Adaptive Non-linear Intensity Modeling in Spatial Point Processes with GAMs

Zahra Maleki

Max Planck Institute of Molecular Cell Biology and Genetics

September 26, 2024

Contents

What is the problem

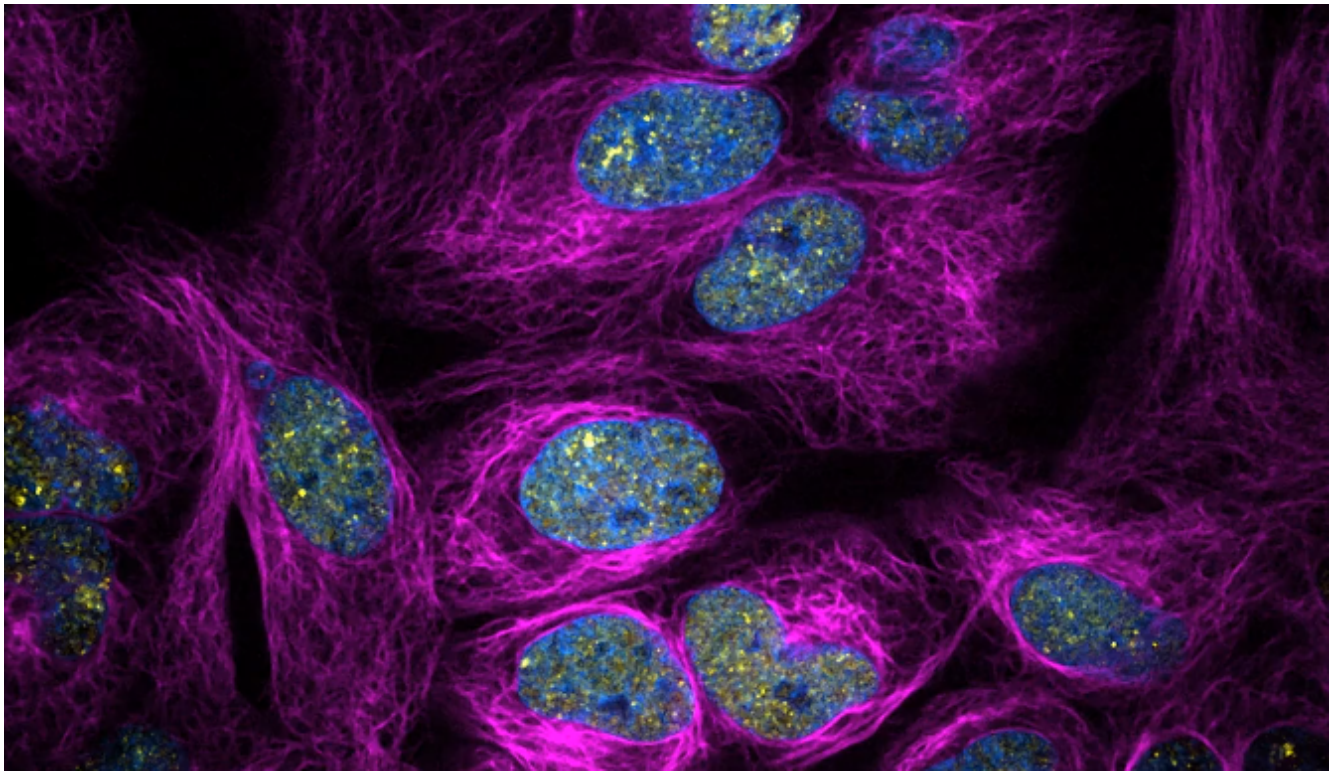

Figure: Bio-molecular condensates forming in cancer cells

In cancer, these condensates group molecules effect tumor growth, affecting how cells grow and survive.

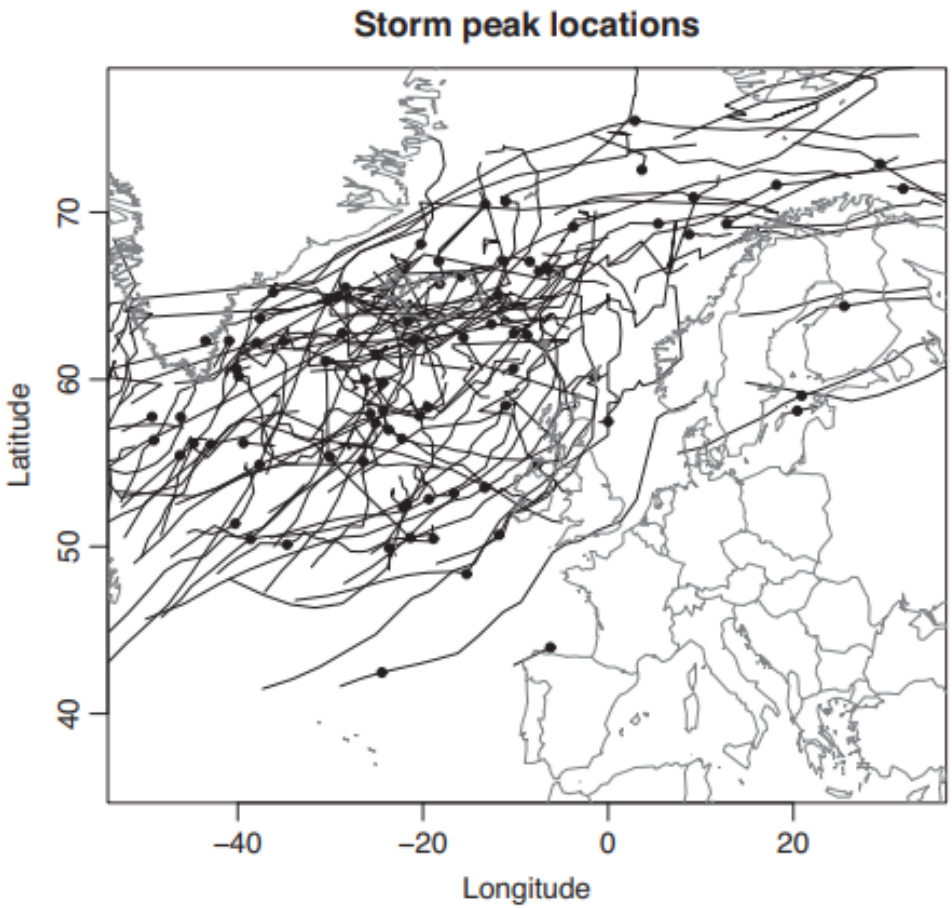

Figure: Storm peak locations across Europe

Understanding where storm peaks occur and how this might vary with time is critical for analyzing seasonal patterns and their influence.

Intensity and Density

The intensity of a spatial point process is the **expected number of events per unit area**.
As process is usually unknown, intensity cannot be directly measured. Instead, the density of the point pattern is used as an estimate of the intensity.

$$\hat{\lambda} = \frac{n}{a}$$

Generalized Linear Models (GLMs)

GLMs assume a linear relationship between the predictors and the response variable

$$g(\mu) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Biological structures often exhibit complex, non-linear relationships.

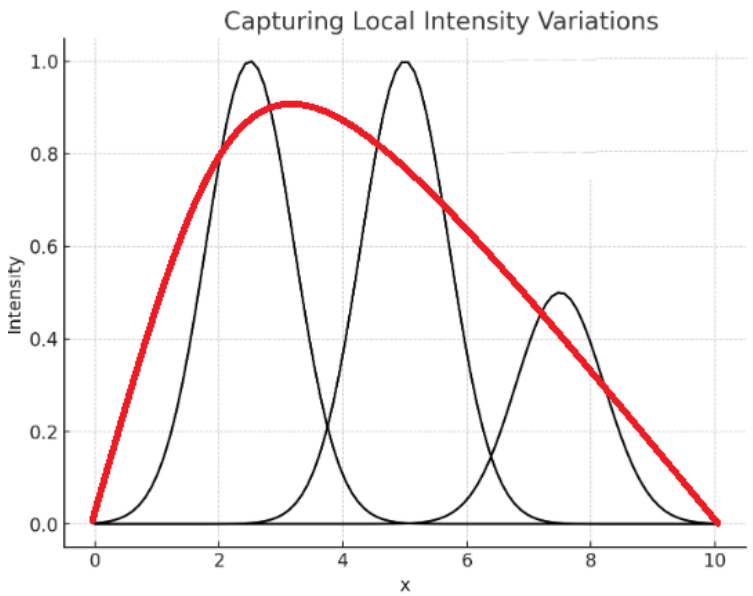

Figure: Capturing Local Intensity Variations

What I learned

Generalized Additive Models (GAMs)

GAMs allow for the automatic detection of non-linear effects.
The model relates a univariate response variable Y to predictor variables x_i and link function g . The expected value of Y is modeled as:

$$g[\mu(X)] = \alpha + f_1(X_1) + \cdots + f_p(X_p)$$

This structure allows for flexible, non-linear relationships between Y and the predictors.

Model Overview

- $g(\mu) = \log(\mu)$: Poisson count data.

The idea is to replace the unknown smooth functions in the model with basis expansions

$$f_j(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x_j)$$

The basis dimension K_j is selected to be sufficiently large

For B-splines of degree 2:

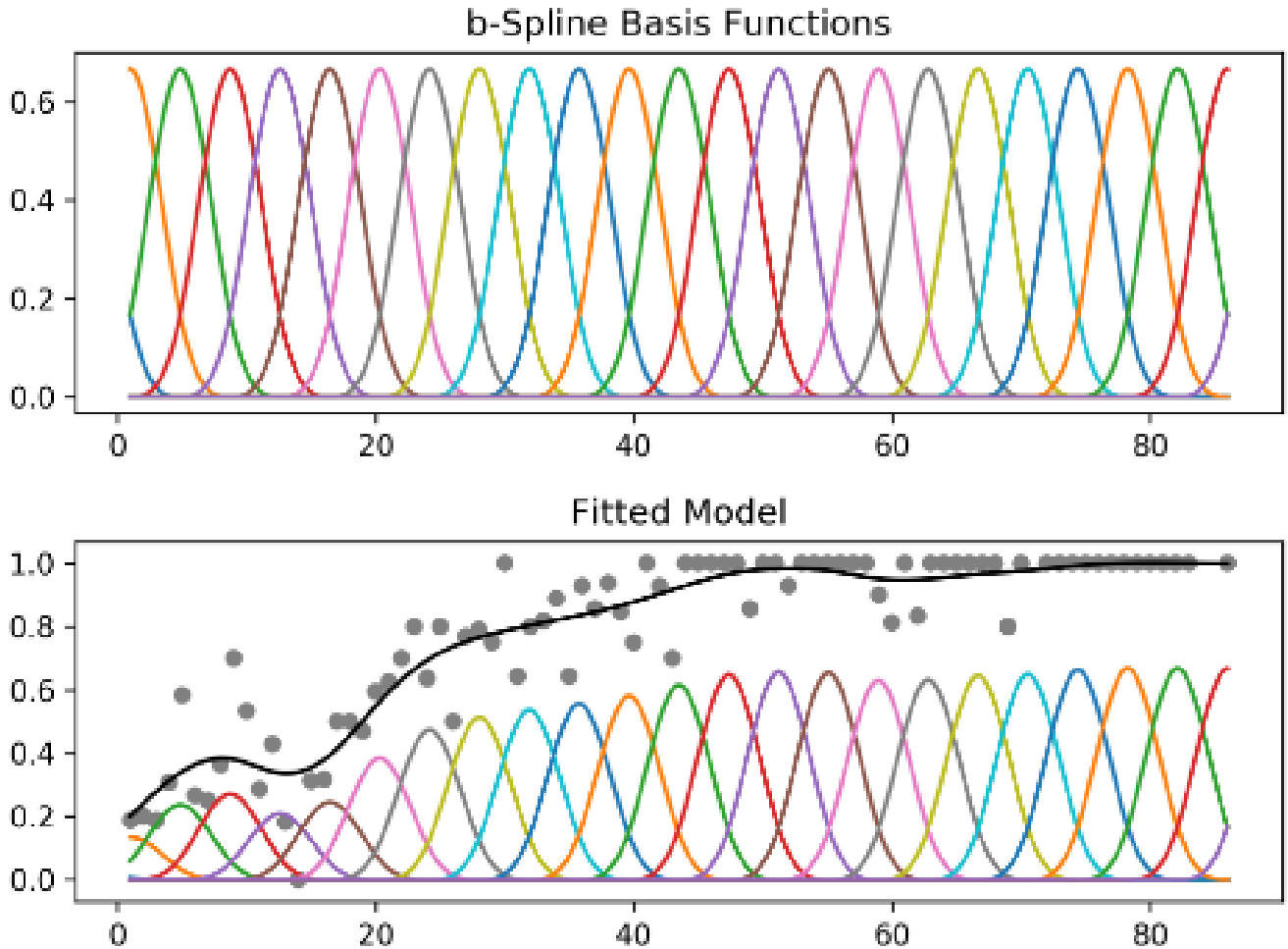

Figure: 1D B-spline Basis Functions

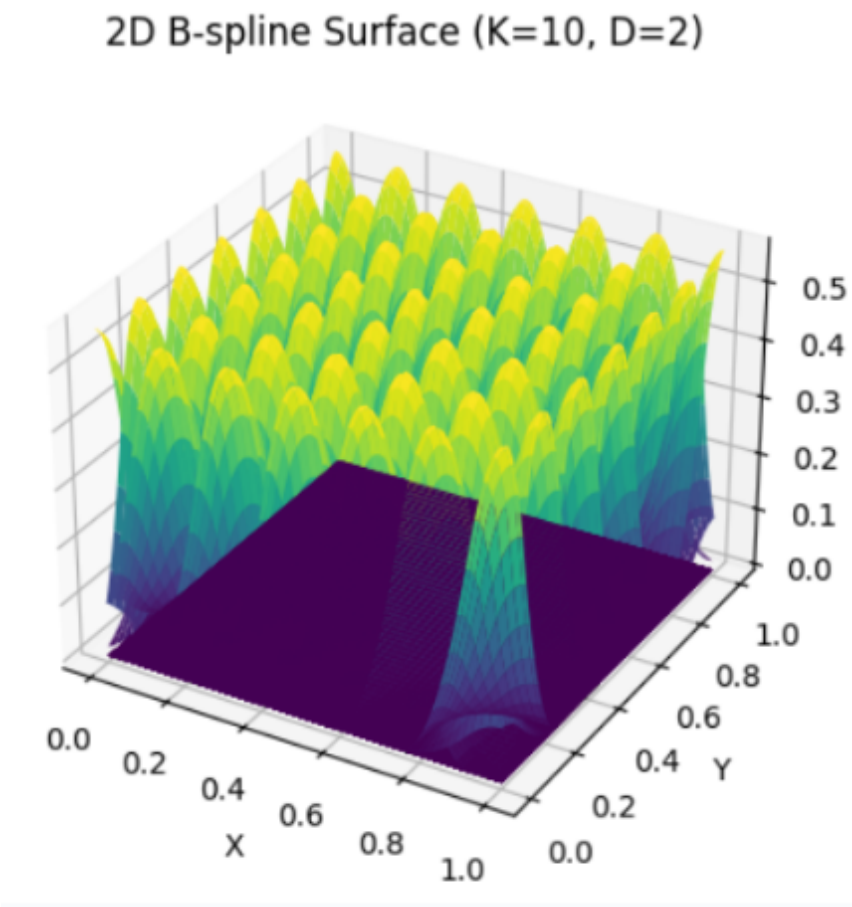

Figure: 2D B-spline Surface

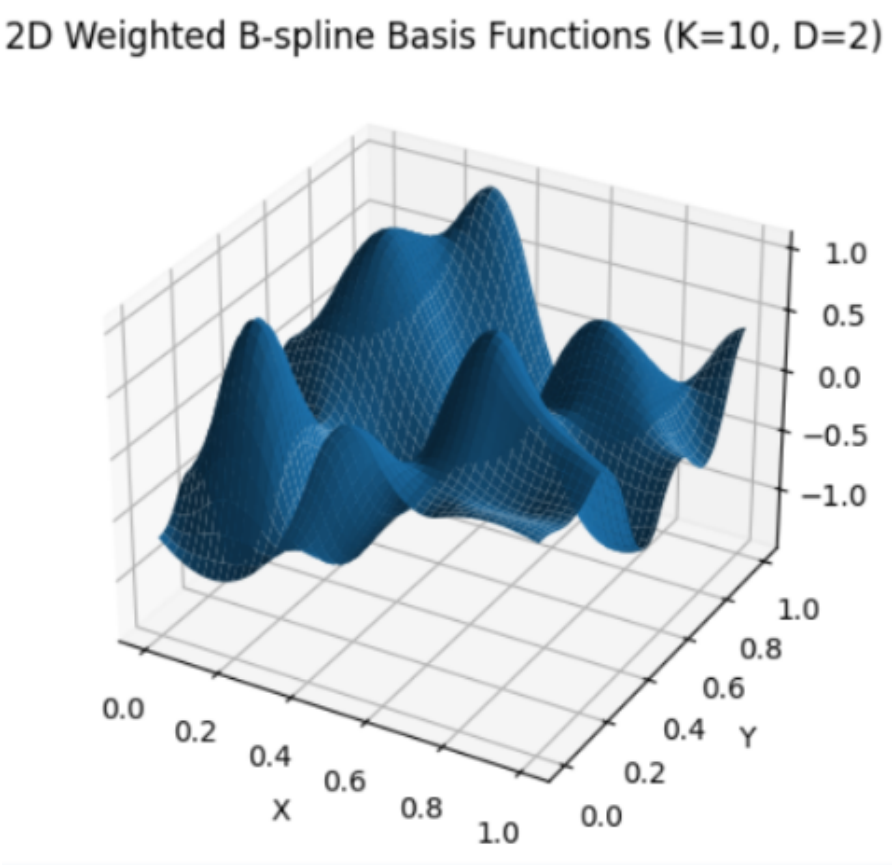

Figure: 2D Weighted B-spline Basis Functions

fitting the model

Due to the large number of basis functions K_j, the model can become over-parameterized.

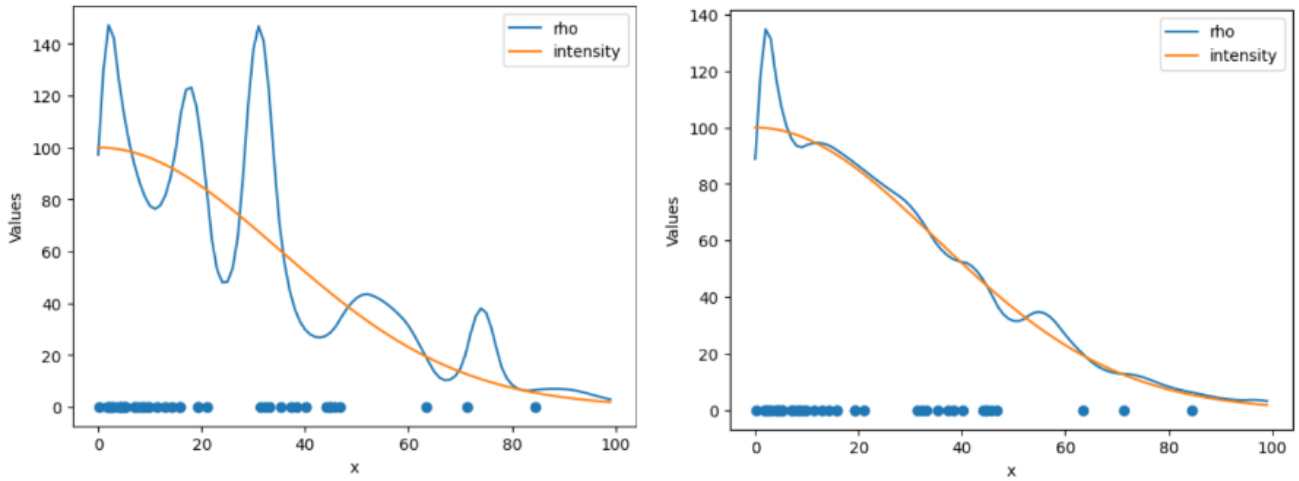

Figure: the effect of increasing K

Solution? a balance between model fit and smoothness

Penalties

The fitting procedure minimizes a penalized version of the deviance.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ D(\boldsymbol{\beta}) + \sum_j \lambda_j \int \left(f_j''(x)\right)^2 dx \right\}$$

The smoothing parameters λ_j control the trade-off between model fit and smoothness

When $\lambda_j \to \infty$, $f_j(x_j)$ becomes a straight line, leading to a linear relationship between the covariate and the response.

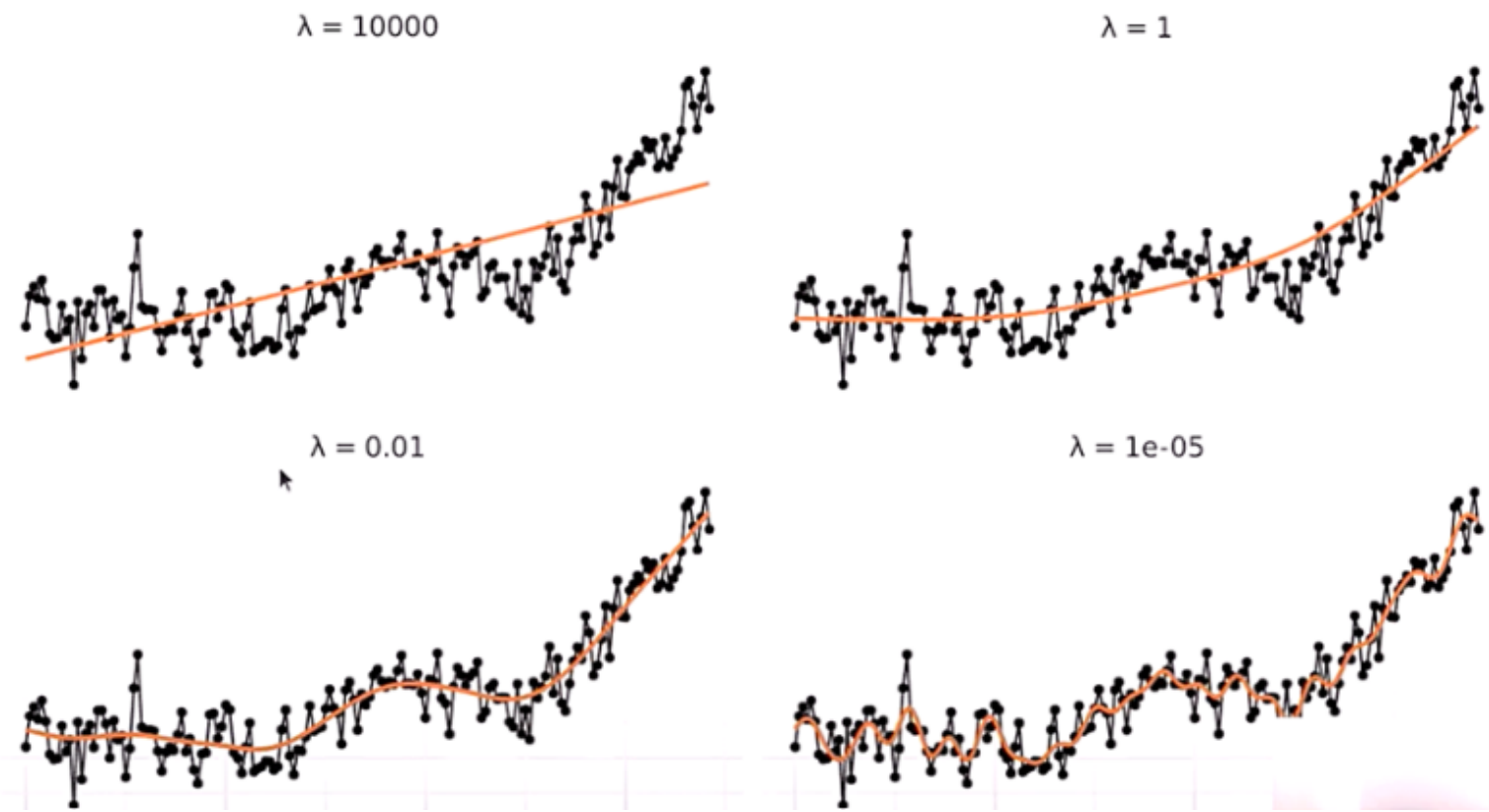

Figure: expressing the effect of λ for global temperature

Quadratic Form

the wiggliness penalties can be expressed as:

$$\int \left(f_j''(x)\right)^2 dx = \boldsymbol{\beta}^T S_j \boldsymbol{\beta}$$

S_j is The penalty matrix is computed from the basis.

$$\{S_j\}_{kl} = \int b_j''(t) b_k''(t) dt$$

where b_j and b_k are the basis functions, and b_j'' denotes the second derivative of the basis function b_j. Final objective function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ D(\boldsymbol{\beta}) + \sum_j \lambda_j \boldsymbol{\beta}^T S_j \boldsymbol{\beta} \right\}$$

Poisson point process

A Poisson point process describes random events occurring independently over space or time, with no interaction between points.

homogeneous

Events occur at a constant rate across the space.

in-homogeneous

Allows the event rate to vary across space or time.

The mean of the Poisson-distributed events:

$$\Lambda(V;\boldsymbol{\theta},\mathrm{z}(V)) = \int_V \lambda(\mathrm{v};\boldsymbol{\theta},\mathrm{z}(\mathrm{v}))d\mathrm{v}$$

The likelihood

Given a dataset of hazard events indexed by $i = 1, \ldots, n$, occurring on V, the likelihood of the Poisson process model is:

$$L(\boldsymbol{\theta}) = \exp\left(|V| - \Lambda(V; \boldsymbol{\theta}, \mathrm{z}(V))\right) \prod_{i=1}^{n} \lambda(\mathrm{v}_i; \boldsymbol{\theta}, \mathrm{z}(\mathrm{v}_i))$$

$$\ell_p(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) - \frac{1}{2} \sum_{k=1}^{K} \rho_k \boldsymbol{\beta}_k^\top \mathrm{S}_k \boldsymbol{\beta}_k$$

What I did

Generating the data

To generating in-homogeneous Poisson point process data:

1. sample points from a homogeneous Poisson process
2. apply a thinning algorithm (acc/reject sampling with $p(v) = \frac{\lambda(v)}{\lambda_{max}}$)
3. Points with higher intensities are more likely to be accepted

100 sets with the same underlying intensity function.

$$\lambda(x, y) = 100 \cdot \exp\left(-\frac{x^2 + y^2}{s^2}\right) \quad \text{where} \quad s = 0.5$$

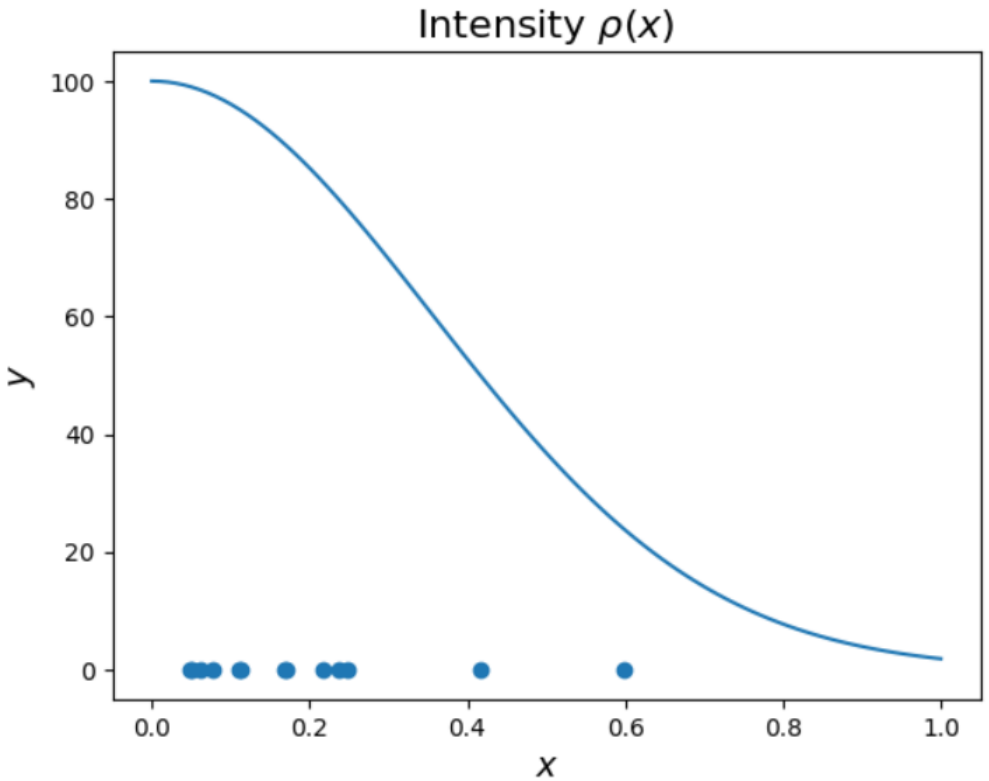

Figure: 1D Intensity Profile

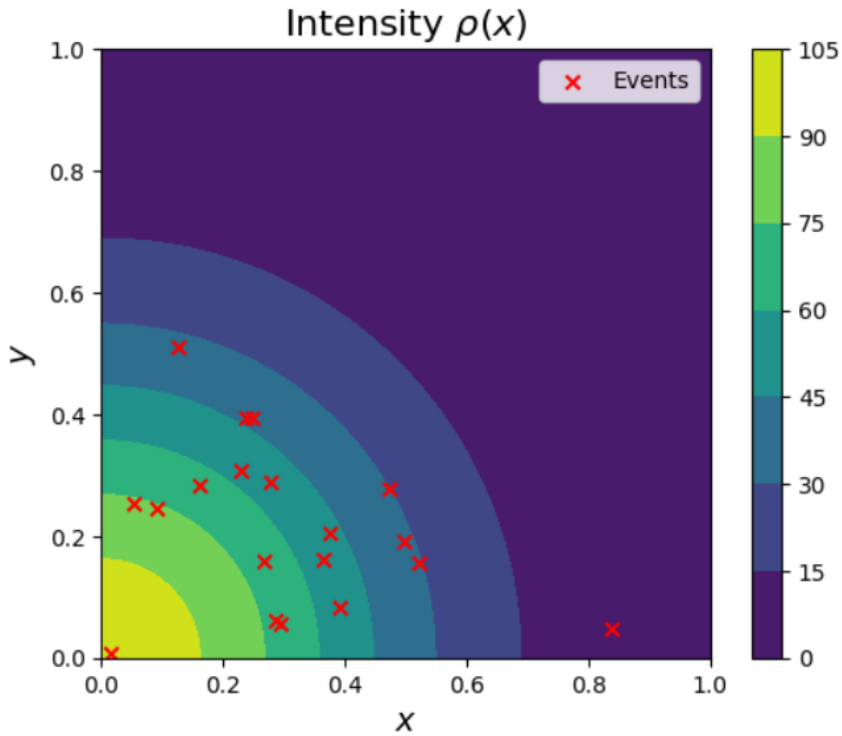

Figure: 2D Intensity Map with Events

GAM on generated 1D data

Applying the Generalized Additive Model (GAM) and learning the vector of coefficients β

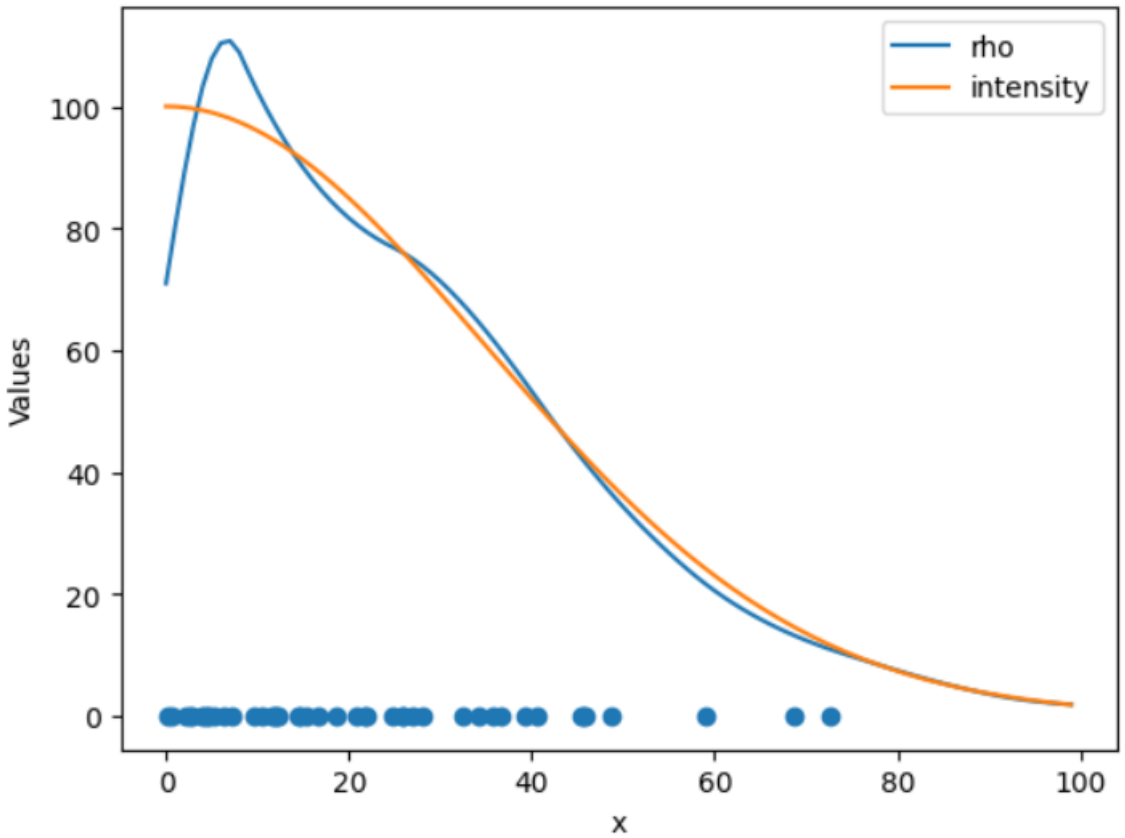

Figure: Comparison of learned intensity vs true intensity for a 1D domain.

- **Mean Squared Error (MSE)**: The average of the squared differences between the predicted and true intensity values.

$$\text{MSE} = 35.8116$$

- **R-squared (R^2) Score**: Tthe proportion of variance in the true intensity that is explained by the learned intensity.

$$R^2 = 0.9507$$

GAM learned intensity explains about 95% of the variance in the true intensity.

GAM on combined datasets

one Advantage of GAMs compared GLMsis their ability to capture the variance and local distribution.

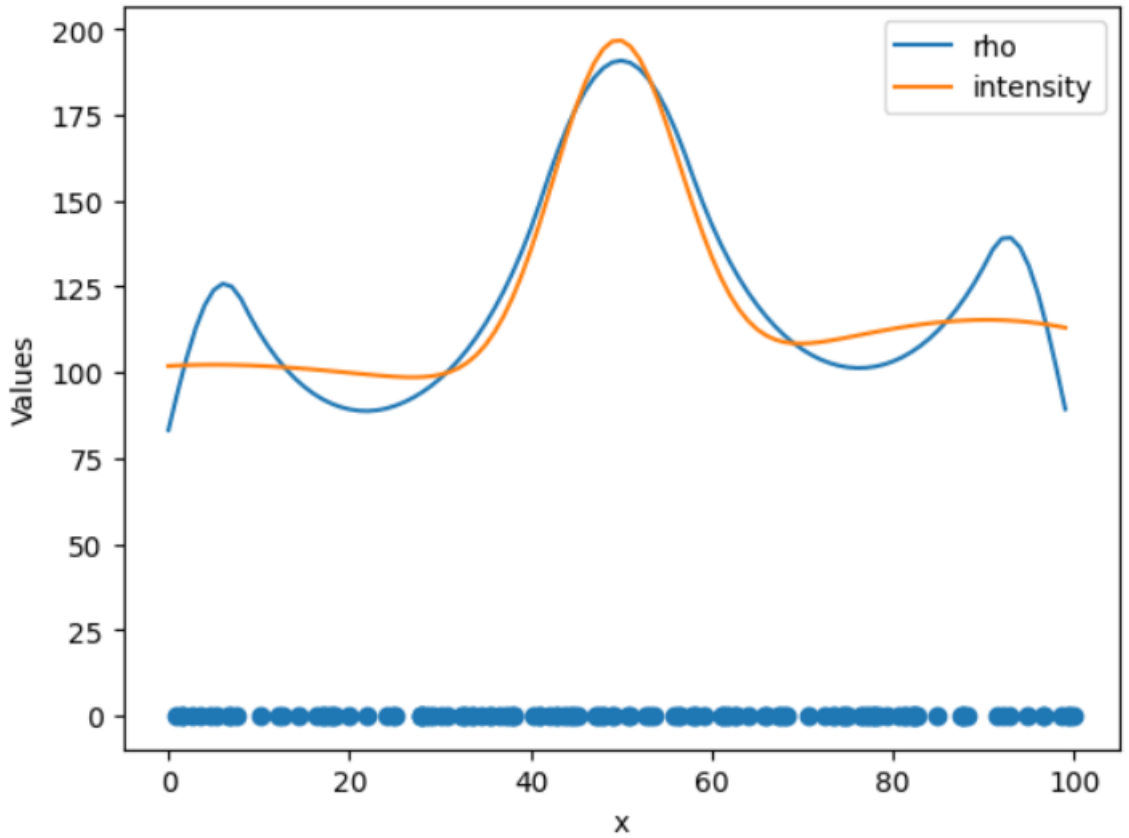

Figure: GAM results for a combination of datasets

Each dataset has a different variance and mean.

- Variance: [0.5, 0.1, 0.3]
- Mean: [0, 0.5, 1]

The results:

- MSE $= 99.33$
- $R^2 = 0.87$

GAMs are suited for understanding patterns that vary across different regions of the data.

Distribution Analysis Based on Distance

Understanding how the presence of a particular data point influences the surrounding spatial distribution.

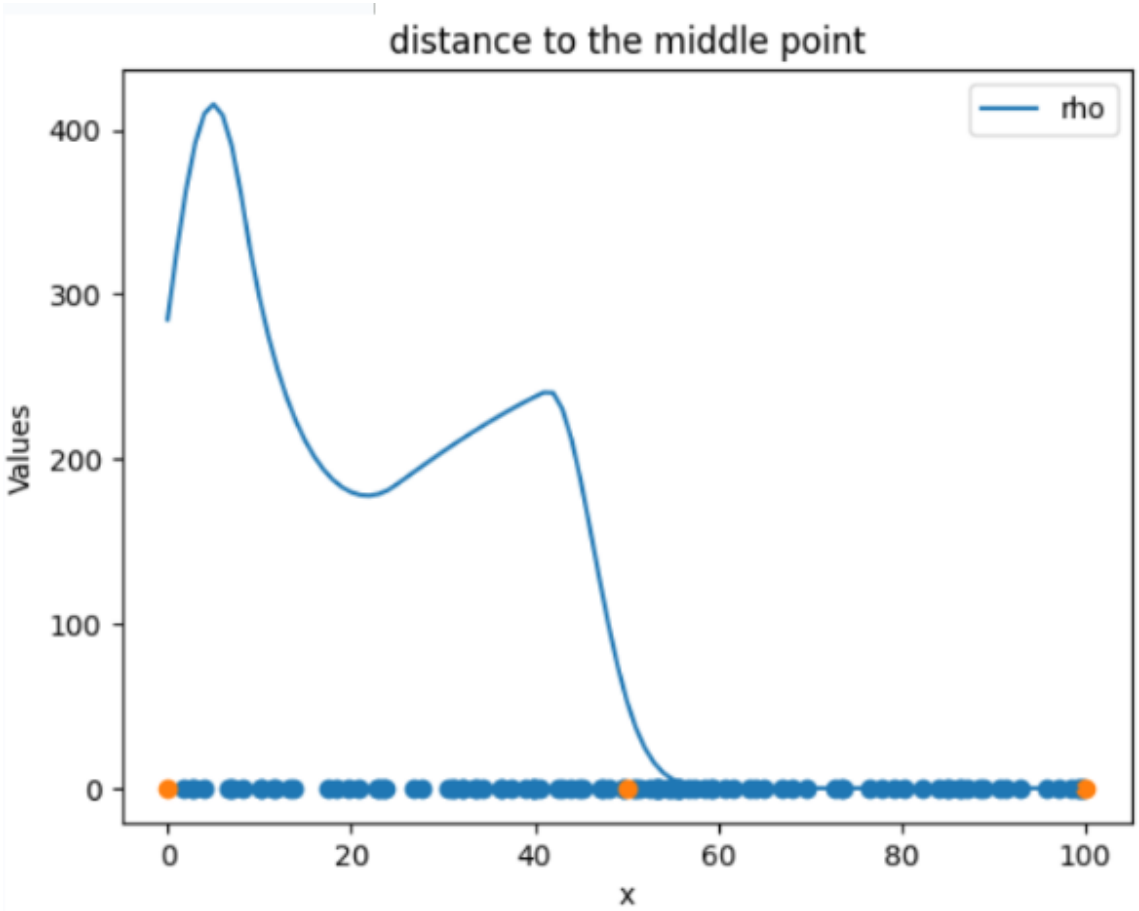

Figure: Intensity based on the distance to the middle point

2D Poisson point process

The evaluated intensity for the generated data:

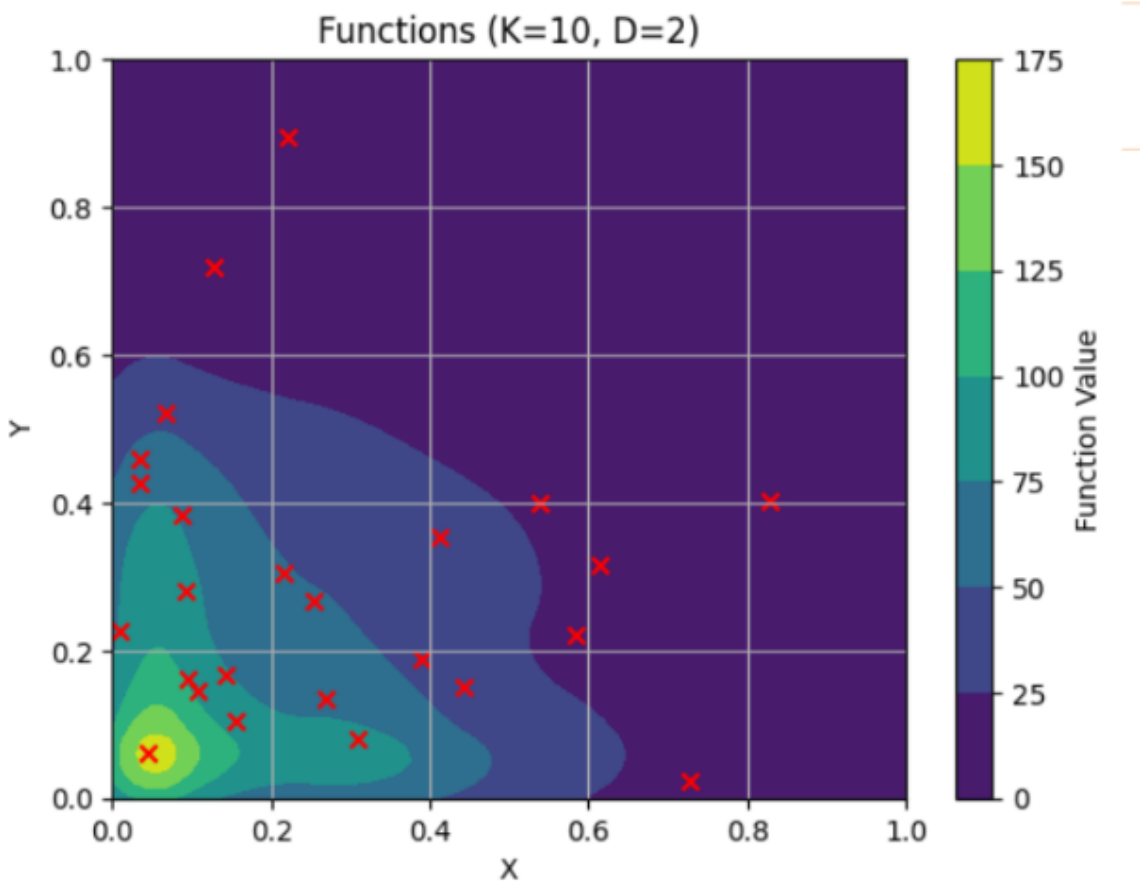

Figure: Evaluated intensity function for the generated data (K=10, D=2).

The performance of the model was evaluated using several metrics:

- MSE $= 38.599$
- $R^2 = 0.8798$
- Absolute Difference Between Integrals $= 0.2038$
- The calculated function integral $= 19.3079$

Smoothing parameter estimation

The original approach : maximizing the restricted log-likelihood

$$\ell_R(\rho) = \ell_p(\hat{\beta}) + \frac{1}{2}\log|\mathrm{S}^{-1}| - \frac{1}{2}\log|\mathrm{H}| + \text{constant}$$

requiring complex calculations is computationally expensive.
my approach: To use cross-validation. By plotting the NLL of the validation set against vs smoothing parameters and finding the **minimum**.

The Estimation

Splitting the data into two equal halves ($p = 0.5$). **P-Thinning Algorithm**
The NLL for the training set is defined as:

$$\log \ell_{\text{train}}(\beta) = \sum_{u \in X_{\text{train}}} \log \lambda(u) - p \int \lambda(u) du - \frac{1}{2} \sum_{k=1}^{K} \rho_k \boldsymbol{\beta}_k^\top \mathrm{S}_k \boldsymbol{\beta}_k$$

The NLL for the validation set is given by:

$$\log \ell_{\text{val}}(\beta) = \sum_{u \in X_{\text{val}}} \log \lambda(u) - (1-p) \int \lambda(u) du$$

Validation plot for 2D data

Using my approach, the plot shows the optimal smoothing parameter for 2D Poisson data is around 10^{-10}, balancing data fit and smoothness.

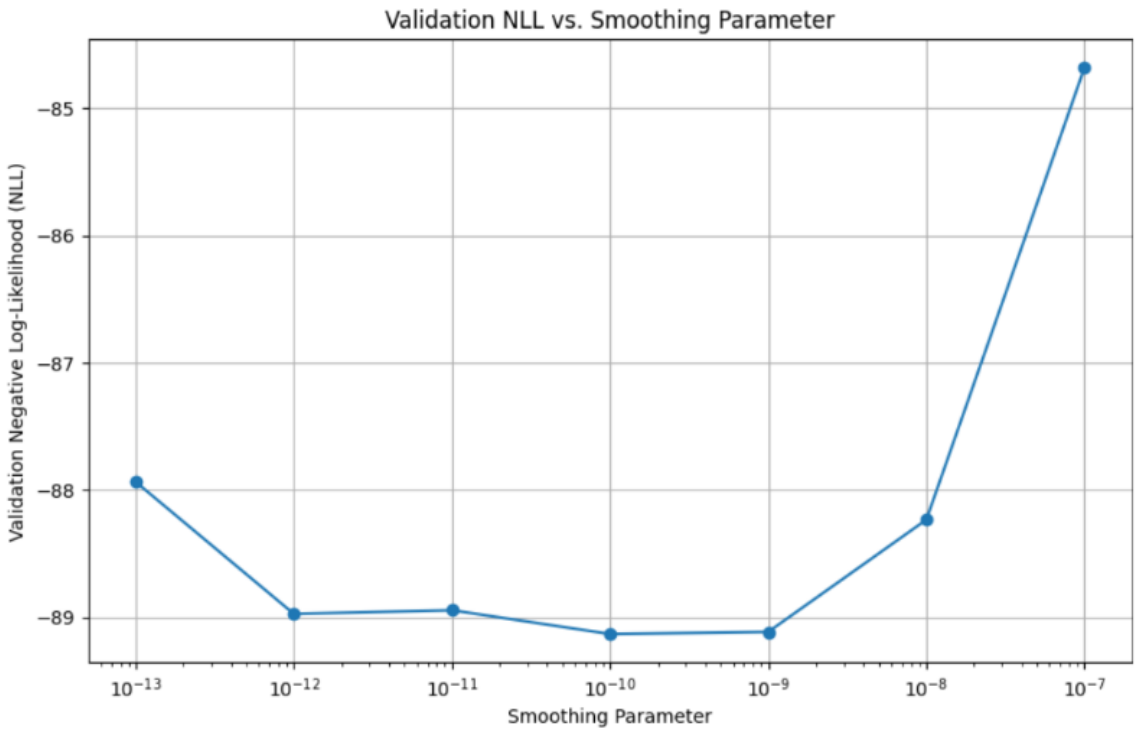

Figure: Validation NLL vs. Smoothing Parameter.

Generating Cox Point Processes

A Cox point process is an extension of the Poisson point process where the intensity function is a realization of a random process. useful for modeling spatial data with dependencies.

simulating a Log-Gaussian Cox Process (LGCP):

1. Simulate a Gaussian random field (GRF) over the spatial domain.
2. Exponentiate the GRF
3. Sample points from the LGCP using this.

LGCP = Log Intensity of a Gaussian Process

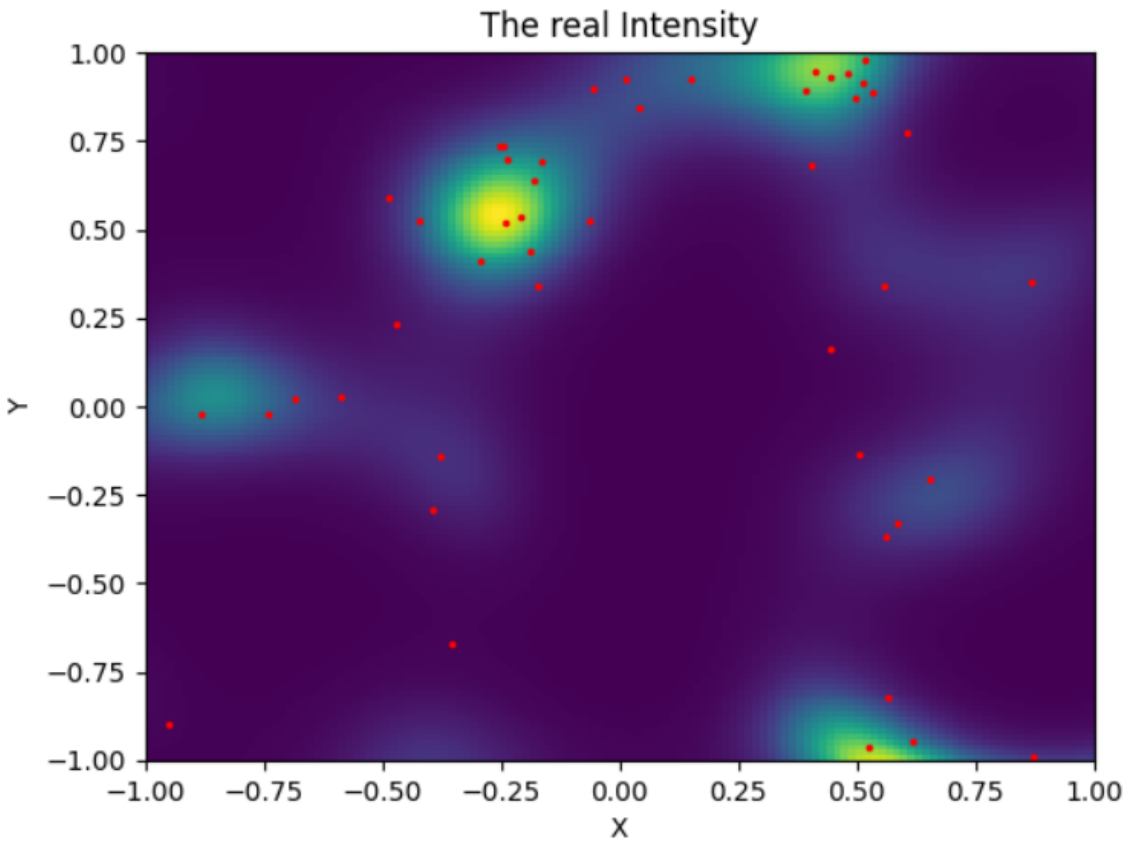

Figure: a sample of Cox generated data

The log-likelihood for the Cox process is the same as for the in-homogeneous Poisson point process.

GAM on Cox data

For one of the samples of this experiment, the best smoothing parameter was found to be smoothing_param $= 0.0088 \times 10^{-8}$.'
The performance of the model:

- MSE $= 1465.31$
- $R^2 = 0.4353$
- Difference in integrals $= 19.451$
- Calculated integral of the actual intensity $= 118.8932$

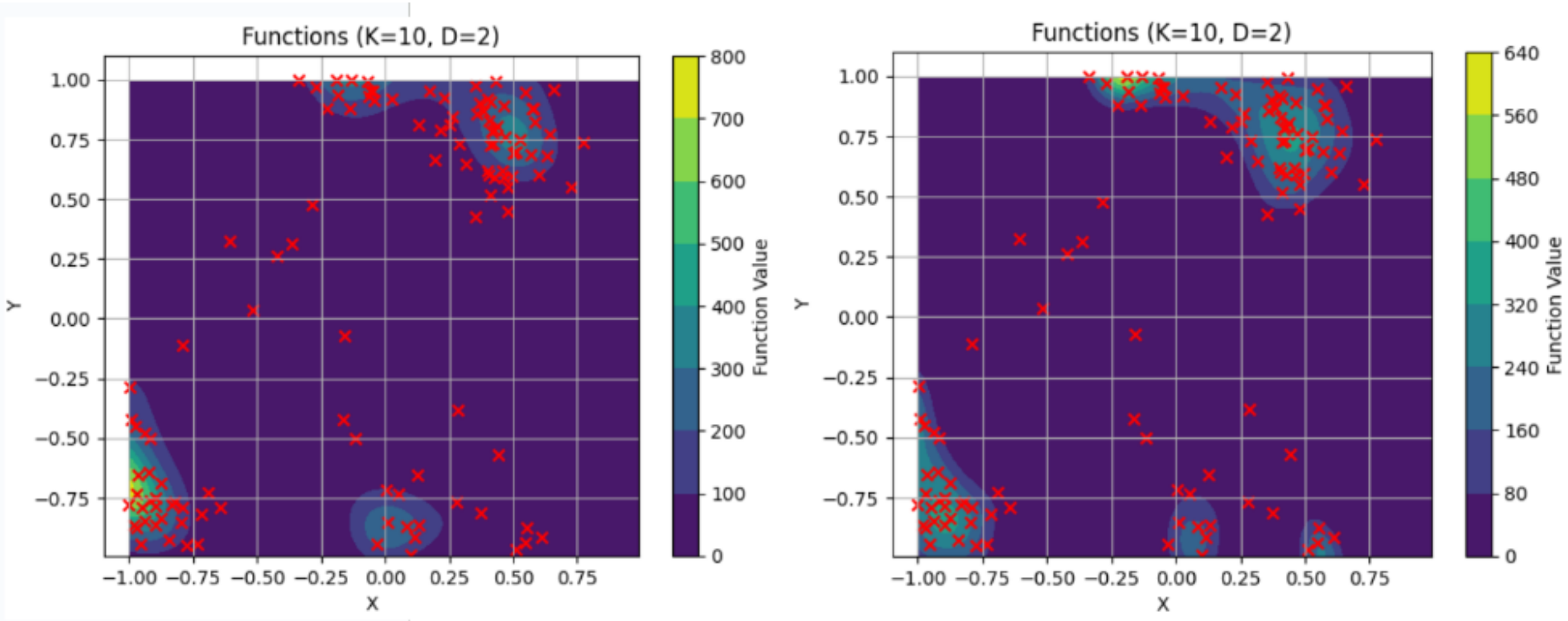

Figure: Left: Actual intensity. Right: Predicted intensity.

The effect of The smoothing Parameter

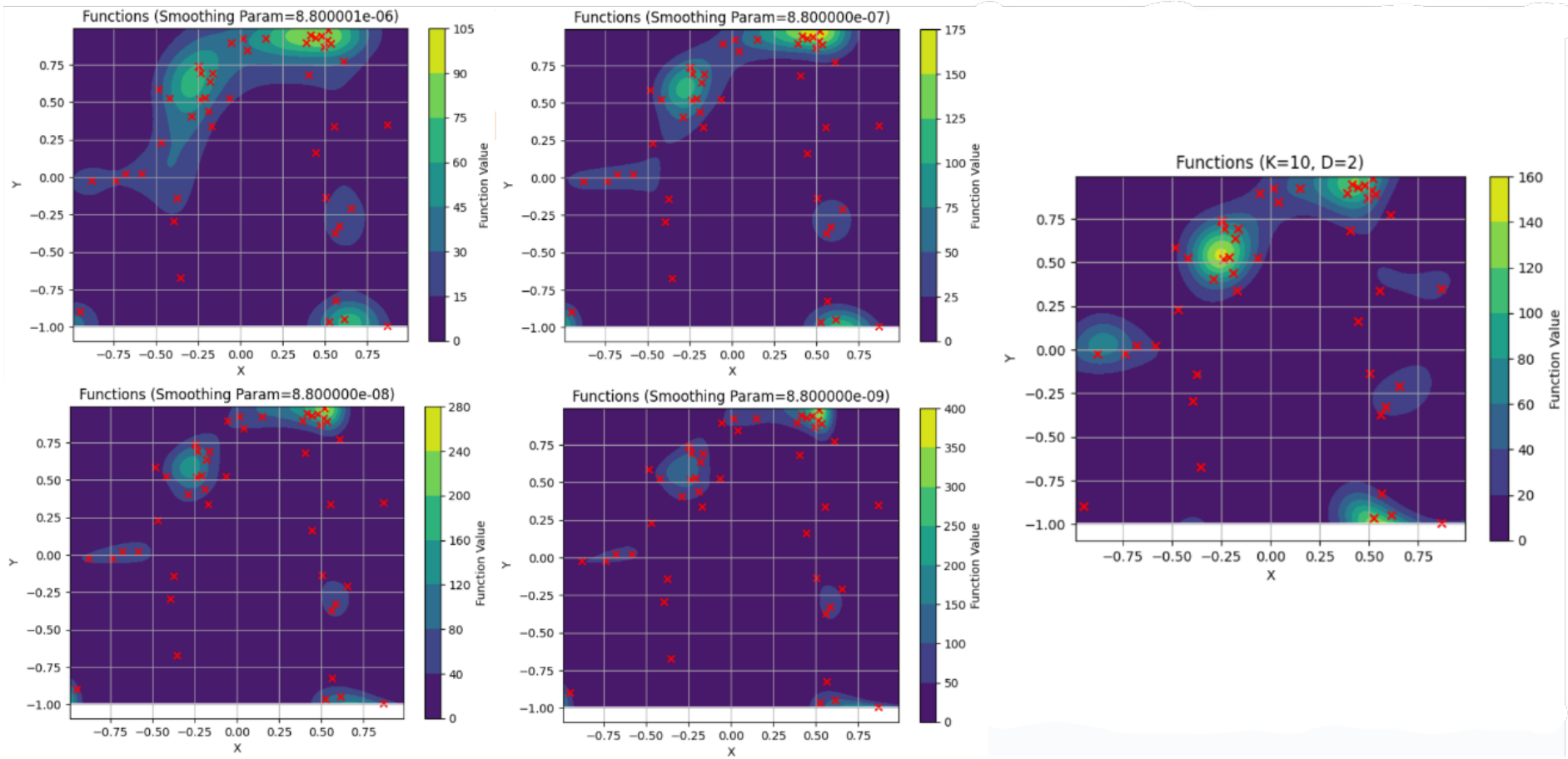

Figure: Effect of smoothing parameter on the learned intensity function. The right plot is the actual intensity.

Estimating the best smoothing Parameter

From the NLL plot, we observe that the minimum NLL occurs at around 10^{-6}, which suggests this is the optimal smoothing parameter.

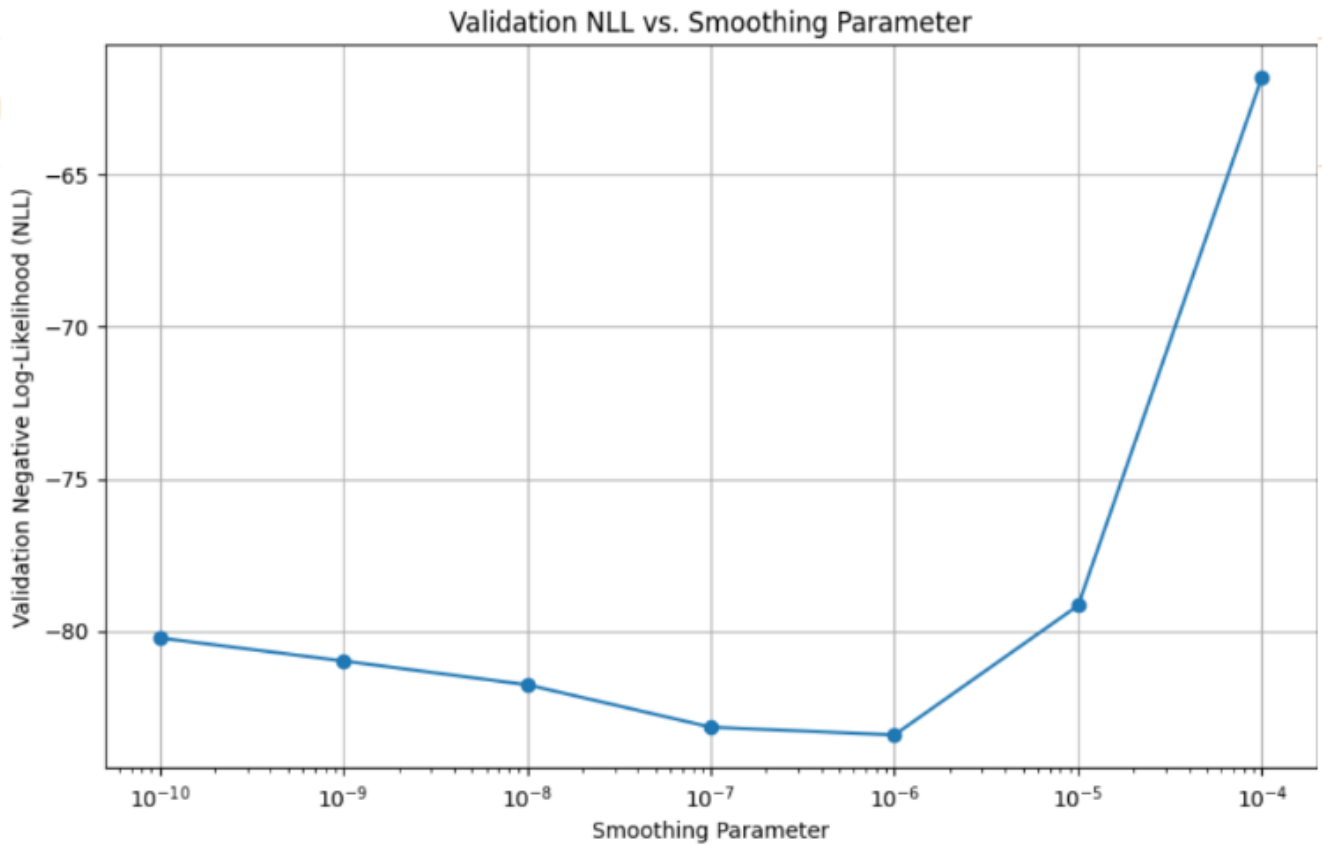

Figure: Validation NLL vs. Smoothing Parameter

Applications

Gorilla Nesting Sites Dataset

The locations of nesting sites of gorilla groups in the Kagwene Gorilla Sanctuary, Cameroon.

- Major class: 350 data points
- Minor class: 297 data points
- Rainy season: 372 data points
- Dry season: 274 data points

This dataset is ideal for showing how GAMs model spatial patterns in ecological data, focusing on how environmental factors influence gorilla nesting site distribution.

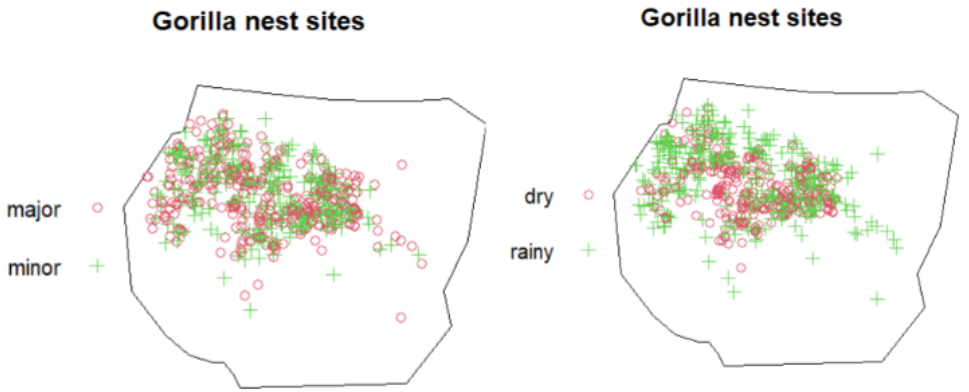

Figure: classes of the datasets

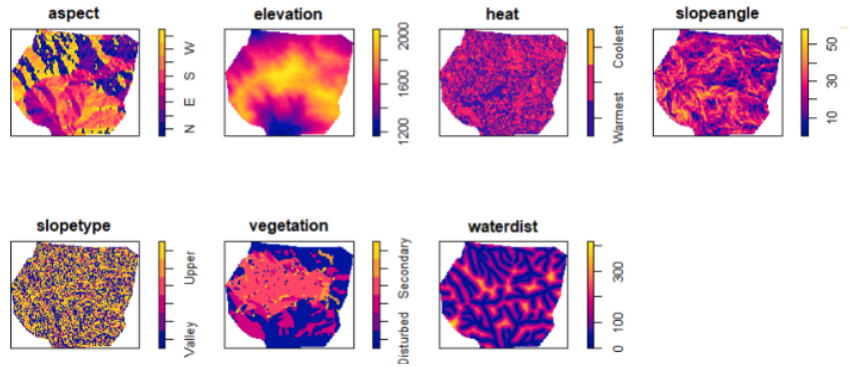

Figure: The spatial covariates of the dataset

GAM on the Location of nesting sites

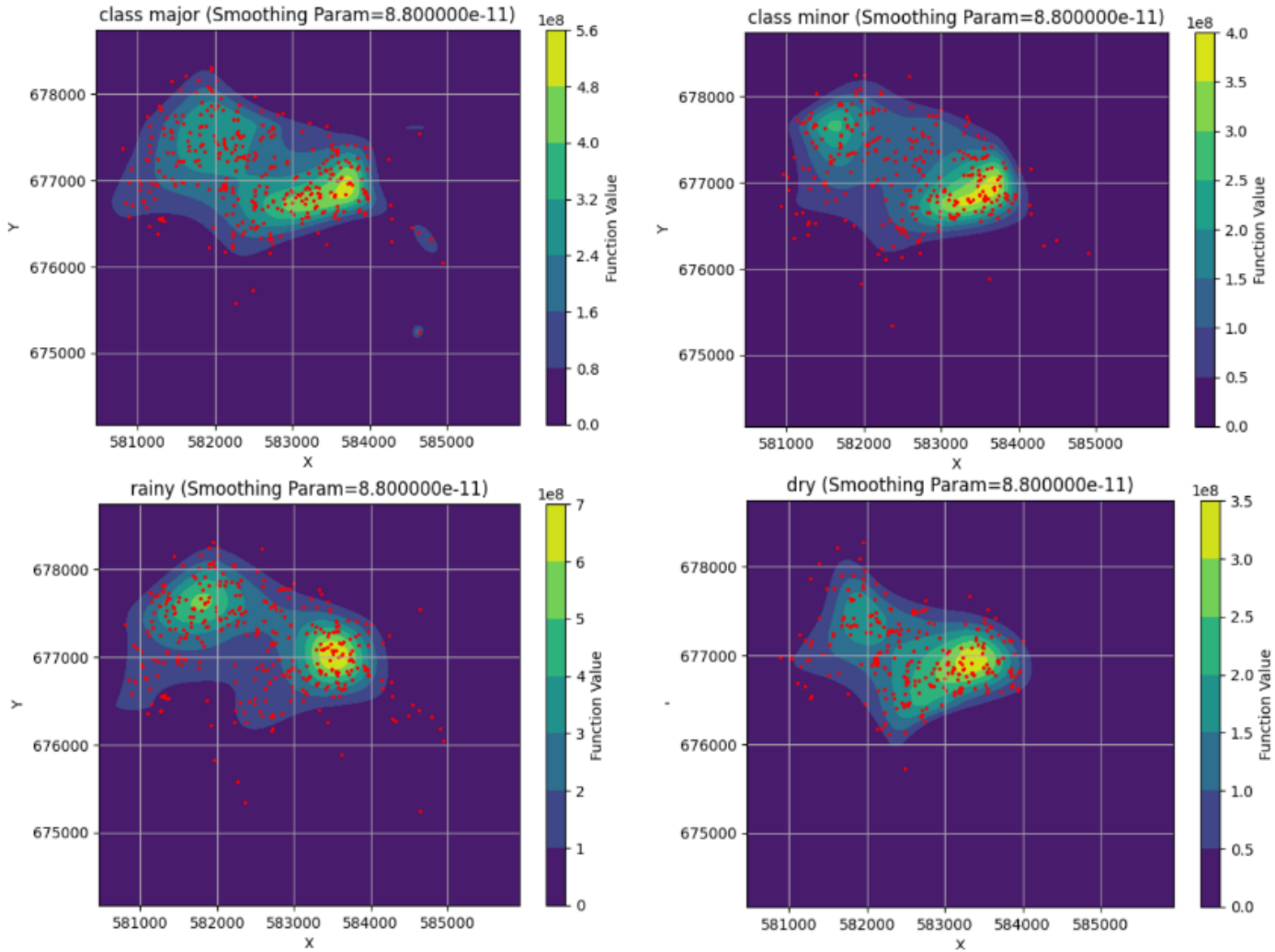

Figure: Intensity estimation for different classes and seasons.

Heat Data Visualization

Heat data is aligned it with the estimated intensity function.
The heat map indicates temperature variations across the National Park, where:

- Red represents the hottest regions,
- Yellow corresponds to moderate temperatures,
- Blue shows the coolest areas, and
- Dark blue represents areas with no available data.

Ratio of points where intensity is above the threshold and Modest $= 0.9414$

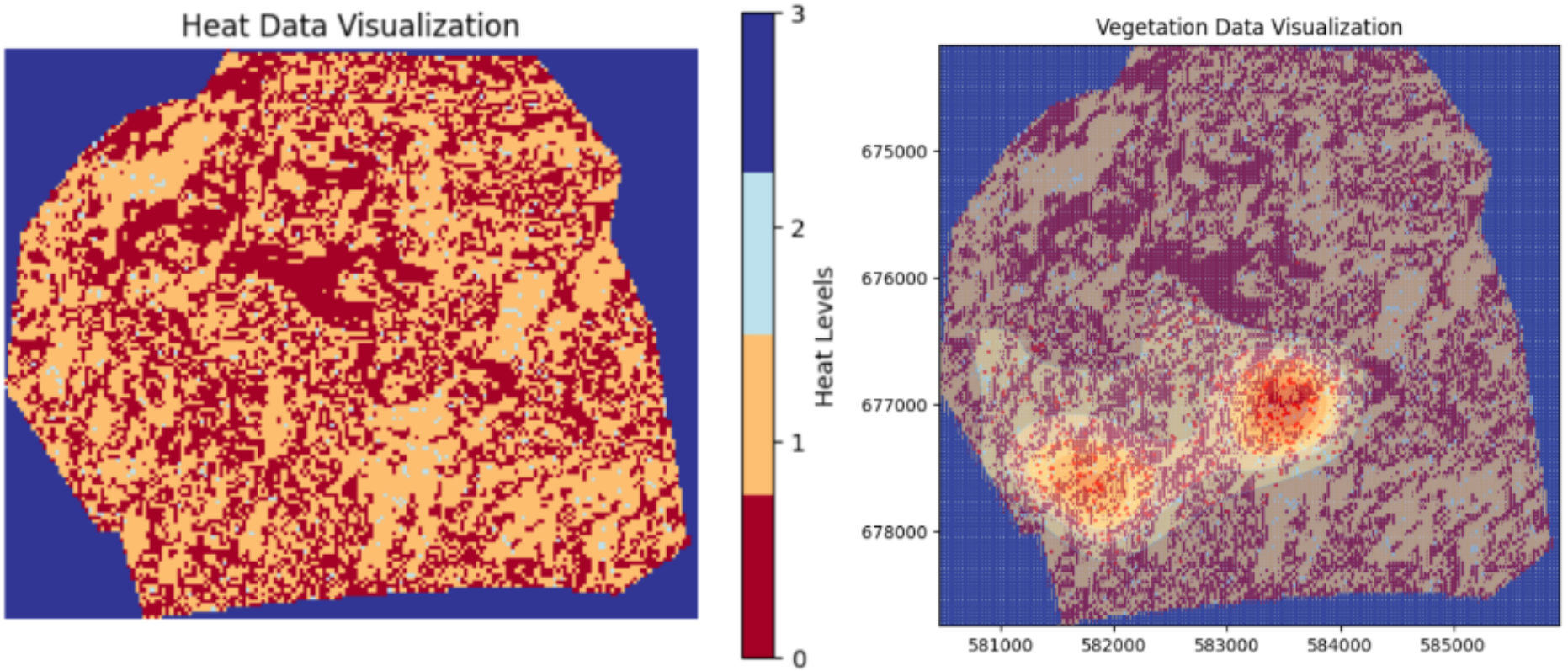

Figure: Heat data visualization (temperature levels) and aliened plot of intensity and heat data

Vegetation and Intensity

Vegetation data is aligned it with the estimated intensity.

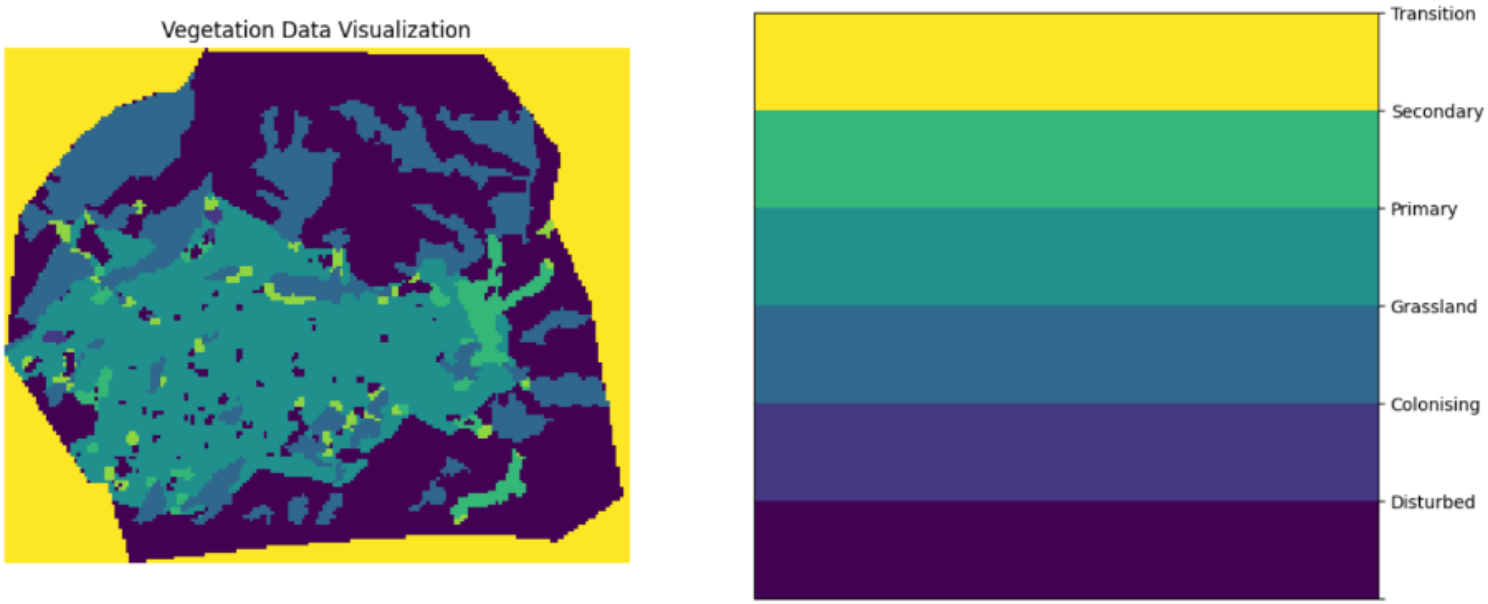

Figure: Vegetation Data Visualization.

Gorillas predominantly chose to nest in areas classified as **primary** and **secondary forests**.
Ratio of points above the threshold and in 'Primary' : 0.94653

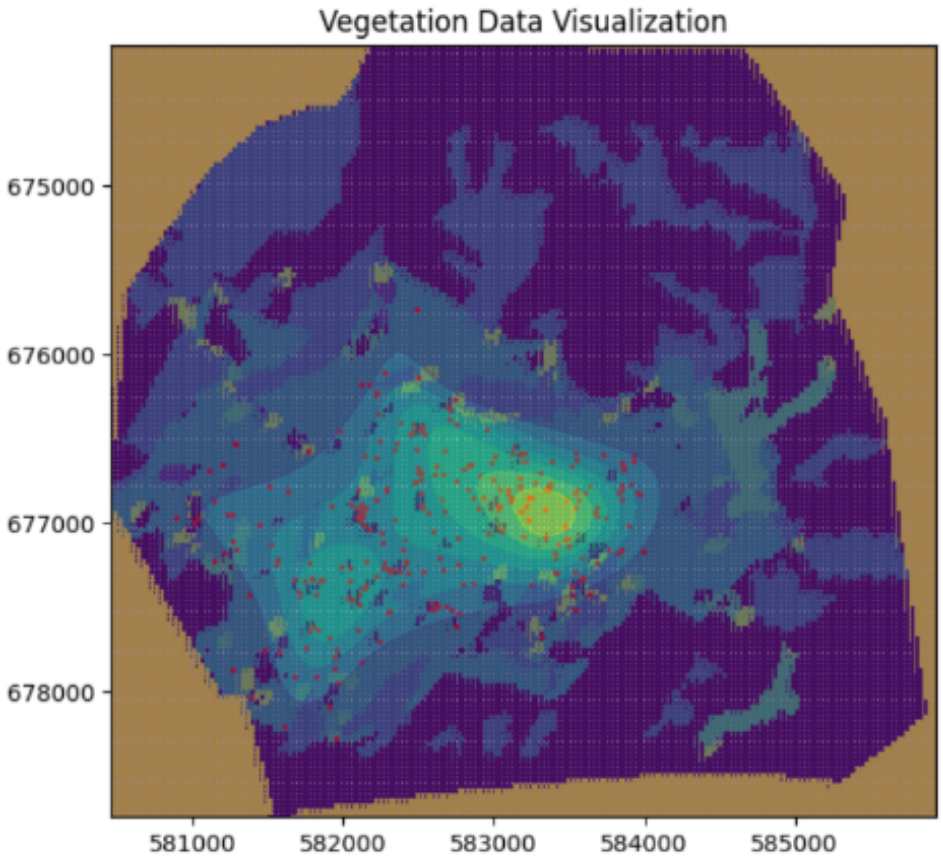

Figure: Vegetation Data Visualization aligned with estimated intensity of gorilla nests.

Distance to Nearest Water

The intensity of gorilla nests based on their distance to the nearest water source.

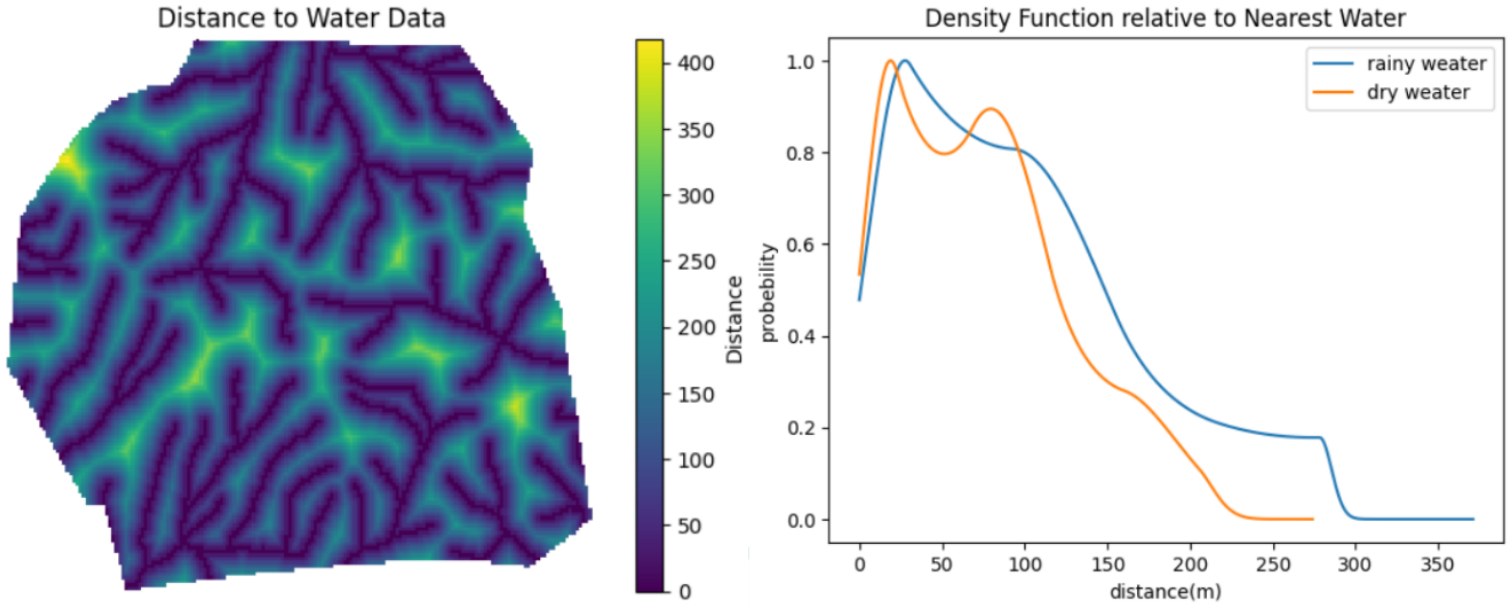

Figure: Left: Distance to water across the study region. Right: Density function of nest distance relative to the nearest water source during the rainy and dry seasons.

The wider curve indicates that gorillas are more flexible in their nesting locations when water is more readily available in the environment.

- Standard deviation in the rainy season: 79.10 meters
- Standard deviation in the dry season: 78.88 meters

Problems and Future works

Problems and Future works

Numerical Optimization with NLL Gradient: Switching from PyTorch to Analytical optimization using the NLL gradient formula.

Extending GAM Applications: Training multiple GAMs on diffrent factors simultaneously to capture varying intensity functions and Superposition.

Chebyshev Nodes for Basis Knots: Using Chebyshev nodes as basis knots helps reduce plot overshooting, particularly when the maximum intensity is at 0.

Appendix

B-splines

For all t between the knots t_p and t_{m-p} The Cox–de Boor recursion formula

$$B_{i,p}(t) = \begin{cases} \text{non-zero} & \text{if } t_i \leq t < t_{i+p+1}, \\ 0 & \text{otherwise.} \end{cases}$$

The higher $(p+1)$-degree B-splines are defined by recursion:

$$B_{i,p}(t) := \frac{t - t_i}{t_{i+p} - t_i} B_{i,p-1}(t) + \frac{t_{i+p+1} - t}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}(t).$$

Thinning Algorithm

Algorithm 1 Thinning Algorithm

function INTENSITY(x, y)
    $s \leftarrow 0.5$
    return $100 \times \exp\left(-\frac{x^2+y^2}{s^2}\right)$
end function
function THINNING_PROB$(x, y, \rho_{\text{max}})$
    return INTENSITY$(x, y)/\rho_{\text{max}}$
end function
Sample $N_{\text{points}} \sim \text{Poisson}(\rho_{\text{max}})$
$x_{\text{samp}} \leftarrow \text{Uniform}(0, 1, N_{\text{points}})$, $y_{\text{samp}} \leftarrow 0$
$p_{\text{acc}} \leftarrow$ THINNING_PROB$(x_{\text{samp}}, y_{\text{samp}}, \rho_{\text{max}})$
$r_{\text{reject}} \leftarrow \text{Uniform}(0, 1, N_{\text{points}}) < p_{\text{acc}}$
$x_{\text{acc}} \leftarrow x_{\text{samp}}[r_{\text{reject}}]$
return x_{acc}

Figure: Thinning Algorithm

Thank you!