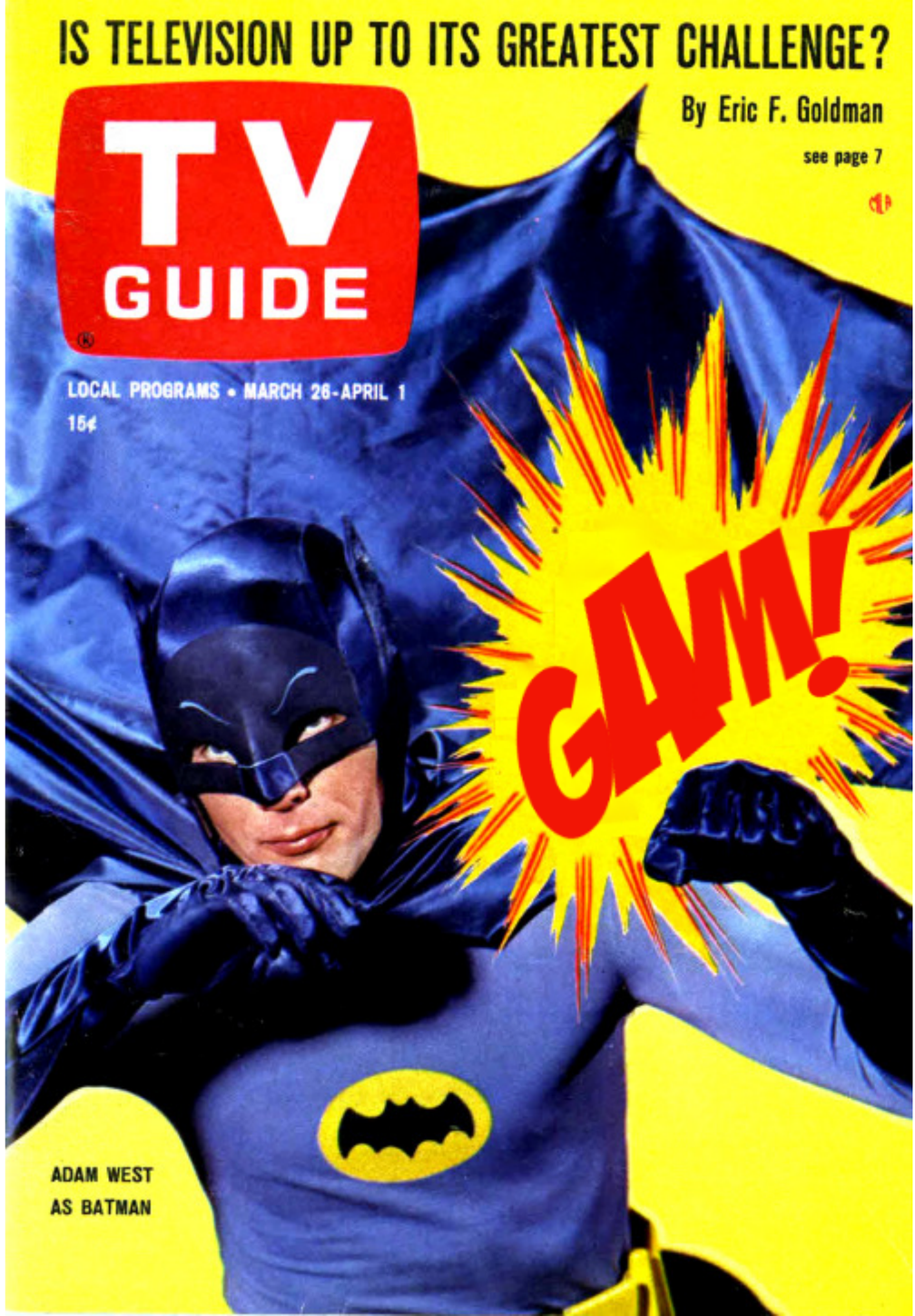
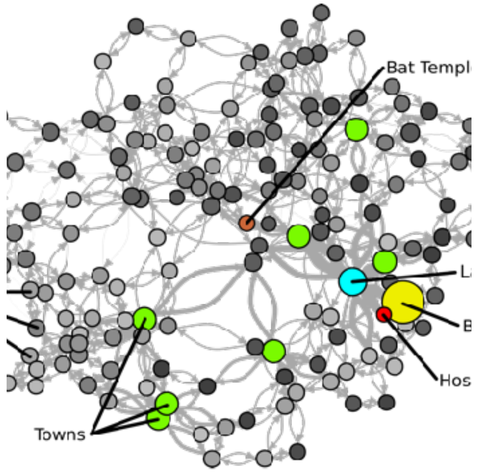
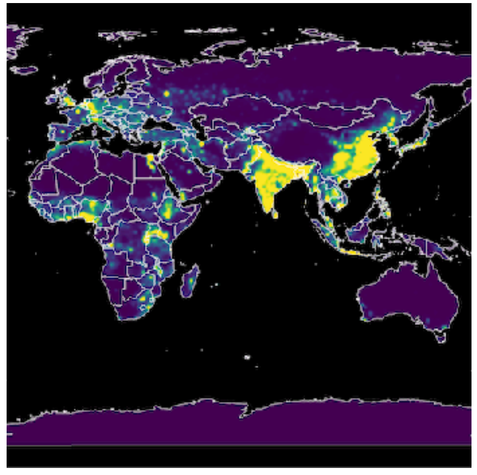
Nonlinear Modeling in R with GAMs:

the magical world of **mgcv**

Noam Ross
@noamross
#nyhackr, 2017-11-15







EcoHealth Alliance

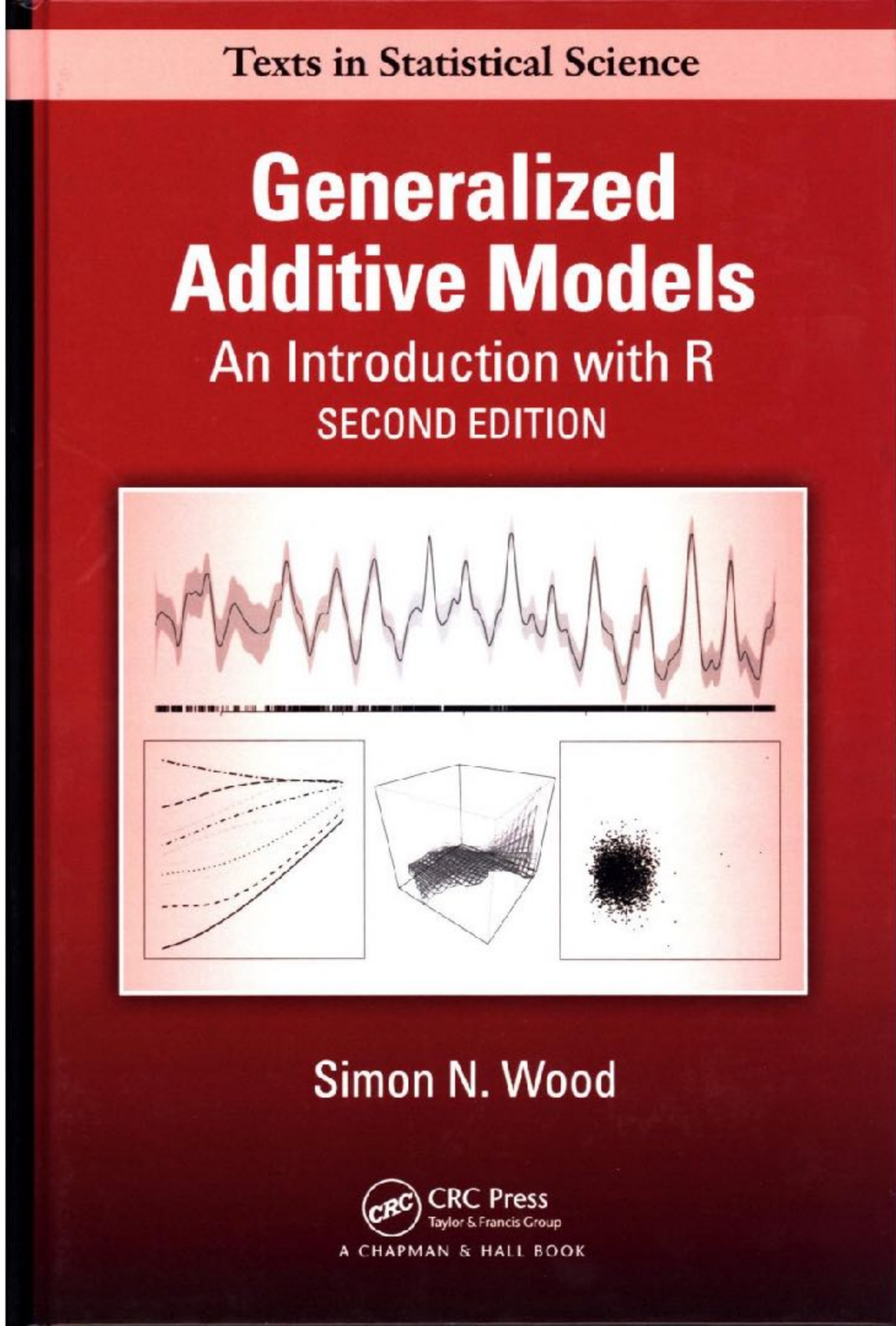
Local conservation. Global health.

Pre-Thanks

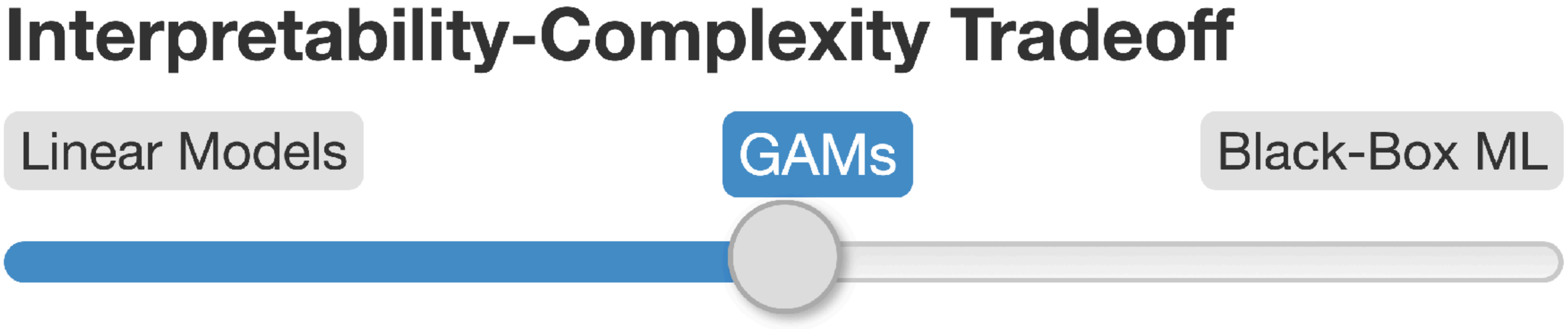
Gavin Simpson
(@ucfagls)

Eric Pedersen
(@ericJpedersen)

David Miller
(@millerdl)



Why Generalized Additive Models?



When to use GAMs

- To predict from complex, nonlinear, possibly interacting relationships
- To understand and make inferences about those relationships
- To control for for those relationships

Not bad at prediction!

Performance in Binary Classification of Direct Mail Customer Acquisition

Model	Validation AUROC	Estimation Time	Scoring Time
Random forest	0.809	6.39	39.38
GAM, lambda=0.6	0.807	3.47	0.52
GAM, estimate lambdas	0.815	42.72	0.29
GAM, estimate lambdas, extra shrinkage	0.814	169.73	0.33
SVM	0.755	13.41	1.12
Linear logit	0.800	0.1	0.006
KNN with K=100	0.800	NA	3.34

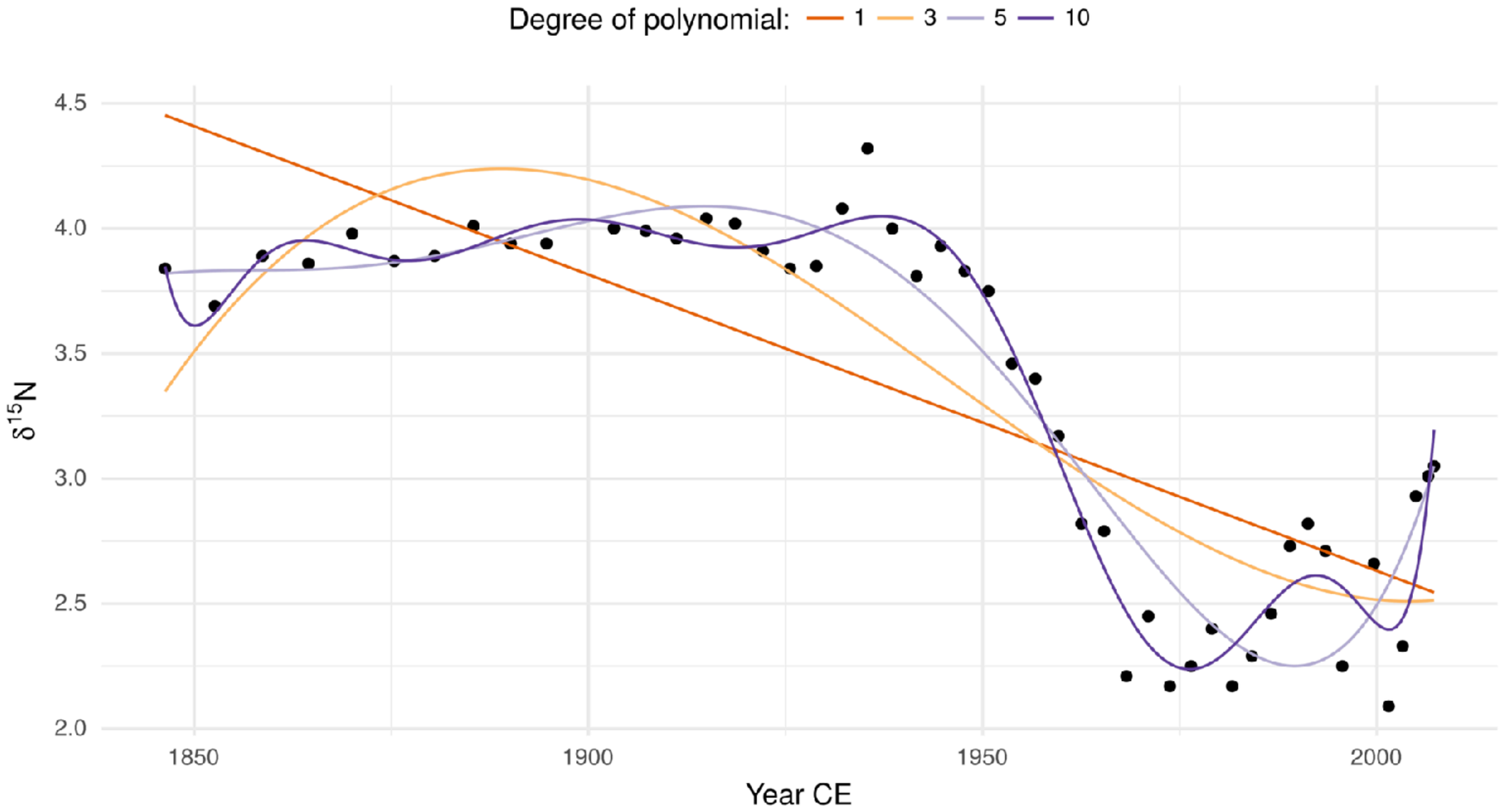
From Kim Larsen @ Stitchfix: https://github.com/klarsen1/gampost

A Thimbleful of Theory

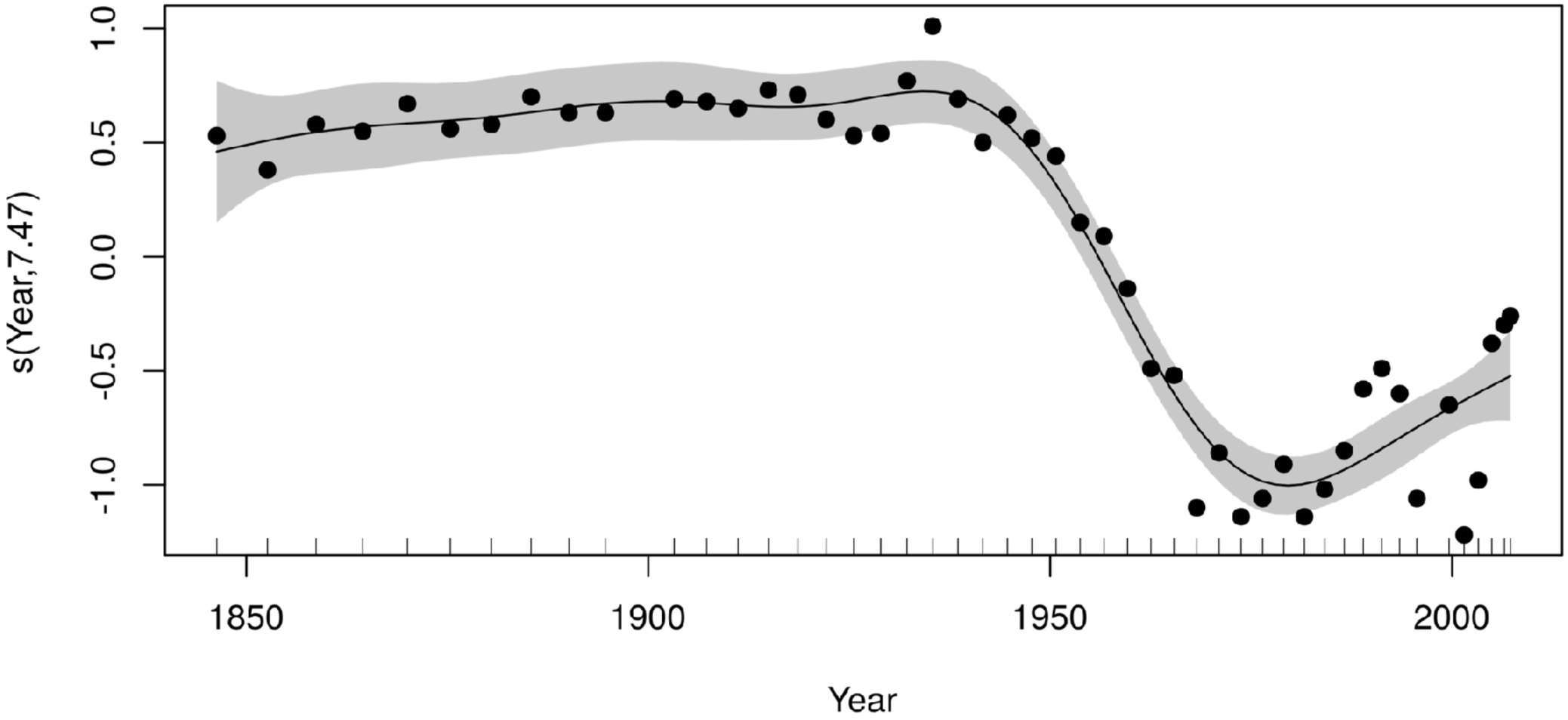
What are GAMs?

- **Generalized:** Can handle many distributions of normal, binomial, count, or other data
- **Additive:** terms simply add together, but terms themselves are not linear
- **Model:** Model

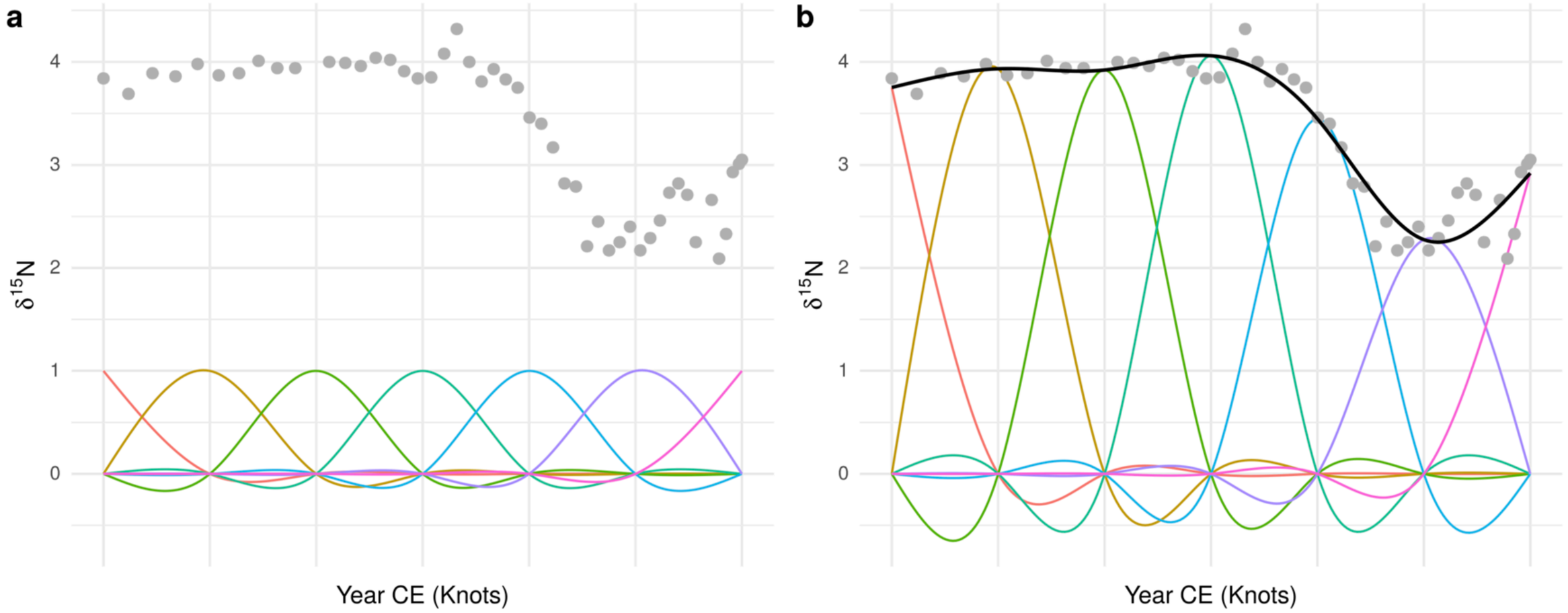
Going from Linear to Additive



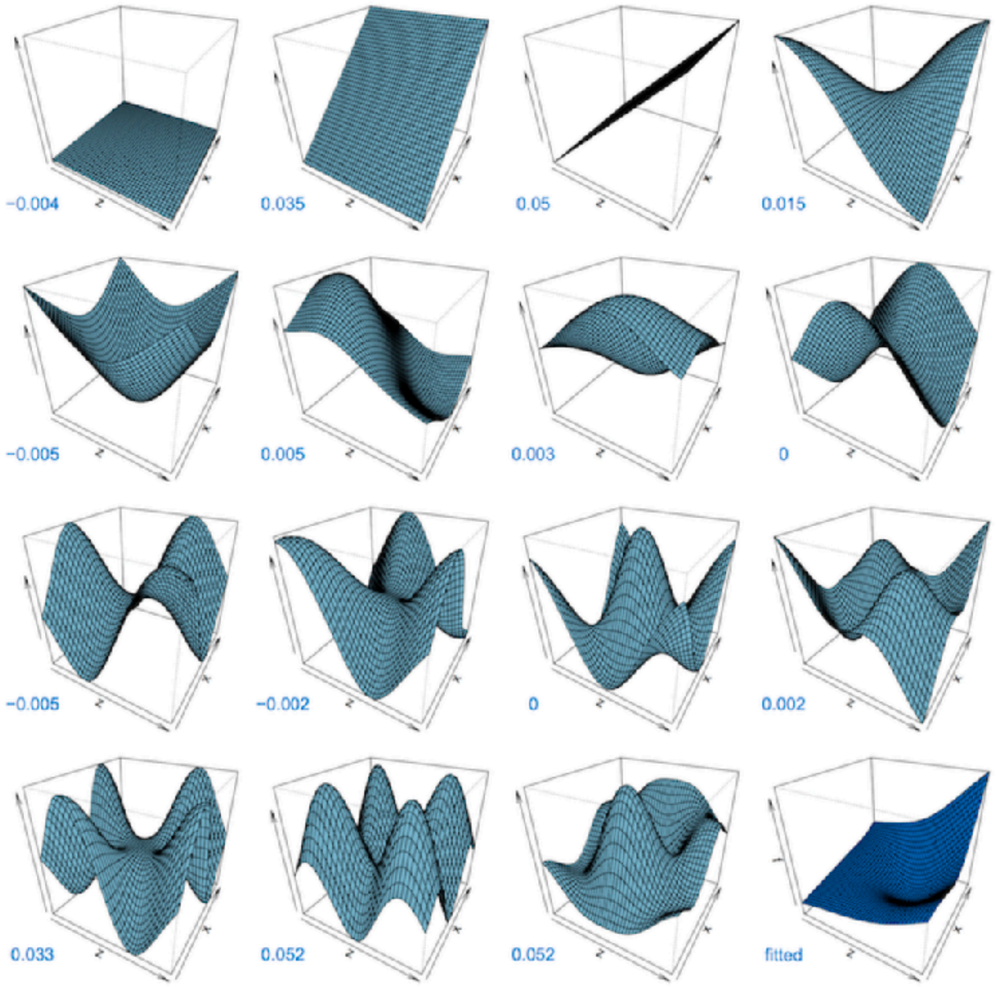
Going from Linear to Additive



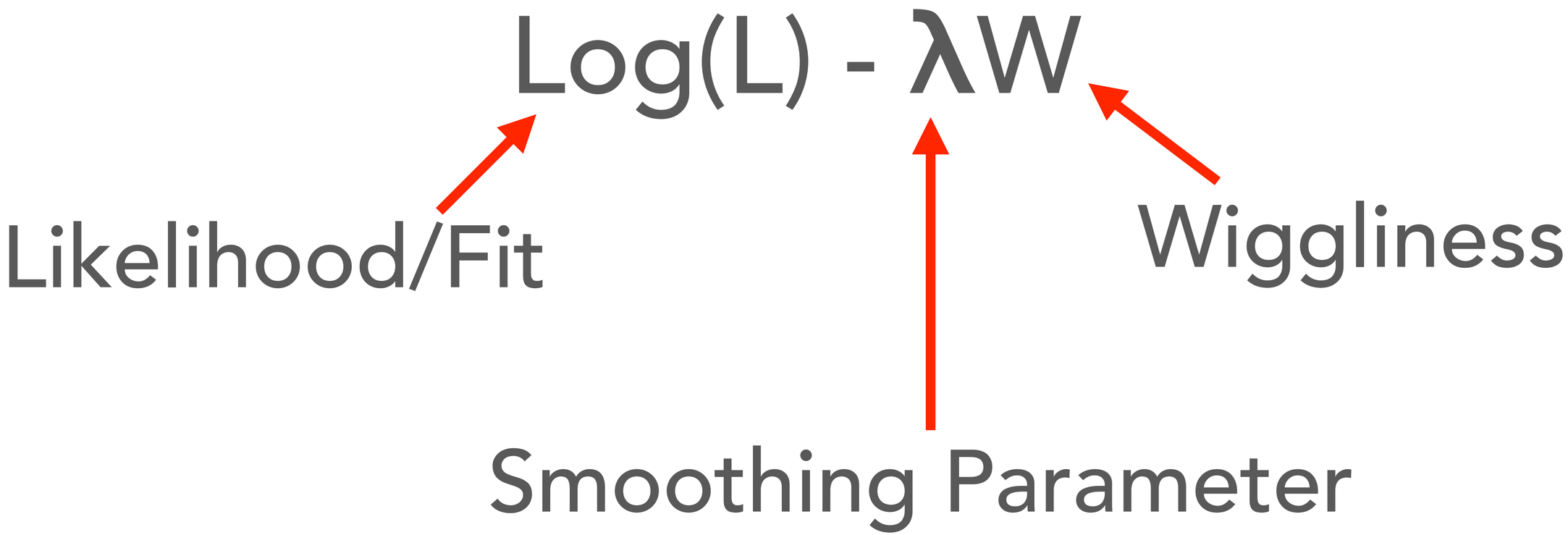
GAM Smooths are made of *basis functions*



Basis functions can have 1, 2, or more dimensions

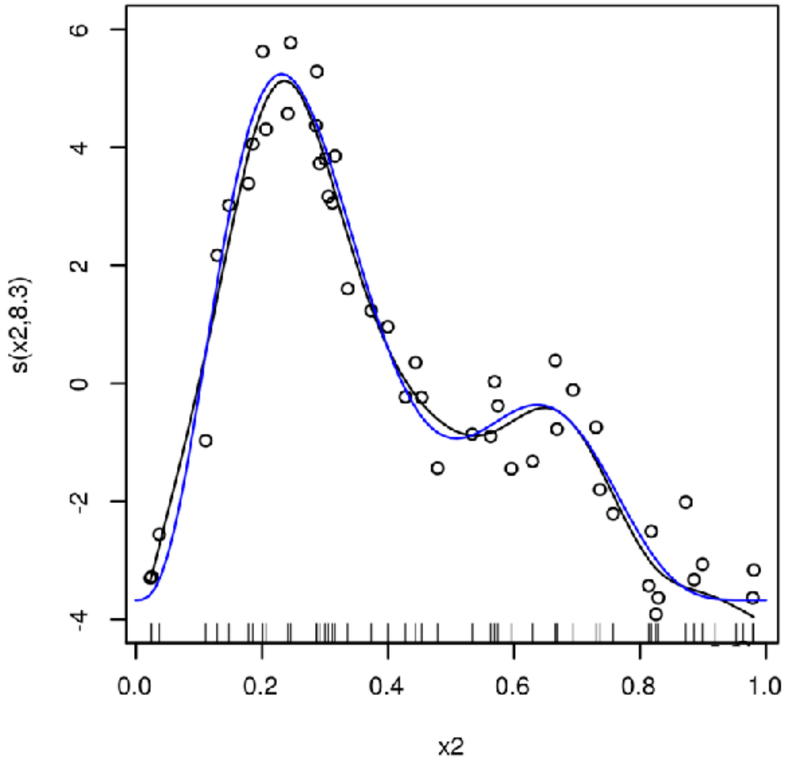


Optimizing Wiggliness

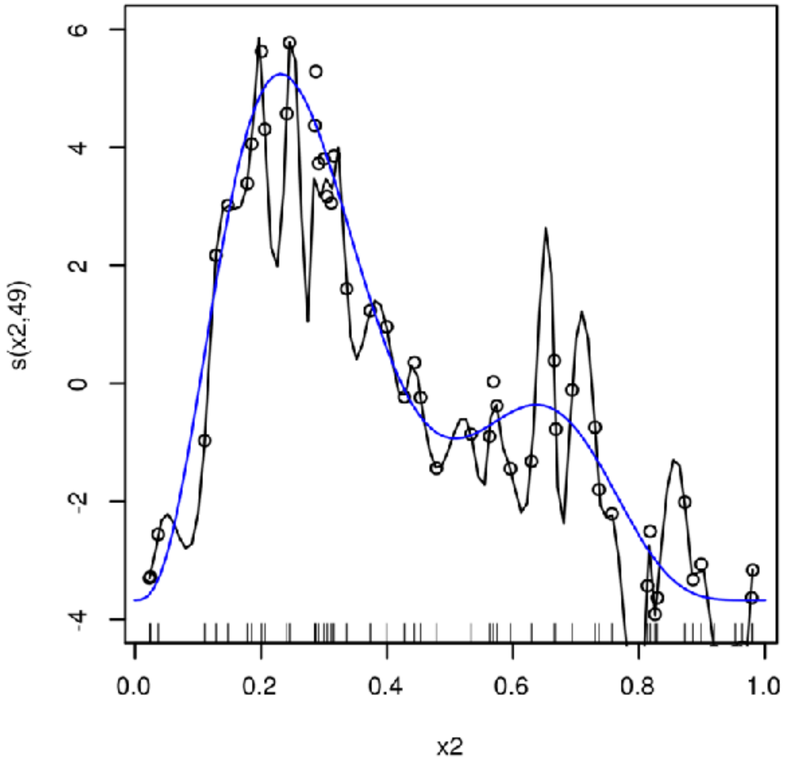


Picking a Smoothing Parameter

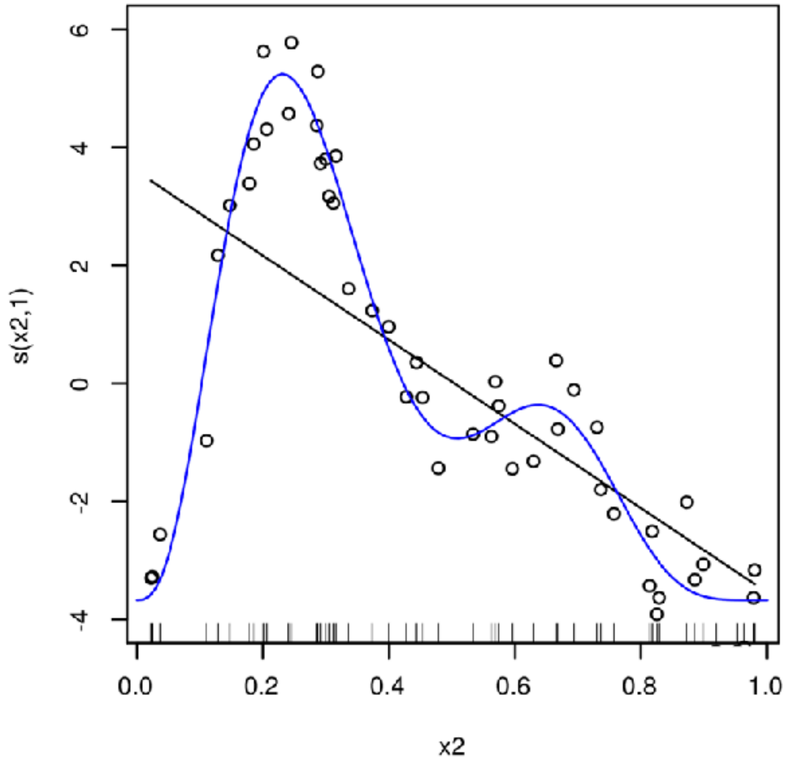
λ = just right



$\lambda = 0$

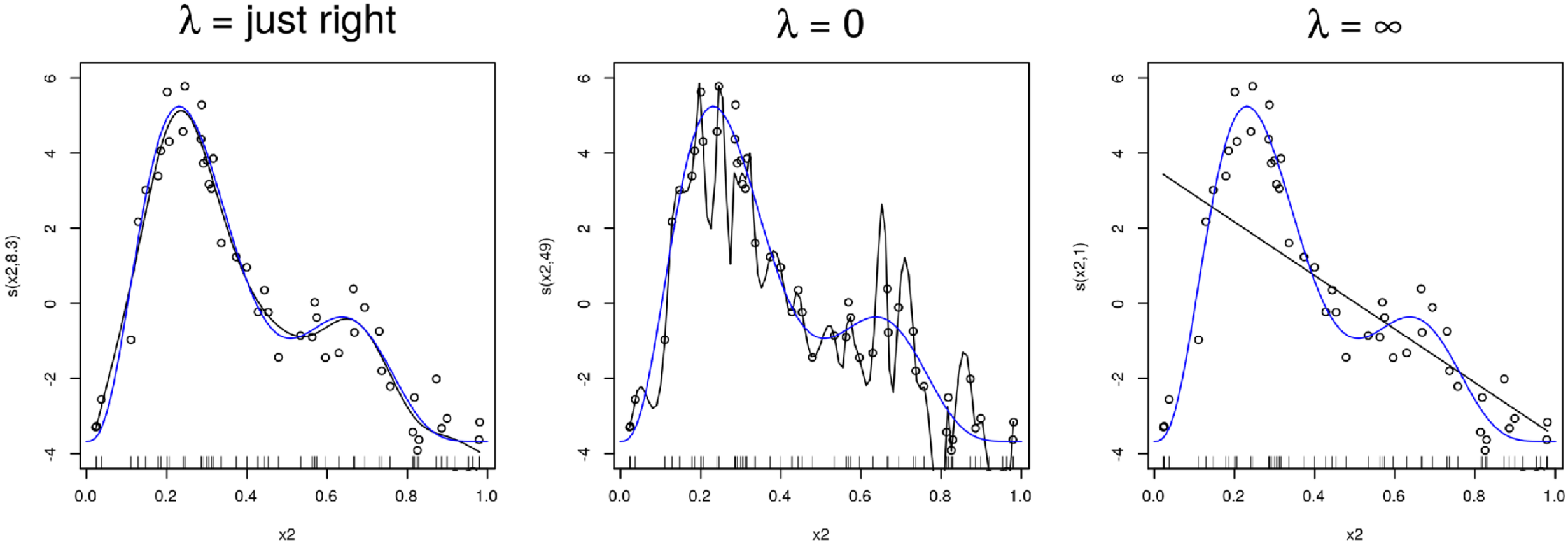


$\lambda = \infty$




More Theory

Picking a Smoothing Parameter



(This is automated in **`mgcv`**, phew!)

A Smidgen of Syntax

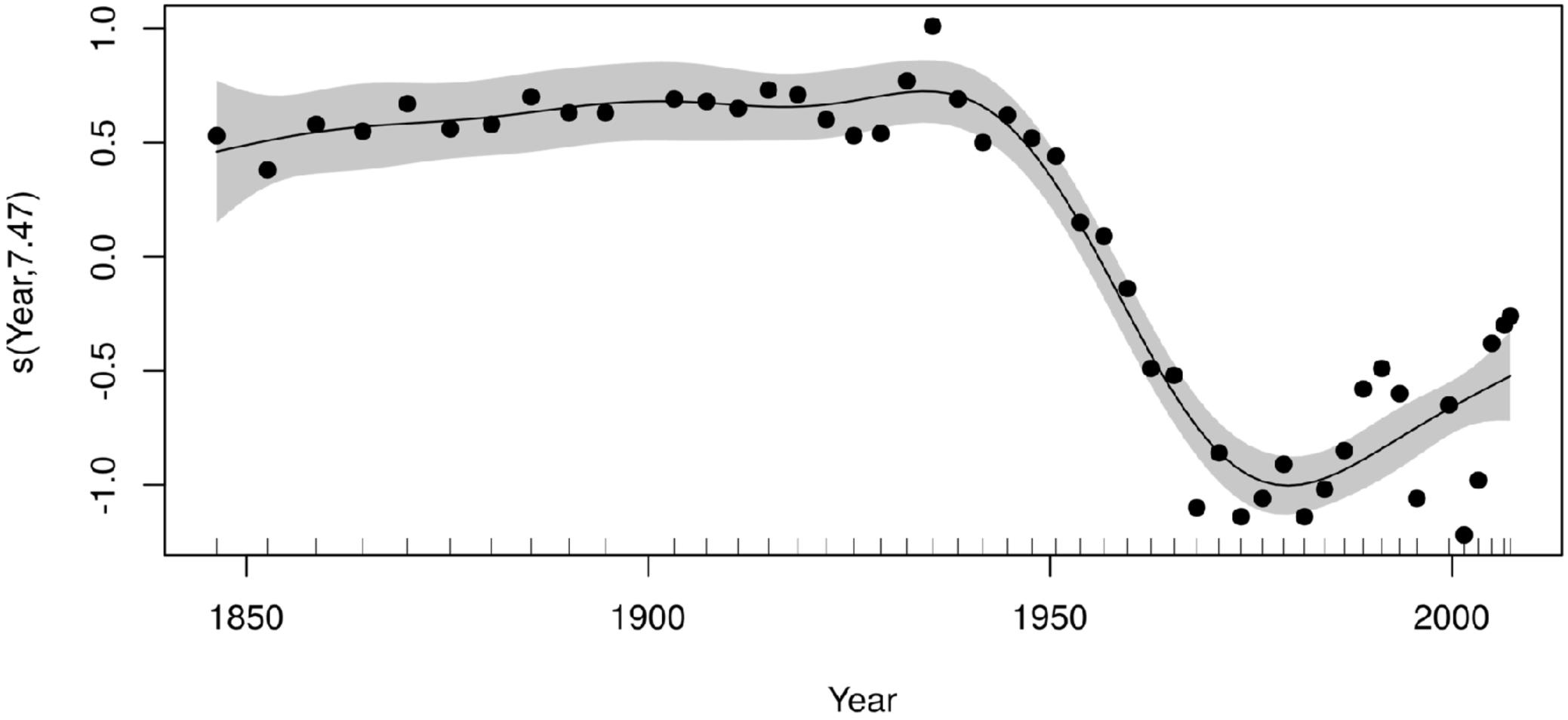
Fitting a GAM in R

```
lm(y ~ x1 + x2, data=data)

glm(y ~ x1 + x2, data=data, family=binomial)

library(mgcv)
gam(y ~ x1 + s(x2),     # model formula
    data=data,          # your data
    family = gaussian   # or something more exotic
    method = "REML")     # how to pick λ
```

The GAM Formula

```
y ~ x1 +                # linear terms
      s(                # smooth terms:
        x2,             #   variable
        bs = "tp",      #   the kind of basis function
        k  = 10,        #   how many basis functions
        ...)            #   other complex and
                        #   basis-specific stuff
```

Going from Linear to Additive



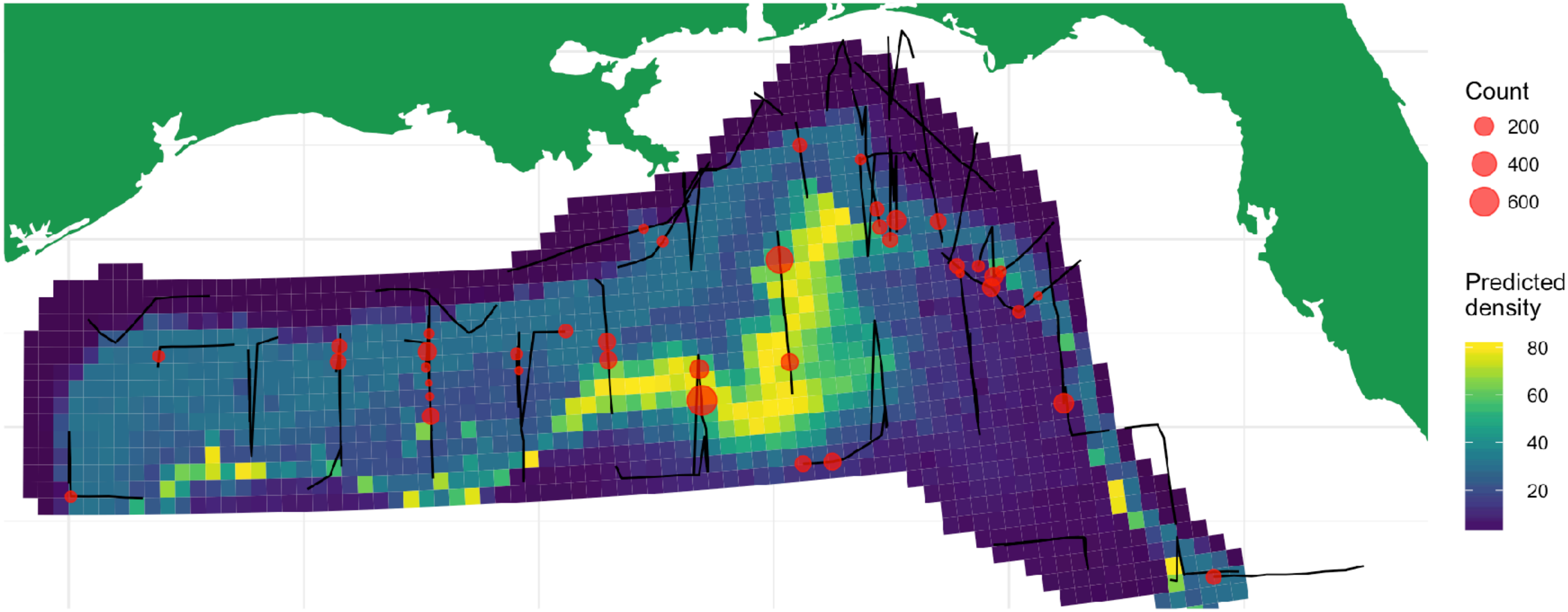
The GAM Formula in 2D

```
y ~ s(x1) + s(x2) # Two additive smooths

y ~ s(x1, x2)     # 2D smooth/interaction

y ~ te(x1, x2)    # 2D smooth, two wigglinesses

y ~ te(x1) + te(x2) + ti(x1, x2)
# 2D smooth, two wigglinesses, interaction as
# a separate term
```


SMOOTHS
IN
SPACE

Smooths in Space

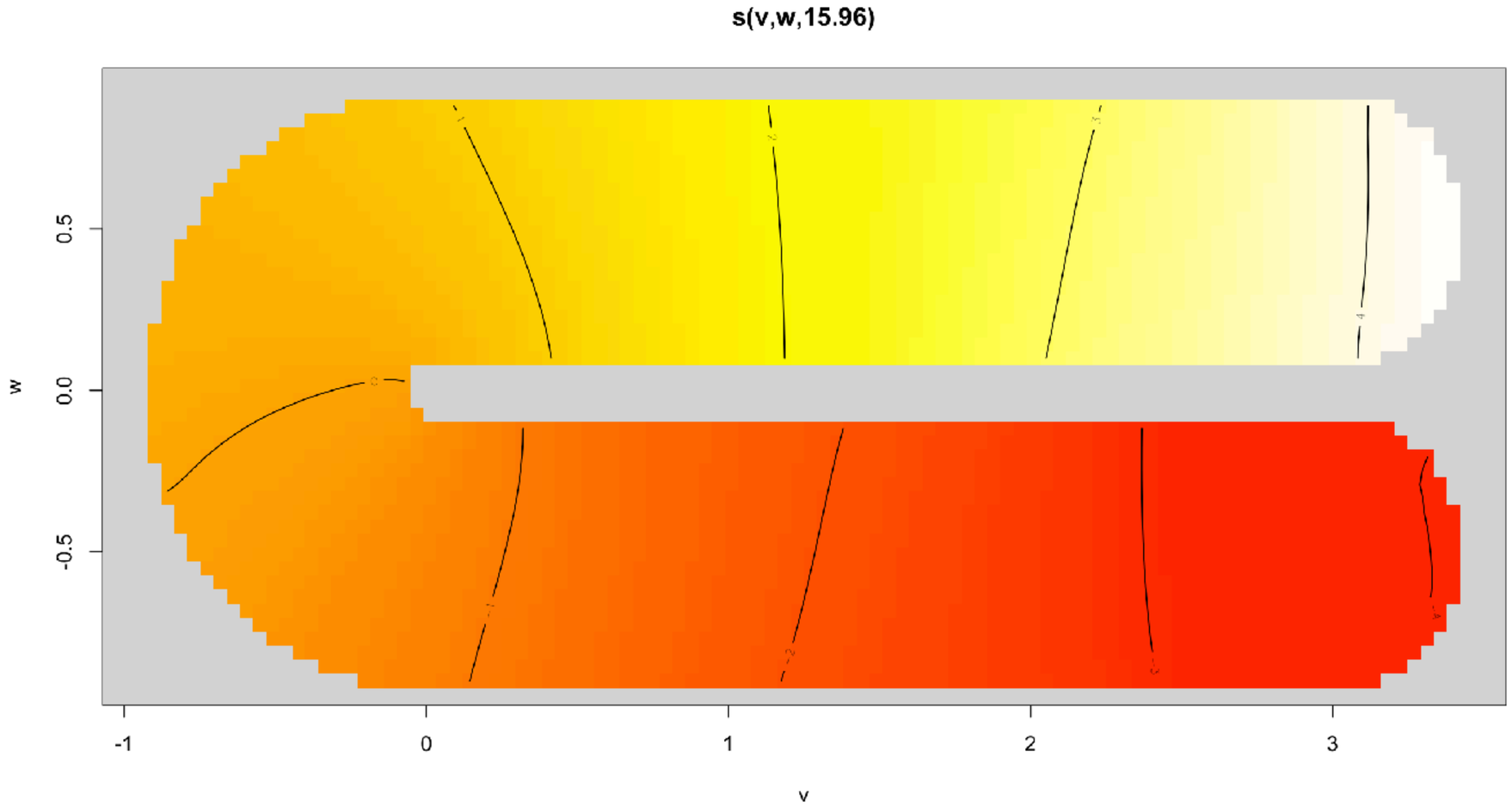
```
gam(d ~ s(x, y) + s(depth), data=dolphin_observations)
```



A Bevy of Basis Functions

Slippery Smooths: "Soap Films"

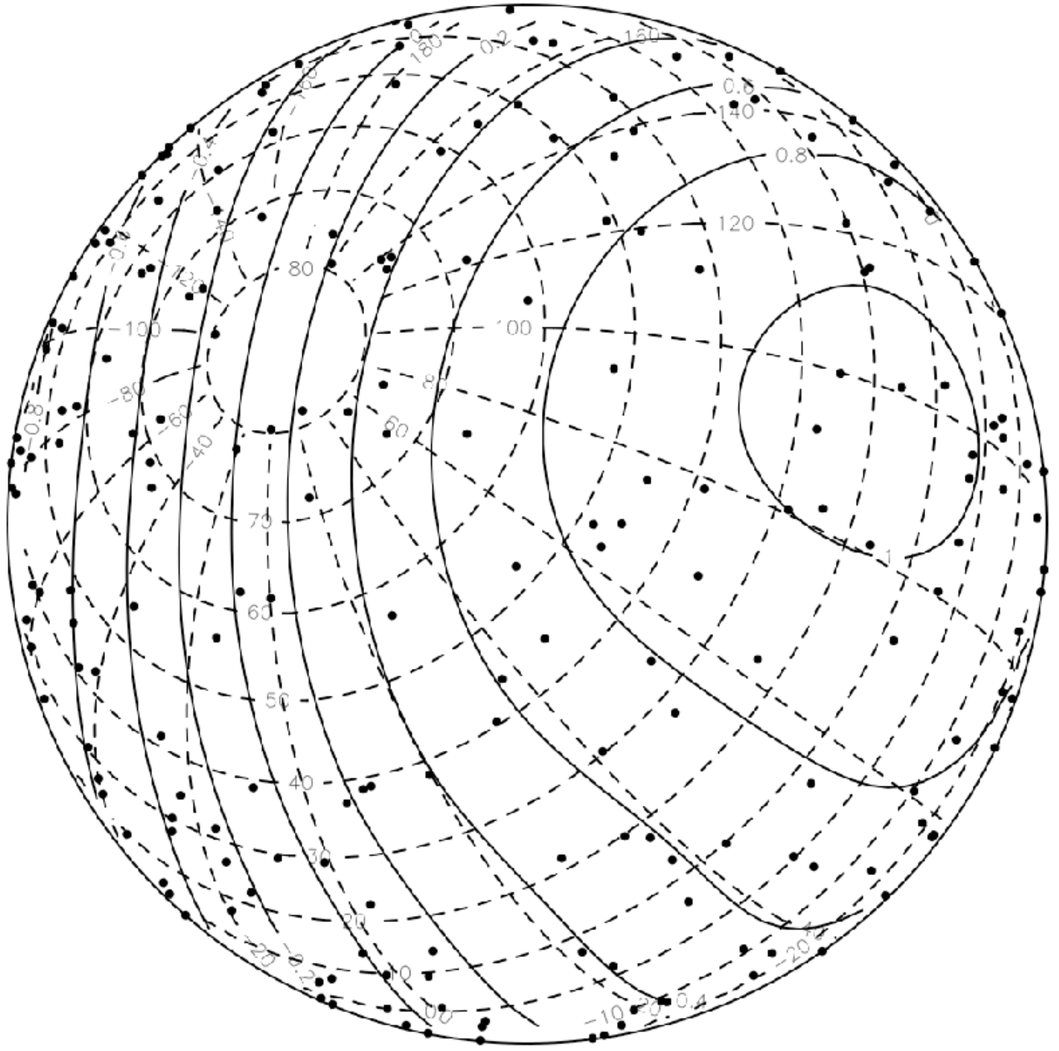
```
gam(d ~ s(x, y, bs="so", xt = list(bnd=my_boundary),
    data=data)
```



Smooths that Make the World Go Round

Spline-on-a-Sphere

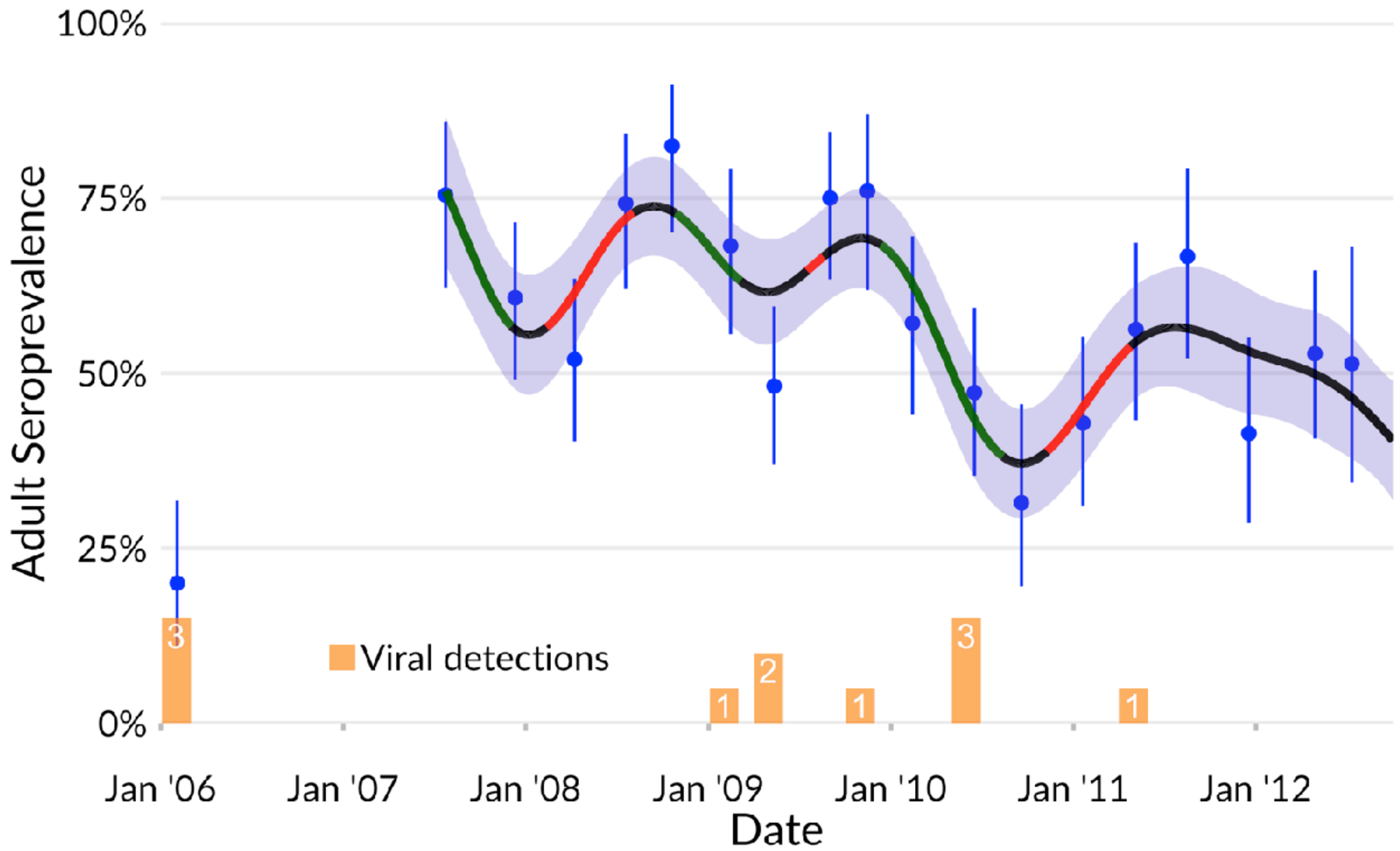
```
gam(y ~ s(latitude,
          longitude,
          bs="sos"),
    data=dat)
```




SMOOTHS
IN
TIME

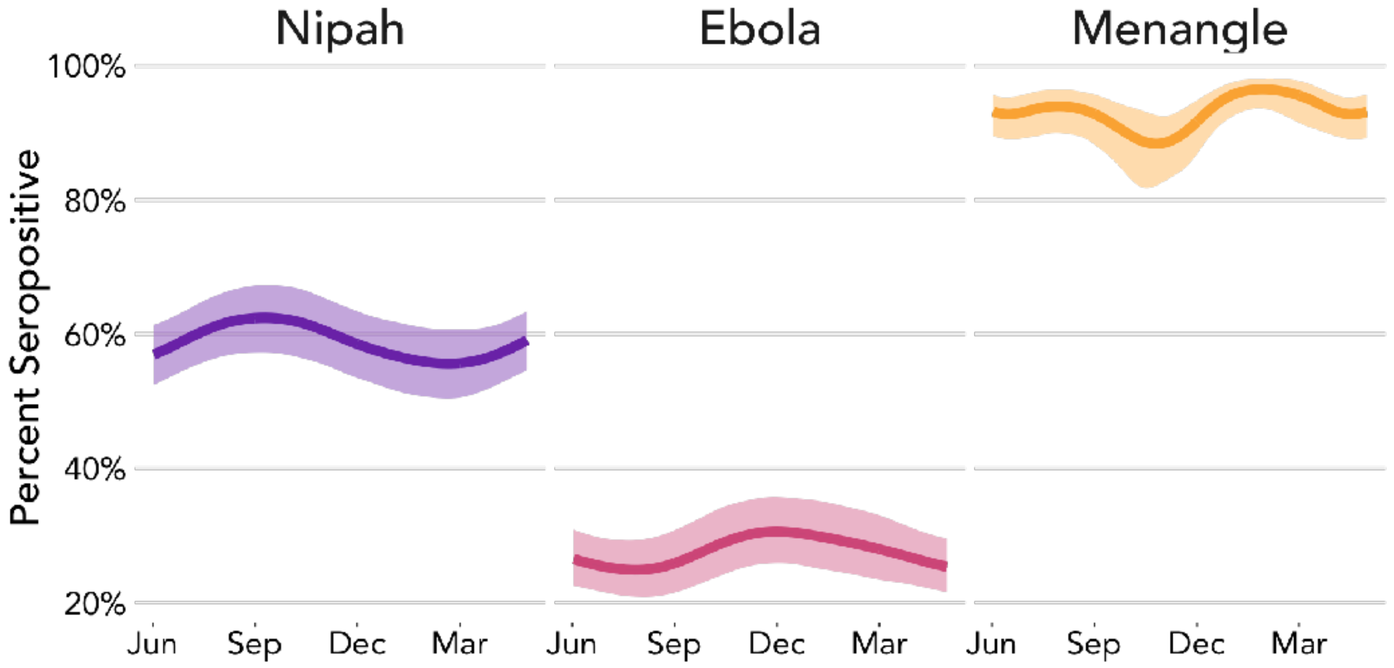
Gaussian Process Smooths

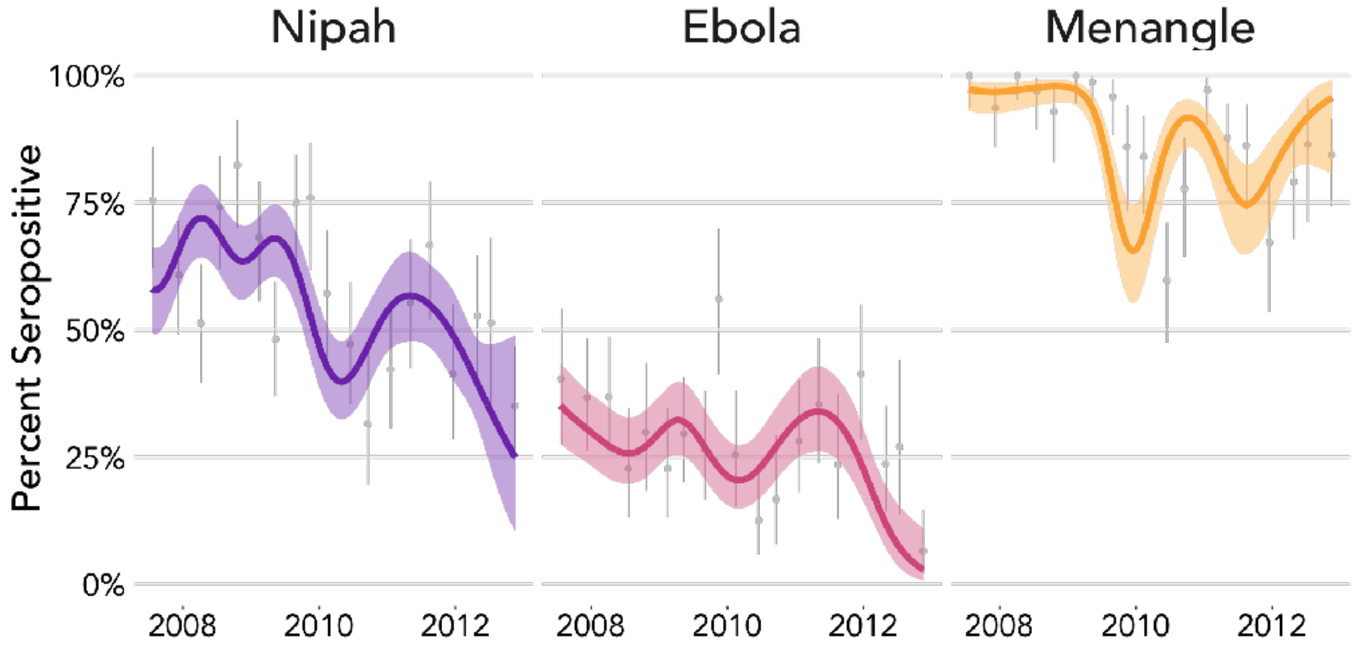
```
gam(y ~ s(time, bs= "gp"), data=bat_antibodies,
```



Cyclic Smooths

```
gam(y ~
      s(time, bs= "gp") +
      s(month, bs = "cc"),
      data=bat_antibodies,
      family = "binomial")
```

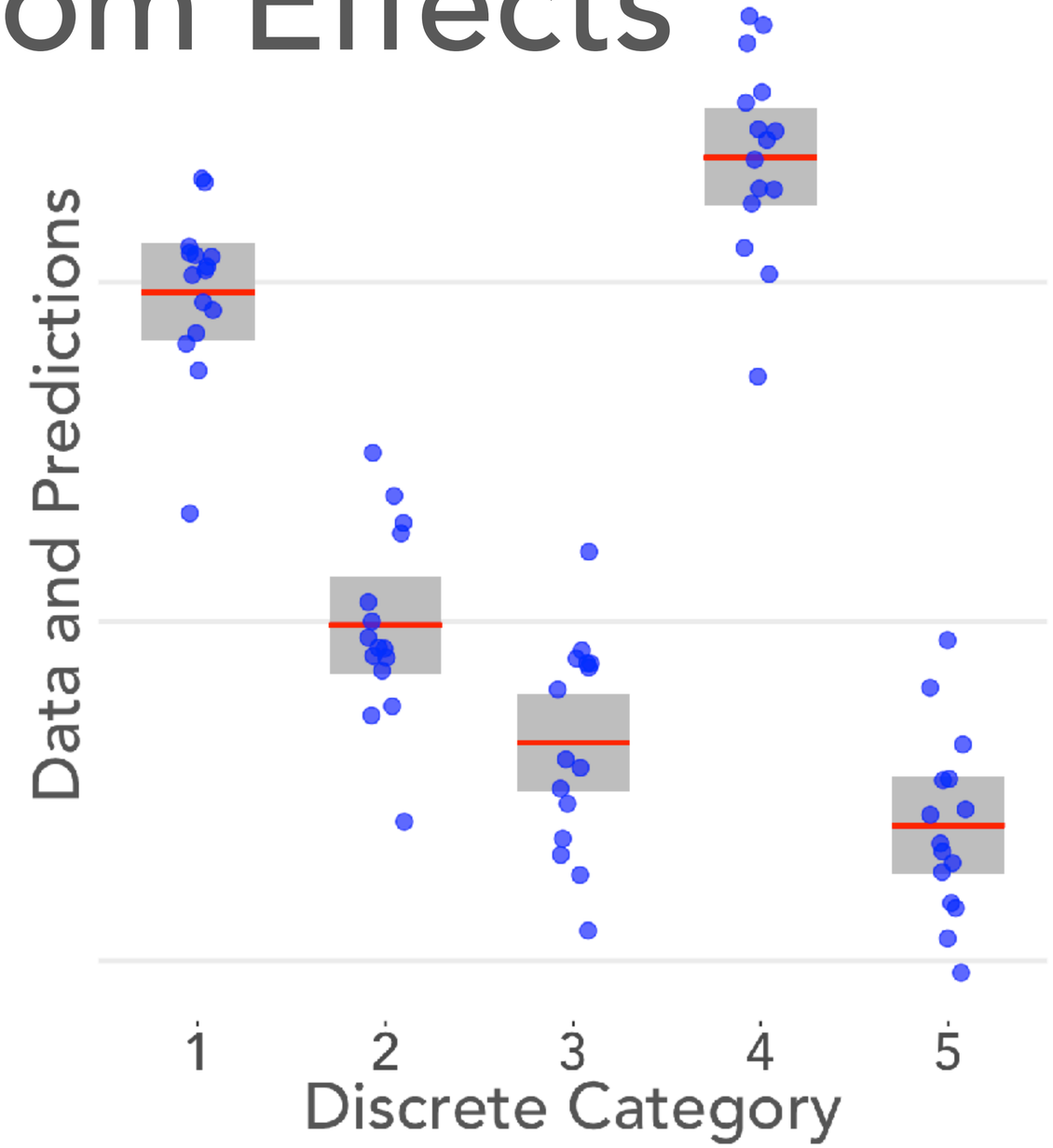




Smooths that Ain't Smooth

Discrete Random Effects

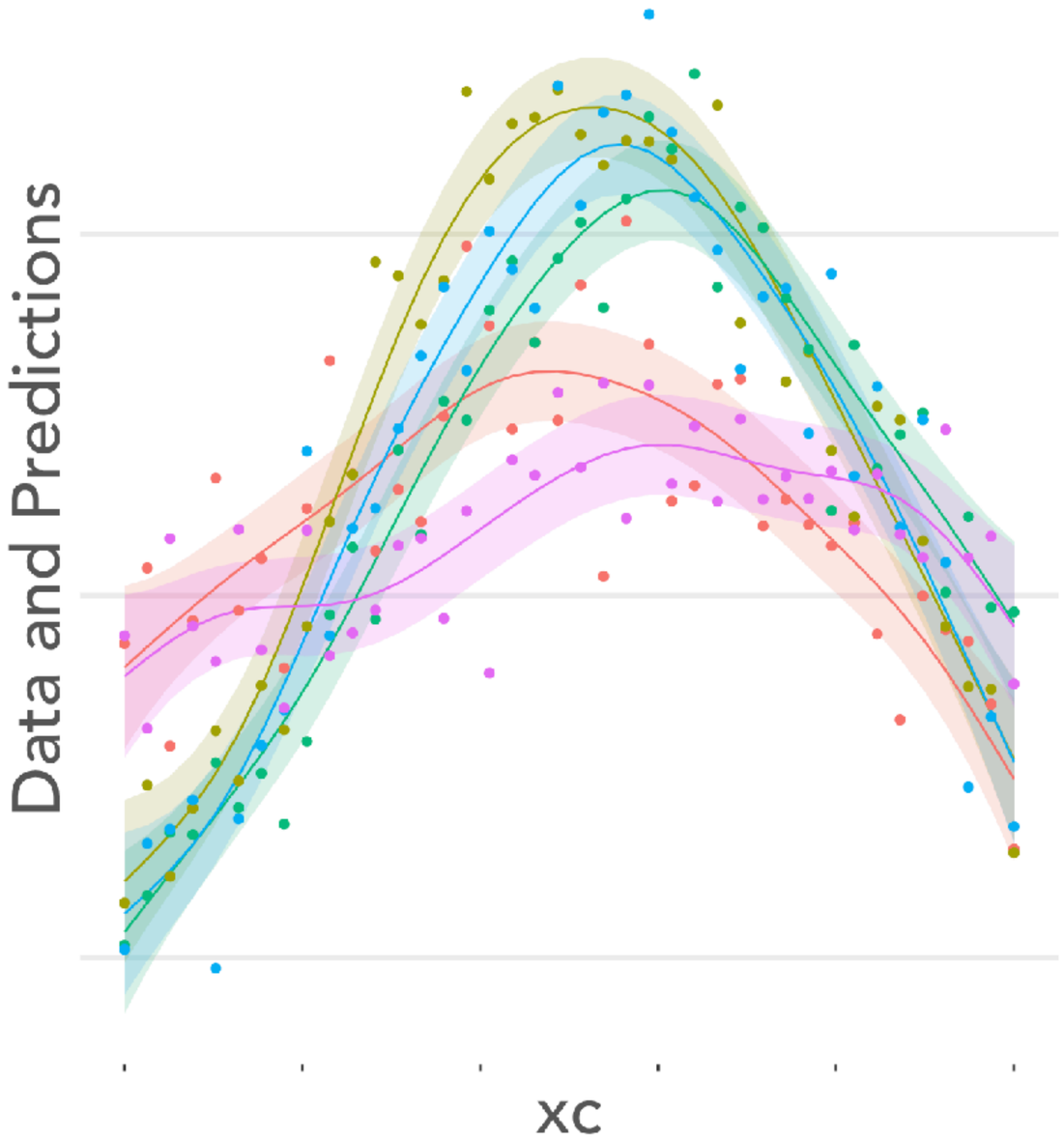
```
gam(y ~ s(x, bs = "re"),
    data=dat)
```



Factor-Smooth Interactions

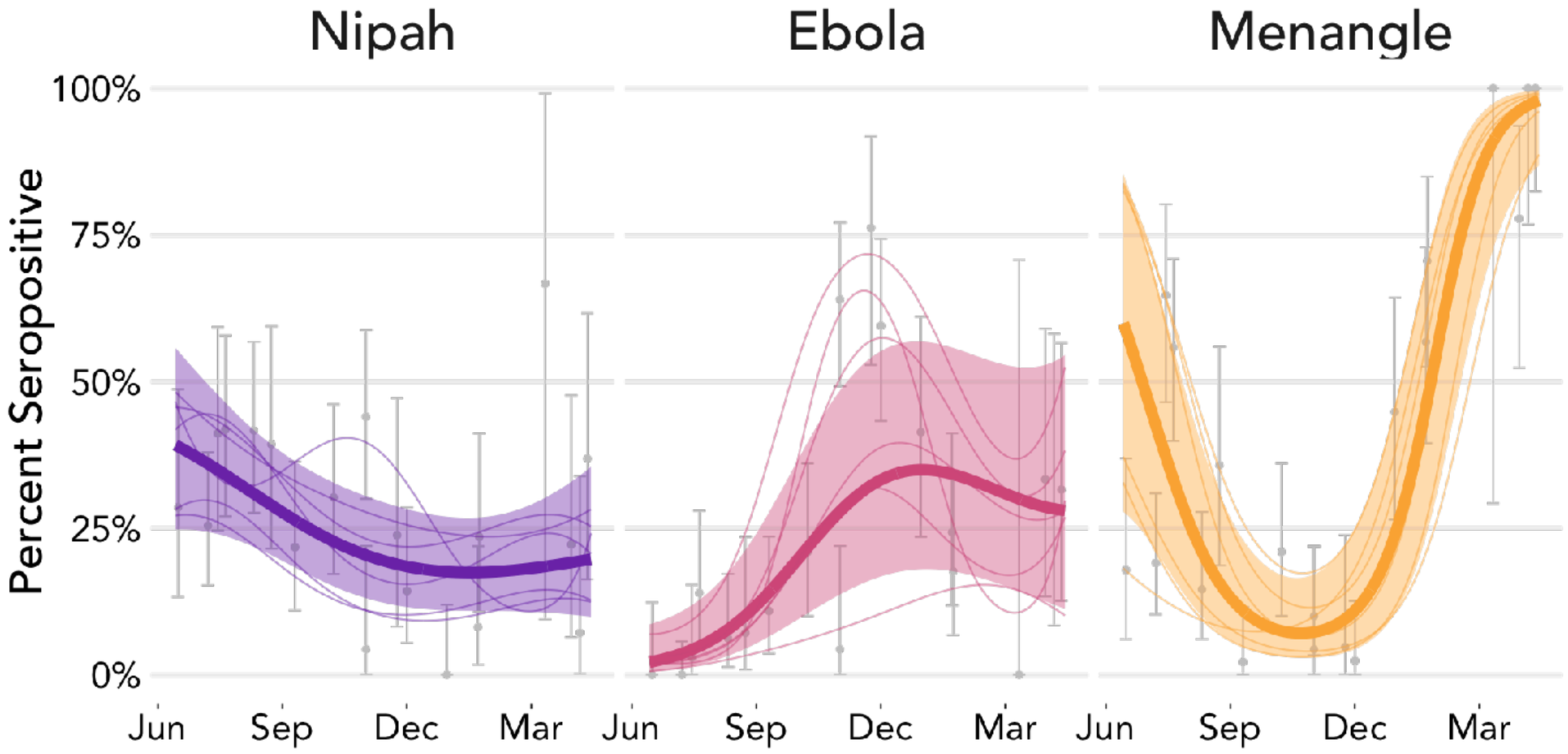
(or , different slopes for different folks)

```
gam(y ~ s(xc, xf, bs = "fs"),
    data=dat)

gam(y ~ te(xc, xf,
           bs = c("tp", "re"),
    data=dat)
```



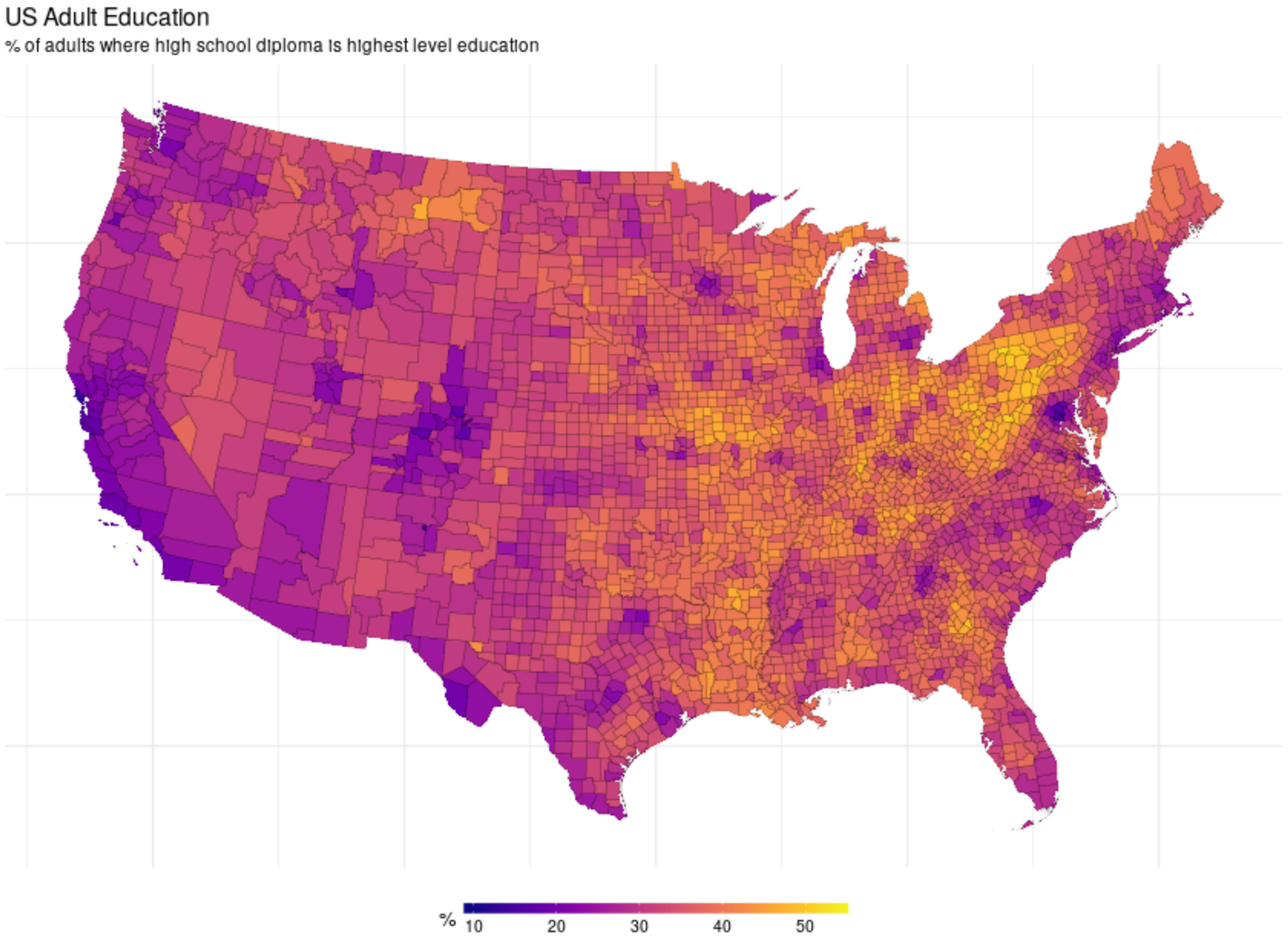
Different Slopes for Different Folks

```
gam(y ~ te(xc, bs="gp") +
        ti(xc, xf, bs = c("gp", "re"), data=dat)
```



Markov Random Fields

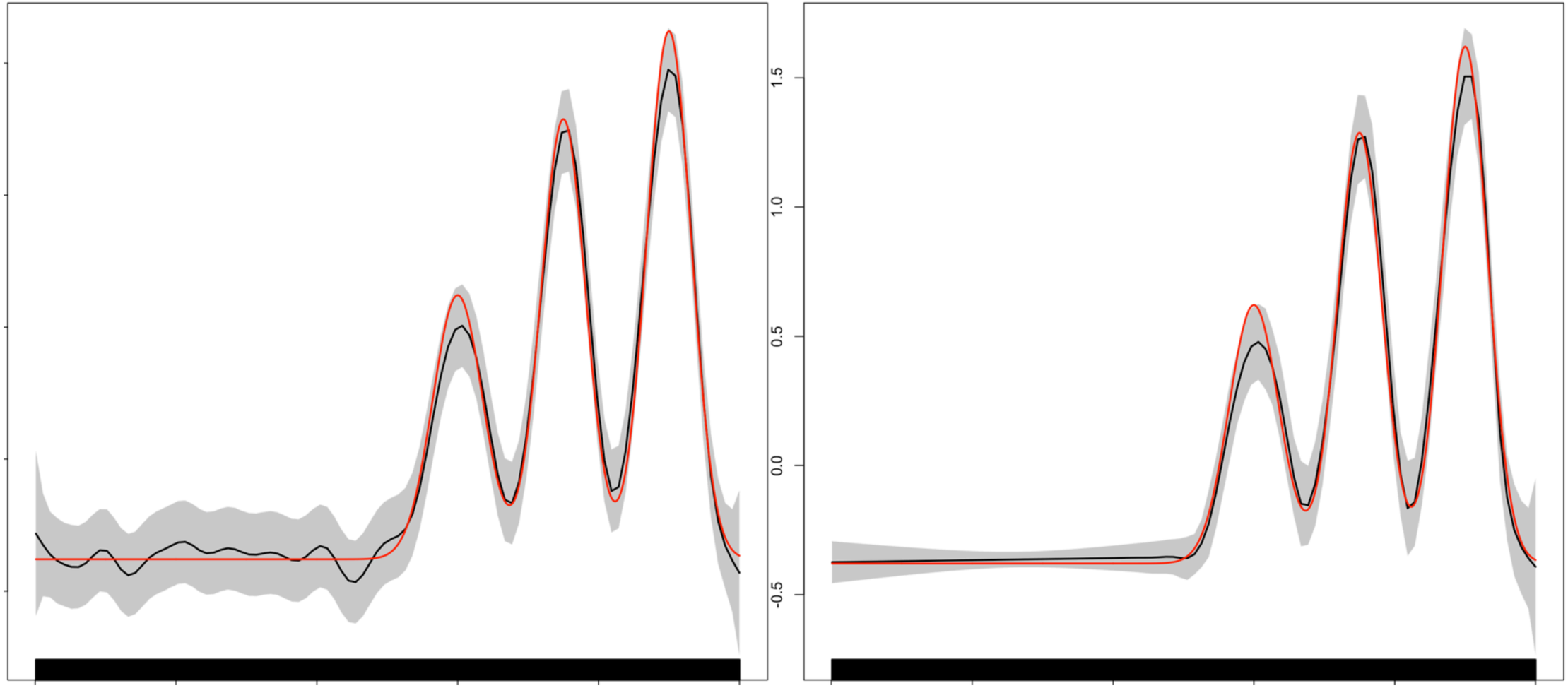
```
gam(y ~ s(x, bs = "mrf",
          xt = list(
            nb = nb
          )),
    data=dat)
```



Adaptive Smooths

(Smooths in your Smooths)

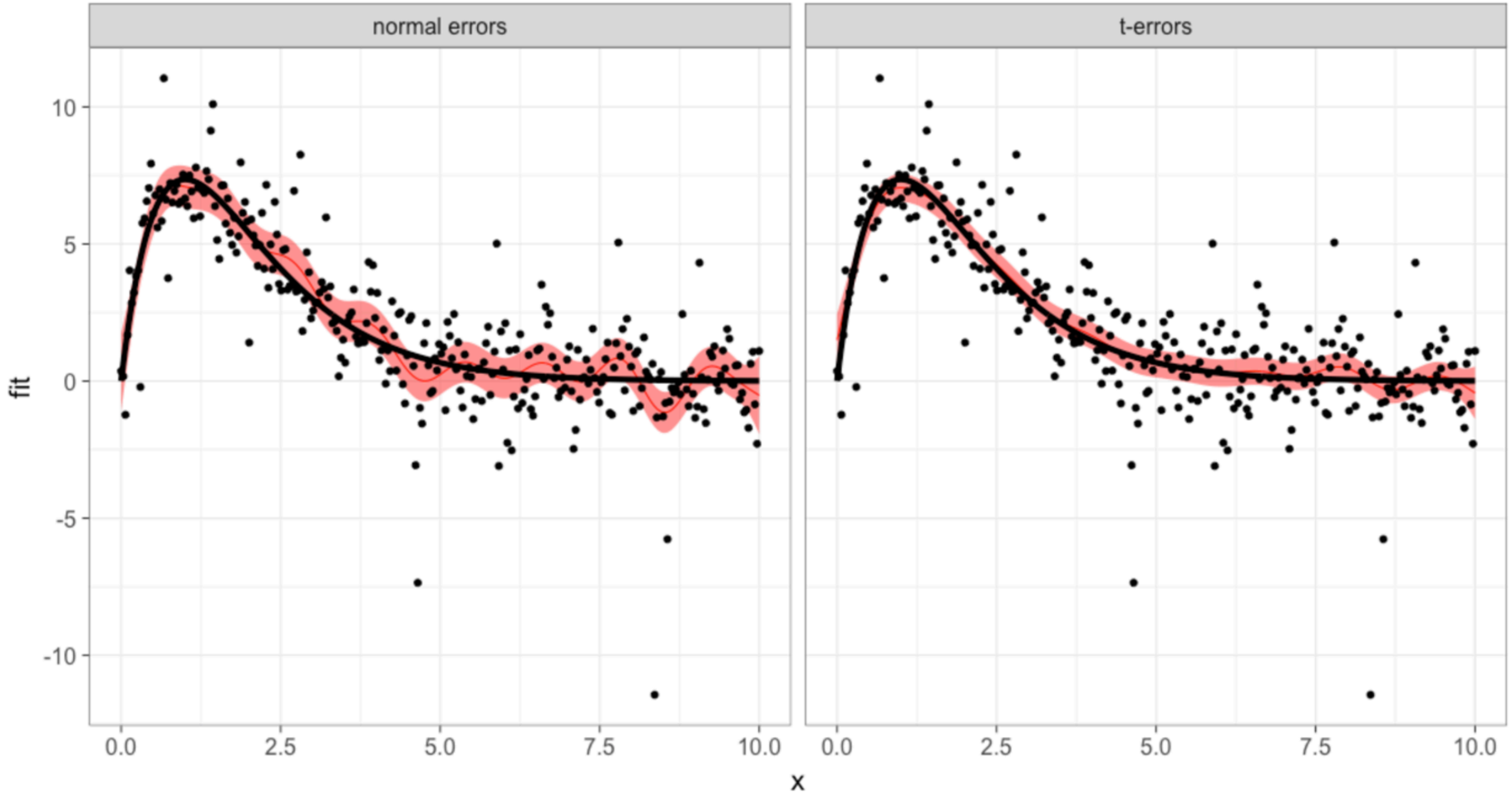
```
gam(y ~ s(x, bs= "ad"), data=data)
```



A Plethora of Probability Distributions

Data with Outliers: Student's T

```
gam(y ~ s(x), data=fat_tailed_data, family = scat)
```

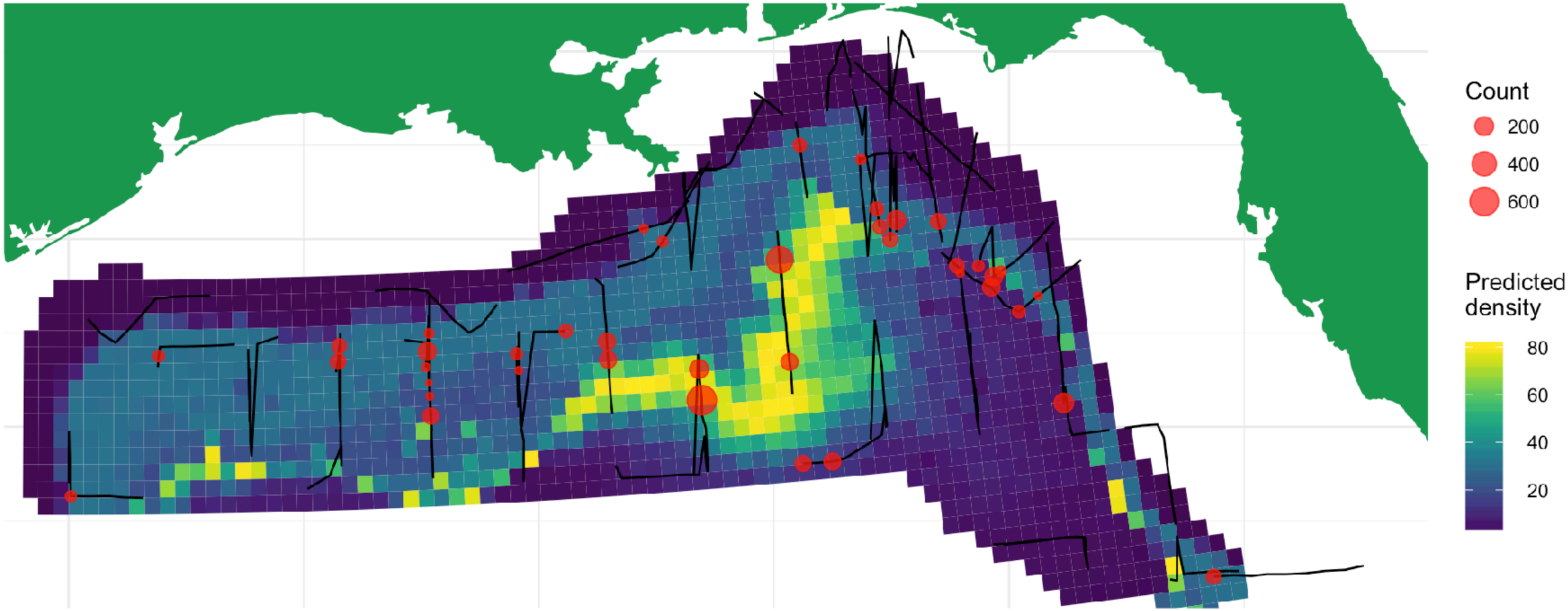


Count Data

```
gam(y ~ x, data=dat,
    family = poisson)

gam(y ~ x, data=dat,
    family = negbin)

gam(y ~ x, data=dat,
    family = tw)
```

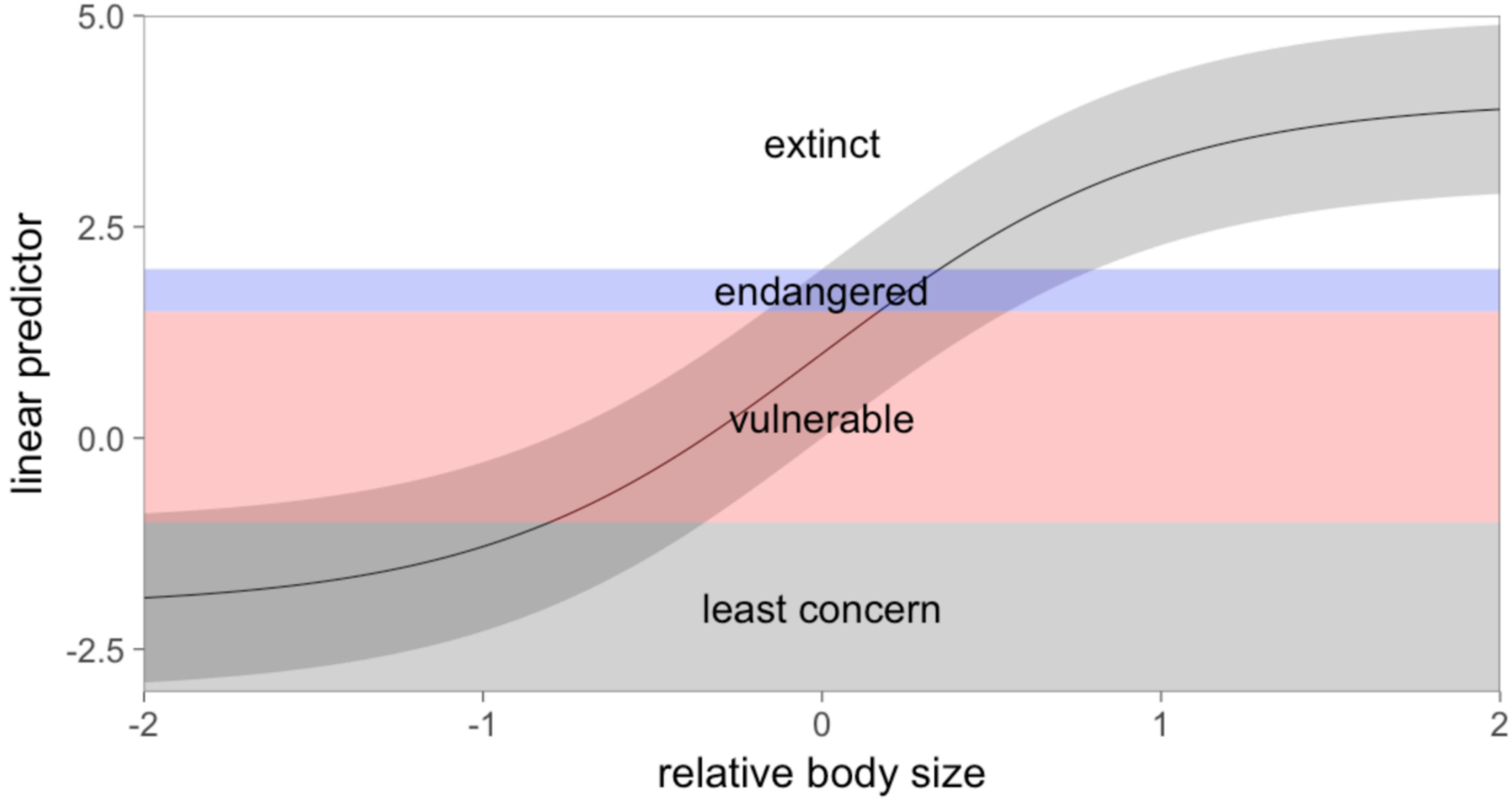
Count Data

```
gam(d ~ s(x, y, bs="tp") + s(depth), data=dolphin_observations,
     family = tw)
```



Ordered Categorical Data

```
gam(ordered_factor ~ s(x), data=data, family = ocat)
```



Multiple Output Variables

Unordered Categories: Multinomial

```
gam(list(category ~ s(x1) + s(x2),
                  ~ s(x1) + s(x2)),
    data= model_dat, family=multinom(K=2))
```

Multiple Continuous Outputs: Multivariate Normal

```
gam(list(category ~ s(x1) + s(x2),
                  ~  s(x1) + s(x3)),
    data= model_dat, family=mvn(K=2))
```

And More!

Survival data: Cox Proportional hazards (`family = **cox.ph**`)

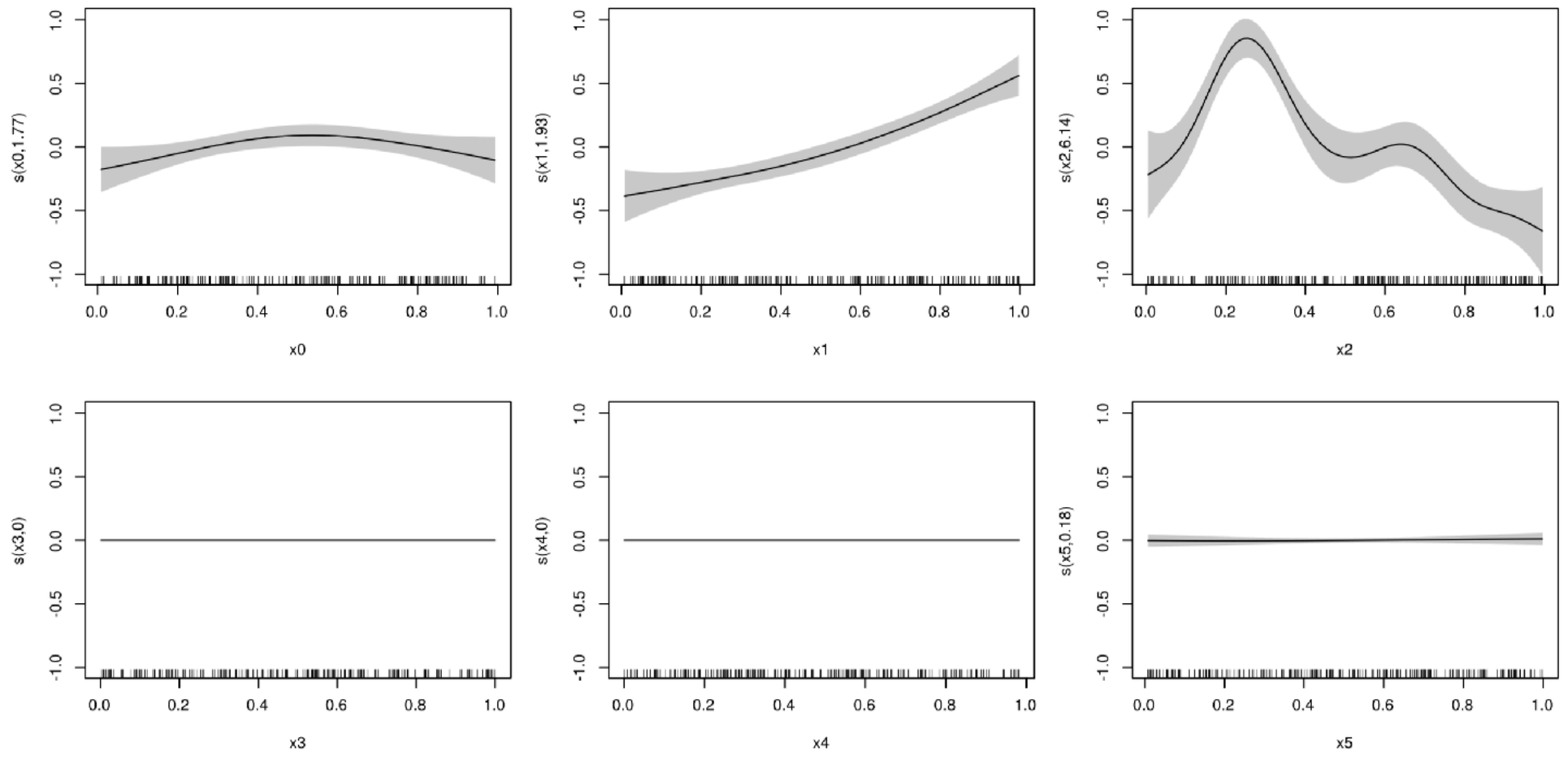
Heteroscedastic data: Gaussian location-scale models (`family = **gaulss**`)

Censored count data: Zero-inflated Poisson (`family = **ziplss**`)

A Few more Features

But I need variable selection

```
gam(y ~ s(x1) + s(x2) + s(x3) + s(x4) + s(x5) + s(x6),
        data=data, family = gaussian, select=TRUE)
```



But my data is biggish

`bam()` is a memory-efficient, high-performance, parallelizable alternative

```
system.time(
  b1 <- gam(y ~ s(x0,bs=bs)+s(x1,bs=bs)+s(x2,bs=bs,k=k),
                  data=dat)
)
```

```
   user  system elapsed
 57.610 259.800  21.673
```

```
system.time(
  b1 <- bam(y ~ s(x0,bs=bs)+s(x1,bs) +s(x2,bs=bs,k=k),
                  data=dat, discrete=TRUE, nthreads=2)
)
```

```
   user  system elapsed
  5.535  33.670   2.532
```

But I have complex hierarchical data

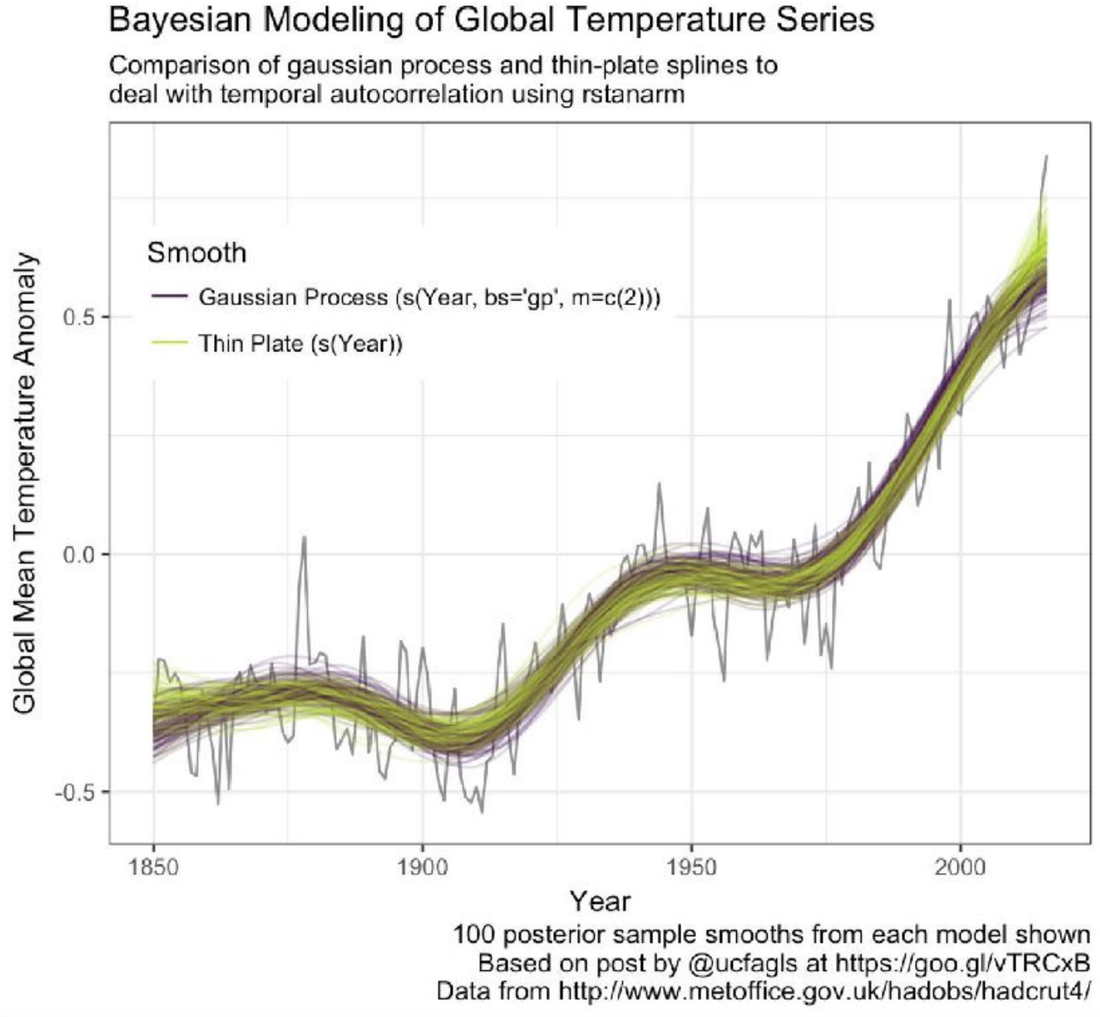
`gamm` OR `gamm4::gamm4` gives you `mgcv + lme4`

```
br <- gamm4(y ~ s(v,w,by=z) +
                s(r,k=20,bs="cr"),
            random = ~ (x+0|g) + (1|g) + (1|a/b))
```

But I want full Bayes!

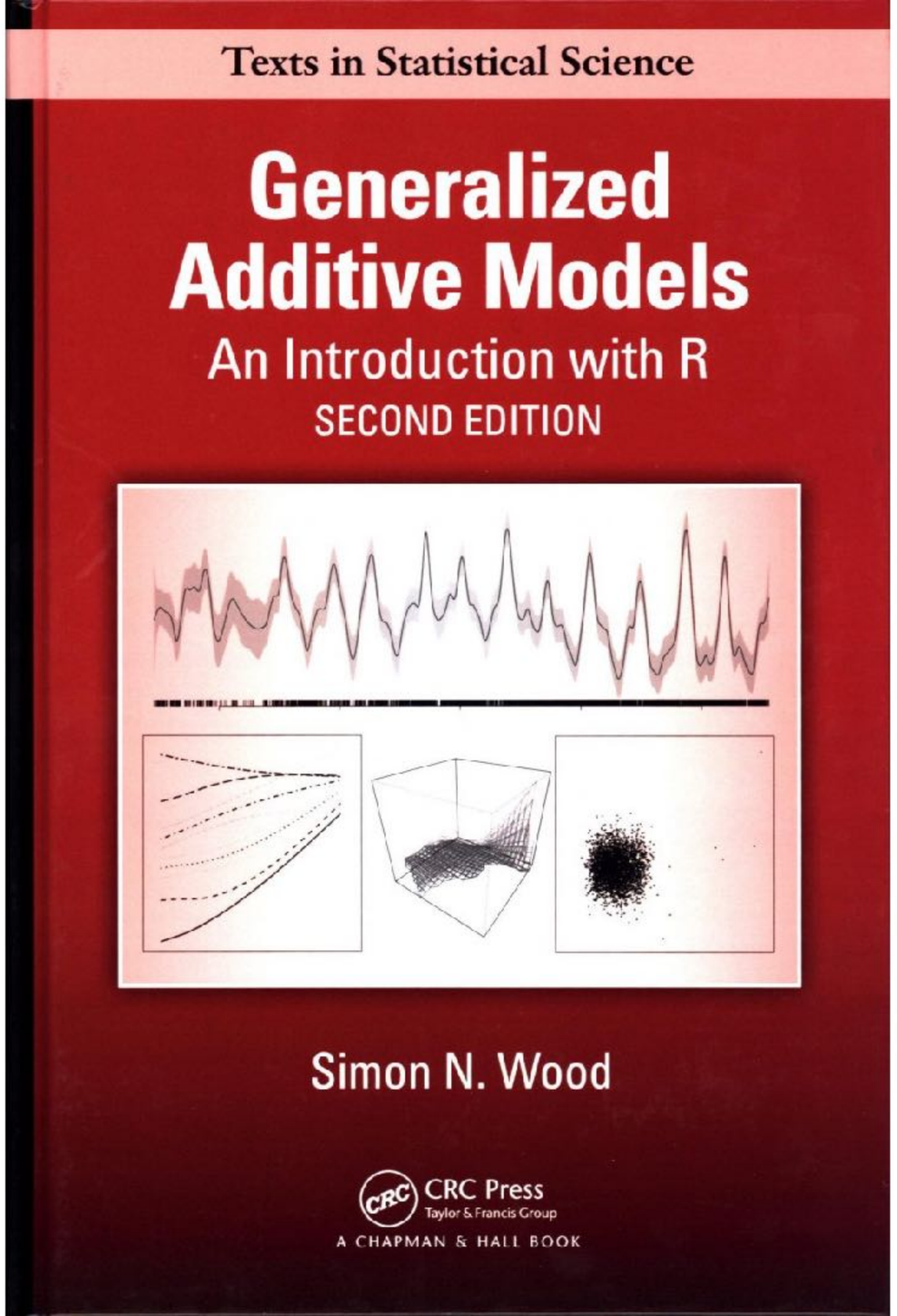
Chill, we've got your back

```
# generates JAGS code
mgcv::jagam()

# mgcv-style GAMs in Stan
rstanarm::stan_gamm4()

# greta/Tensorflow GAMs
# (very in-development by @millerdl)
gretaGAM::jagam2greta()
```



A Roundup of Resources

```
help(package="mgcv")

?smooth.terms
?missing.data
?gam.selection
```



From the bottom of the heap the musings of a geographer

First steps with MRF smooths

19 October 2017 /posted in: R

One of the specialist smoother types in the **mgcv** package is the Markov Random Field (MRF) smooth. This smoother essentially allows you to model spatial data with an intrinsic Gaussian Markov random field (GMRF). GRMFs are often used for spatial data measured over discrete spatial regions. MRFs are quite flexible as you can think about them as representing an undirected graph whose nodes are your samples and the connections between the nodes are specified via a neighbourhood structure. I've become interested in using these MRF smooths to include information about relationships between species. However, these smooths are not widely documented in the smoothing literature so working out how best to use them to do what we want has been a little tricky once you move beyond the typical spatial examples. As a result I've been fiddling with these smooths, fitting them to some spatial data I came across in a tutorial Regional Smoothing in R from The Pudding. In this post I take a quick look at how to use the MRF smooth in **mgcv** to model a discrete spatial data set from the US Census Bureau.

Read on »

Comparing smooths in factor-smooth interactions I by-variable smooths

10 October 2017 /posted in: R

One of the really appealing features of the **mgcv** package for fitting GAMs is the functionality it exposes for fitting quite complex models, models that lie well beyond what many of us may have learned about what GAMs can do. One of those features that I use a lot is the ability to model the smooth effects of some covariate (x) in the different levels of a factor. Having estimated a separate smoother for each level of the factor, the obvious question is, which smooths are different? In this post I'll take a look at one way to do this using `by` -variable smooths.

Read on »

Fitting count and zero-inflated count GLMMs with mgcv

04 May 2017 /posted in: R

Social

ucfagls@gmail.com

@ucfagls

ucfagls

gavinsimpson

ORCID iD

Impactstory profile

#365papers

Subscribe

Blogroll

- Down With time
- The Contemplative Mammoth
- Dynamic Ecology
- Jabberwocky Ecology
- Recology
- R Bloggers
- Andrew Barr's Ancient Eco
- Methods in Ecology & Evolution
- Musings on Quantitative Palaeoecology

fromthebottomoftheheap.net

Trend in irregular time series data

3

1

I have a dataset of water temperature measurements taken from a large waterbody at irregular intervals over a period of decades. (Galveston Bay, TX if you're interested)

Here's the head of the data:

	STATION_ID	DATE	TIME	LATITUDE	LONGITUDE	YEAR	MONTH	DAY	SEASON	MEASUREMENT
1	13296	6/20/91	11:04	29.50889	-94.75806	1991	6	20	Summer	28.0
2	13296	3/17/92	9:30	29.50889	-94.75806	1992	3	17	Spring	20.1
3	13296	9/23/91	11:24	29.50889	-94.75806	1991	9	23	Fall	26.0
4	13296	9/23/91	11:24	29.50889	-94.75806	1991	9	23	Fall	26.0
5	13296	6/20/91	11:04	29.50889	-94.75806	1991	6	20	Summer	28.0
6	13296	12/17/91	10:15	29.50889	-94.75806	1991	12	17	Winter	13.0

(MEASUREMENT is the temperature measurement of interest.)

The full set is available here: https://github.com/jscarlton/galvBayData/blob/master/gbtemp.csv

I would like to remove the effects of seasonal variation to observe the trend (if any) in the temperature over time. Is a time series decomposition the best way to do this? How do I handle the fact that the measurements were not taken at a regular interval? I'm hoping there is an R package for this type of analysis, though Python or Stata would be fine, too.

(Note: for this analysis, I'm choosing to ignore the spatial variability in the measurements. Ideally, I'd account for that as well, but I think that doing so would be hopelessly complex.)

asked 1 year ago

viewed 324 times

active 1 month ago

BLOG

Why Channels?

HOT META POSTS

3 "Benchmark" tag is ambiguous

5 Should the default tag synonyms be in the full form or abbreviated?

6 What to do about [parameter-optimization] tag?

Linked

33 Is there any gold standard for modeling irregularly spaced time series?

C'mon, I really don't think GAMs can do that.

Trolling-Driven Data Science

Get Gavin to Do Your Work

O RLY?

Noam Ross

Information for the mgcv workshop hosted at the Ecological Society of America Annual Meeting 2017.

Links

- Download R
- `mgcv` on CRAN

Welcome to the `mgcv` course webpage.

A course! To be given at the Ecological Society of America conference in Fort Lauderdale, Saturday August 5th 8am-5pm, 2017. Program link here.

This site contains slides, exercises and other materials for the course.

Course overview

To address the increase in both quantity and complexity of available data, ecologists require flexible, robust statistical models, as well as software to perform such analyses. This workshop will focus on how a single tool, the mgcv package for the R language, can be used to fit models to data from a wide range of sources.

mgcv is one of the most popular packages for modelling non-linear relationships. However, many users do not know how versatile and powerful a tool it can be. This workshop will focus on teaching participants how to use mgcv in a wide variety of situations (including spatio-temporal, zero-inflated, heavy-tailed, time series, and survival data) and advanced use of mgcv (fitting smooth interactions, seasonal effects, spatial effects, Markov random fields and varying-coefficient models).

The workshop will give paricipants an understanding of:

- practical elements of smoothing theory, with a focus on why they would choose to use different types of smoothers
- model checking and selection
- the range of modelling possibilities using mgcv.

Participants will be assumed to be familiar with the basics of R (loading/manipulating data, functions, and plotting) and regression in R (`lm()` and `glm()`). The organizers have extensive practical experience with ecological statistics and modelling using `mgcv` .

https://noamross.github.io/mgcv-esa-workshop/

Coming this spring...

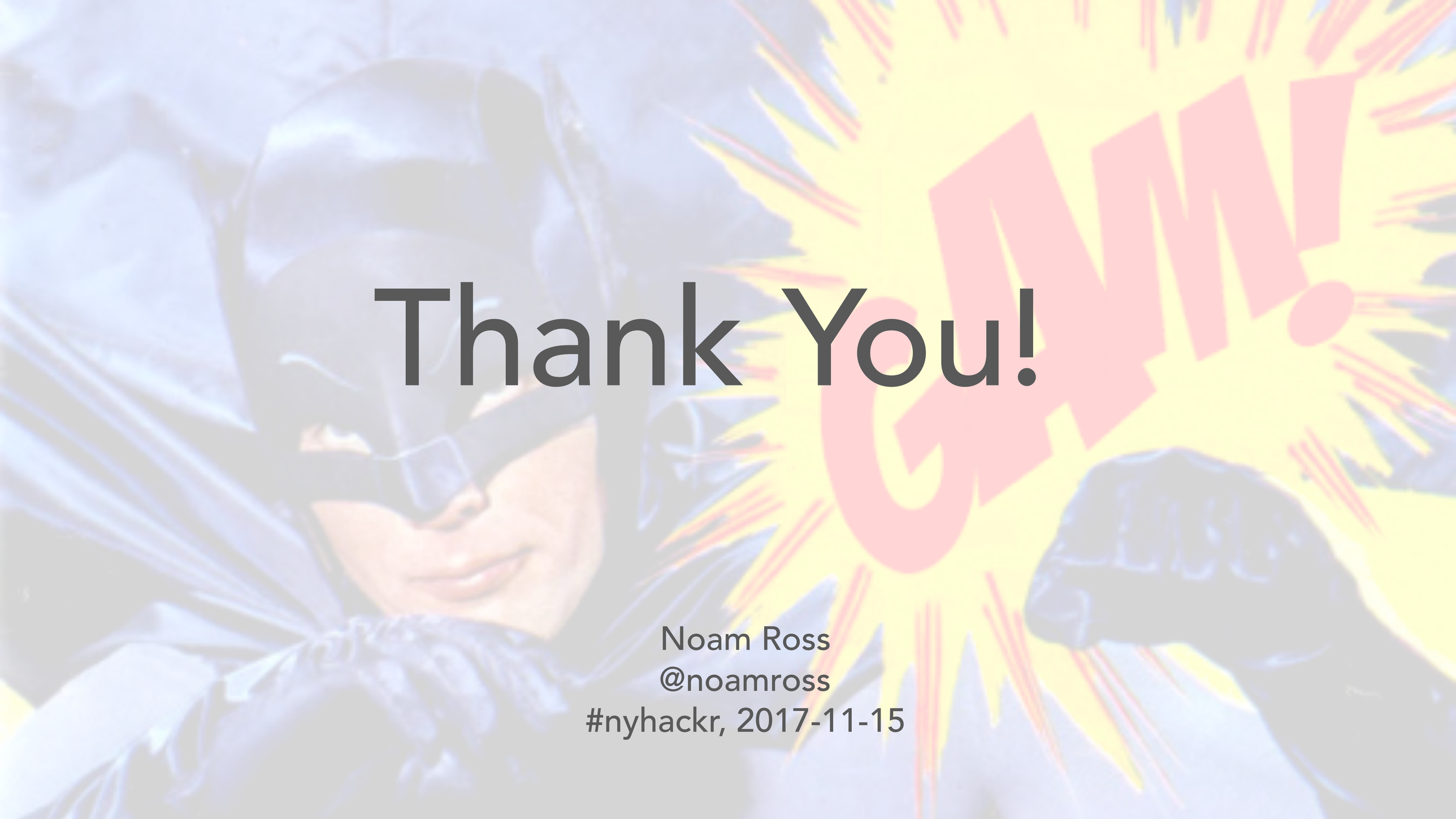
DataCamp

Generalized Additive Models in R

Learn how to fit complex, nonlinear models to data and make predictions using **mgcv** package.

4 hours **Play preview**

NOAM ROSS
Senior Research Scientist at EcoHealth Alliance


Thank You!
Noam Ross
@noamross
#nyhackr, 2017-11-15