EVI vs Temperature through OLS and Gaussian Processes

Daijun Xuan, with guidance from Torey Hilbert

Department of Mathematics, The Ohio State University

DEPARTMENT OF MATHEMATICS

Abstract

This project explores the relationship between the Enhanced Vegetation Index (EVI) and average land surface temperature across Ohio. EVI enhances vegetation signal in remote sensing by reducing atmospheric and background noise. Temperature, a key environmental factor, is measured across spatial grid cells. We compare three regression approaches—naive OLS, a spatially corrected version, and an optimal estimator using generalized least squares and Gaussian processes. Results highlight the importance of accounting for spatial correlation to improve prediction and uncertainty estimation.

Gaussian Processes (GP)

What is a Gaussian Process?

Gaussian Process (GP) is a collection of random variables, any finite number of which have a joint Gaussian distribution. It defines a distribution over functions and is fully specified by a mean function $m(x)$ and a covariance function (kernel) $k(x, x')$:

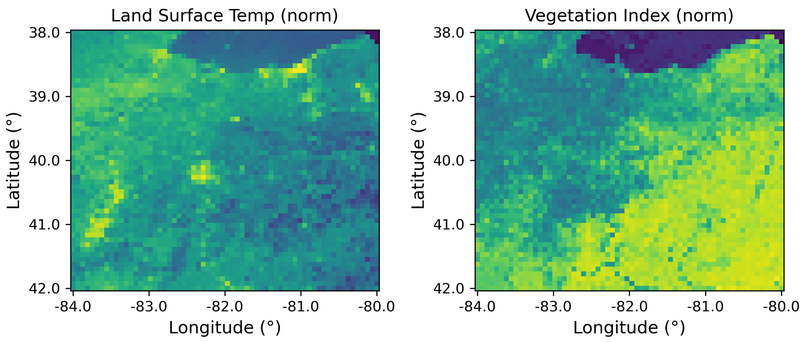
$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

Covariance Function (Kernel):

$$k(x, x') = \exp\left(-\frac{(x - x')^2}{2\ell^2}\right)$$

- x, x': Input values (e.g., temperature at different locations)
- ℓ: Length-scale parameter controlling how quickly correlation decays with distance
- $k(x, x')$: Covariance (similarity) between inputs x and x'

Gaussian Process (GP) regression captures spatial dependence between inputs—like temperatures at nearby locations—using a covariance function (kernel).



Remote sensing data from MODIS over Ohio on 06/11/2020. Fitting the above Gaussian Process to the EVI data, we find that the bandwidth is approximately 200 km, or approximately 2°.

Three Estimators for Spatial Linear Regression

1. Naive OLS (Incorrect Variance)

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

- β: Regression coefficient vector (the parameters to be estimated)
- ε: Error term, incorrectly assumed to be uncorrelated.
- $\hat{\beta}$: Estimated coefficients using Ordinary Least Squares (OLS)
- X: Design matrix, with rows of the form $\begin{bmatrix} 1 & \text{Temp}_{s_i} \end{bmatrix}_{s_i \in S}$.

When this correlation is ignored, the model treats spatially clustered patterns as random noise, leading to underestimated standard errors, overconfident predictions, and potentially misleading inferences.

2. Corrected Variance OLS

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \Sigma)$$

Corrected Variance:

$$\left(X^T X\right)^{-1} X^T \Sigma X \left(X^T X\right)^{-1}$$

- Σ: Spatial covariance matrix (known or estimated via a model)

3. Optimal Estimator (Best Linear Unbiased Estimator)

We want to find a weight vector $\boldsymbol{c} \in \mathbb{R}^n$ such that:

$$\hat{\beta}_1 = \boldsymbol{c}^T \boldsymbol{y}$$

subject to the constraints:

- $\mathbb{E}[\hat{\beta}_1] = \beta_1$
- $\text{Var}(\hat{\beta}_1)$ is minimized

Using Lagrange multipliers, the optimal solution is:

$$\boldsymbol{c} = \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Thus, the optimal estimator is:

$$\hat{\beta}_1 = \begin{bmatrix} 0 & 1 \end{bmatrix} (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \boldsymbol{y}$$

The optimal weights c depend on the spatial covariance matrix Σ, ensuring that the estimator accounts for spatial correlation and yields more accurate results than standard OLS.

Simulation Results

To study how each of these methods might perform on the real EVI and Temperature data, we simulate Gaussian Processes on a 2D grid of longitude and latitude values, with parameters fitted to be similar to the real data. In particular, on a 54x64 grid of lon/lat pairs in the range [-84, 80]x[38, 42], we use the kernel

$$\text{Cov}(\text{EVI}_{s_1}, \text{EVI}_{s_2}) = 1.81^2 \exp\left(\frac{-||s_1 - s_2||^2}{2 \cdot (2.068)^2}\right) + 0.05.$$

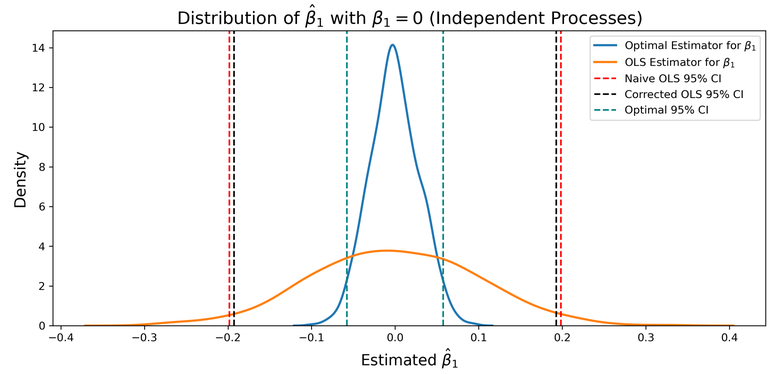
We make two *independent* draws from the above process, and then attempt the linear regression:

$$\text{EVI} = \beta_0 + \beta_1 \cdot \text{Temp} + \epsilon.$$

Note that here, Temp and EVI were genuinely independently generated, and so $\beta_1 = 0$ by design. For each of the three previous methods, we generate 95% confidence intervals for β_1 by

$$w = 4 \cdot \sqrt{\text{Var}(\hat{\beta}_1)}, \quad \text{and} \quad 95\%\ CI: \ \left[\hat{\beta}_1 - \frac{w}{2},\ \hat{\beta}_1 + \frac{w}{2}\right]$$

As expected, we find that the coverage of 95% confidence intervals for both the Corrected Variance estimator and the optimal estimator were indeed 95%. Meanwhile, the Naive OLS intervals cover the true parameter ($\beta_1 = 0$) 24% of the time.



Conclusion

This project examines the relationship between the Enhanced Vegetation Index (EVI) and temperature using OLS regression and Gaussian Processes (GP). Naive OLS underestimates uncertainty by ignoring spatial dependence. By incorporating spatial correlation—via corrected OLS and the Best Linear Unbiased Estimator (BLUE)—we improve accuracy and reliability. GP further offers a flexible, non-parametric approach for modeling spatial uncertainty. Our results highlight the importance of accounting for spatial structure in environmental data analysis.

Reference: NASA. (n.d.). MODIS: Moderate Resolution Imaging Spectroradiometer. NASA. https://modis.gsfc.nasa.gov/