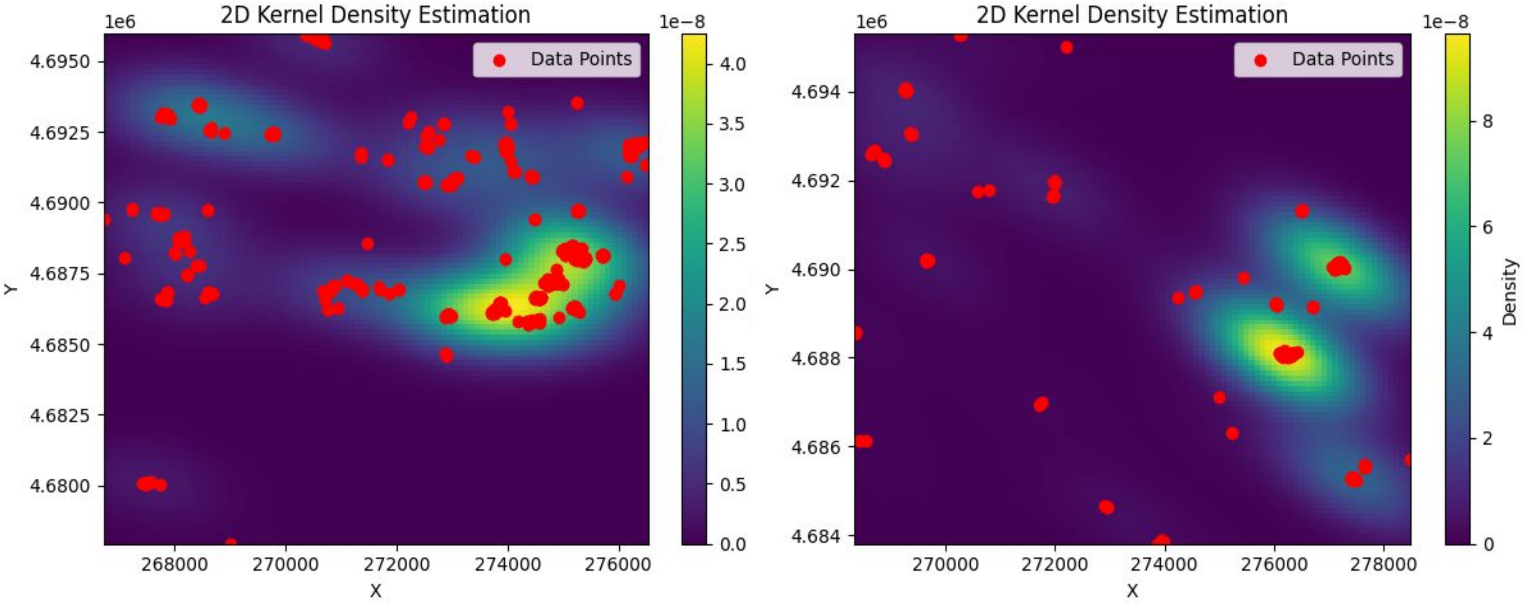
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Red Maple Species: 2012 & 2013



2D Kernel Density Estimation

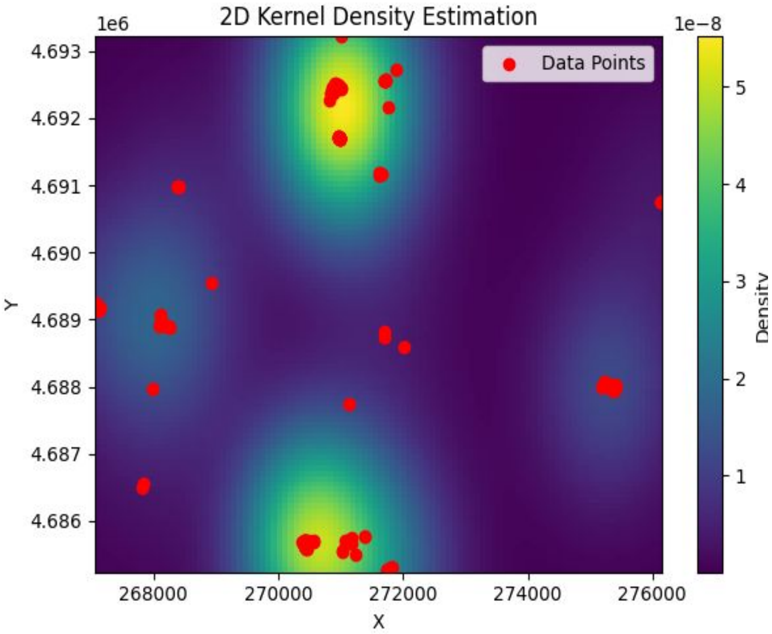
Red Maple Species: 2014 & 2015

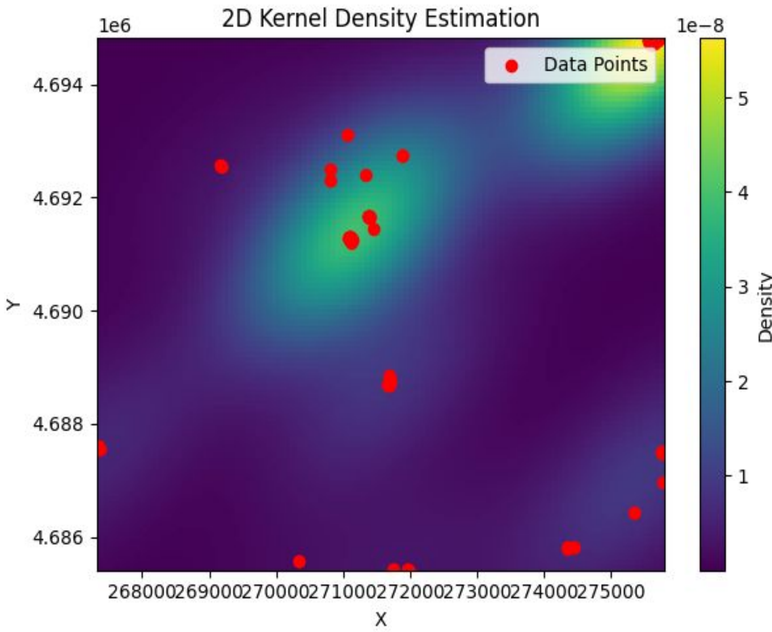
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps





2D Kernel Density Estimation

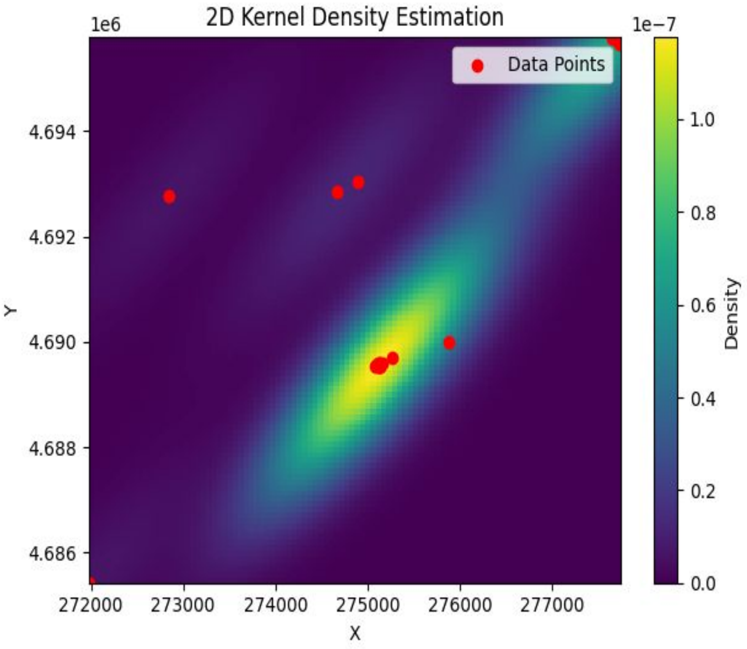
Red Maple Species: 2016 & 2017

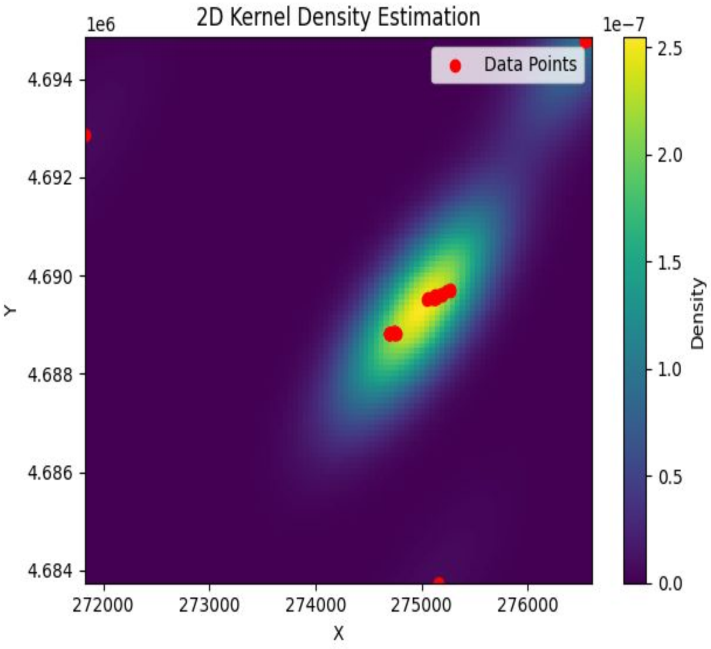
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps





Overview

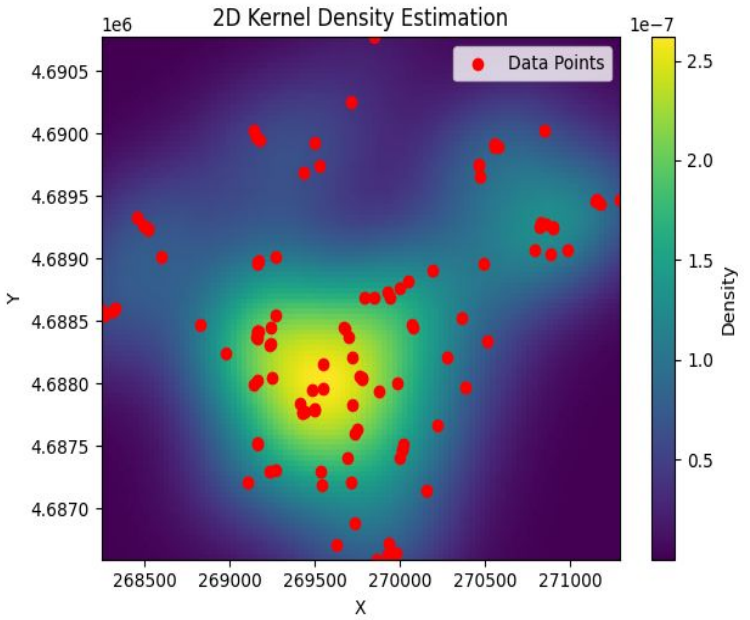
Motivation

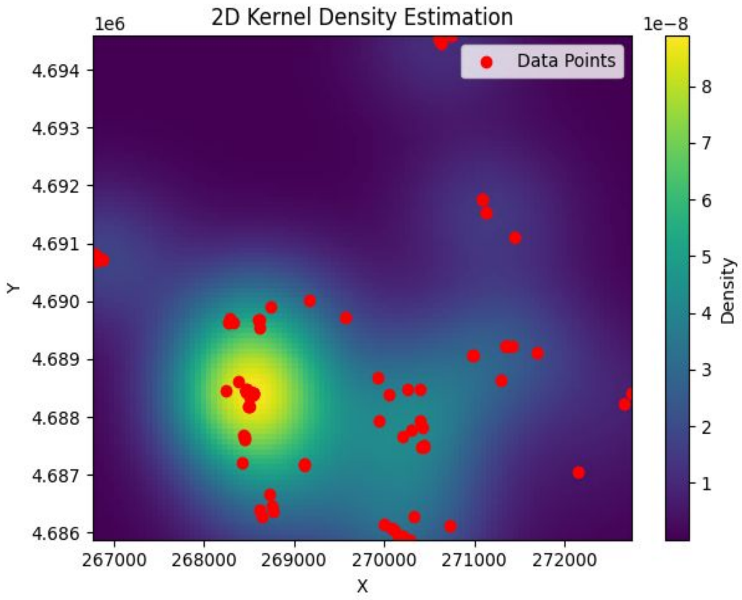
Progress
- **Data Analysis**
- Model Construction

Next Steps

Progress - Data Analysis

2D KDE (Sugar maple: 2008 & 2009 data)





Progress - Data Analysis

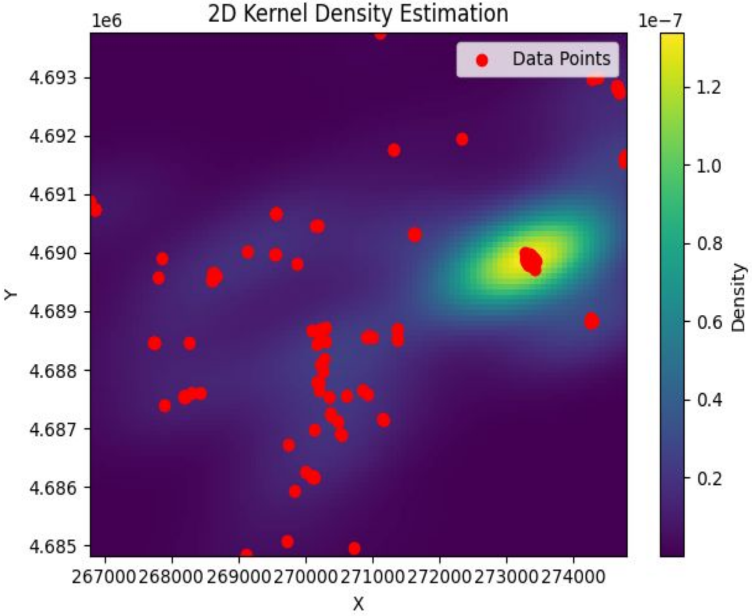
Overview

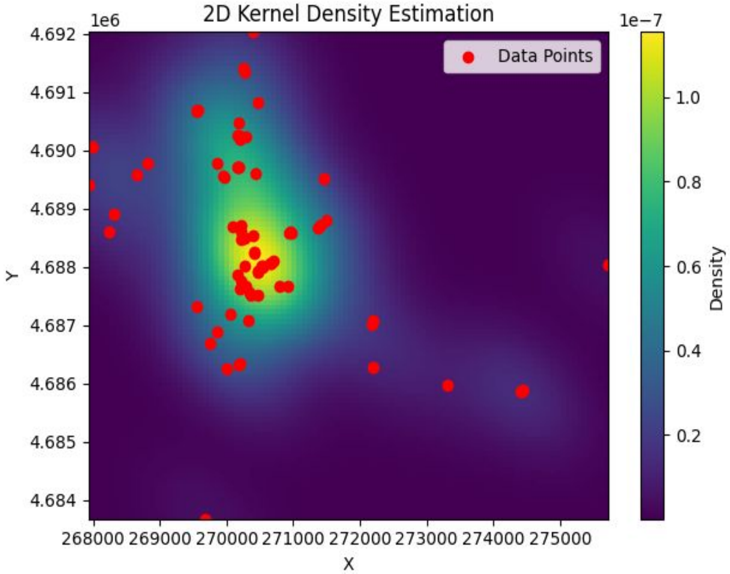
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

2D KDE (Sugar maple: 2010 & 2011 data)





Progress - Data Analysis

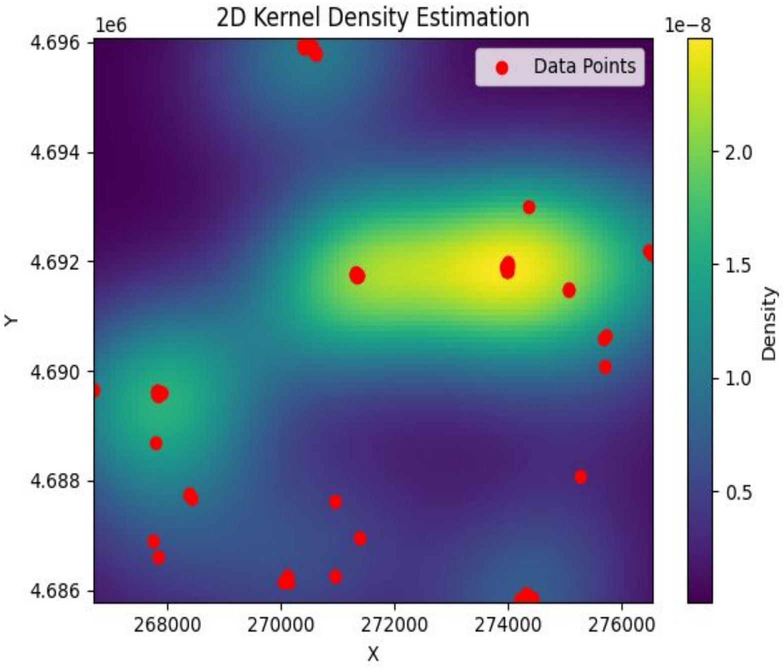
Overview

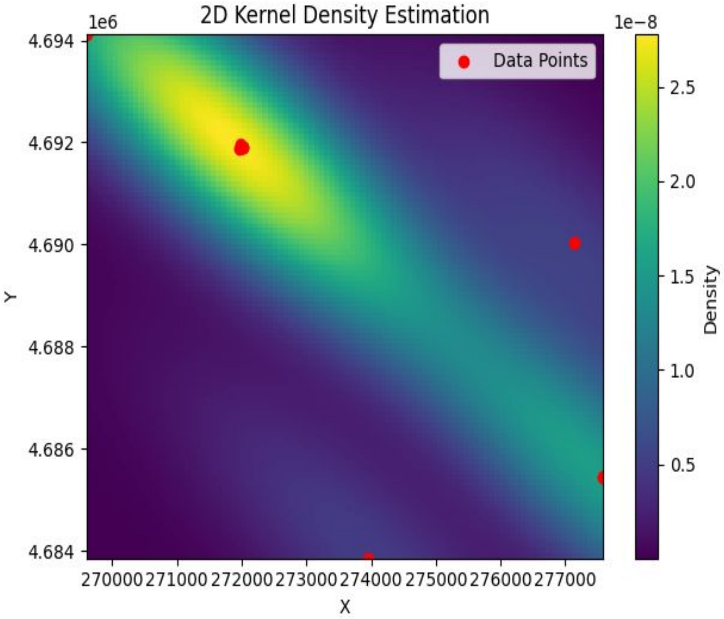
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

2D KDE (Sugar maple: 2012 & 2013 data)





Progress - Data Analysis

2D KDE (Sugar maple: 2014 & 2015 data)

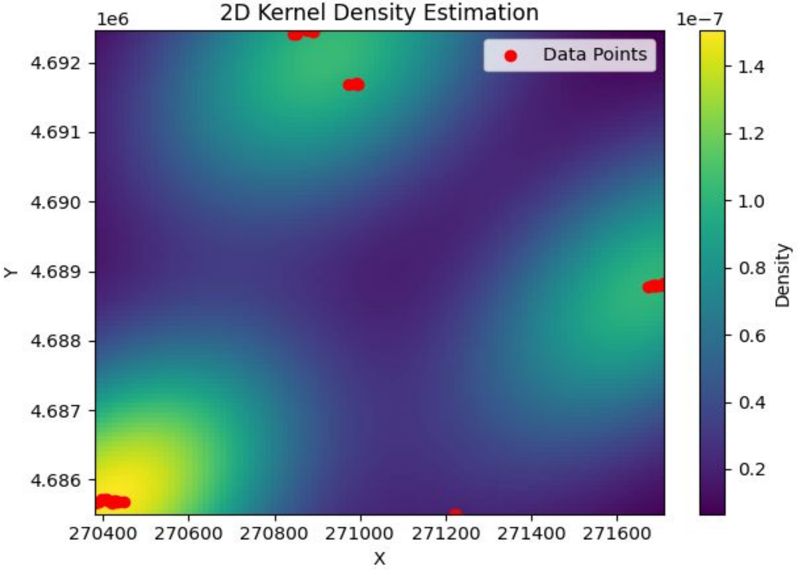
Overview

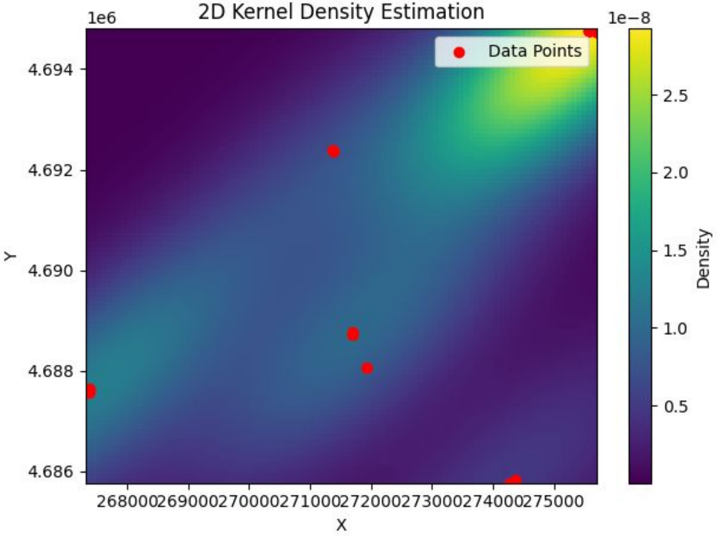
Motivation

Progress

- **Data Analysis**
- Model Construction

Next Steps





Progress - Data Analysis

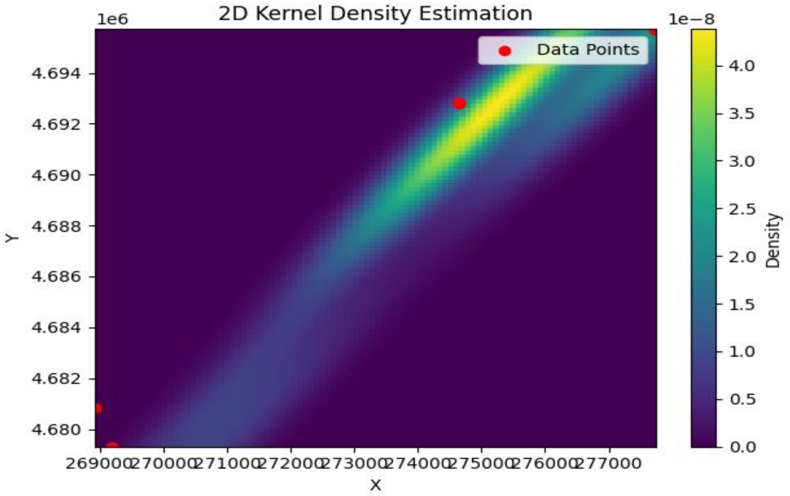
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

2D KDE (Sugar maple: 2016 data)



Progress - Data Analysis

Overview

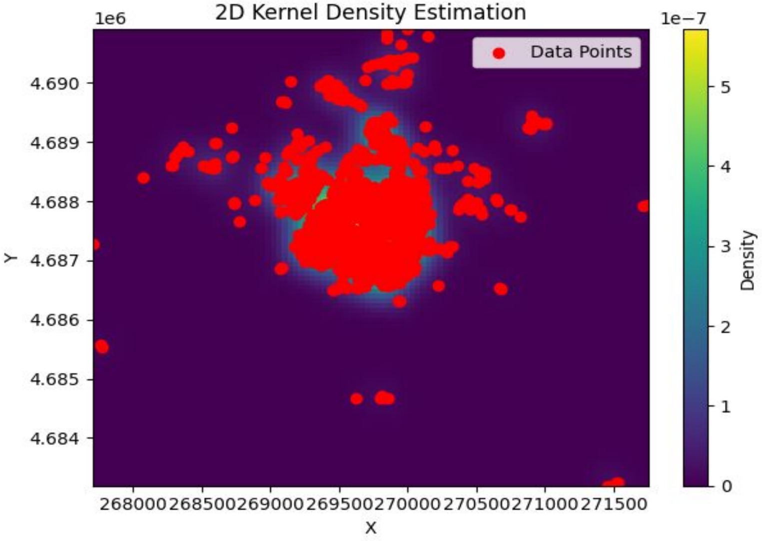
Motivation

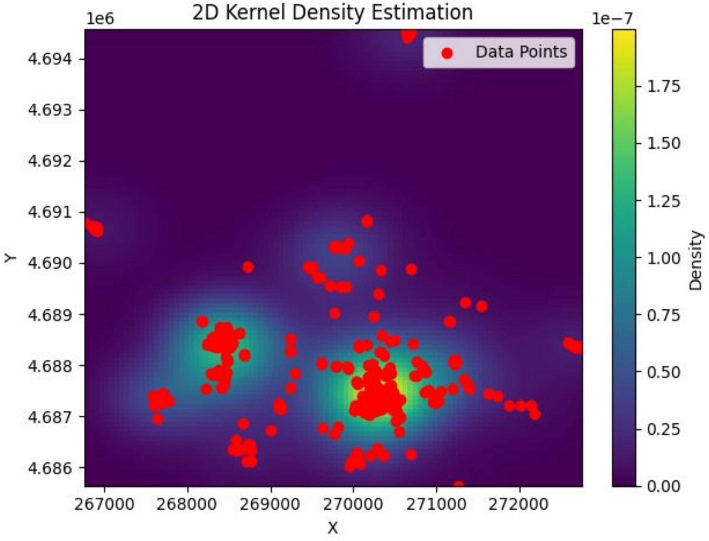
Progress

- **Data Analysis**
- Model Construction

Next Steps

2D KDE (Norway maple: 2008 & 2009 data)





Overview

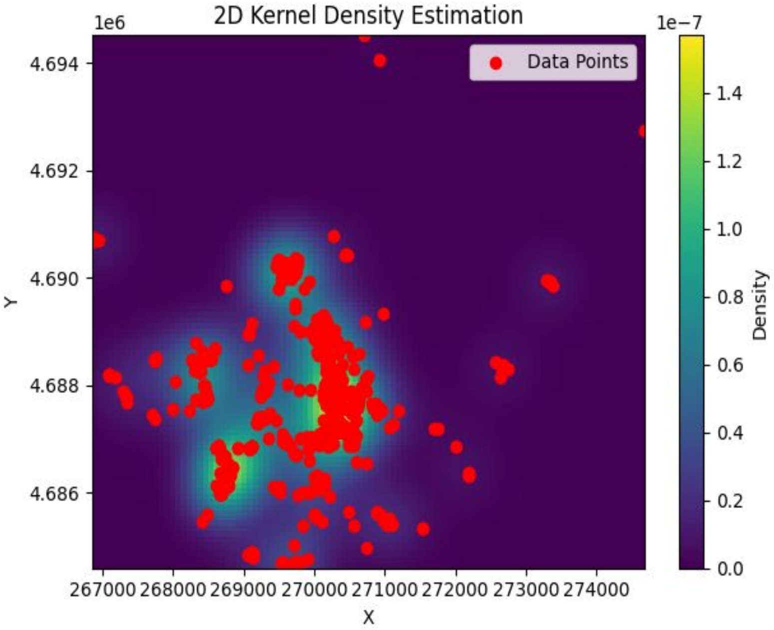
Motivation

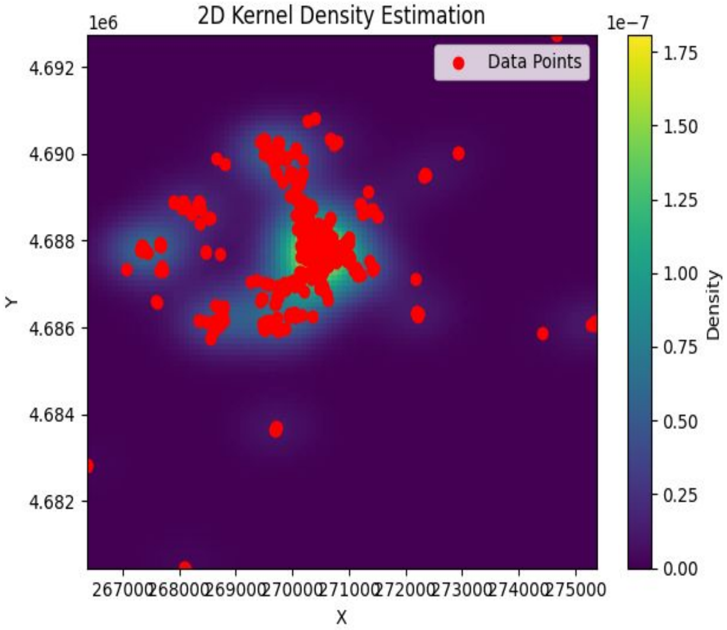
Progress
- **Data Analysis**
- Model Construction

Next Steps

Progress - Data Analysis

2D KDE (Norway maple: 2010 & 2011 data)





Progress - Data Analysis

2D KDE (Norway maple: 2012 & 2013 data)

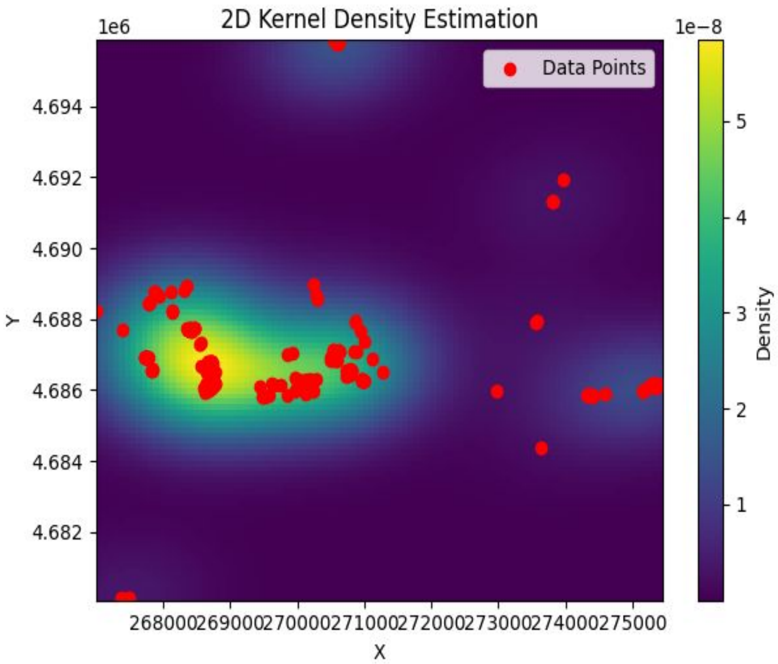
Overview

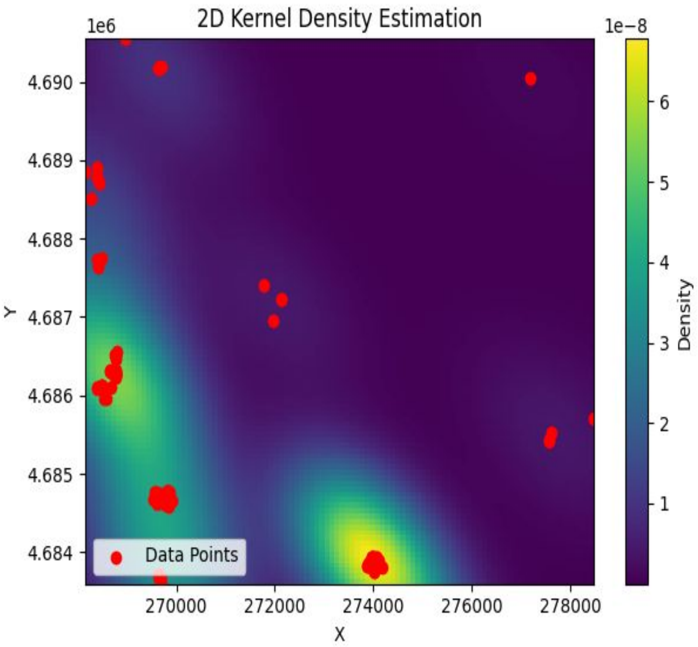
Motivation

Progress

- **Data Analysis**
- Model Construction

Next Steps





Progress - Data Analysis

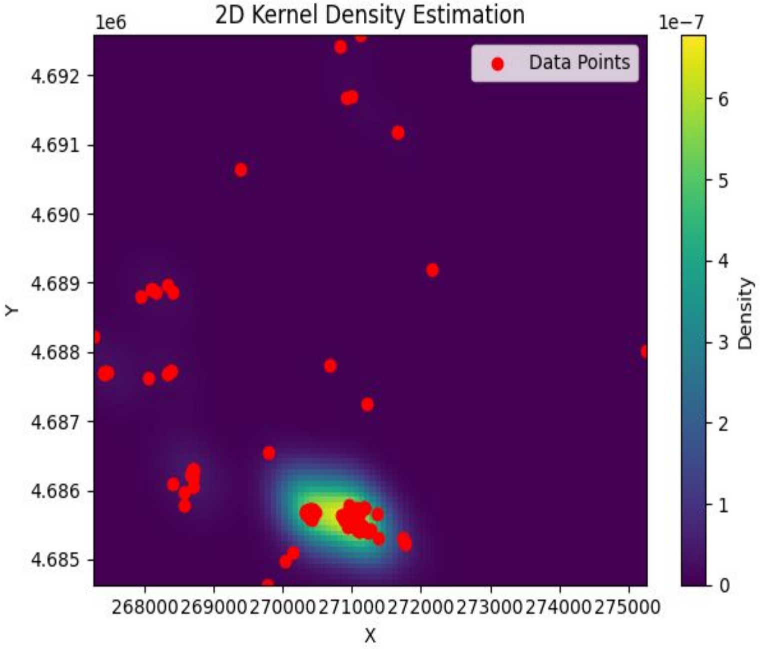
Overview

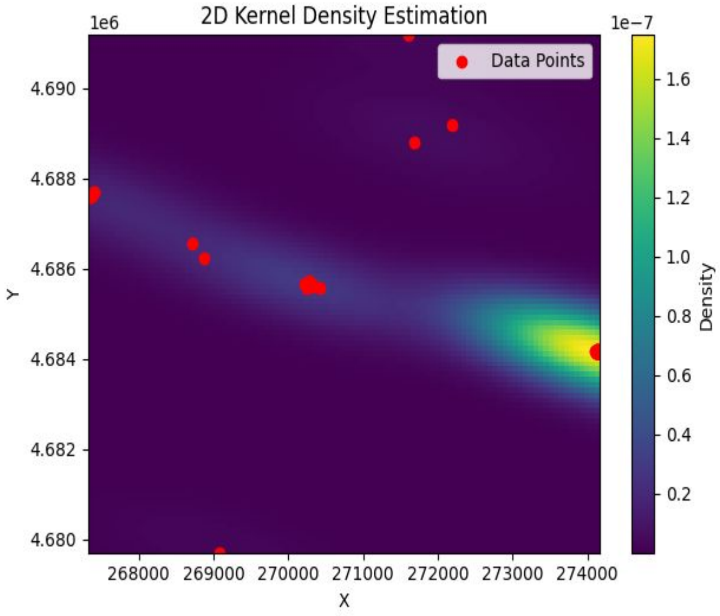
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

2D KDE (Norway maple: 2014 & 2015 data)





Overview

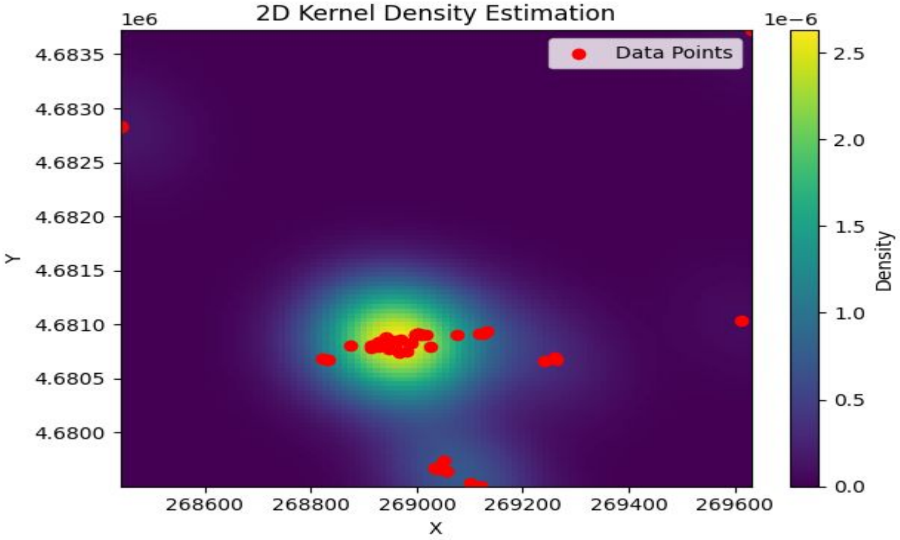
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

Progress - Data Analysis

2D KDE (Norway maple: 2016)



Progress - Data Analysis

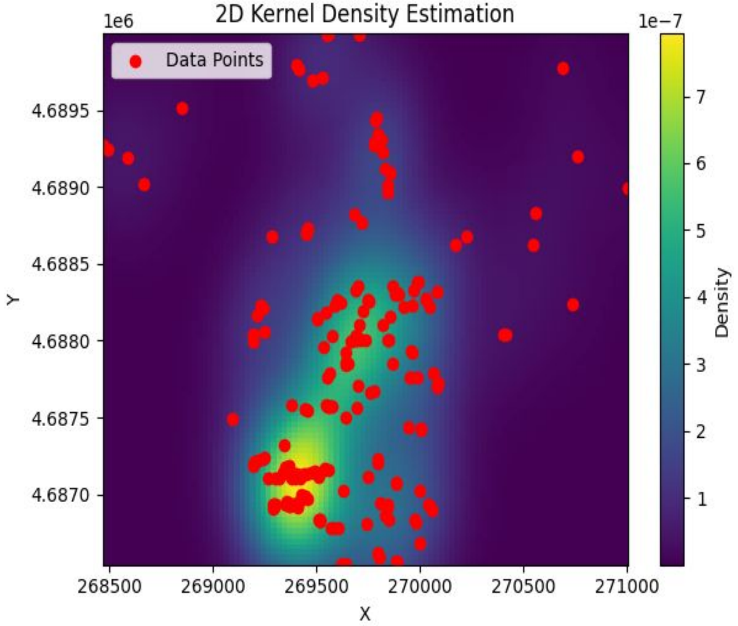
2D KDE (Silver maple: 2008 & 2009 data)

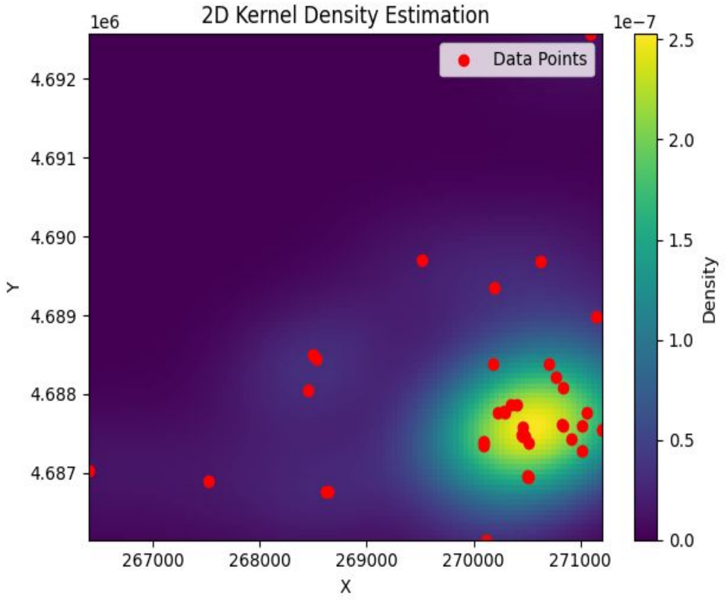
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps





Progress - Data Analysis

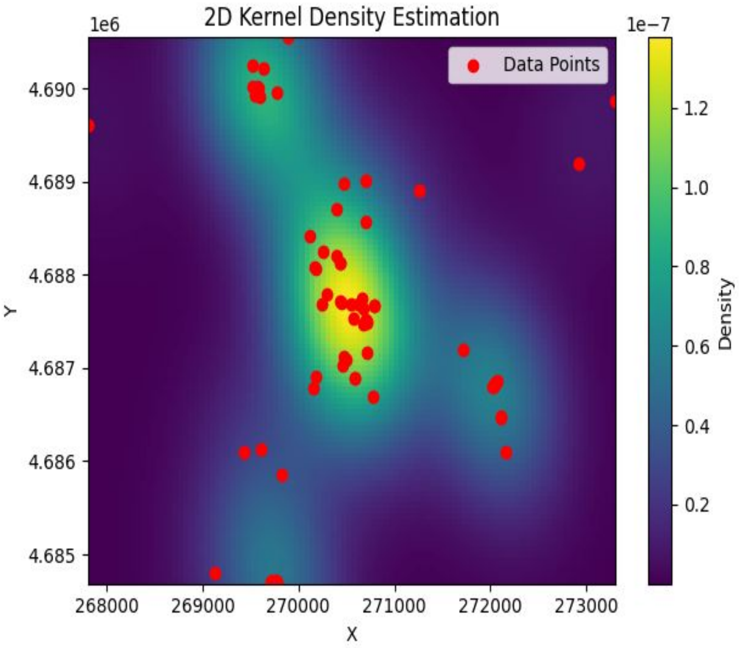
Overview

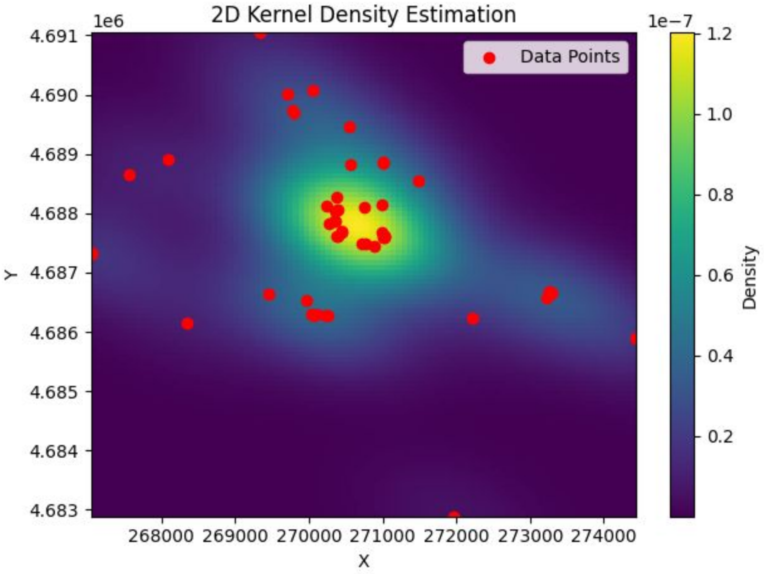
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

2D KDE (Silver maple: 2010 & 2011 data)





Progress - Data Analysis

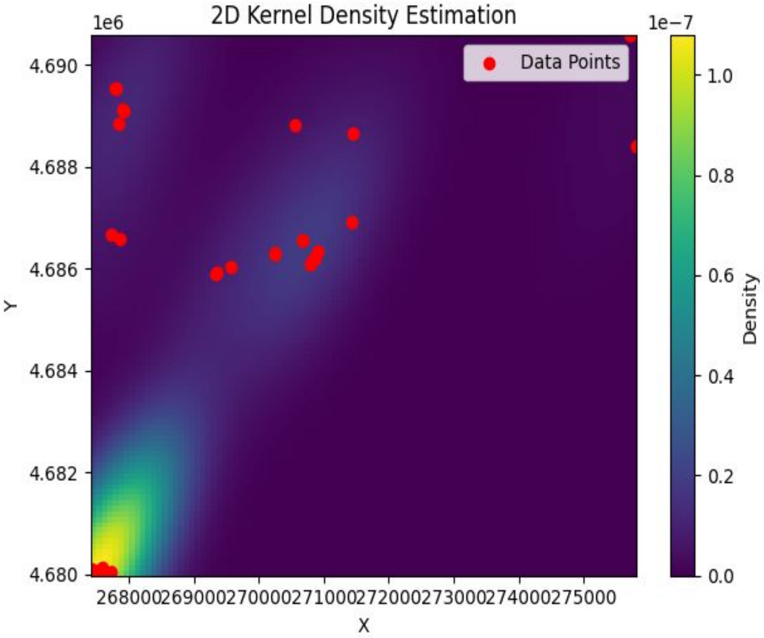
Overview

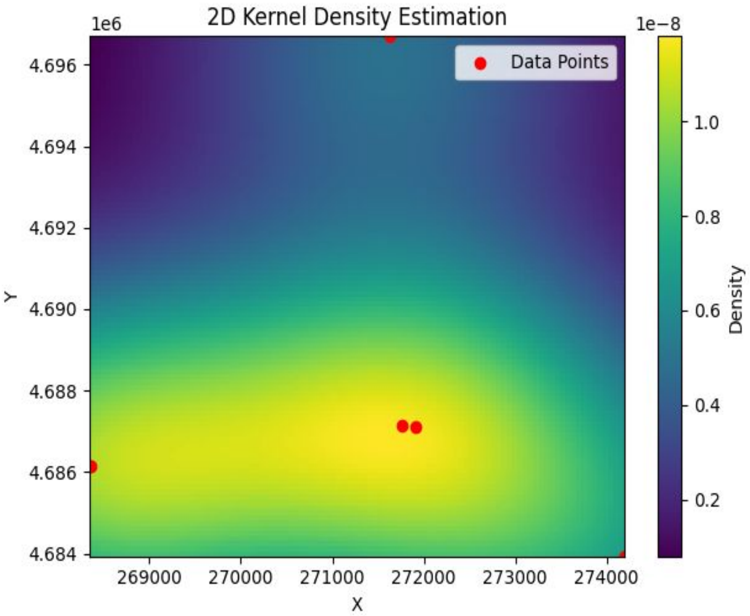
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

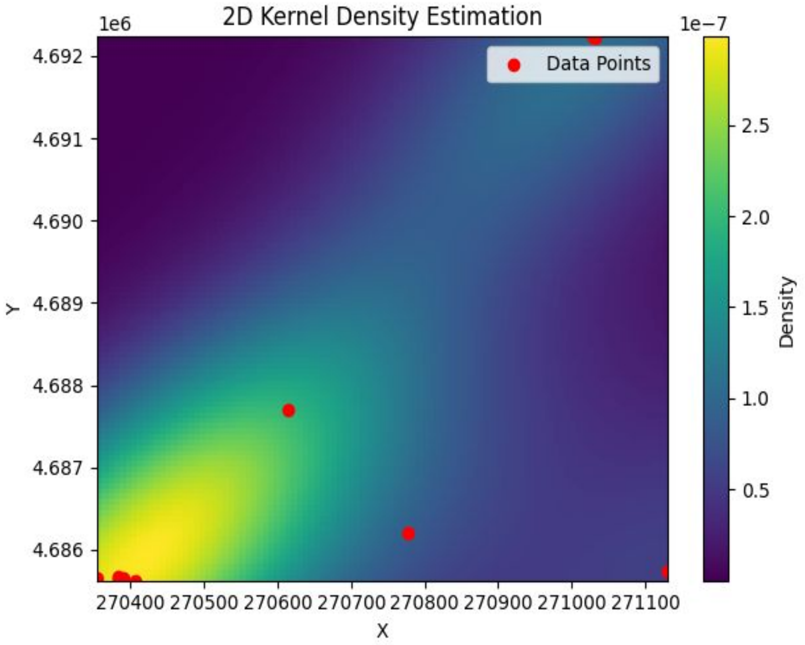
2D KDE (Silver maple: 2012 & 2013 data)





Progress - Data Analysis

2D KDE (Silver maple: 2014 data)



Overview

Motivation

Progress

- **Data Analysis**
- Model Construction

Next Steps

Application of the KDE

Kernel Density Estimations (KDE)

- We can approximate a 1-D or 2-D distribution as a Gaussian distribution.
- Take a histogram of the data and fit a normal curve to each bin. Then, adding all curves together will give us our approximate underlying distribution.
- **The peak of the summed curve should indicate the peak of densities, corresponding to potential clusters.**

Detecting Clusters

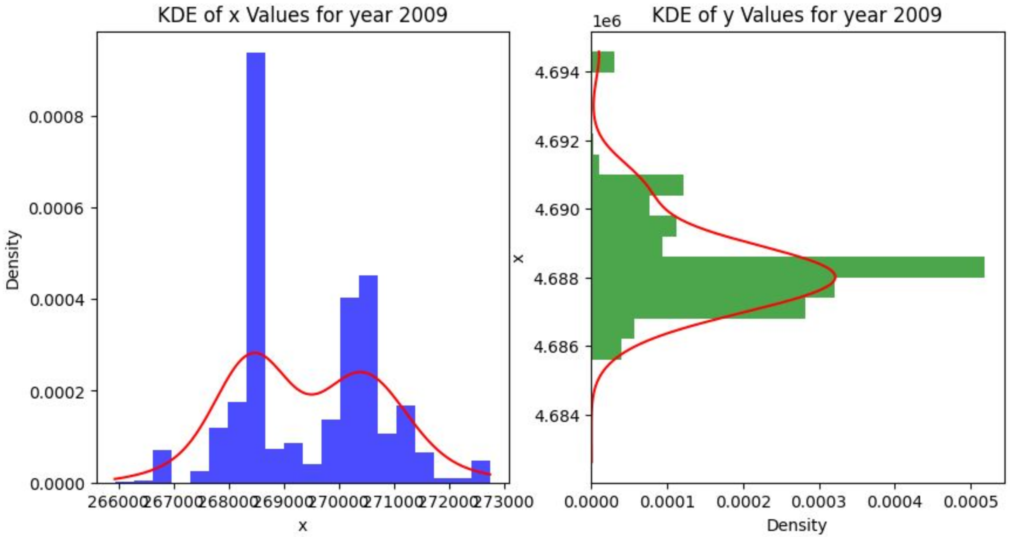
Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

One-Dimensional KDE: 2009 Data



Detecting Clusters

Overview

Motivation

Progress

- **Data Analysis**
- Model Construction

Next Steps

In theory...

- Use 1-D KDE to pinpoint the 2-D clusters

In reality...?

Detecting Clusters

Overview

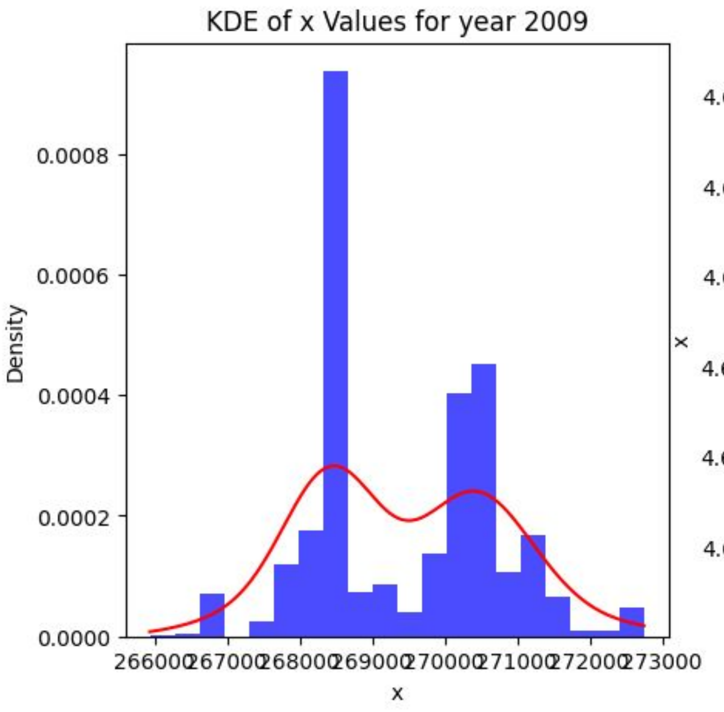
Motivation

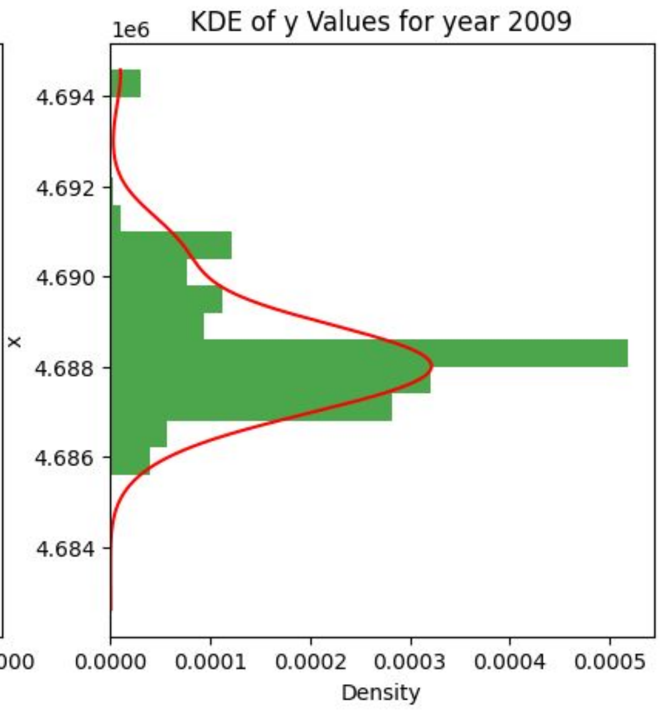
Progress

- **Data Analysis**
- Model Construction

Next Steps

One-Dimensional KDE: 2009 Data





Detecting Clusters

Overview

Motivation

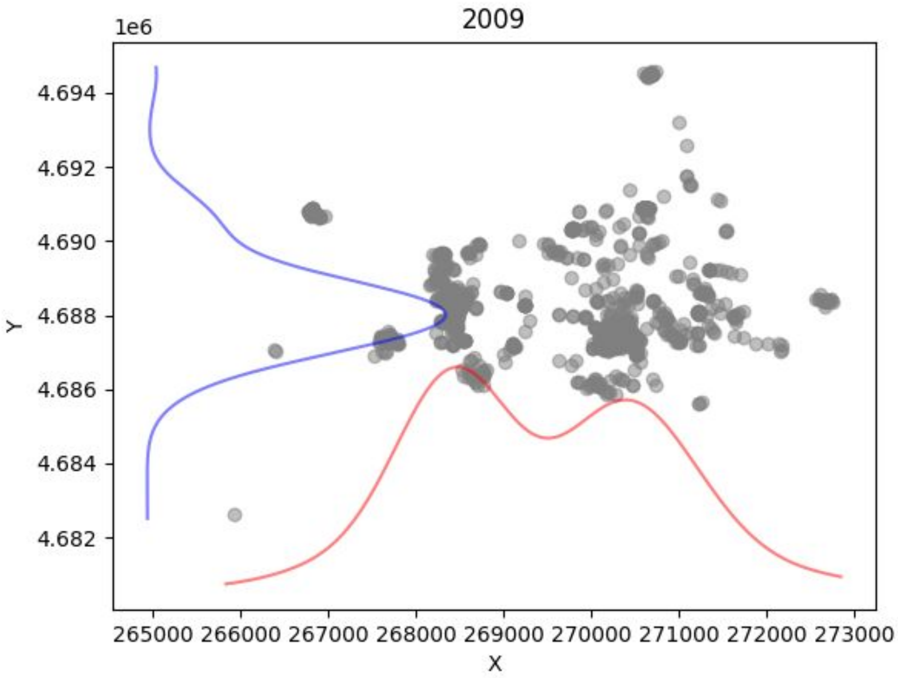
Progress

- **Data Analysis**
- Model Construction

Next Steps

In theory...

- Use 1-D to pinpoint the 2-D clusters



Detecting Clusters

Overview

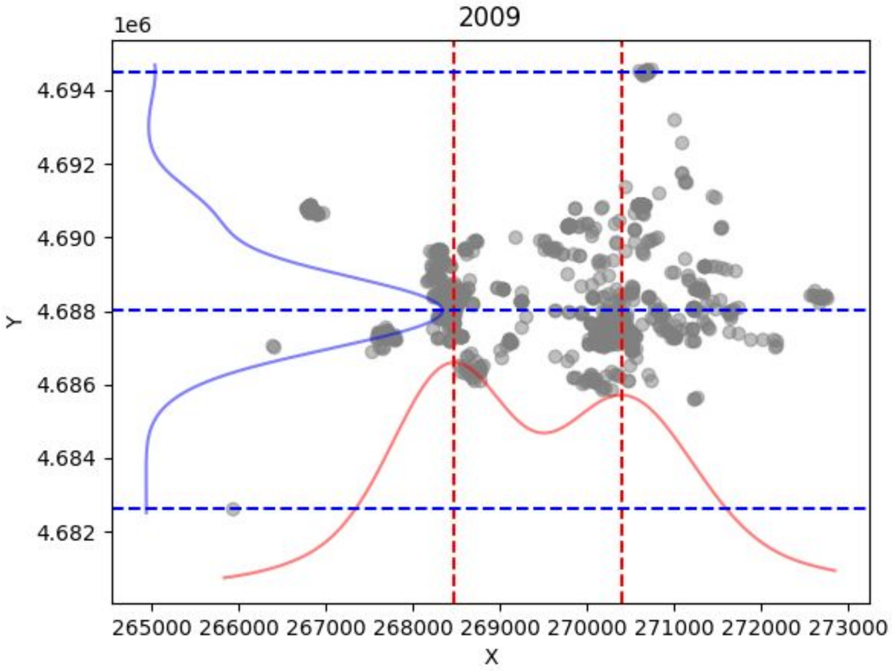
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

In theory...

- Use 1-D to pinpoint the 2-D clusters



Detecting Clusters

Overview

Motivation

Progress

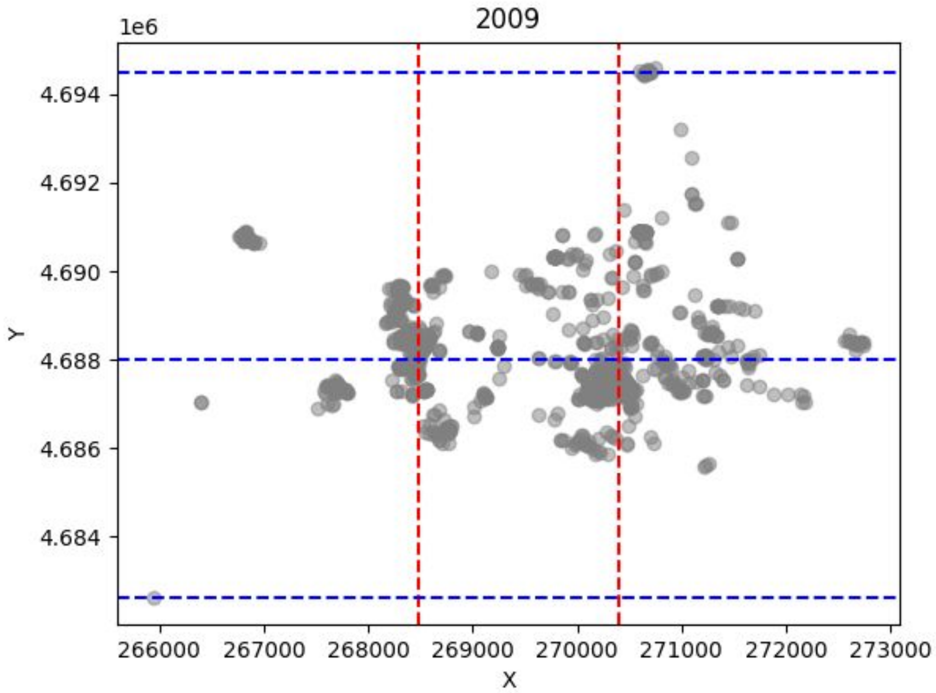
- **Data Analysis**
- Model Construction

Next Steps

In practice...

- Found Cluster!

Or did we?



Detecting Clusters

What about the other years?

Detecting Clusters

Overview

Motivation

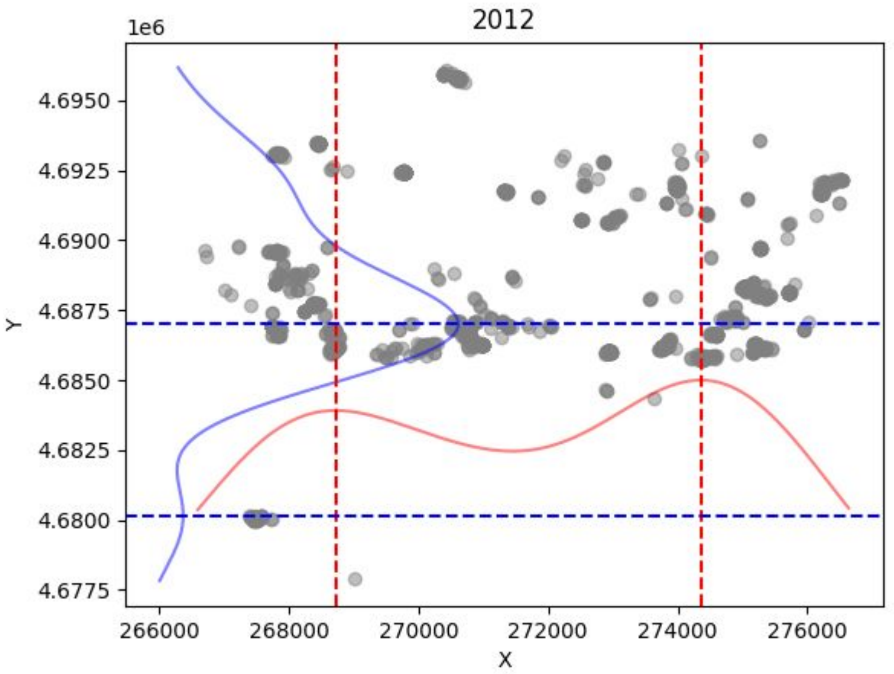
Progress

- **Data Analysis**
- Model Construction

Next Steps

In theory...

- Use 1-D to pinpoint the 2-D clusters



Detecting Clusters

Overview

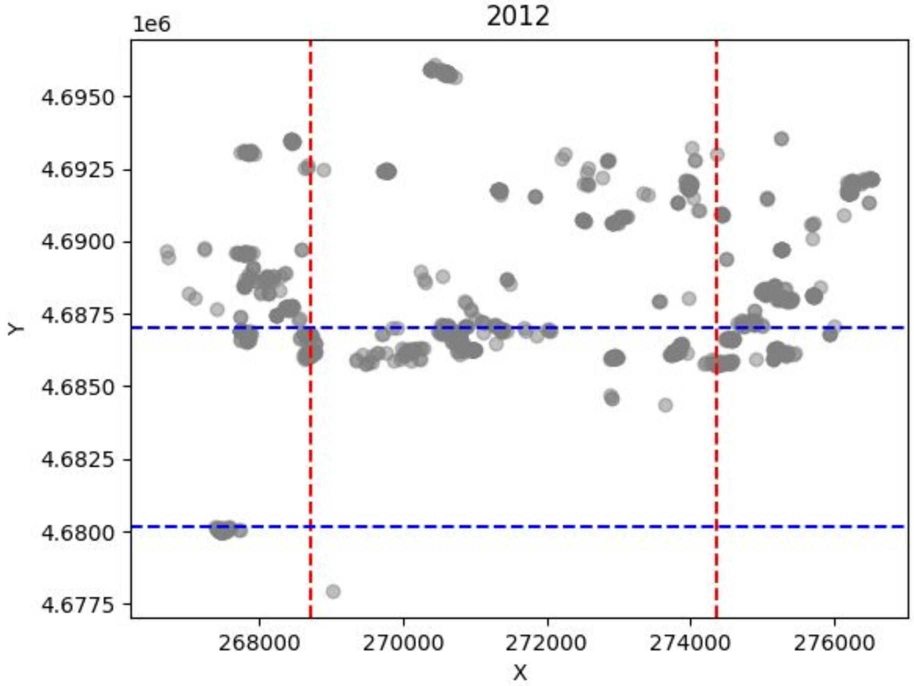
Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps

In reality...

- Not so nice anymore :(



Detecting Clusters

Overview

Motivation

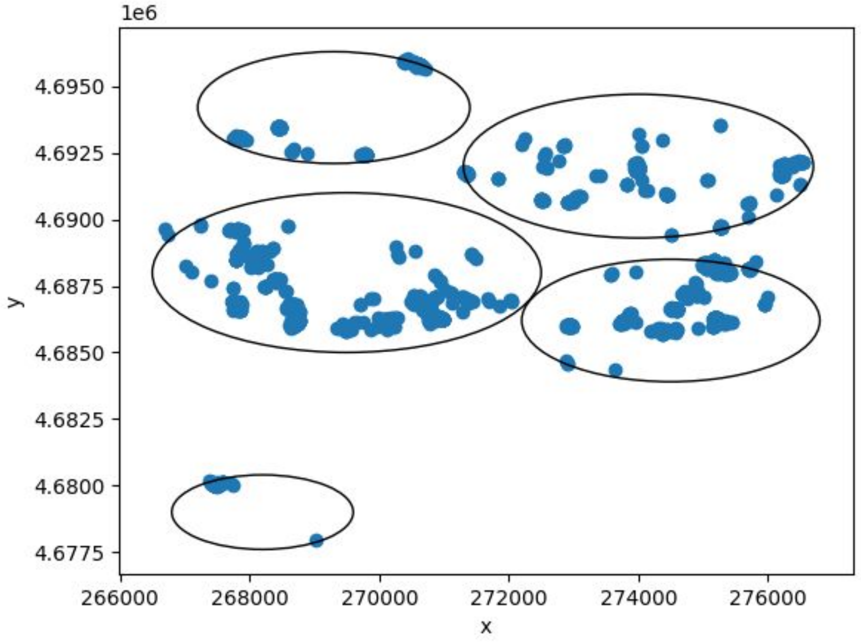
Progress

- **Data Analysis**
- Model Construction

Next Steps

In reality...

- Structures cannot be determined as good as eyeballing the data directly.



Overview

Motivation

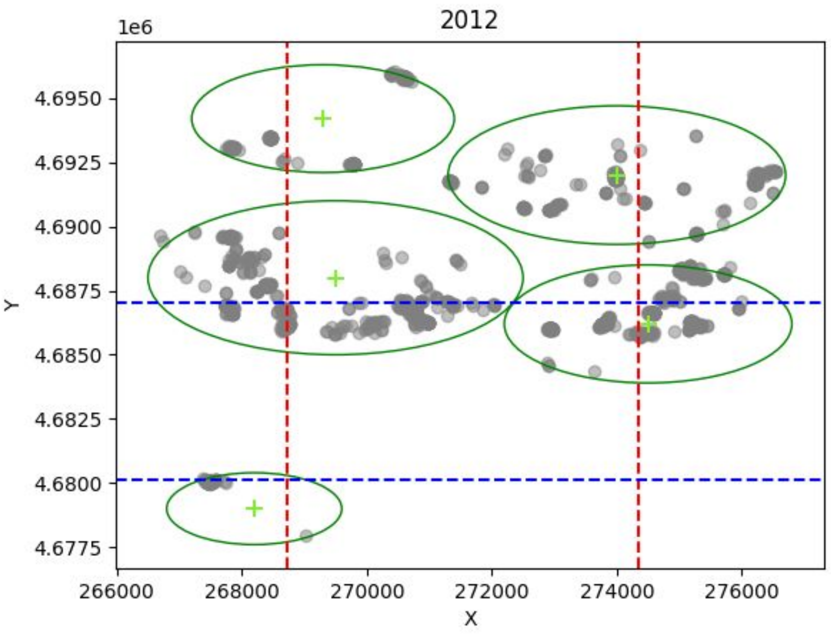
Progress
- **Data Analysis**
- Model Construction

Next Steps

Detecting Clusters

In reality...

- Structures cannot be determined as good as eyeballing the data directly.



Here the green circles/centers are results of eyeballing with help from 2D KDEs.

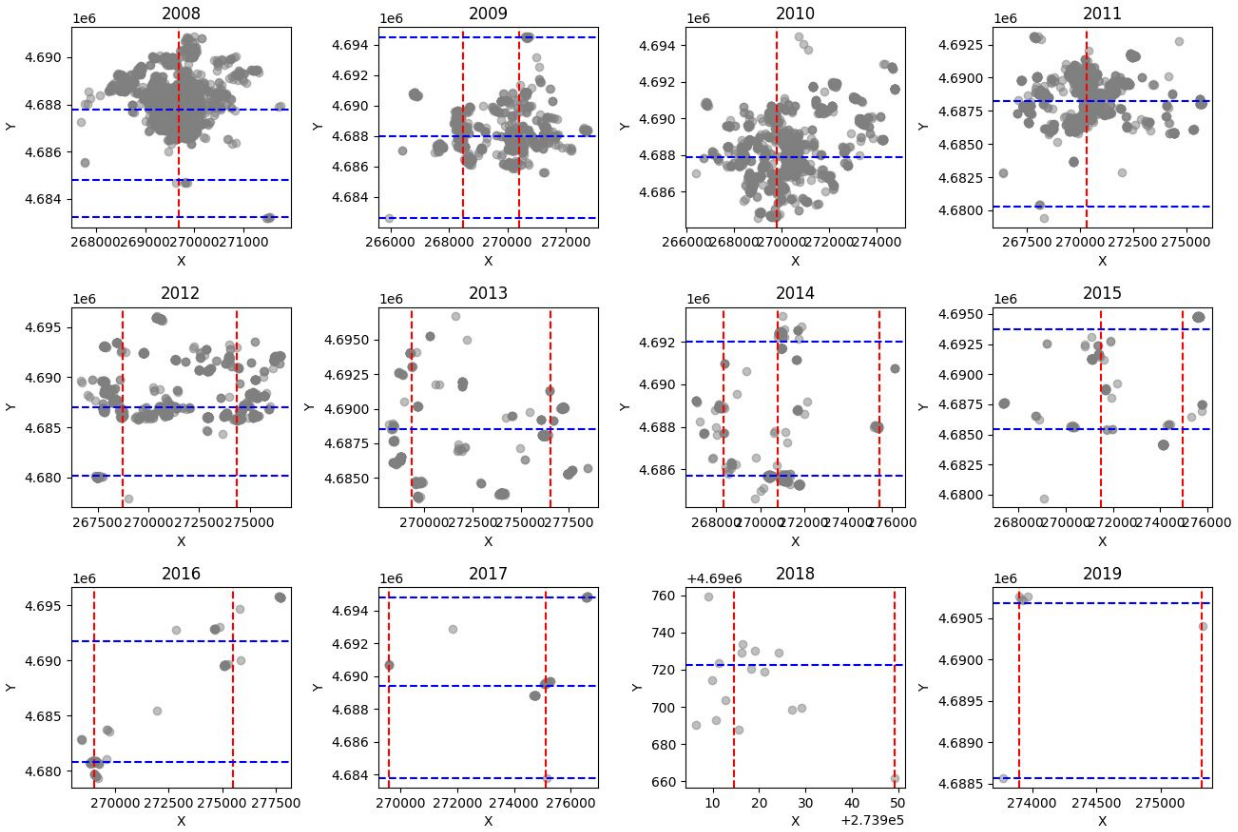
Detecting Clusters

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



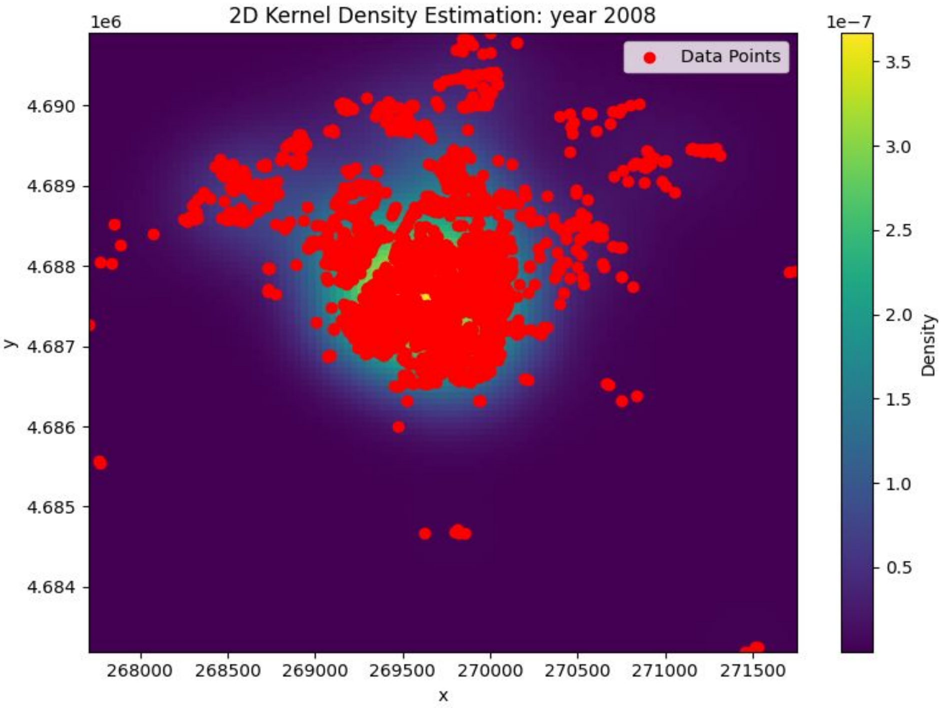
Detecting Clusters

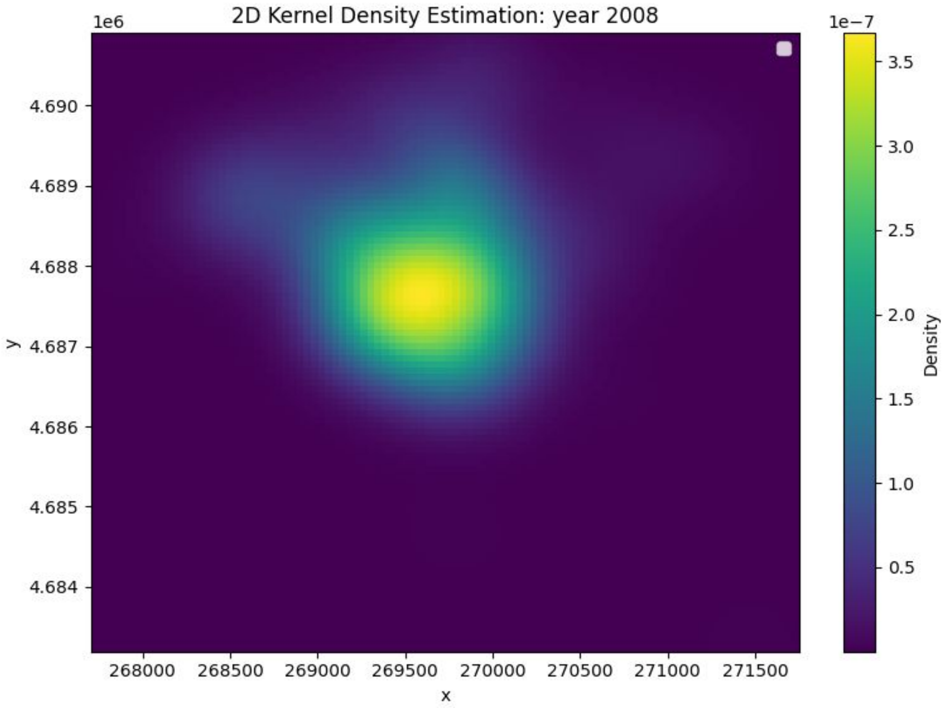
In reality...

- Structures cannot be determined as good as eyeballing the data directly.
  - Resolving to determining clusters by eye

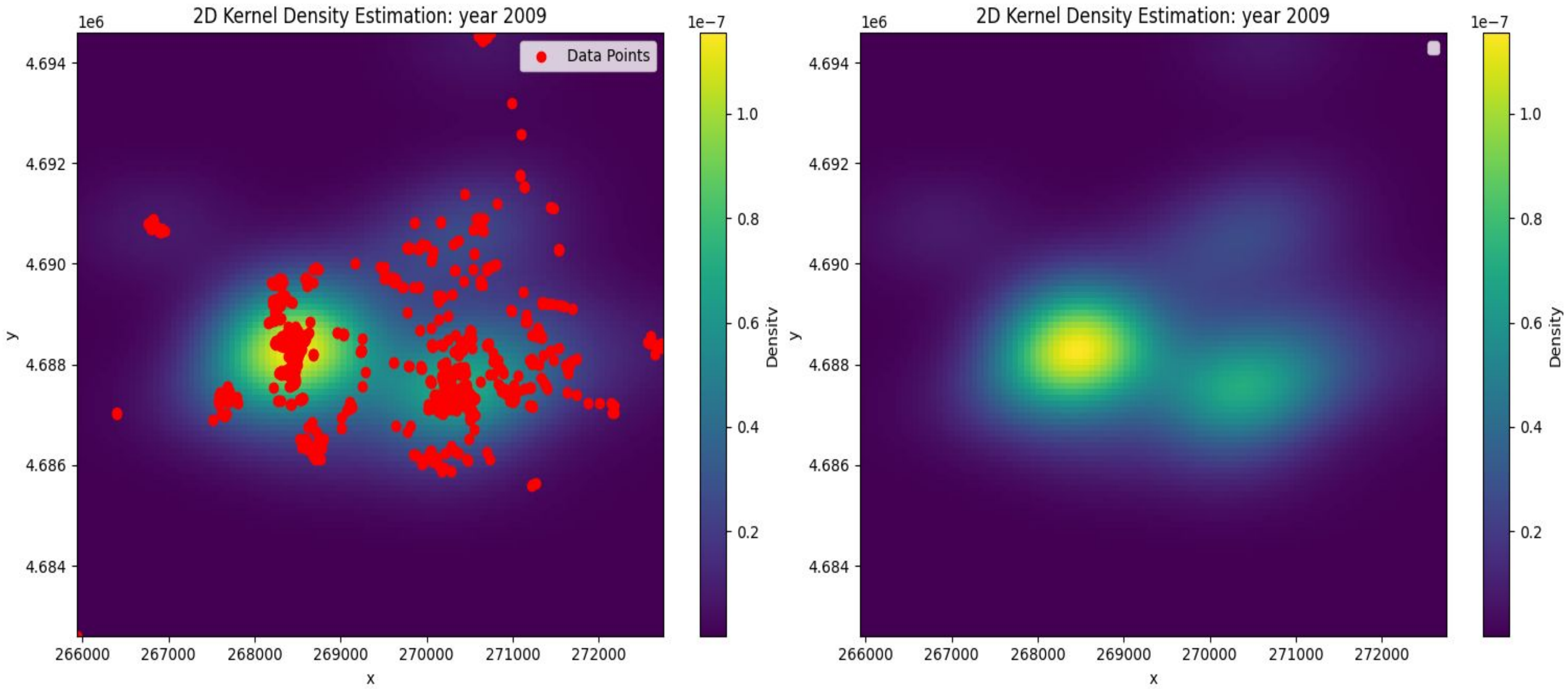
- Currently an open question!

2D KDE for Year 2008





2D KDE for Year 2009



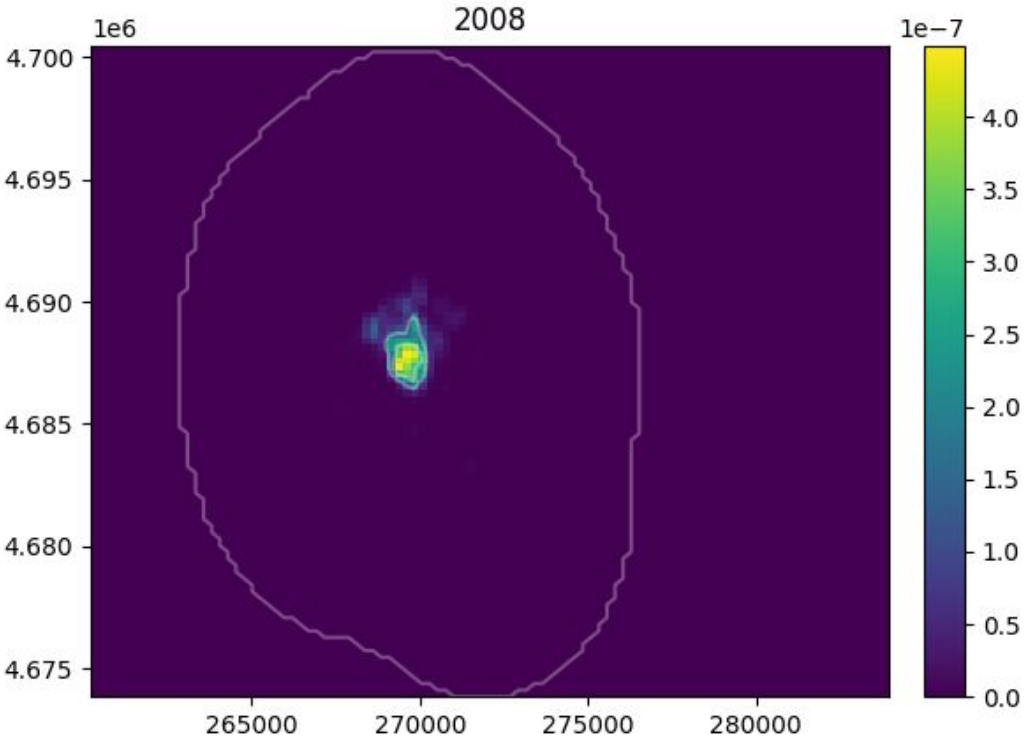
Temporal Shift by Year - 2008

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



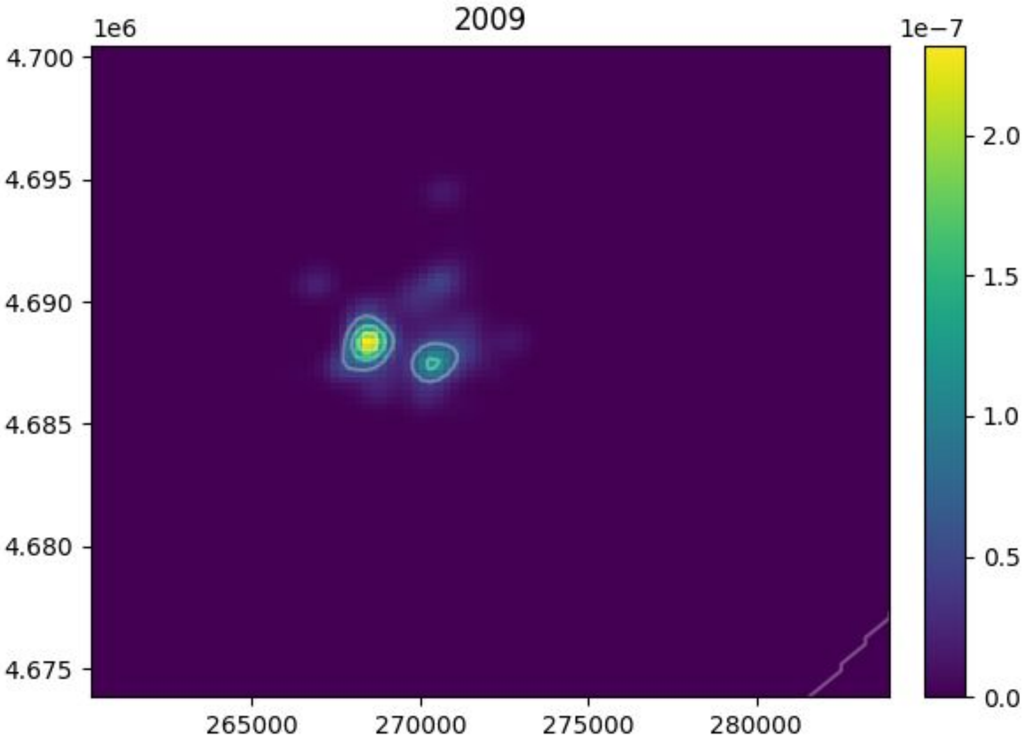
Temporal Shift by Year - 2009

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



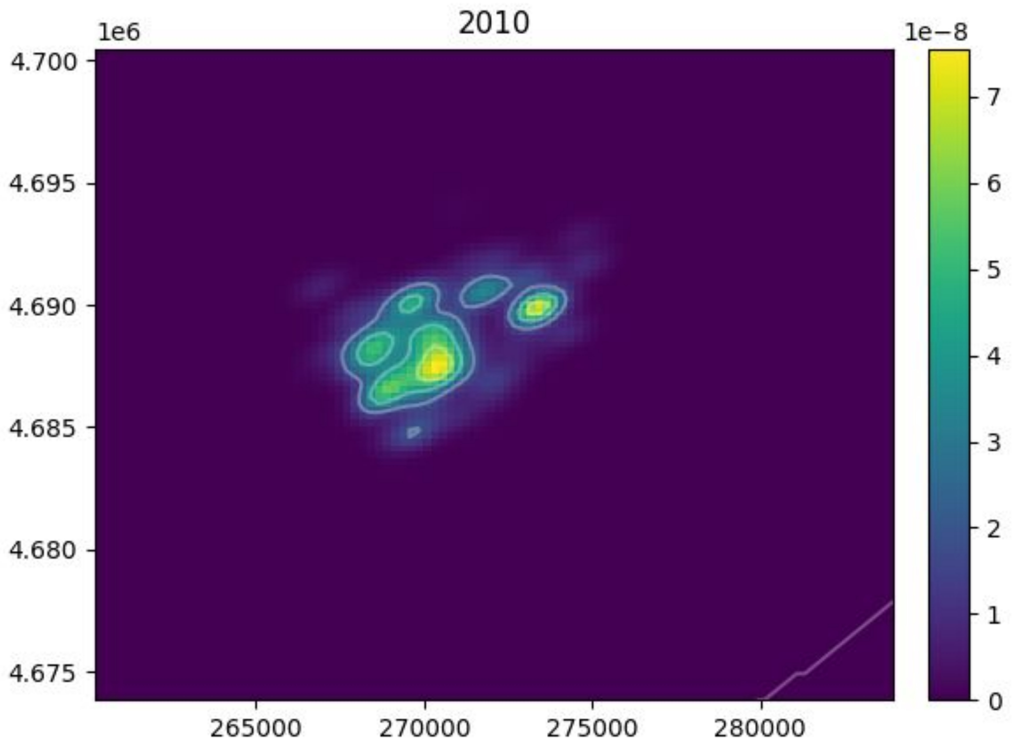
Temporal Shift by Year - 2010

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



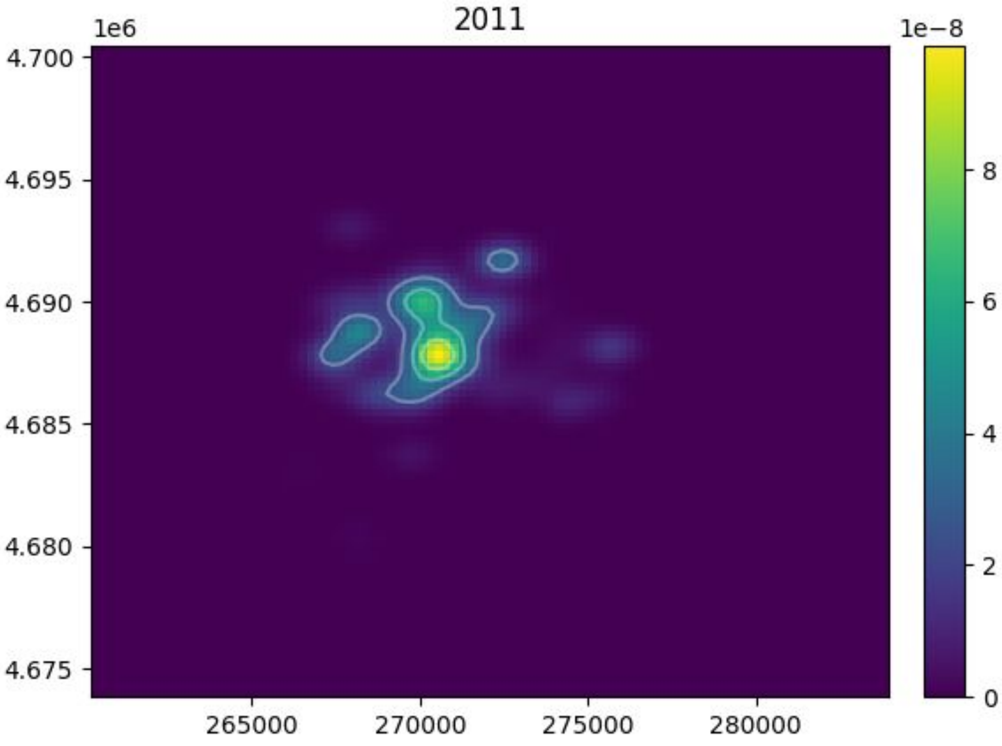
Temporal Shift by Year - 2011

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



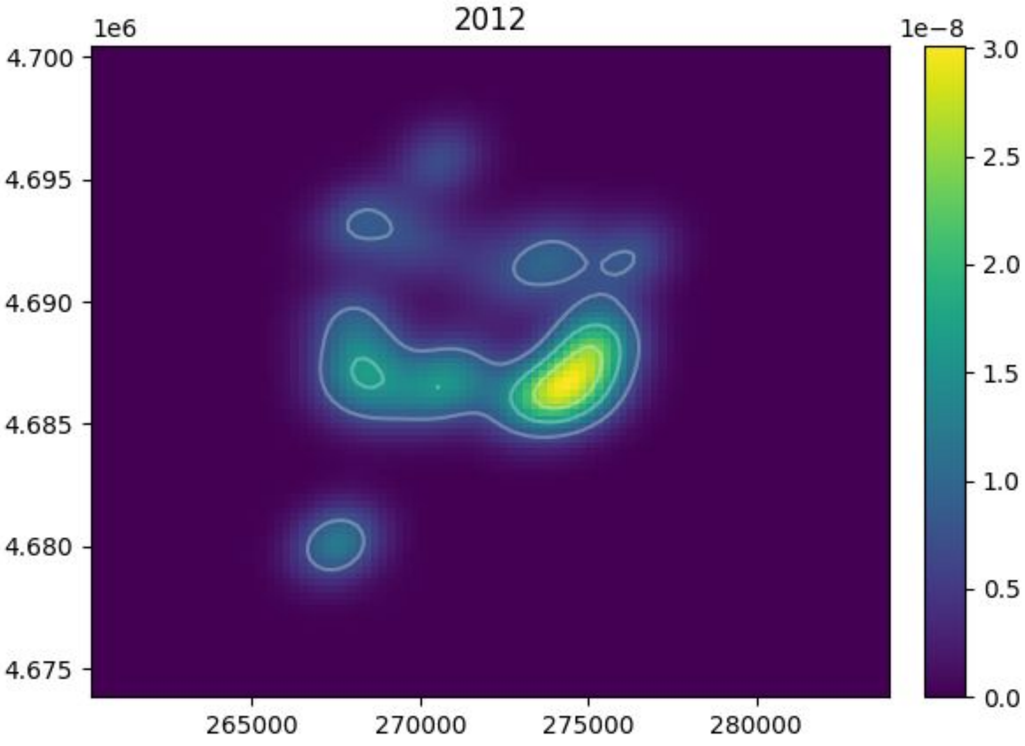
Temporal Shift by Year - 2012

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



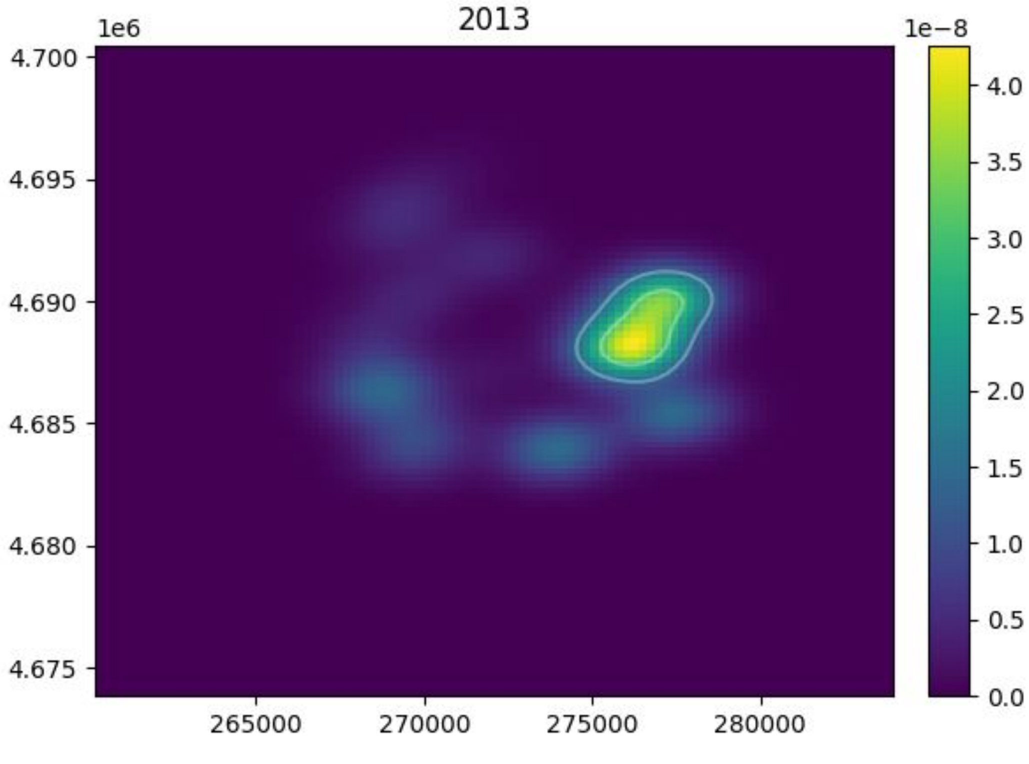
Temporal Shift by Year - 2013

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



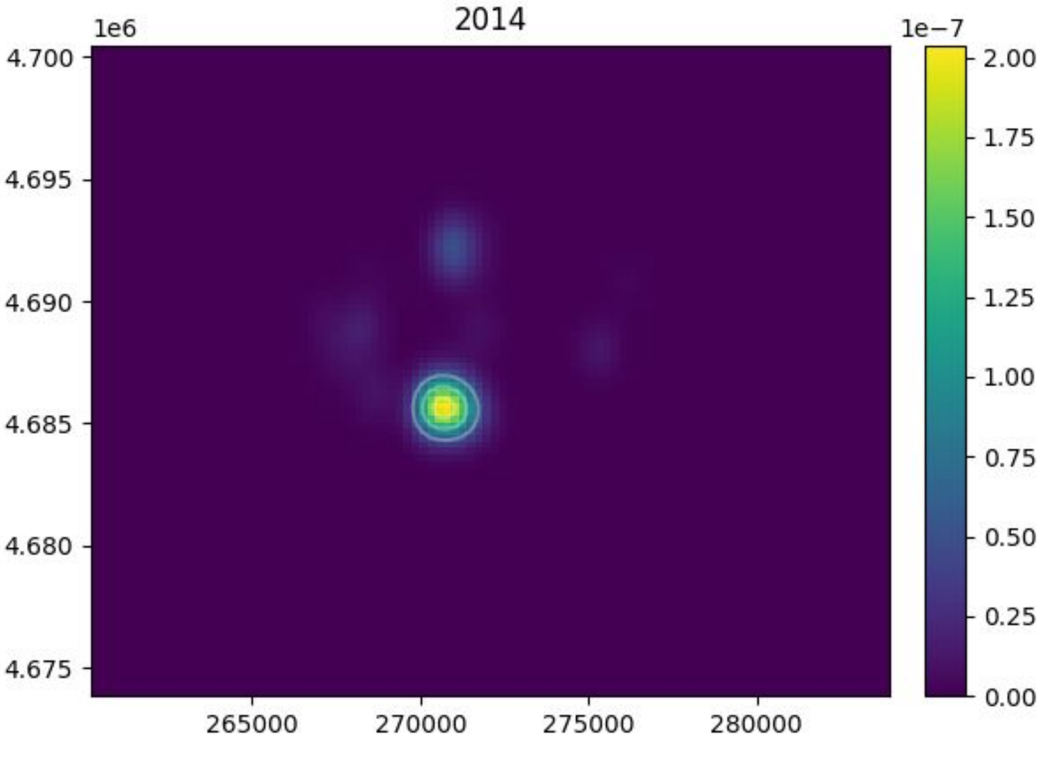
Temporal Shift by Year - 2014

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



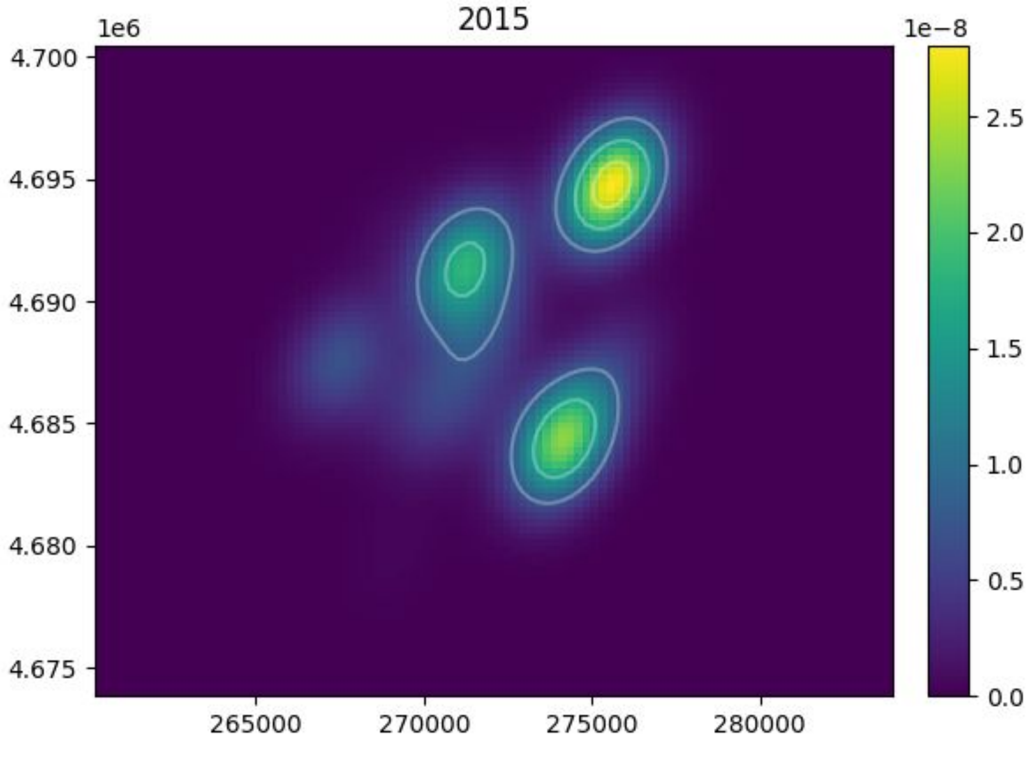
Temporal Shift by Year - 2015

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



Temporal Shift by Year - 2016

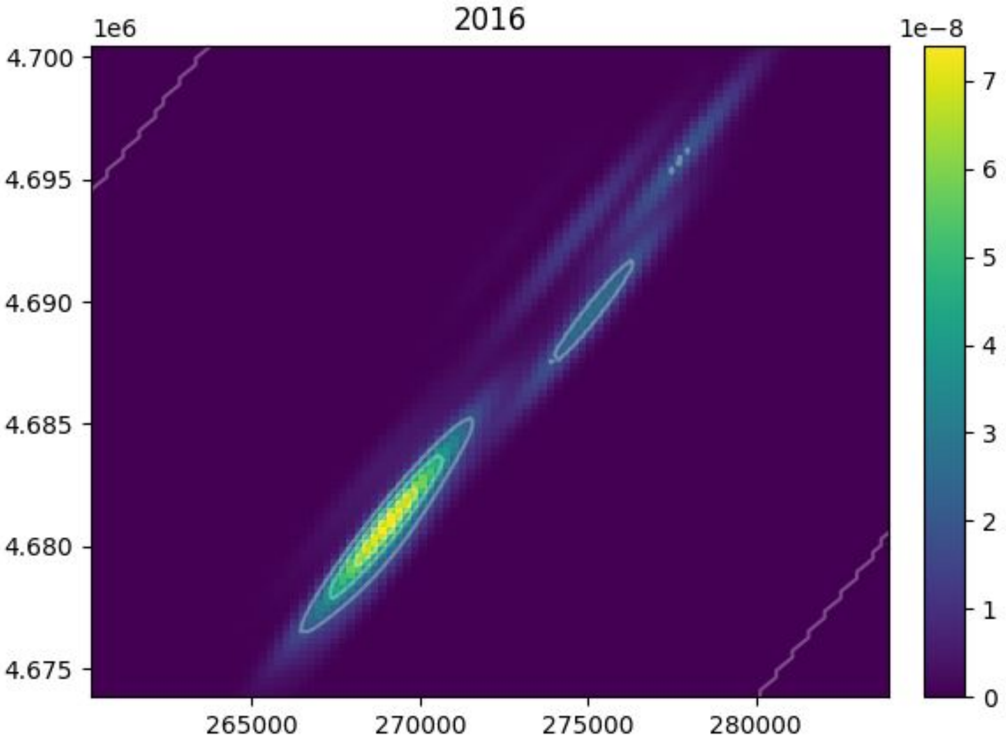
Overview

Motivation

Progress

- **Data Analysis**
- Model Construction

Next Steps



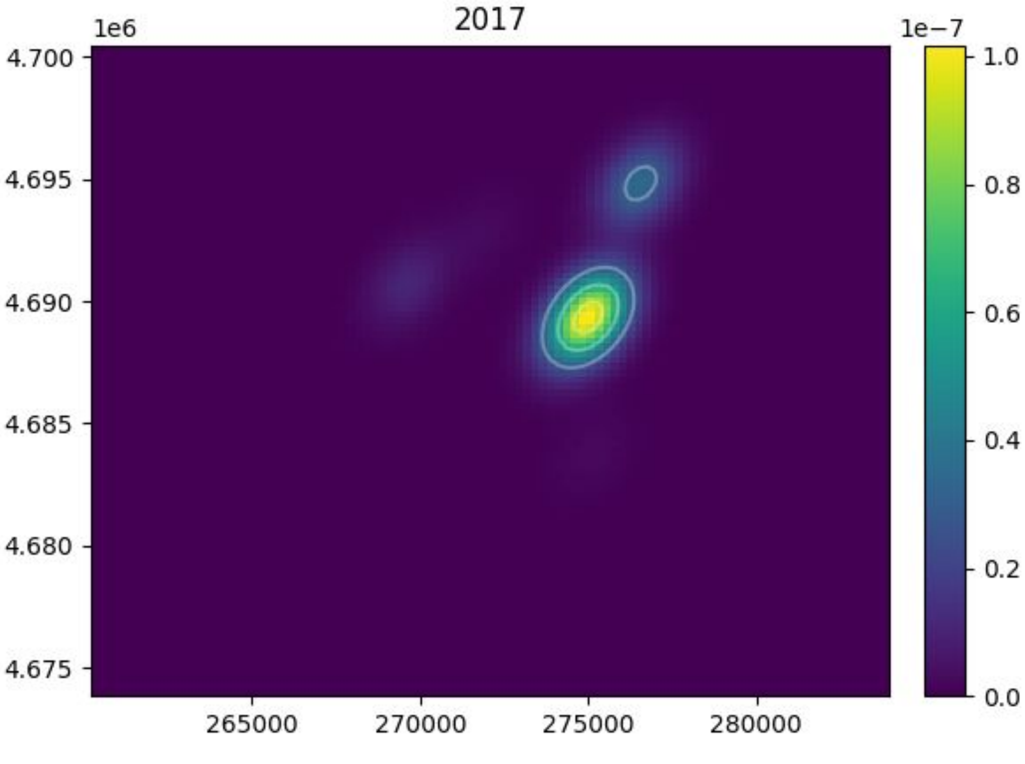
Temporal Shift by Year - 2017

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



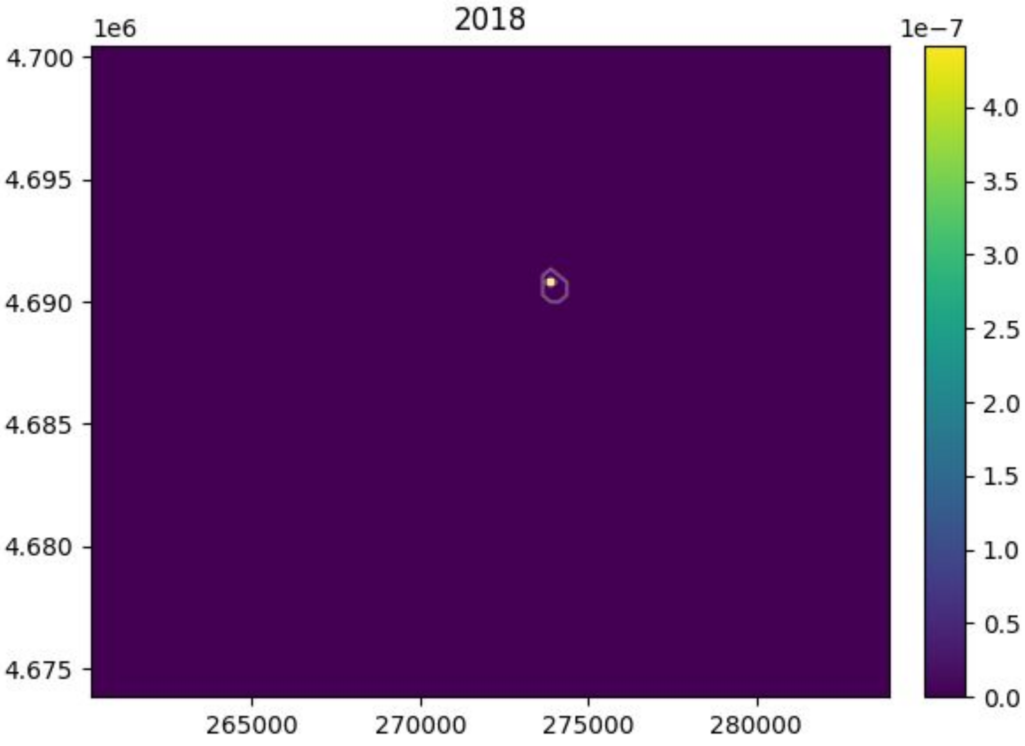
Temporal Shift by Year - 2018

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



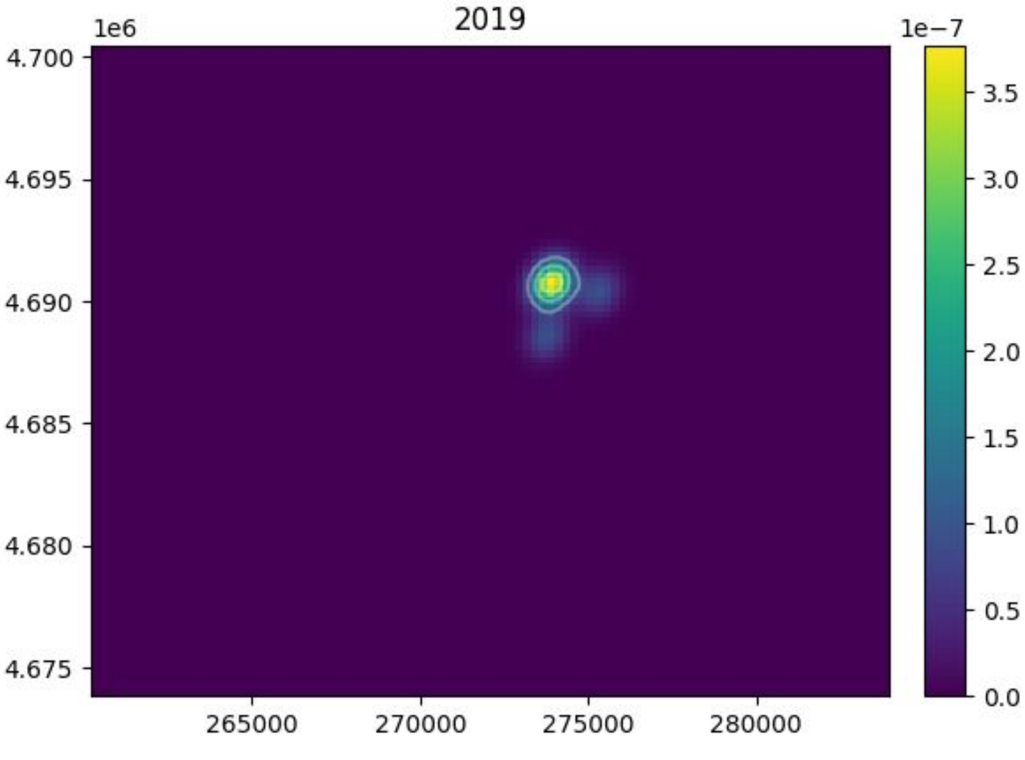
Temporal Shift by Year - 2019

Overview

Motivation

Progress
- **Data Analysis**
- Model Construction

Next Steps



Model Construction

Constructing a Model for Our Data

- Important quantities to model:

Model Construction

Constructing a Model for Our Data

- Important quantities to model:
  - The **number of points** in the grid

Model Construction

Constructing a Model for Our Data

- Important quantities to model:
    - The **number of points** in the grid
    - The **spread of the points** within the grid

Model Construction

Constructing a Model for Our Data

- Important quantities to model:
    - The **number of points** in the grid
    - The **spread of the points** within the grid

- We must create a model that **fits both of these parameters.**

Spatial Poisson Point Process

- We model the **number of points** λ as a *Poisson random variable.*
  - This distribution "counts" the number of events of the process in our grid.

Spatial Poisson Point Process

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

- We model the **number of points** λ as a *Poisson random variable.*
  - This distribution "counts" the number of events of the process in our grid.
- The probability of k events occurring is given by

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!}$$

- The parameter λ is both the **mean** and **variance.**

Non-Homogeneous Point Process

- For a *homogeneous* point process, λ remains constant along the entire grid.

Non-Homogeneous Point Process

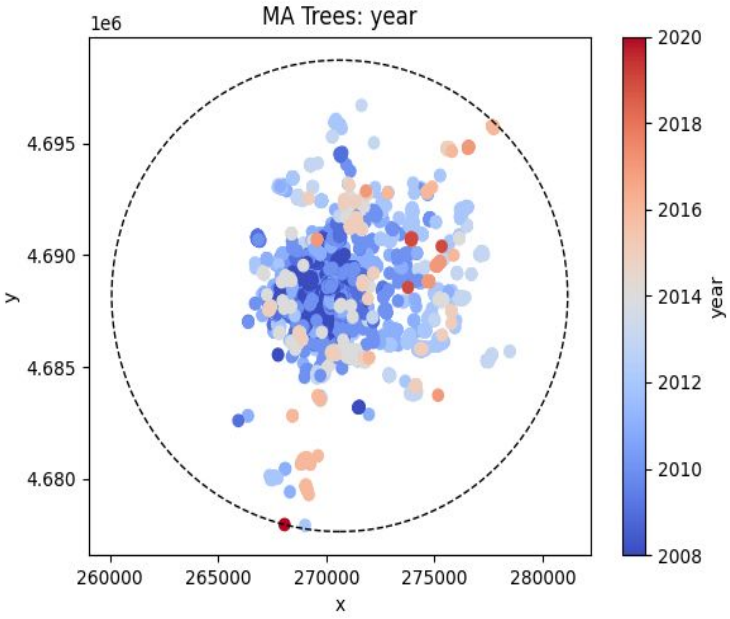
Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

- For a *homogeneous* point process, λ remains constant along the entire grid.
- However, our data is **non-homogeneous:** $\lambda(x)$ depends on the point x in the grid.



Estimating the Values of $\lambda(x)$

Overview

Motivation

Progress

- Data Analysis
- **Model Construction**

Next Steps

- We estimate the values of $\lambda(x)$ by first counting the points in our data set:

Estimating the Values of $\lambda(x)$

- We estimate the values of $\lambda(x)$ by first counting the points in our data set:

  1. Draw circles $C_1, \ldots, C_N$ around point clusters.

Estimating the Values of $\lambda(x)$

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

- We estimate the values of $\lambda(x)$ by first counting the points in our data set:

1. Draw circles $C_1, \ldots, C_N$ around point clusters.
2. Find the number of data points $\hat{\lambda}(C_i)$ in each circle.

Estimating the Values of $\lambda(x)$

- We estimate the values of $\lambda(x)$ by first counting the points in our data set:

    1. Draw circles $C_1, \ldots, C_N$ around point clusters.
    2. Find the number of data points $\hat{\lambda}(C_i)$ in each circle.
    3. Divide by the area of the circle:

$$\hat{\lambda}_i = \frac{\hat{\lambda}(C_i)}{area(C_i)}$$

Estimating the Values of $\lambda(x)$

- This gives us an estimate of the λ value within each circle C_i.
    - We want a formula for each point $x \in C_i$.

Estimating the Values of $\lambda(x)$

- This gives us an estimate of the λ value within each circle C_i.
  - We want a formula for each point $x \in C_i$.
- To do this, we use a formula that decreases with distance:

$$\hat{\lambda}_i(x) = \hat{\lambda}_i e^{-dr^2(x)}$$

  - $r(x)$ is the distance from the circle's center
  - *d > 0* is the decay rate (Default: 0.01)

Bayesian Inference

- Inside each circle C_i, we have some M_i data points

$$X_i = \{x_1, \ldots, x_{M_i}\}$$

which we interpret as random samples.

Bayesian Inference

- Inside each circle C_i , we have some M_i data points

$$X_i = \{x_1, \ldots, x_{M_i}\}$$

which we interpret as random samples.

- For each circle, we aim to calculate the *posterior distribution*

$$p_i(\lambda_i | x)$$

Bayesian Inference

- If we make an initial assumption about the distribution of the (random) values $\lambda = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N$, then we can use *Bayesian Inference* to approximate the true distribution:

$$p_i(\lambda_i|x) = p_i(x|\lambda_i) f_i(\lambda_i)$$

Bayesian Inference

- If we make an initial assumption about the distribution of the (random) values $\lambda = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N$, then we can use *Bayesian Inference* to approximate the true distribution:

$$p_i(\lambda_i|x) = p_i(x|\lambda_i)f_i(\lambda_i)$$

- We assume a *prior distribution* $f_i(\lambda_i)$ for these parameters.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Bayesian Inference

- We use a **Gamma distribution** as our prior:

$$\lambda_i \sim f_i(\lambda_i) = Gamma(a_{\lambda_i}, b_{\lambda_i})$$

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Bayesian Inference

- We use a **Gamma distribution** as our prior:

$$\lambda_i \sim f_i(\lambda_i) = Gamma(a_{\lambda_i}, b_{\lambda_i})$$

- A gamma distribution describes the spread of points with a *shape* and a *rate* a_{λ_i} . b_{λ_i}
  - These can be thought of as **additional model parameters**.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Bayesian Inference

- The *likelihood distribution* can be calculated from our data points:

$$p_i(x|\lambda_i) = \prod_{j=1}^{M} p_i(x_j|\lambda_i(x_j))$$

Bayesian Inference

- The *likelihood distribution* can be calculated from our data points:

$$p_i(x|\lambda_i) = \prod_{j=1}^{M} p_i(x_j|\lambda_i(x_j))$$

- Each of these functions is given by the Poisson distribution.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Bayesian Inference

- The prior distribution can be estimated from our (sampled) points:

$$f_i(\lambda_i) = \prod_{j=1}^{M} f_i(\lambda_i(x_j))$$

Bayesian Inference

- The prior distribution can be estimated from our (sampled) points:

$$f_i(\lambda_i) = \prod_{j=1}^{M} f_i(\lambda_i(x_j))$$

  - Recall that $f_i(\lambda_i)$ is a Gamma distribution.

Sampling from a Posterior

Markov Chain Monte Carlo (MCMC) Model

- From the paper "Bayesian Nonparametric Nonhomogeneous Poisson Process..." (arXiv:1907.03186):

$$
\begin{aligned}
&k \sim p(\cdot), \text{where } p(\cdot) \text{ is a p.m.f on } \{1, 2, \ldots\} \\
&\lambda_r \overset{\text{ind}}{\sim} \text{Gamma}(a, b), \quad r = 1, \ldots, k, \\
&P(z_i = j \mid \pi, k) = \pi_j, \quad j = 1, \ldots, k, \ i = 1, \ldots, n, \\
&\pi \mid k \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \\
&N(A_i) \mid z, \lambda, k \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_{z_i}), \quad i = 1, \ldots, n,
\end{aligned}
$$

Estimating Parameter Space

Maximum Likelihood Estimate

- Our goal is to **maximize the posterior distribution**

$$\tilde{\lambda}_i = \underset{\lambda_i}{\operatorname{argmax}}\{p_i(\lambda_i|x)\}$$

Estimating Parameter Space

Maximum Likelihood Estimate

- Our goal is to **maximize the posterior distribution**

$$\tilde{\lambda}_i = \underset{\lambda_i}{\operatorname{argmax}}\{p_i(\lambda_i|x)\}$$

$$= \underset{\lambda_i}{\operatorname{argmin}}\{-\log p_i(\lambda_i|x)\}$$

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Estimating Parameter Space

Maximum Likelihood Estimate

- Using Bayesian Inference,

$$\tilde{\lambda}_i = \underset{\lambda_i}{\operatorname{argmin}}\{-\log p_i(x|\lambda_i) - \log f_i(\lambda_i)\}$$

Estimating Parameter Space

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Maximum Likelihood Estimate

- Using Bayesian Inference,

$$\tilde{\lambda}_i = \underset{\lambda_i}{\operatorname{argmin}}\{-\log p_i(x|\lambda_i) - \log f_i(\lambda_i)\}$$

$$= \underset{\lambda_i}{\operatorname{argmin}}\left\{-\sum_{j=1}^{M}\Big(\log p_i(x_j|\lambda_i(x_j)) + \log f_i(\lambda_i(x_j))\Big)\right\}$$

Algorithm

Step 1

- Take the set of data points and draw circles $C_1, \ldots, C_N$ around the **point clusters**.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Algorithm

Step 1

- Take the set of data points and draw circles $C_1, \ldots, C_N$ around the **point clusters**.
  - Clusters should be at the center of the circle and the data points become more sparse as distance increases from the center.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Algorithm

Step 1

- Take the set of data points and draw circles $C_1, \ldots, C_N$ around the **point clusters**.
    - Clusters should be at the center of the circle and the data points become more sparse as distance increases from the center.
- In our case, this was done by hand.
    - There is no ideal algorithm to capture all clusters.

Algorithm

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Step 2

- Estimate the $\lambda(x)$ values for our data points inside the circle:

$$\hat{\lambda}_i(x_j) = \hat{\lambda}_i e^{-dr^2(x_j)}$$

Algorithm

Step 2

- Estimate the $\lambda(x)$ values for our data points inside the circle:

$$\hat{\lambda}_i(x_j) = \hat{\lambda}_i e^{-dr^2(x_j)}$$

- Recall that $\hat{\lambda}_i$ is the estimated value for the circle C_i.

Algorithm

Step 2

- Estimate the $\lambda(x)$ values for our data points inside the circle:

$$\hat{\lambda}_i(x_j) = \hat{\lambda}_i e^{-dr^2(x_j)}$$

- Recall that $\hat{\lambda}_i$ is the estimated value for the circle C_i.
- This gives us an initial **expected** number of points around x_j.

Algorithm

Step 3

- Next, use the values $\hat{\lambda}_i(x_j)$ to generate the (initial) **number of observed events** k_j for each point x_j.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Algorithm

Step 3

- Next, use the values $\hat{\lambda}_i(x_j)$ to generate the (initial) **number of observed events** k_j for each point x_j.
  - This is done by creating a Poisson random variable with mean $\hat{\lambda}_i(x_j)$.

Algorithm

Step 3

- Next, use the values $\hat{\lambda}_i(x_j)$ to generate the (initial) **number of observed events** k_j for each point x_j.
    - This is done by creating a Poisson random variable with mean $\hat{\lambda}_i(x_j)$.
    - It gives an idea of the number of data points we should find around x_j.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Algorithm

Step 4

- Use our calculations for the MLE to complete the optimization

$$\tilde{\lambda}_i = \underset{\lambda_i}{\operatorname{argmax}}\{p_i(\lambda_i|x)\}$$

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Algorithm

Step 4

- Use our calculations for the MLE to complete the optimization

$$\tilde{\lambda}_i = \underset{\lambda_i}{\operatorname{argmax}}\{p_i(\lambda_i|x)\}$$

  - This also gives us the values $a_{\tilde{\lambda}_i}, b_{\tilde{\lambda}_i}$ for the prior (Gamma) distribution.

Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Algorithm

Step 5

- To create simulations, start with a **uniformly distributed** set of data points $Y = \{y_1, \ldots, y_S\}$ inside our circle.

Algorithm

Step 5

- To create simulations, start with a **uniformly distributed** set of data points $Y = \{y_1, \dots, y_S\}$ inside our circle.
    - The number of points should be **proportional to the optimal value** $\tilde{\lambda}_i$.

Algorithm

Step 5

- To create simulations, start with a **uniformly distributed** set of data points $Y = \{y_1, \ldots, y_S\}$ inside our circle.
    - The number of points should be **proportional to the optimal value** $\tilde{\lambda}_i$.
- Use the optimal values $a_{\tilde{\lambda}_i}, b_{\tilde{\lambda}_i}$ for the Gamma distribution to generate samples from our prior distribution

$$\lambda_k \sim Gamma(a_{\tilde{\lambda}_i}, b_{\tilde{\lambda}_i})$$

Algorithm

Step 5

- These new λ_k samples are used to calculate the distance-adjusted $\lambda_k(y_k)$ values for our new points.

Algorithm

Step 5

- These new λ_k samples are used to calculate the distance-adjusted $\lambda_k(y_k)$ values for our new points.
    - **Remove** the new points y_k whose lambda value $\lambda_k(y_k)$ is below some tolerance (default: 0.01).

Algorithm

Step 5

- These new λ_k samples are used to calculate the distance-adjusted $\lambda_k(y_k)$ values for our new points.
    - **Remove** the new points y_k whose lambda value $\lambda_k(y_k)$ is below some tolerance (default: 0.01).
- The remaining points will (approximately) live in the distribution of our data.

Applying the model to the data

Overview

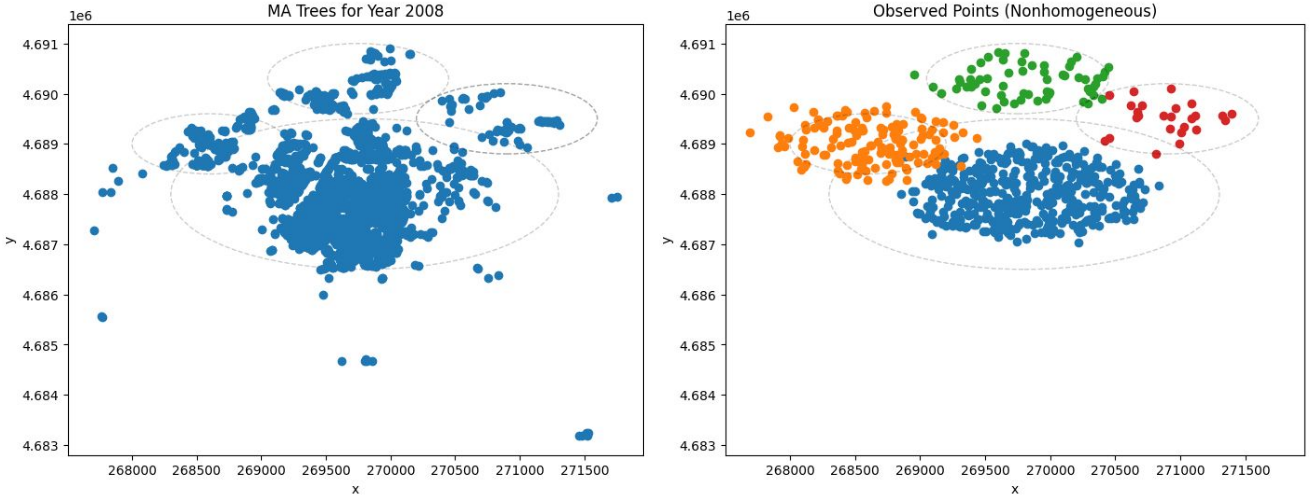
Motivation

Progress

- Data Analysis
- **Model Construction**

Next Steps

Results: 2008 Trees (*decay rate = 1e-5*)



Applying the model to the data

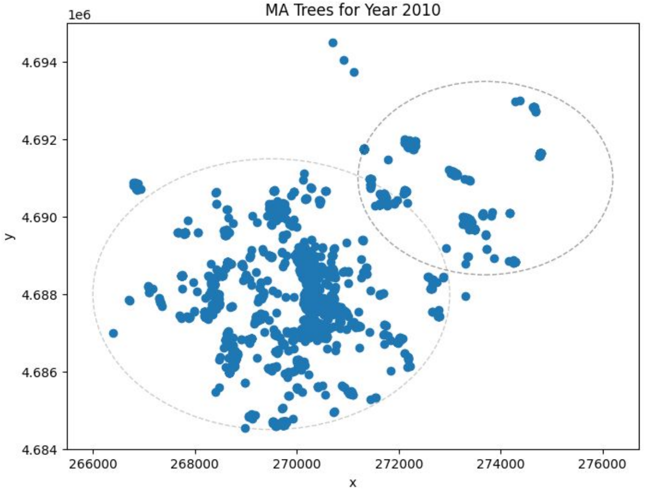
Overview

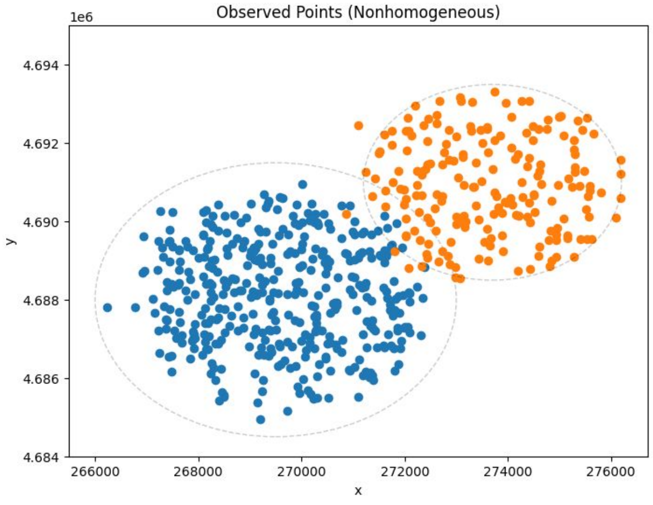
Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Results: 2010 Trees (*decay rate = 1e-6*)





Applying the model to the data

Overview

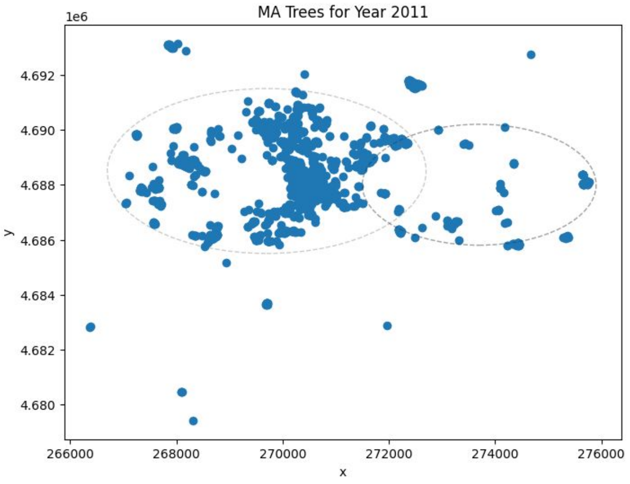
Motivation

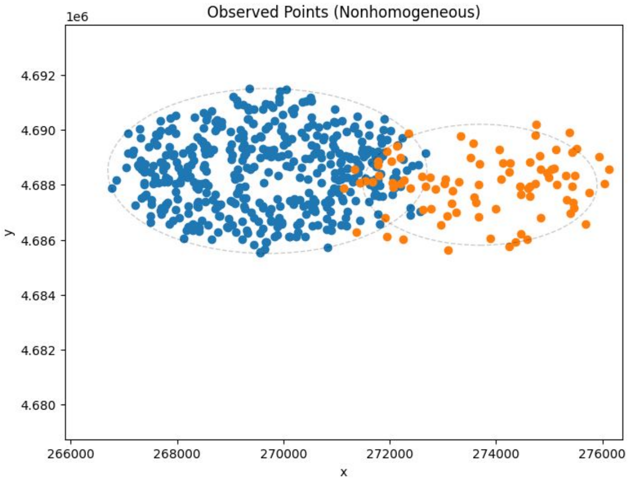
Progress

- Data Analysis
- **Model Construction**

Next Steps

Results: 2011 Trees (*decay rate = 1e-6*)





Applying the model to the data

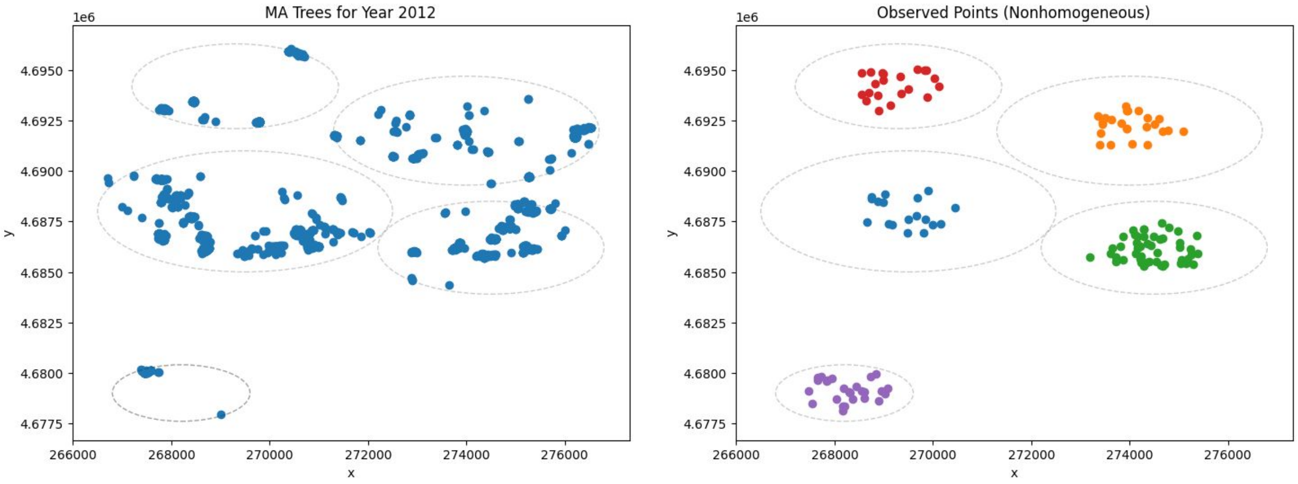
Overview

Motivation

Progress
- Data Analysis
- **Model Construction**

Next Steps

Results: 2012 Trees (*decay rate = 5e-6*)



Model Pitfalls

Ambiguity in the Algorithm

- There is no set way to identify the point clusters and create circles around them.

Model Pitfalls

Ambiguity in the Algorithm

- There is no set way to identify the point clusters and create circles around them.
- The decay rate *d* may need to be different for each circle.

Model Pitfalls

Overview

Motivation

Progress

- Data Analysis
- **Model Construction**

Next Steps

Ambiguity in the Algorithm

- There is no set way to identify the point clusters and create circles around them.
- The decay rate *d* may need to be different for each circle.
- The number of simulated points in each region can be better defined. We currently use

$$S = \hat{\lambda}_i$$

Next Steps

Overview

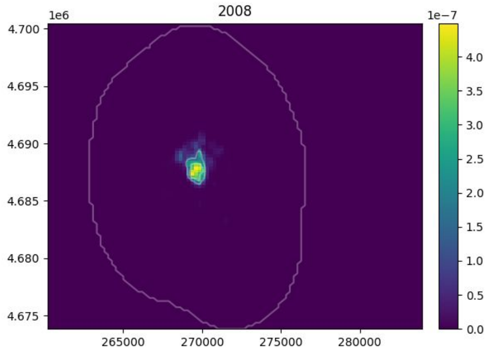
Motivation

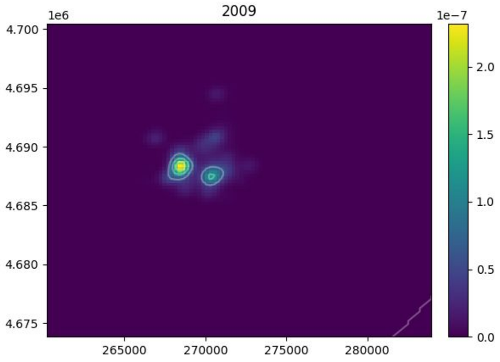
Progress
- Data Analysis
- Model Construction

Next Steps

KDE for Cluster Identification

- To identify the best places for circles, we can use the 2-D KDE estimation heat maps.





Next Steps

Optimal Decay Rate and Number of Points

- We would like to find the best values to use for the decay rate d and the number of points S to generate for our simulations.

Next Steps

Optimal Decay Rate and Number of Points

- We would like to find the best values to use for the decay rate d and the number of points S to generate for our simulations.
    - A larger decay rate d will mean more points are removed as the data moves further from the center.

Next Steps

Optimal Decay Rate and Number of Points

- We would like to find the best values to use for the decay rate *d* and the number of points *S* to generate for our simulations.
    - A larger decay rate *d* will mean more points are removed as the data moves further from the center.

- We want to find the **ideal balance** between these values.

Thank you!

Next Steps

Better Model

- From the paper "Bayesian Nonparametric Nonhomogeneous Poisson Process..." (arXiv:1907.03186):

$$
\begin{aligned}
&k \sim p(\cdot), \text{where } p(\cdot) \text{ is a p.m.f on } \{1, 2, \ldots\} \\
&\lambda_r \overset{\text{ind}}{\sim} \text{Gamma}(a, b), \quad r = 1, \ldots, k, \\
&P(z_i = j \mid \pi, k) = \pi_j, \quad j = 1, \ldots, k, \; i = 1, \ldots, n, \\
&\pi \mid k \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \\
&N(A_i) \mid z, \lambda, k \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_{z_i}), \quad i = 1, \ldots, n,
\end{aligned}
$$