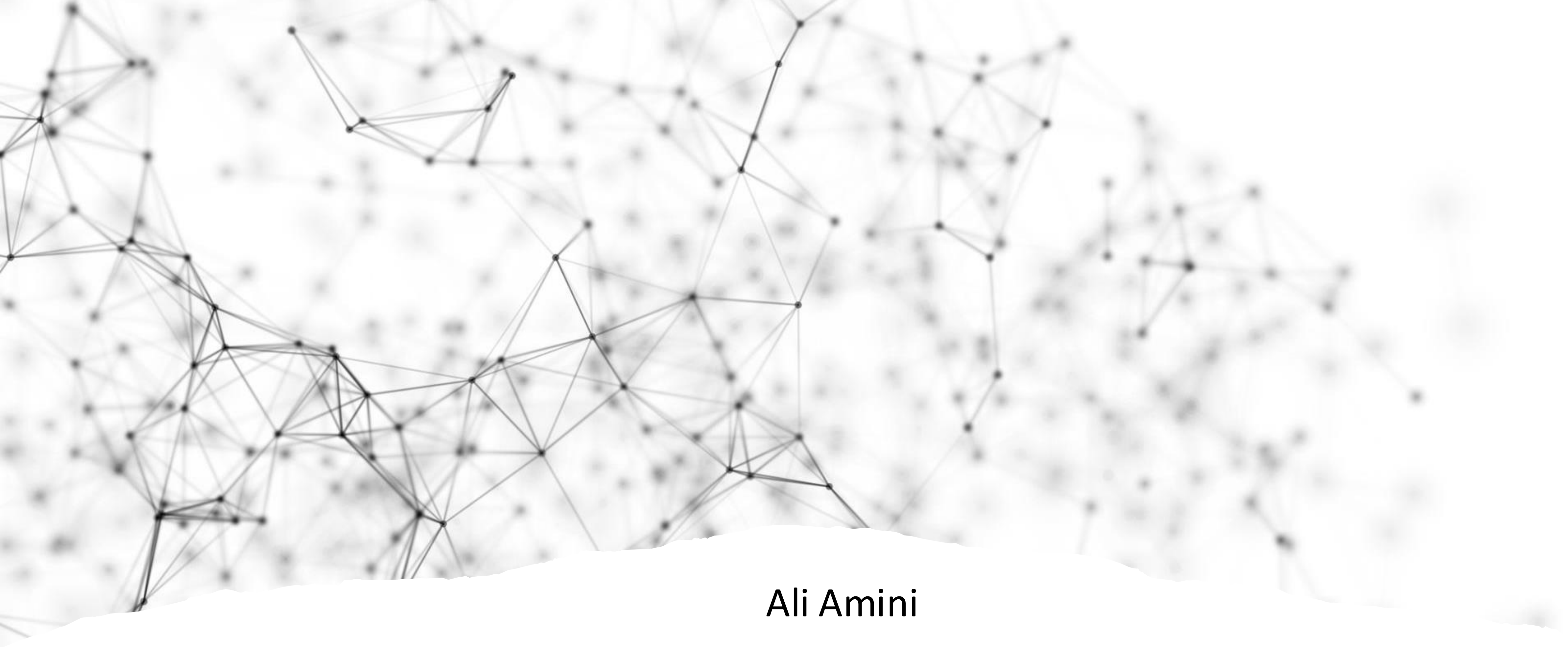
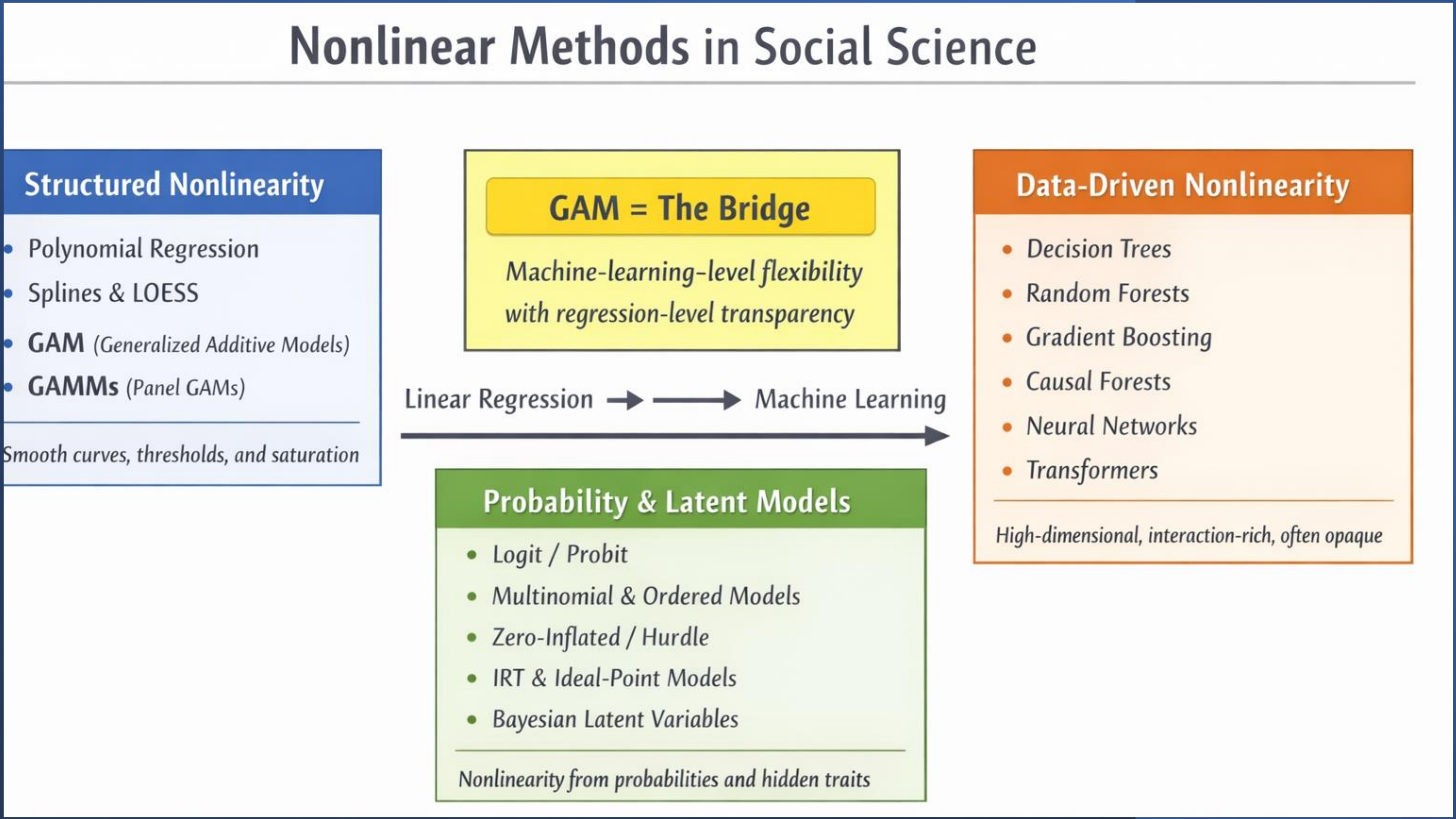


Ali Amini

Nonlinear Methods for Social Science:

Why We (Often) Must Move Beyond Linearity


Nonlinear Methods in Social Science
Structured Nonlinearity
Polynomial Regression
Splines & LOESS
GAM (Generalized Additive Models)
GAMMs (Panel GAMs)
Smooth curves, thresholds, and saturation
GAM = The Bridge
Machine-learning-level flexibility with regression-level transparency
Linear Regression → → Machine Learning
Probability & Latent Models
Logit / Probit
Multinomial & Ordered Models
Zero-Inflated / Hurdle
IRT & Ideal-Point Models
Bayesian Latent Variables
Nonlinearity from probabilities and hidden traits
Data-Driven Nonlinearity
Decision Trees
Random Forests
Gradient Boosting
Causal Forests
Neural Networks
Transformers
High-dimensional, interaction-rich, often opaque

Overview

WHAT IS GAM?

WHY GAM?

HOW GAM? [CODE AND APPLICATIONS]

What is GAM?

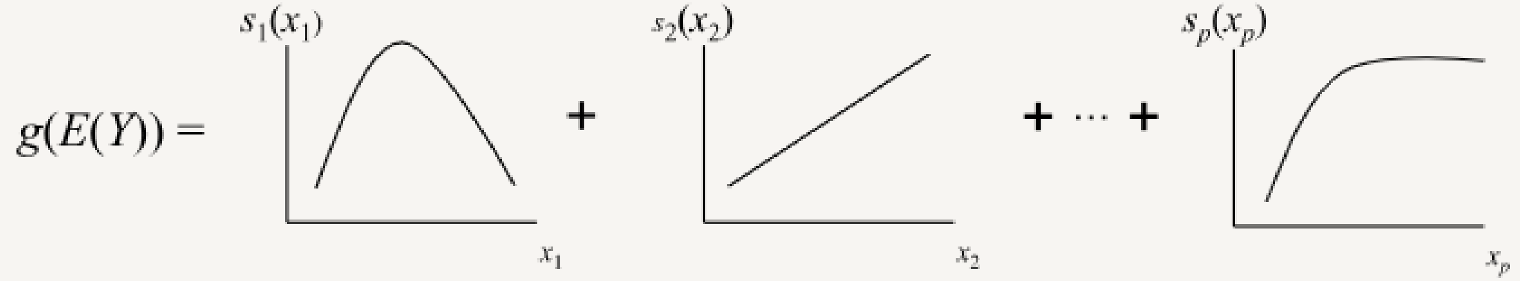
- Developed by Trevor Hastie and Robert Tibshirani in **1986**
- Extends linear models to account for non-linear relationships between predictors and response variables
- Utilizes smoothing functions to model non-linearities effectively

Statistical Science
1986, Vol. 1, No. 3, 297–318

Generalized Additive Models

Trevor Hastie and Robert Tibshirani Important scholars--> Google them!

Abstract. Likelihood-based regression models such as the normal linear regression model and the linear logistic model, assume a linear (or some other parametric) form for the covariates $X_1, X_2, \cdots, X_p$. We introduce the class of *generalized additive models* which replaces the linear form $\sum \beta_j X_j$ by a sum of smooth functions $\sum s_j(X_j)$. The $s_j(\cdot)$'s are unspecified functions that are estimated using a scatterplot smoother, in an iterative procedure we call the *local scoring* algorithm. The technique is applicable to any likelihood-based regression model: the class of *generalized linear models* contains many of these. In this class the linear predictor $\eta = \sum \beta_j X_j$ is replaced by the additive predictor $\sum s_j(X_j)$; hence, the name generalized additive models. We illustrate the technique with binary response and



We can write the GAM structure as:

$$g(E(Y)) = \alpha + s_1(x_1) + \cdots + s_p(x_p),$$

What is GAM?

- Mathematical Insight:
  - Relationship between Y and each S(i) follows a smooth pattern, which can be either linear or non-linear
  - Replaces X(i) with S(Xi) :
  - S(Xi) is a smoothing function that adapts to the data
  - Captures the underlying trend over a specified span, allowing for both linear and non-linear patterns

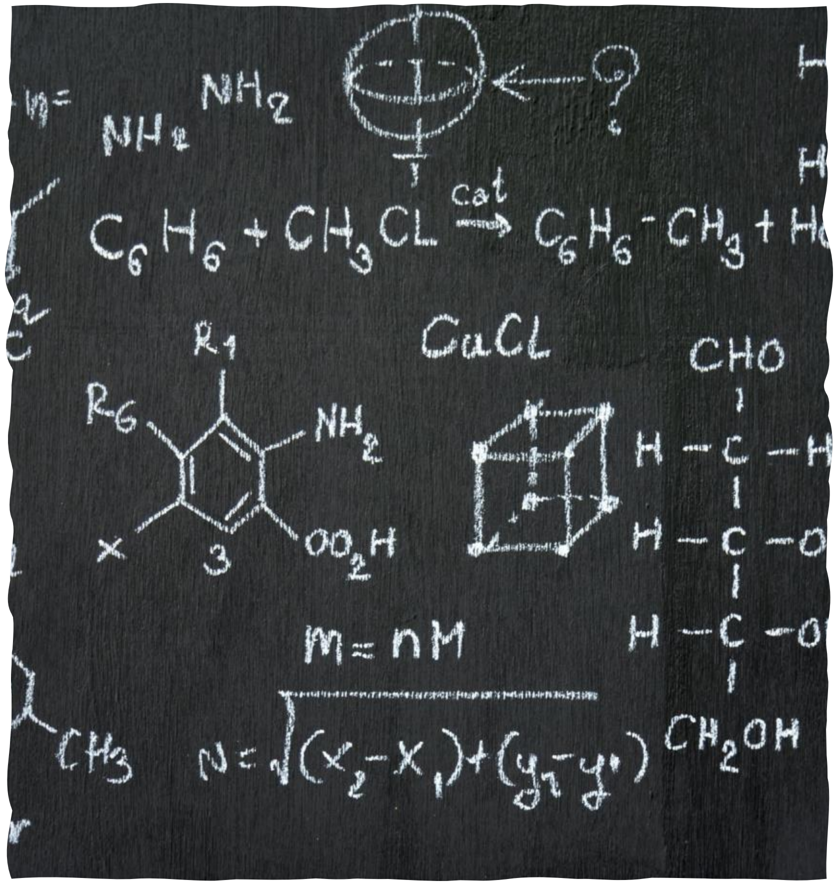
Why GAM?

- Powerful and yet simple technique → Easy to interpret
- Many real-world relationships are non-linear → uncover hidden trends
- Fancy to model non-linear effects→ Nature is not [always] linear!

Ordinary Least Square

- What is OLS? Great invention because of beautiful interpretation
- Finds coefficients that minimize the sum of squared residuals
- Limitations of OLS? Assumes linear relationship between predictors and response
- OLS may not be suitable for data with non-linear relationships
- GAMs provide more flexibility to overcome this limitation

Math: remember only x → s(x)



- **OLS regression model** $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$

- **GAM model** $y = \beta_0 + S_1(x_1) + \ldots + S_p(x_p) + \varepsilon$
- **Key differences:**
  -Sj(xj)are smoothing functions, not linear terms
  - Allow for non-linear relationships between predictors and response

- Properties of smoothing functions:
  - Flexible
  - can take on variety of shapes
  - avoid overfitting data
  - Estimated in data-driven way from the data
  -GAMs estimate the smoothing functions to capture non-linearities

LM → GAM → ML

Model Comparison Table

Model	Functional Form
Linear Model (LM)	$y = \beta_0 + \beta_1 x$
GAM	$y = \beta_0 + f(x)$
Machine Learning	$y = f(x_1, x_2, x_3, \ldots)$

Intuition

Think of a predictor like **income → vote probability**.

- **LM** forces it to be a straight line
- **ML** lets it be anything (wiggly, interacting, opaque)
- **GAM** lets it be **a smooth curve you can see and interpret**

ELSEVIER

Journal of Criminal Justice

Volume 102, January–February 2026, 102554

Help-seeking behaviors of stalking victims: Integrating machine learning and regression approaches to examine how victimization consequences shape victims' decisions

Auzeen Shariati [a], Fariba Allahyoorti Dehaghi [b], Ali Amini [c]

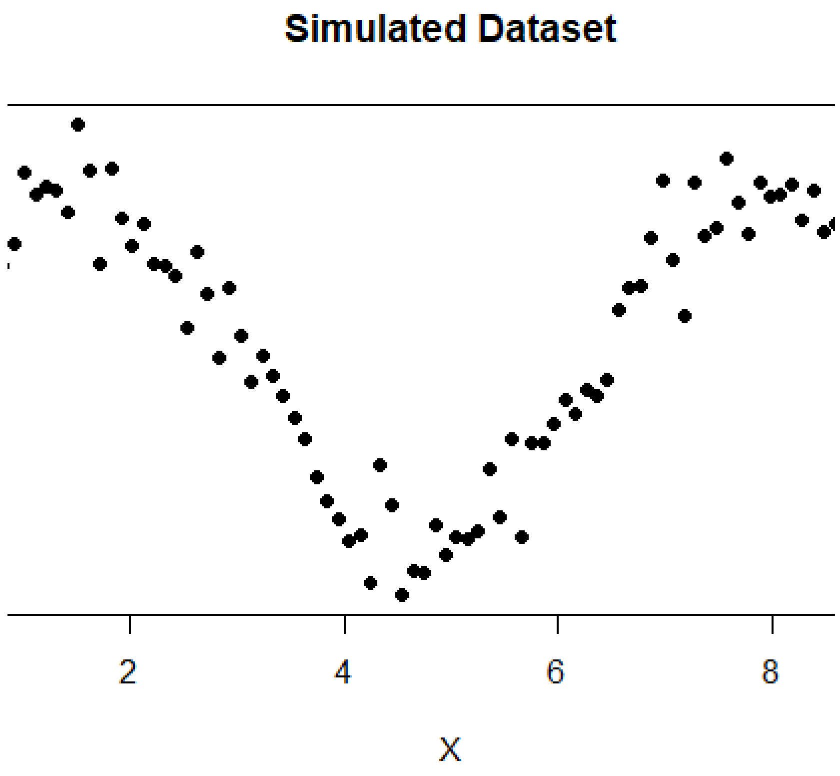
Abstract

We examined how the consequences of stalking victimization shape victims' help-seeking behaviors, using the 2019 National Crime Victimization Survey, Supplemental Victimization data. We analyzed three distinct help-seeking outcomes: (a) reporting to police, (b) help-seeking from victim-serving agencies, and (c) help-seeking from persona networks. Logistic regression models assessed the statistical significance of individual predictors, while our non-parametric Machine Learning approach evaluated their predictive power and captured non-linear patterns. Regression results revealed that substantial emotional the likelihood of all three help-seeking behaviors. He ms increased the likelihood of network help-seeking, sociated with lower odds of police reporting. Machine learning identified financial problems, emotional distress, and healt problems as the most predictive features for police reporting, agency help-seeking, and network help-seeking, respectively. These findings underscore the multidimensional nature of victimization consequences and the value of combining traditional statistical inference with machine learning to better understand victim decision-making.

GAM : Machine-learning–level flexibility with regression-level transparency

HOW GAM?(1) R codes- Simulated data

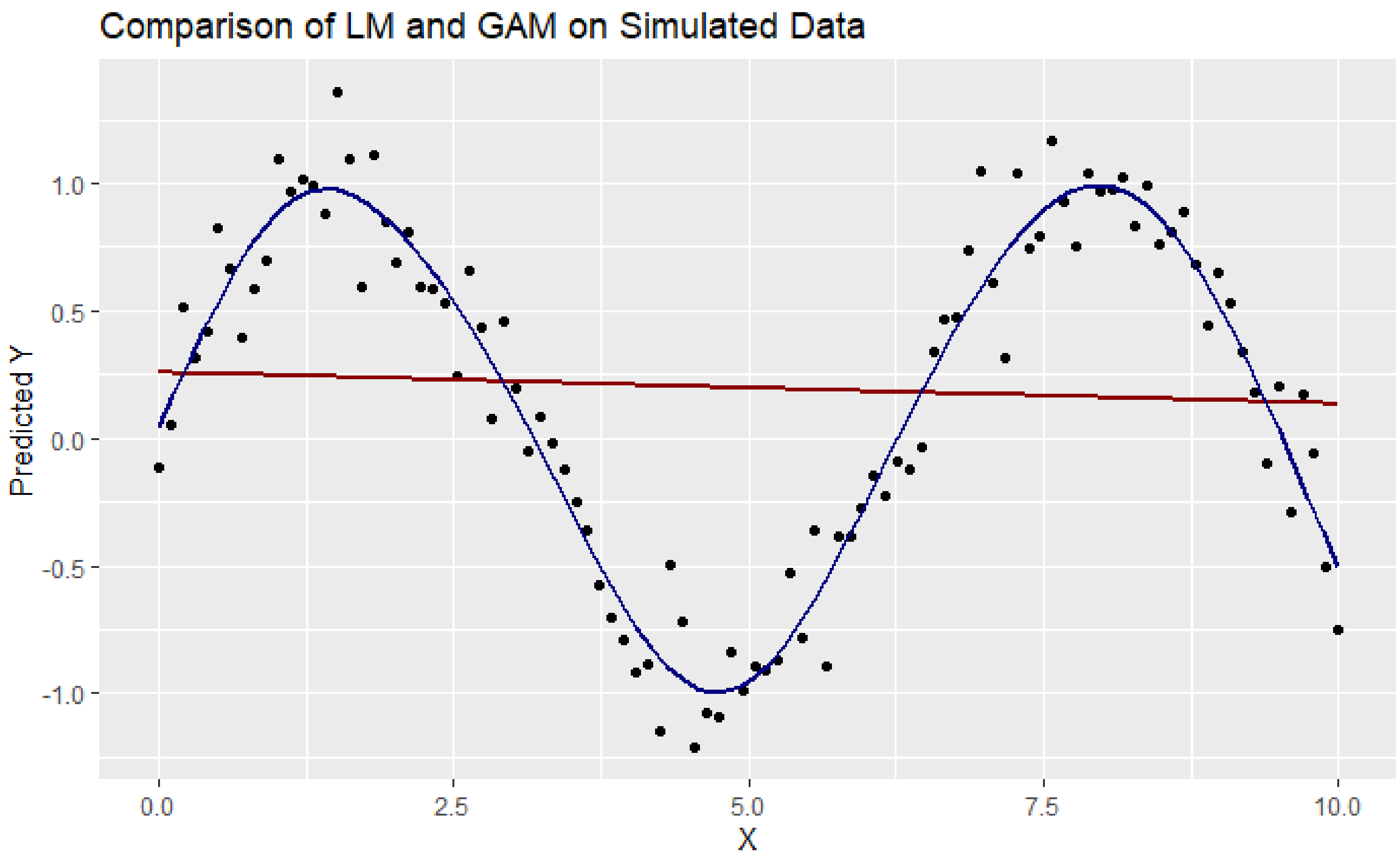
```
# Library loading
library(mgcv)
library(ggplot2)
# Simulating a non-linear dataset
set.seed(123) # Setting a seed for reproducibility
x <- seq(0, 10, length.out = 100)
[redacted] + rnorm(100, sd = 0.2) # Adding some noise
# Creating a data frame
data <- data.frame(x, y)
# Visualizing the dataset
plot(data$x, data$y, main = "Simulated Dataset", xlab =
"X", ylab = "Y", pch = 19)
```



HOW GAM?(1) run GAM vs OLS Model

```
# Fit a linear model
lm_model_simulated <- lm(y ~ x, data = data)
# Fit a GAM model
gam_model_simulated <- gam(y ~ s(x), data = data)
# Create a plot with both fits
ggplot(data, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", colour = "darkred", se = FALSE) +
  geom_smooth(method = "gam", formula = y ~ s(x), colour = "navyblue", se = FALSE) +
  labs(title = "Comparison of LM and GAM on Simulated Data",
       x = "X",
       y = "Predicted Y")
```

How GAM(1) Result; GAM vs OLS Simulated data



lm(y ~ x, data = data)→ Model Driven

vs

gam(y ~ s(x), data = data)→ Data Driven

Model	R-squared	RMSE	p-value	AIC	Signif. smooth terms
LM	0.0028	0.6806	0.5986	210.8	NA
GAM	0.927	0.175	<2e-16	-44.97	s(x) p<2e-16

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

$$AIC = 2k - 2\ln(\hat{L})$$

AIC = Akaike information criterion

k = number of estimated parameters in the model

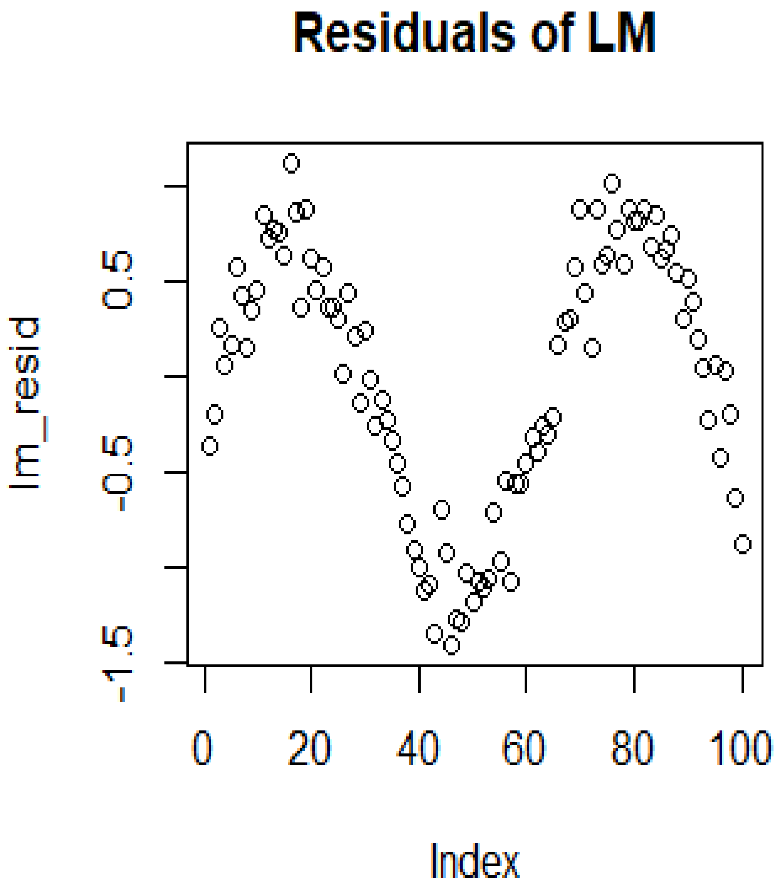
$\hat{L}$ = maximum value of the likelihood function for the model

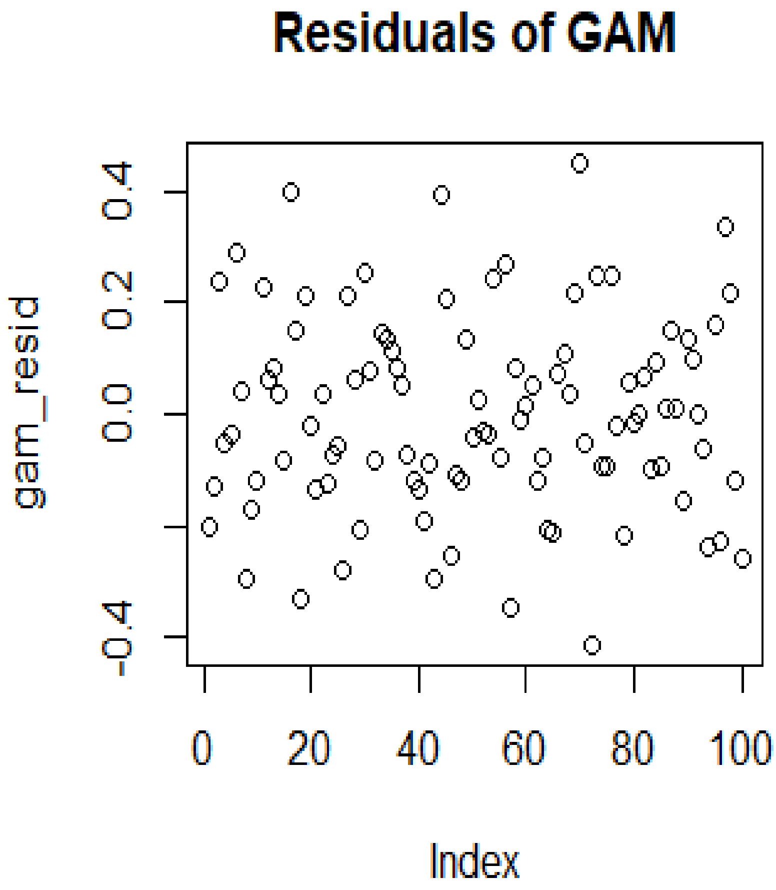
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Model caparison
key model performance metrics

Linear vs. GAM Model Comparison
Residuals

- The sin wave pattern in the LM residuals is → the linear model cannot fit the true nonlinear (sinusoidal) relationship well → There is still nonlinearity that the model misses.
- The random scatter of the GAM residuals → flexibly fitting the nonlinearity in the data-→ There is no structure left that the model has not captured.



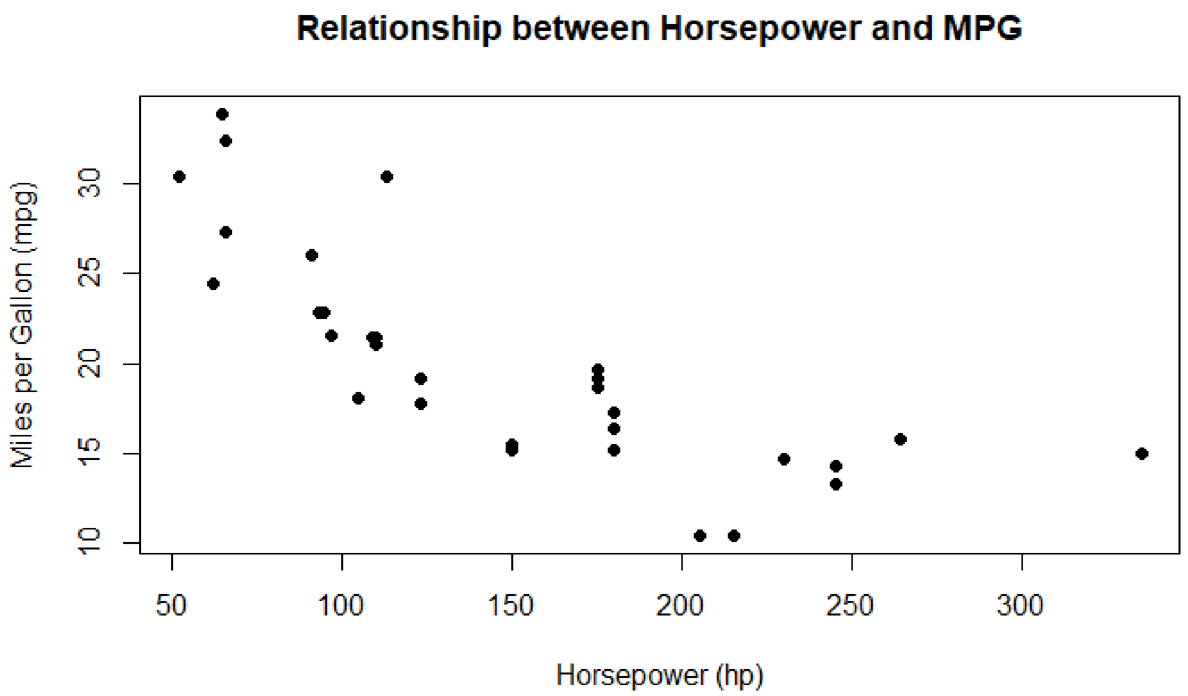


Linear vs. GAM Model Comparison
Anova test

```
anova(lm_model_simulated, gam_model_simulated, test="F")
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ s(x)
  Res.Df    RSS     Df Sum of Sq      F    Pr(>F)
1 98.000 45.401
2 91.204  3.070 6.7964    42.331 185.04 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

HOW GAM?(2)R codes &datasets(mtcars)

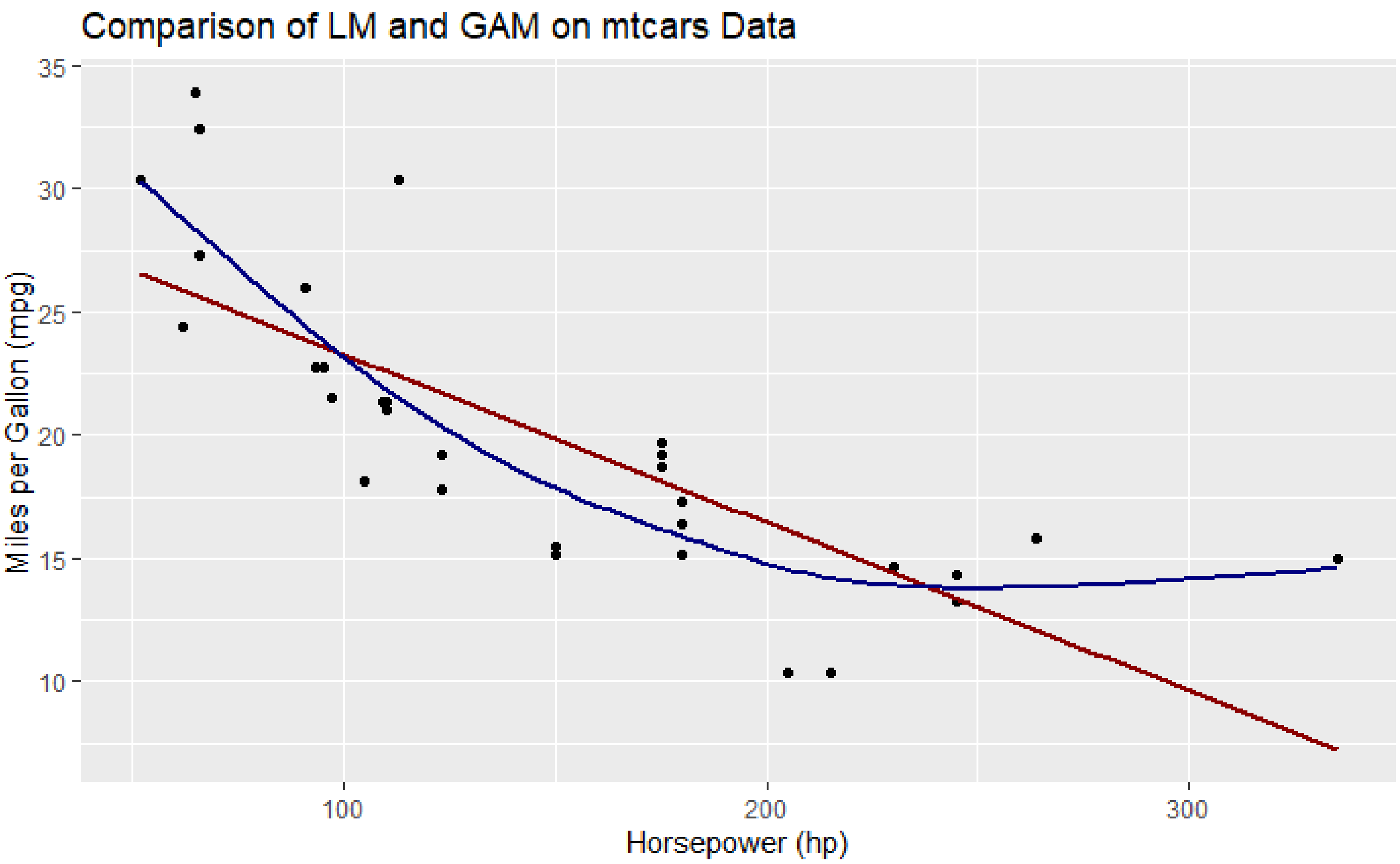
```
# Loading the mtcars dataset

data(mtcars) #The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles

# variables of interest,mpg (Miles/(US) gallon) and hp (Gross horsepower)

# Exploring the relationship between 'mpg' and 'hp'

plot(mtcars$hp, mtcars$mpg, main = "Relationship between Horsepower and MPG", xlab = "Horsepower (hp)", ylab = "Miles per Gallon (mpg)", pch = 19)
```



HOW GAM?(2) run GAM vs OLS Model R data sets(mtcars)

```
# Loading necessary libraries
library(mgcv)
library(ggplot2)
# Fit a linear model
lm_model_mtcars <- lm(mpg ~ hp, data = mtcars)
# Fit a GAM model
gam_model_mtcars <- gam(mpg ~ s(hp) , data = mtcars)
# Create a plot with both fits
ggplot(mtcars, aes(x = hp, y = mpg)) +
  geom_point() +
  geom_smooth(method = "lm", colour = "darkred", se = FALSE) +
  geom_smooth(method = "gam", formula = y ~ s(x), colour = "navyblue", se = FALSE) +
  labs(title = "Comparison of LM and GAM on mtcars Data",
       x = "Horsepower (hp)",
       y = "Miles per Gallon (mpg)")
```

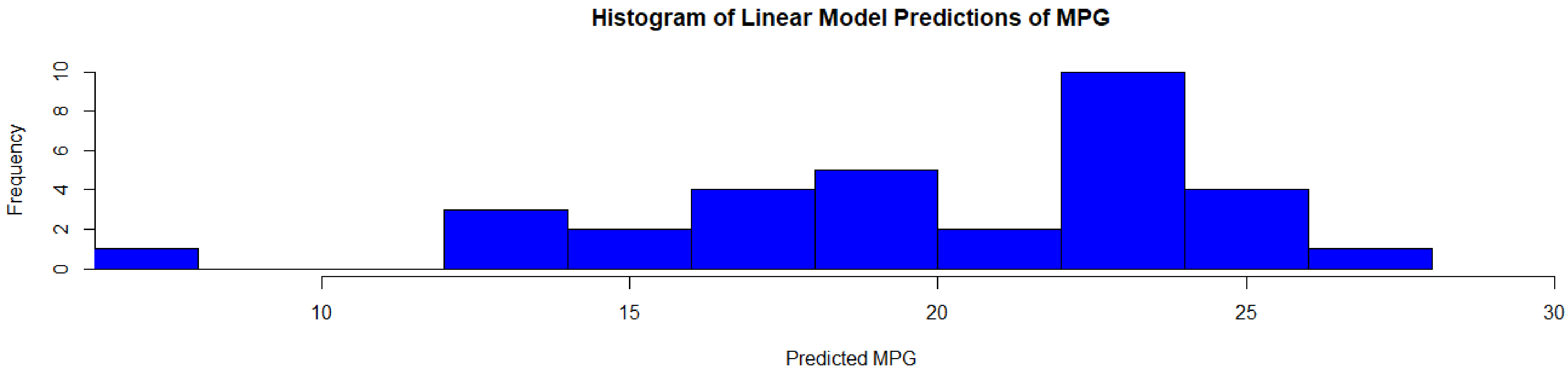
How GAM(2) Result; GAM vs OLS mtcars data

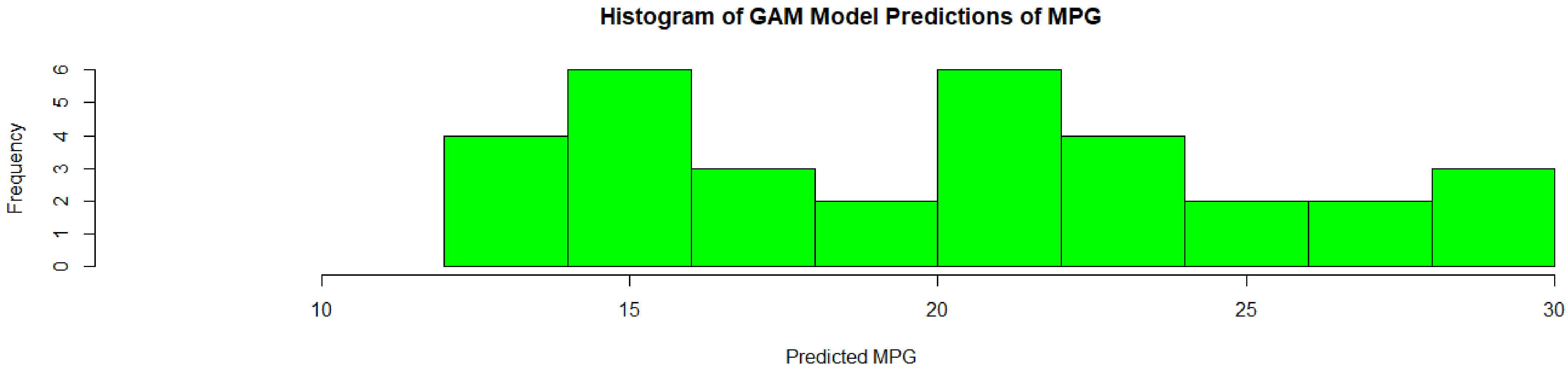


lm(mpg ~ hp, data = mtcars)→ Model Driven

vs

gam(mpg ~ s(hp) , data = mtcars)→ Data Driven


Histogram of Linear Model Predictions of MPG
Frequency
0
2
4
6
8
10
10
15
20
25
30
Predicted MPG


Histogram of GAM Model Predictions of MPG
Frequency
0
1
2
3
4
5
6
10
15
20
25
30
Predicted MPG

Real World application: *model→*

$$T_{\mathrm{anom}} \sim s(\mathrm{TCF}) + s(\mathrm{IMP}) + ti(\mathrm{TCF}, \mathrm{IMP}) + s(\mathrm{ELEV}) + s(\mathrm{ST-WS}, \mathrm{by} = \mathrm{ST-WD}) + s(\mathrm{LON}, \mathrm{LAT})$$

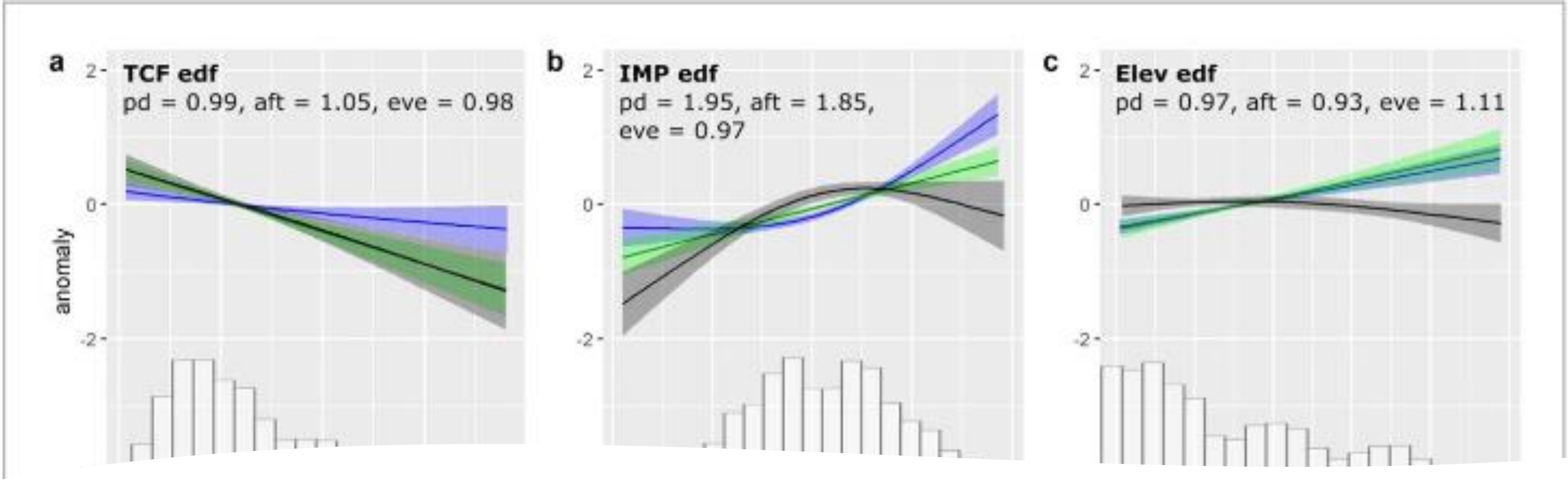
2.3. Analysis using generalized additive models (GAMs)

The relationship between some biophysical variables—most notably tree canopy cover—and air or LST can be nonlinear in nature (Ziter *et al* 2019, Logan *et al* 2020). GAMs are a nonparametric technique that can fit smooth curves between predictor and response variables using penalized regression splines (Pedersen *et al* 2019). In this study, we used the *gam* function in the R package 'mgcv' (version 1.8.31) and fit the models using fast restricted maximum likelihood.

Table 1. Data descriptions.

Variable name	Short name	Description
Vegetation		
Tree canopy	TCF	1 m tree canopy map derived from 2018 City of DC lidar data
Soft canopy	SCF	Tree canopy that does not overhang impervious surface
Hard canopy	HCF	Tree canopy that overhangs impervious surface
Canopy patches	PATCH	Soft canopy patches large enough to have cores (MSPA)
Distributed canopy	DISTRB	Soft canopy, connected or unconnected, no core (MSPA)
Pervious-open	PV-O	Area that is neither soft canopy nor impervious surface
Built environment		
Impervious surface	IMP	Impervious surface from City of DC planimetric data
Building height (sum)	BH	Building heights summed in area (DC building footprints and lidar data)
Building height (IMP norm)	BH-norm	Building heights as above but normalized by IMP to decorrelate
Skyview factor	SVF	Skyview factor calculated using DC lidar data in SAGA GIS
Physiographic		
Elevation	ELEV	City of DC lidar Digital Terrain Model (2018)
Quantile elevation	Q-ELEV	Quantile (local) elevation within 300 m radius
Distance from water	DIST-W	Euclidean distance from Potomac and Anacostia rivers
Car data		
Spatial coordinates	LON, LAT	Temperature measurement locations geographic coordinates
Mobile temperature	MBL-T	Temperature measurements (celsius)
Miles per hour	MPH	Car travel speed
Station data		
Station temperature	ST-T	Temperature (celsius) averaged across four downtown DC stations
Station wind speed	ST-WS	Wind speed at one representative station
Station wind direction	ST-WD	Wind direction at one representative station
Station solar radiation	ST-SR	Solar radiation at one representative station

Michael Alonzo et al 2021

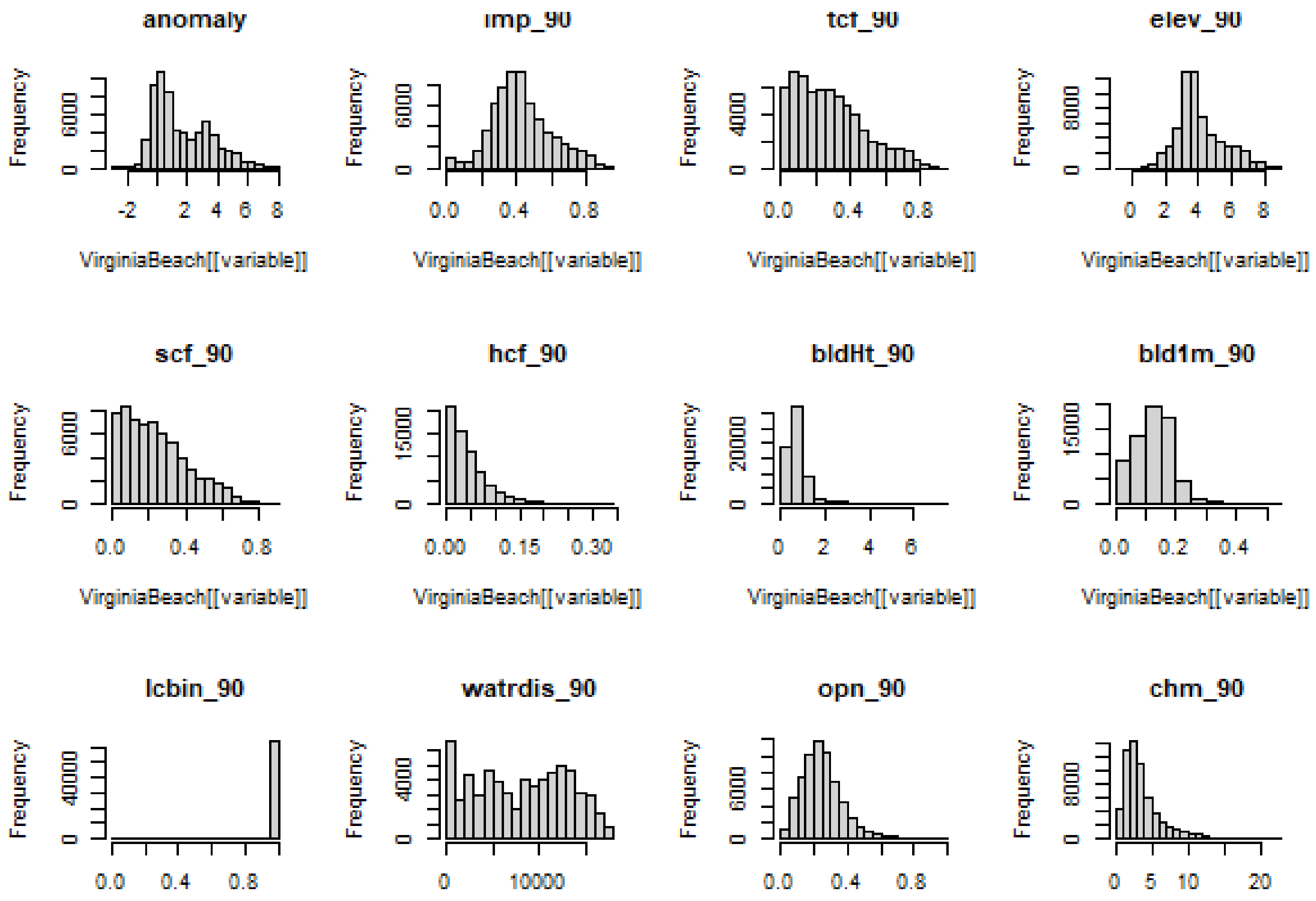


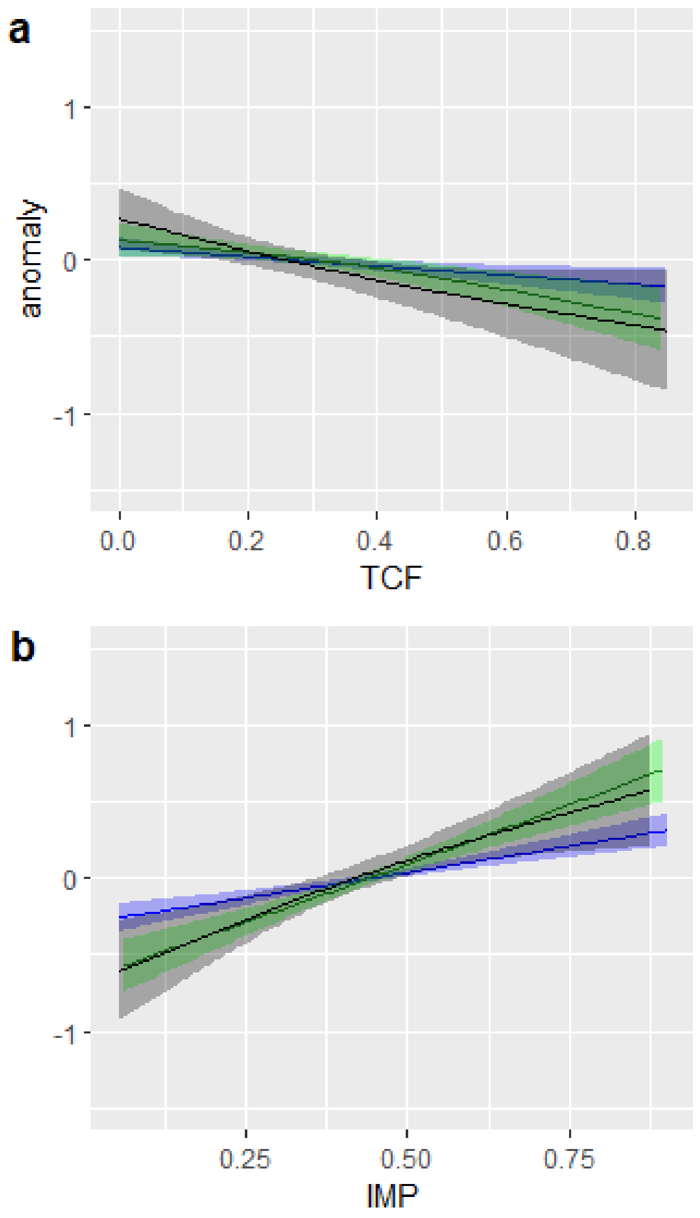
- Michael Alonzo et al 2021

Virginia Beach, July 15th, 2021

How GAM(3) Real world application; Tree Canopy and Temperature N ~ 62,000

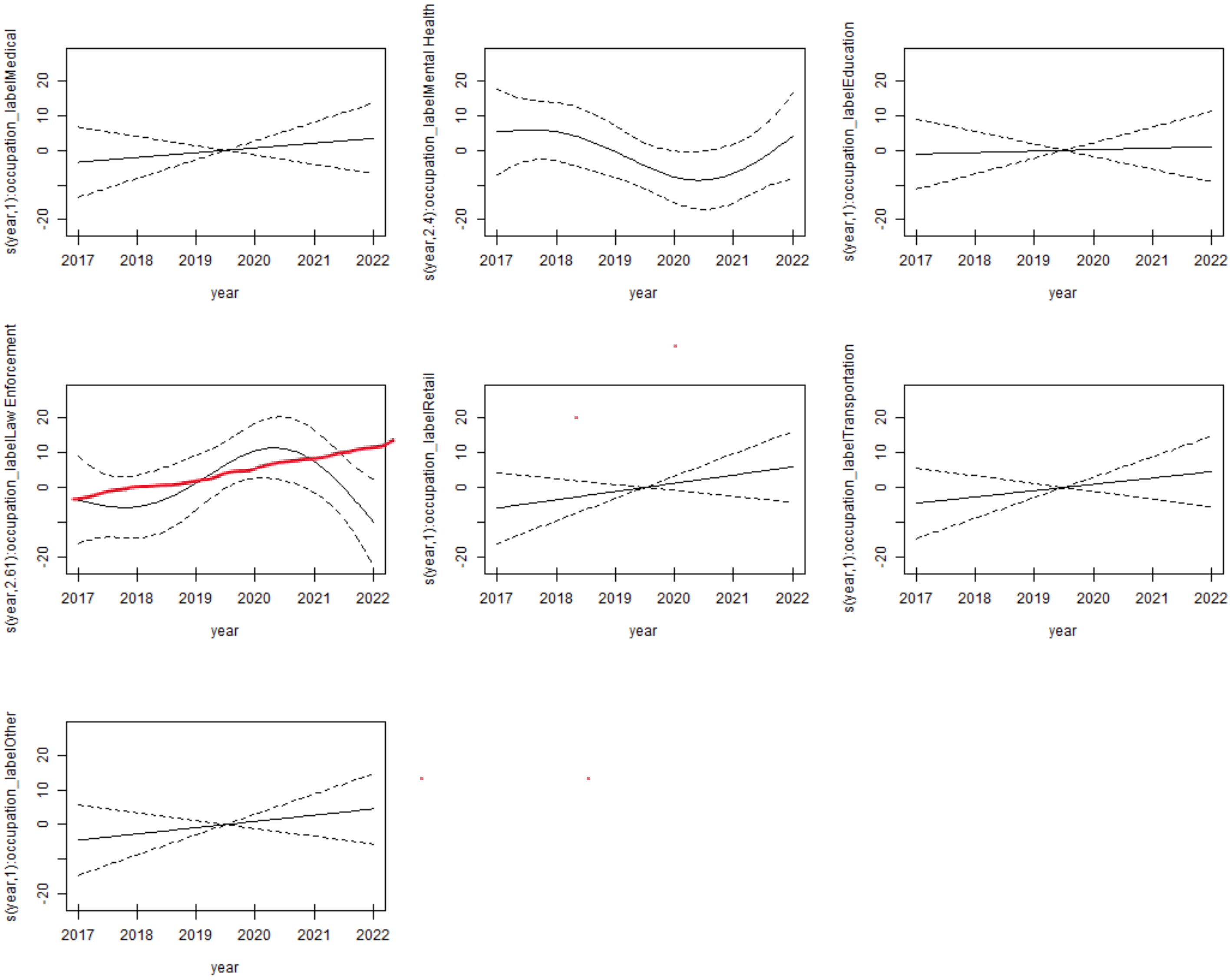
Morning= 23166
Afternoon=17396
Evening=20698


anomaly
imp_90
tcf_90
elev_90
scf_90
hcf_90
bldHt_90
bld1m_90
lcbin_90
watrdis_90
opn_90
chm_90
Frequency
VirginiaBeach[[variable]]

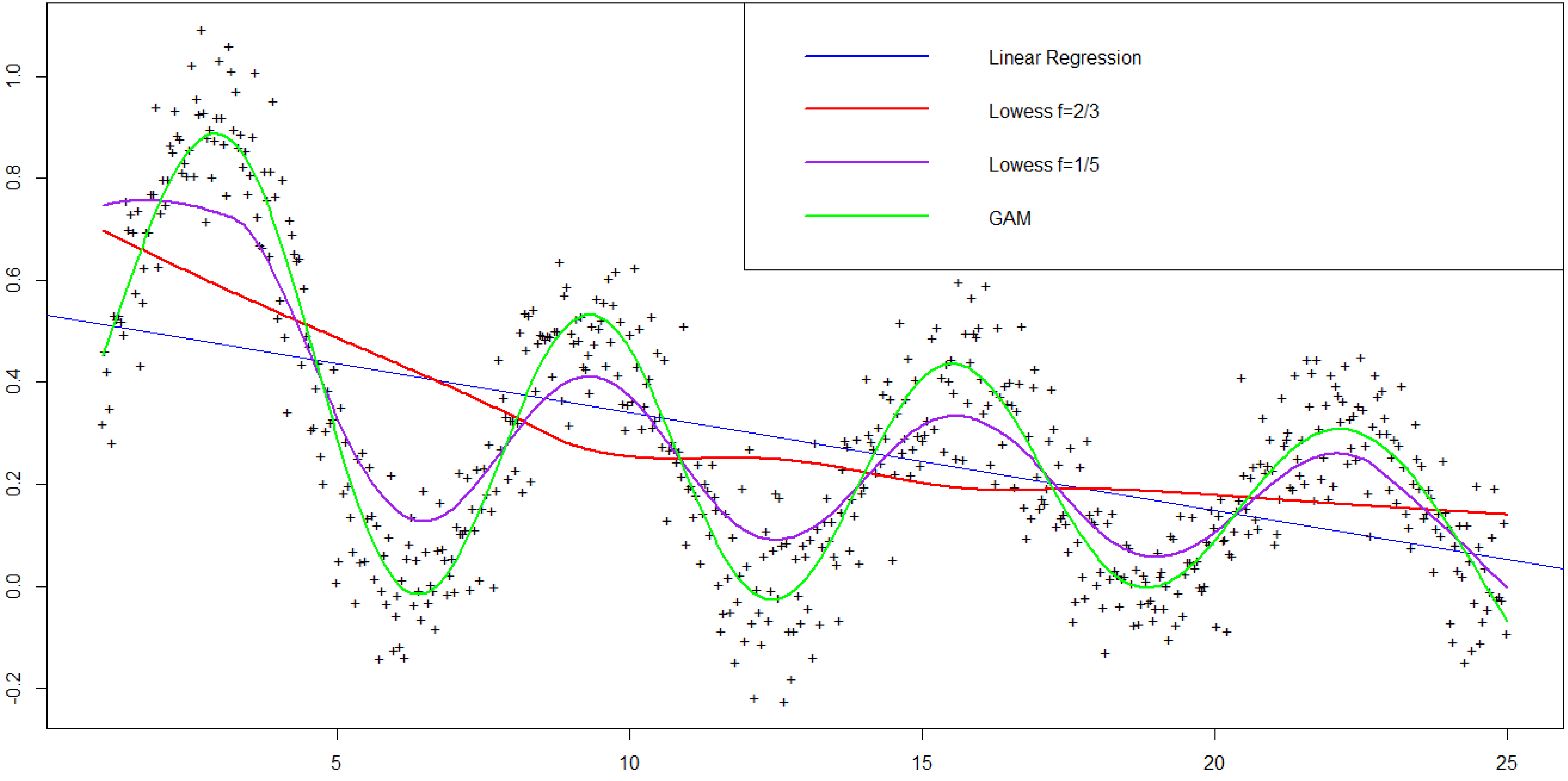

a
anomaly
TCF
b
IMP

Social Science Application

- Workplace violence
- Big data
- Non-linear over time



Jeff's slide 2 - Running Lowess




Thank You!

References

[1] GAM: The Predictive Modeling Silver Bullet Blog post by Kim Larsen: https://multithreaded.stitchfix.com/blog/2015/07/30/gam/

[2] Hastie, Trevor and Tibshirani, Robert. (1986), Generalized Additive Models, Statistical Science, Vol. 1, No 3, 297-318.

[3] Wood, S. N. (2006), Generalized Additive Models: an introduction with R, Boca Raton: Chapman & Hall/CRC

[4] Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. Journal of the American Statistical Association 99, 673–686

[5] Marx, Brian D and Eilers, Paul H.C. (1998). Direct generalized additive modeling with penalized likelihood, Computational Statistics & Data Analysis 28 (1998) 193-20

[6] Sinha, Samiran, A very short note on B-splines, http://www.stat.tamu.edu/~sinha/research/note1.[PDF](/assets/files/gam.pdf)

[7] Michael Alonzo et al 2021 Environ. Res. Lett. Spatial configuration and time of day impact the magnitude of urban tree canopy cooling

[8] https://jeffgill.org/classes/

[9] Openai & ClaoudAI