Here's a data analysis problem:

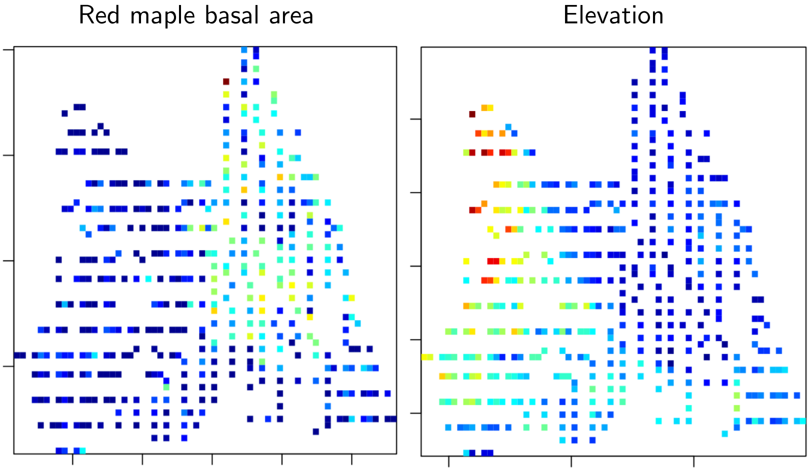
For the 2002 forest inventory data (Finley et al 2008; BEF.dat, spBayes).

Problem: Replace a laborious outcome measurement with a function of predictors measured by satellites.

- Outcome: red maple total basal area (metric tons biomass).
- Predictors: Elevation, slope, brightness (TC1), greenness (TC2), wetness (TC3).
- 437 observations.

This problem could have been on the MS exam I wrote in 1982 . . .

EXCEPT these observations are spatially referenced



Here's a standard model for analyzing data like this

At spatial locations indexed by s, model outcome $y(s)$ as

$$y(s) = x(s)\beta + w(s) + \epsilon(s)$$

- $x(s)$ are covariates, including an intercept
- $w(s)$ is a stationary GP, mean 0, covariance function $\sigma_{\mathfrak{s}}^2 \; K(\rho)$
- $\epsilon(s)$ is iid N with mean 0, variance $\sigma_{\mathfrak{e}}^2$
- Unknown parameters: β, $\sigma_{\mathfrak{s}}^2$, ρ, and $\sigma_{\mathfrak{e}}^2$.

. . . giving this mixed linear model and restricted likelihood

For observations at $\{s_1, \ldots, s_n\}$, the mixed linear model is

$$\mathbf{y} = X\beta + \mathbf{I}_n\mathbf{w} + \epsilon$$

- X's rows are the $x(s_i)$
- $\mathbf{w} = (w(s_1), \ldots, w(s_n))' \sim N(0, G) \quad \text{for} \quad G = \sigma^2_{\mathfrak{s}}\ K(\{s_i\}; \rho)$
- $\epsilon \sim N(0, R) \quad \text{for} \quad R = \sigma^2_e\ I$

$\sigma^2_{\mathfrak{s}}$, ρ, σ^2_e can be estimated by maximizing the log restricted likelihood

$$-\log|V| - \log|X'V^{-1}X| - \mathbf{y}'[V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}]\mathbf{y}$$

where $V = G + R$, a dense matrix.

For models like this, we don't have LM-quality tools

The RL doesn't have a closed form; effects of data features are obscure.

Variograms are useful but

- don't give specific info about how the data determine $\hat{\sigma}^2_{\mathfrak{s}}, \hat{\rho}, \hat{\sigma}^2_e$
- aren't much help in assessing non-stationarity.

The usual residuals are problematic:

- This model can fit any dataset arbitrarily well.
- If the model smooths much, residuals are biased.
- Residuals don't tell us how the data determine $\hat{\sigma}^2_{\mathfrak{s}}, \hat{\rho}, \hat{\sigma}^2_e$.

Bose, Hodges, Banerjee *Biometrics* 2018

BHB *Biometrics* (2018) is Step 1 (maybe) in filling this gap.

This talk emphasizes <u>*ideas*</u> using a simplified problem:
data collected on a 1-D regular grid with no fixed effects.

I'll mention how we've addressed these simplifications.

The three ideas that make this tractable

1. Approximate the GP; transform the data.

2. The resulting (approximate) restricted likelihood has a simple form; use that form to understand how the data determine the fit.

3. Extend tools from linear models.

Idea #1: Spectral approximation for a stationary GP

Data taken at locations $s_j \in \{0, 1, ..., M-1\}$, M a multiple of 2.

Frequencies $\omega_m \in \{0, \frac{1}{M}, ..., \frac{1}{2}, -\frac{1}{2} + \frac{1}{M}, ..., -\frac{1}{M}\}$, $m = 0, 1, ..., M-1$.

Approximate the GP $w(s_j)$ by

$$g(s_j) = a_0 + 2 \sum_{m=1}^{\frac{M}{2}-1} \left(a_m \cos(2\pi\omega_m s_j) - b_m \sin(2\pi\omega_m s_j)\right) + a_{\frac{M}{2}} \cos(2\pi\omega_{\frac{M}{2}} s_j)$$

a_m, b_m have independent mean zero Gaussian priors with variances proportional to $\sigma^2_{\mathfrak{s}}\ \phi(\omega_m; \rho)$, the spectral density of the GP covariance.

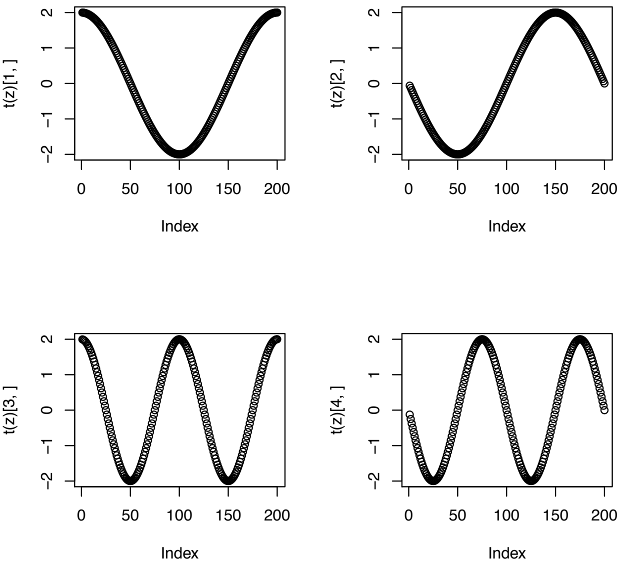
Based on Paciorek (2007), Wikle (2002).

With the approximation $g(s_j)$, the model becomes

$$\mathbf{y} = X\beta + Zu + \epsilon$$

- Assuming no fixed effects: X is a vector of 1's.
- Omit a_0 to avoid an identification problem.
- Z's columns are sin/cos functions and do not depend on unknowns.
- $Z'Z \;=\; Diag(1/2M, 1/2M, ..., 1/M); \quad Z'1 = \mathbf{0}$.
- $u \sim N(0, G),\;\; G \;=\; \sigma^2_{\mathfrak{s}}\,\text{Diag}(\; \frac{1}{2M}\phi(\omega_{m(j)};\rho),\; \frac{1}{M}\phi(\omega_{M/2};\rho)\;)$
- $\epsilon \sim N(0, R),\;\; R \;=\; \sigma^2_e I$

Columns 1 to 4 of the random effects design matrix Z



Idea #1: Transform the data so the log RL is simple

Using the spectral approximation to the GP, the model is

$$\mathbf{y} = X\beta + Zu + \epsilon$$

Pre-multiply this equation by $(Z'Z)^{-0.5}Z'$ to give:

$$v \;=\; (Z'Z)^{-0.5}Z'\mathbf{y} \;=\; (Z'Z)^{0.5}u \;+\; (Z'Z)^{-0.5}Z'\epsilon$$

Then

$$\begin{aligned} \mathrm{E}(v) &= 0 \\ \mathrm{Cov}(v) &= \sigma_{\mathfrak{s}}^2 Diag(a_j(\rho)) + \sigma_e^2 I \\ a_j(\rho) &= \phi(\omega_{m(j)};\rho) \end{aligned}$$

Idea #2: The (approximate) RL has a simple form.

$$v \sim N(\ 0,\ \sigma_{\mathfrak{s}}^2 Diag(a_j(\rho)) + \sigma_e^2 I\ )$$
$$v = (Z'Z)^{-0.5} Z'\mathbf{y}$$

The (approximate) log RL for $(\sigma_{\mathfrak{s}}^2, \rho, \sigma_e^2)$ is the likelihood arising from v:

$$-\frac{1}{2}\sum_{j=1}^{M-1}[\ \log(\sigma_{\mathfrak{s}}^2 a_j(\rho) + \sigma_e^2) + \ v_j^2/\ (\sigma_{\mathfrak{s}}^2 a_j(\rho) + \sigma_e^2)\ ].$$

The keys to understanding this (approximate) RL as a function of the data are the transformed data v_j and the $a_j(\rho)$.

Given ρ, the v_j^2 follow a gamma-errors GLM

The log RL has the form of the likelihood from a gamma-errors GLM with the identity link:

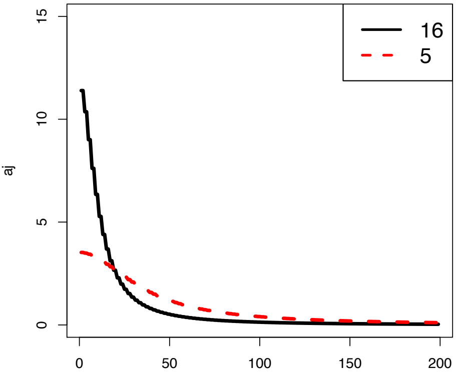
$$-\frac{1}{2}\sum_{j=1}^{M-1}\left[\ \log(\sigma_{\mathfrak{s}}^2 a_j(\rho) + \sigma_e^2) + \ v_j^2 / \ (\sigma_{\mathfrak{s}}^2 a_j(\rho) + \sigma_e^2)\ \right]$$

- The v_j^2 are the gamma-distributed "data".
- $1/2$ is the gamma's shape parameter.
- $E(v_j^2) = \sigma_{\mathfrak{s}}^2 a_j(\rho) + \sigma_e^2$
- $\text{Var}(v_j^2) = 2\left(\sigma_{\mathfrak{s}}^2 a_j(\rho) + \sigma_e^2\right)^2$

How does $a_j(\rho)$ change with ρ?

Exponential covariance function K: $a_j(\rho)$ for $\rho = 5$ and 16.

The horizontal axis is j.



For larger ρ, the $a_j(\rho)$ start higher and decline to zero more sharply.

Idea #2: Use this (approximate) RL to understand the fit

The approximate RL uses the data *only* through the v_j,
projections of $\mathbf{y}$ onto sin/cos functions of different frequencies.

The projections v_j don't depend on any unknowns.

The projections v_j are the same for all GP covariance functions.

Using different GP models for the random effect
$\Leftrightarrow$ fitting different gamma regressions to the same transformed data.

The model for the v_j^2 is a GLM with 3 parameters $\Rightarrow$ no overfitting.

How do the data determine parameter estimates?

The "observations" are the v_j^2; parameters are fit such that:

$$E(\ v_j^2 \mid \sigma_{\mathfrak{s}}^2, \rho, \sigma_e^2\ ) \ =\ \sigma_{\mathfrak{s}}^2\ a_j(\rho)\ +\ \sigma_e^2.$$

The $a_j(\rho)$ are non-increasing in j and for large j,

$$E(\ v_j^2 \mid \sigma_{\mathfrak{s}}^2, \rho, \sigma_e^2\ ) \ \approx\ \sigma_e^2.$$

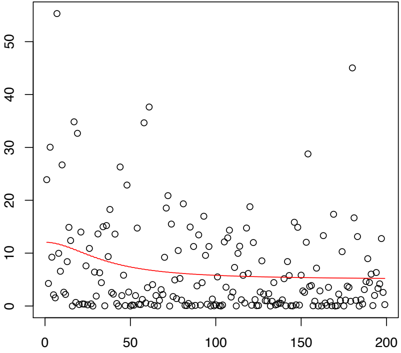
Loosely,

- $\hat{\sigma}_e^2$ is "in the middle" of the v_j^2 for large j.
- $\hat{\rho}$ fits the rate at which the v_j^2 decline for "small" j.
- $\hat{\sigma}_{\mathfrak{s}}^2$ makes $\hat{\sigma}_{\mathfrak{s}}^2 a_j(\hat{\rho}) + \hat{\sigma}_e^2$ go through the middle of the v_j^2 for "small" j.

Examples of v_j^2 vs. j, with fits

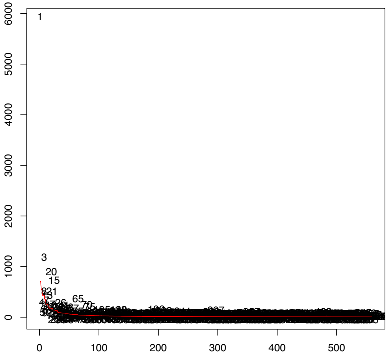
Simulated: $\sigma_{\mathfrak{s}}^2 = 2, \rho = 5, \sigma_e^2 = 5$

$\sigma_{\mathfrak{s}}^2$ smaller than σ_e^2



Red maple basal area

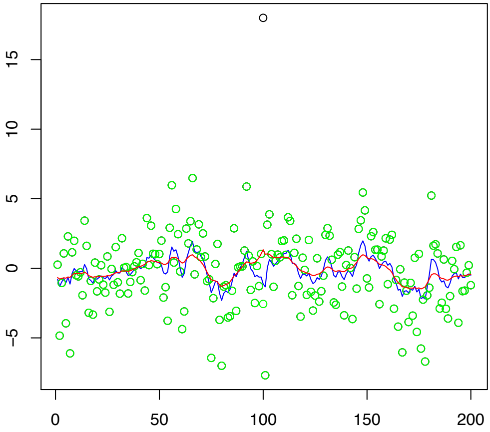
Whopping non-stationarity



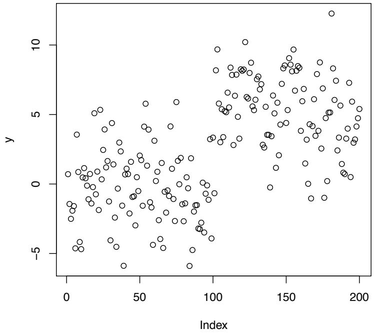
Some conjectures about how the data determine estimates

An outlier inflates the v_j^2 for large j's (high frequencies) $\Rightarrow$ inflated $\hat{\sigma}_e^2$.

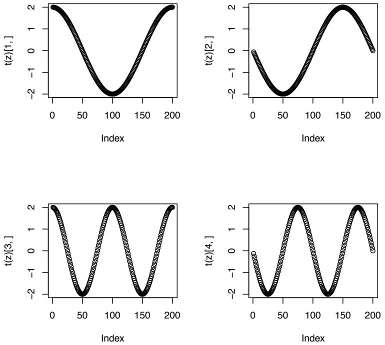
Little effect on v_j^2 for "small" j (low frequencies) and thus on $\hat{\sigma}_{\mathfrak{s}}^2$ and $\hat{\rho}$.



Data contaminated with shift



(a) data simulated from GP with $\sigma_{\mathfrak{s}}^2$=2, σ_e^2=5 and ρ=5 with mean shift from 0 to 5 midway



(b) first four columns of Z, the spectral basis matrix, on the domain [1,2,...,199,200]

Hypothesize: how does this shift in mean affect the estimates of the GP parameters ?

1. v_2^2 will be inflated, this will lead to an inflated value of the estimate of $\sigma_{\mathfrak{s}}^{\mathbf{2}}$.
2. So to capture the sharp decline in v_j^2's the estimate of ρ will be inflated too.
3. v_j^2's for larger j's broadly unaffected, hence the estimate of $\sigma_{\mathbf{e}}^{\mathbf{2}}$ will remain almost the same.

Parameter estimates (SE): data with mean shift

	exact RL		
	$\sigma^2_{\mathfrak{s}}$	σ^2_e	ρ
actual values	2	5	5
uncontaminated	2.29 (0.11)	4.75 (0.11)	6.89 (0.82)
contaminated	11.50 (0.36)	5.61 (0.06)	106.48 (6.06)
actual values	10	0.1	16.67
uncontaminated	9.99 (0.34)	0.11 (0.01)	17.29 (0.67)
contaminated	17.96 (0.77)	0.16 (0.01)	29.99 (1.32)

The data contain little information about ν

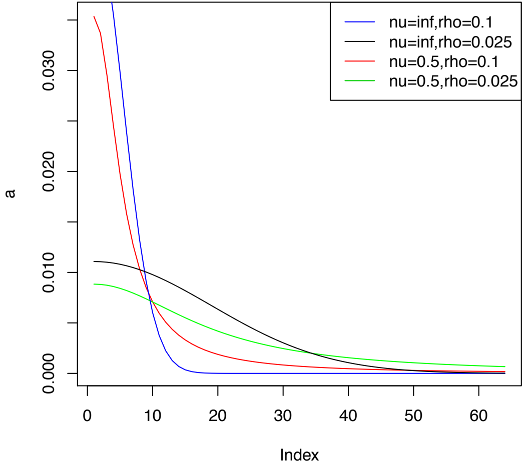


Figure: a_j's for Matérn (ν=0.5) and Matérn (ν=∞) for the spectral approximation on the domain [0,1,...,62,63] on the horizontal axis.

Idea #3: Extend tools from linear models and GLMs

Plot the v_j^2 and fitted values $\hat{\sigma}_{\mathfrak{s}}^2 a_j(\hat{\rho}) + \hat{\sigma}_e^2$ vs. j.

This shows the data and model fit corresponding to the RL.

It's a direct look at the "signal" for non-stationarity.

Added variable plots show how the data produce a fixed effect's estimate.
An AVP can be done in both the

Observation domain ($\mathbf{y}$) and

Spectral domain (the v_j)

Added variable plot in observation domain

Adding predictor C to the model $y = X\beta + C\alpha + u + \epsilon$,
X contains predictor already in the model.

Multiply both sides of the model equation by $\hat{V}^{-0.5}$, ^ denotes estimates from fitting a model with X.

Then multiply both sides by
$\hat{P} = I - \hat{V}^{-0.5} X (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-0.5}$

- Plot $\hat{P}\hat{V}^{-0.5}y$ vs $\hat{P}\hat{V}^{-0.5}C$.

Added variable plot in the spectral domain

Adding predictor C to the model $y = X\beta + C\alpha + u + \epsilon$,
X contains predictors already in the model.

Multiply both sides by $(I - P_X)$,
then multiply by $(Z'Z)^{-0.5}Z'$, to get

$v^* = (Z'Z)^{-0.5}Z'(I - P_X)y$: the v_j from the residual y and
$v_C^* = (Z'Z)^{-0.5}Z'(I - P_X)C$: the v_j from the residual C.

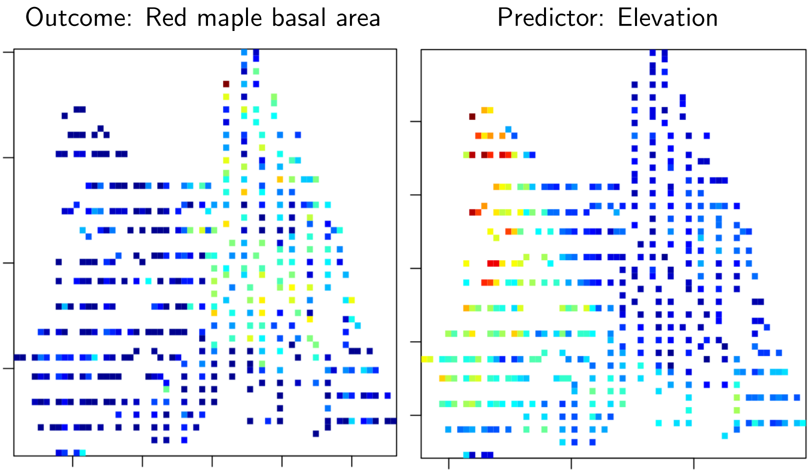
- Plot $\hat{D}v^*$ vs $\hat{D}v_C^*$,

where $\hat{D} = Diag(1/\sqrt{\hat{\sigma}_{\mathfrak{s}}^2 a(\hat{\rho}) + \hat{\sigma}_e^2})$, $\hat{}$ denotes estimates obtained from fitting a model with only X, no C.

Added variable plots (AVPs) in the two domains

- The AVPs in the spectral domain and in the observation domain estimate the same coefficient for a particular predictor.
- The spectral domain AVP highlights particular large scale trends in the data. The observation domain AVP highlights particular localized details.
- The spectral domain AVP involves the spectral approximation which the observation domain AVP does not.

Back to the forest inventory data



Undoing the simplifying assumptions

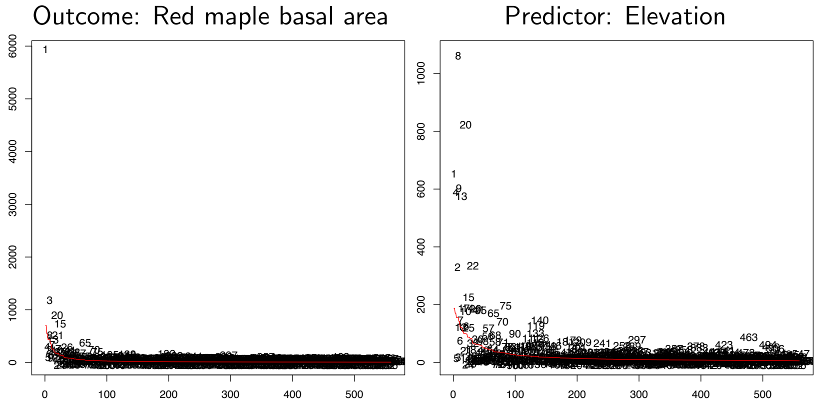
The paper and supplements discuss these at length.

• Spectral approximation in 2-D: Each design-matrix column corresponds to a pair of frequencies, one in each dimension (Paciorek 2007).

• Data not on a grid: Pre-smooth to a grid. ($\exists$ a better way?)

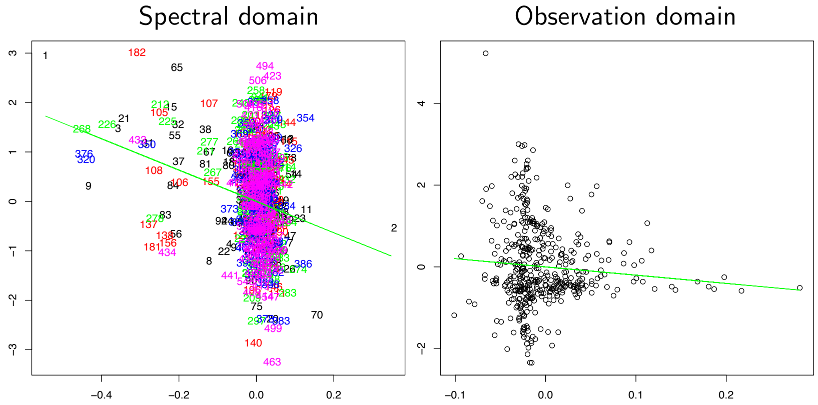
• Fixed effects: Regress them out and apply the spectral approximation to the residuals.

Plots of v_j^2 vs. j for outcome and predictor



The horizonal axis is $j \Rightarrow$ low frequencies at left, high frequencies at right

AVPs for elevation



Info from the AVPs vs. the actual fits

	Elevation			From the RL		
	Estimate	SE	P-value	$\hat{\sigma}_{\mathfrak{s}}^2$	$\hat{\rho}$	$\hat{\sigma}_e^2$
Intercept-only	–	–	–	29.6	6.0	16.2
AVP, Spectral	-3.17	0.51	10^{-10}	–	–	–
AVP, Observation	-2.02	1.09	0.07	–	–	–
Real fit	-2.52	0.29	tiny	22.0	2.9	13.8

Focus on the ideas, not on our specific choices

The important thing is the 3 ideas:

1. Approximate the GP; transform the data.
2. $\Rightarrow$ simpler forms, which makes the fit understandable.
3. Extend tools from linear models and GLMs.

All the specific choices we've made could be replaced (I think).