Biometrics 74, 863–873
September 2018

DOI: 10.1111/biom.12848

Toward a Diagnostic Toolkit for Linear Models with Gaussian-Process Distributed Random Effects

Maitreyee Bose,[1,*] James S. Hodges,[2,] and Sudipto Banerjee[3,***]**

[1]Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
[2]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
[3]Department of Biostatistics, University of California, Los Angeles, California 90095, U.S.A.
[*]*email:* bosem2@uw.edu
[**]*email:* hodge003@umn.edu
[***]*email:* sudipto@ucla.edu

Summary. Gaussian processes (GPs) are widely used as distributions of random effects in linear mixed models, which are fit using the restricted likelihood or the closely related Bayesian analysis. This article addresses two problems. First, we propose tools for understanding how data determine estimates in these models, using a spectral basis approximation to the GP under which the restricted likelihood is formally identical to the likelihood for a gamma-errors GLM with identity link. Second, to examine the data's support for a covariate and to understand how adding that covariate moves variation in the outcome y out of the GP and error parts of the fit, we apply a linear-model diagnostic, the added variable plot (AVP), both to the original observations and to projections of the data onto the spectral basis functions. The spectral- and observation-domain AVPs estimate the same coefficient for a covariate but emphasize low- and high-frequency data features respectively and thus highlight the covariate's effect on the GP and error parts of the fit, respectively. The spectral approximation applies to data observed on a regular grid; for data observed at irregular locations, we propose smoothing the data to a grid before applying our methods. The methods are illustrated using the forest-biomass data of Finley et al. (2008).

Key words: Added variable plot; Gaussian process; Lack of fit; Linear mixed model; Missing predictor; Spectral approximation.

1. Introduction

Gaussian processes (GPs) are widely used in longitudinal, functional, and spatial data analysis because the properties they inherit from the normal distribution make them easy to work with. Fitting a GP to data involves estimating the process parameters, most commonly the process variance and range, along with an error variance. One way to do this is by writing the GP as a component of a linear mixed model and maximizing the restricted likelihood, which is identical to the marginal posterior from a Bayesian analysis with particular priors. It is, however, unclear how the resulting parameter estimates are influenced by features in the data like outliers or non-stationarities in the mean or covariance function. Fuglstad et al. (2014) argue that even if non-stationarity is present, it is difficult to model properly and fitting a stationary model usually gives satisfactory predictions. Also, GPs are now easily accessible to non-specialists (e.g., in SAS Inc.'s JMP package), so it is useful to know how a given form of non-stationarity affects the fit of a stationary isotropic GP.

To elucidate further, consider an example from Finley et al. (2008). Data on forest biomass and some covariates were available over a specific region; prediction of forest biomass at unmeasured locations was of interest. For over 30 years, tools have been available that completely characterize independent-errors linear models fit to such datasets, but analogous tools do not exist for models with spatially correlated random effects, which are commonly modeled using GPs. To better understand fits of the latter models, we need a simple, interpretable form of the restricted likelihood. This article proposes such a form, which leads to tools for examining fits of linear mixed models with GP-distributed random effects and for choosing covariates. Section 7 demonstrates the tools using this example.

This article addresses two challenging and hitherto untackled problems. First, using the spectral approximation to a GP, we propose tools to help understand exactly *how* the data determine estimates of variance-structure parameters in a mixed linear model with a GP-distributed random effect (MLM/GP). Phenomena like spatial confounding (e.g., Paciorek, 2010; Hodges and Reich, 2010) make it clear that we cannot simply assume that a model, in its role as a likelihood, behaves according to its face-value interpretation as a probability model; rather, our tools must directly display the influence of data on estimates, as do tools for linear models. Some methods exist for MLM/GPs but they are weak. One approach, popular in geostatistics, is informally examining residuals using exploratory tools such as variograms (see, e.g., Chiles and Delfiner, 2009; Banerjee et al., 2014; Cressie, 2015), which describe the degree of dependence (spatial range) and extent of variability (sill and nugget) in the data.

Variograms are useful but do not provide specific information about how functions of the data determine estimates. Also, variograms are most useful for stationary, even isotropic, processes and will not help much in ascertaining the effects of nonstationarity on stationary isotropic GP fits. Exploratory analysis of residuals themselves generally does not provide specific information about how functions of the data determine estimates, and residuals in MLM fits are biased (e.g., Hodges 2014, Chapter 8) with the largest bias in parts of the fit most affected by shrinkage/smoothing.

Our second objective is to help analysts understand how adding a fixed effect to an MLM/GP moves variation in the outcome *y into* the fixed-effect part of the fit and *out of* the GP and error parts of the fit. As we will see, the GP and error variance fits are determined mostly by, respectively, low- and high-frequency data features not captured in fixed effects. Adding a fixed effect to a MLM/GP can "take" variation mostly from the GP part of the fit, mostly from the error part of the fit, or substantially from both. The methods developed to understand an MLM/GP fit suggest using a diagnostic tool from linear models, added variable plot (AVPs), which show the data's information, observation-by-observation, about the coefficient of a fixed effect, enabling a modeler to understand whether the information about that fixed effect's coefficient is broadly distributed through the data or arises from a few functions of the data. The spectral basis we use permits an AVP on the spectral scale; one can also make an AVP using the original observations. The two AVPs estimate the same coefficient for the added variable (modulo the spectral approximation) but the spectral-domain AVP emphasizes the contribution of low-frequency data features and thus highlights the effect a candidate predictor will have on the GP part of the fit, while the observation-domain AVP gives more emphasis to the contribution of high-frequency data features (while avoiding the spectral approximation). These two AVPs thus adapt a linear-model diagnostic to MLM/GPs in a way that provides information about why the GP and error parts of the fit change the way they do when a fixed effect is added.

In pursuing these goals, we hew to Weisberg's (1983) principles for diagnostics, in particular, looking at the data as directly as possible and providing a plot to go with each diagnostic so the effect of individual observations can be assessed. To meet our objectives, we need a tractable form of the restricted likelihood, which we obtain using a basis approximation to the GP-distributed random effect, the spectral approximation (Wikle, 2002; Paciorek, 2007). With this choice, the restricted likelihood for the MLM/GP's variance structure unknowns becomes formally identical to the likelihood arising from a gamma-errors GLM with identity link, so familiar data-analytic intuition and tools can be brought to bear. The spectral approximation applies to data observed on a regular grid; for data observed at irregular locations, we will assume the data have been mapped to a regular grid (Paciorek, 2007; Reich et al. 2011). We do not see this as a major drawback because our focus is understanding GP fits rather than enhancing the model's richness and flexibility.

We emphasize that the choice of the spectral approximation is predicated on properties of its basis functions that serve our purposes. This article is *not* about the spectral approximation *per se* or its properties *as a model*. Others have proposed the spectral approximation to speed computation (e.g., Fuentes 2006, Paciorek 2007), but that is also not our purpose. We know of no other approximation (e.g., Karhunen–Loève expansions, wavelet basis, kernel convolutions, or predictive processes) that would work in the subsequent development.

The rest of this section describes an approach to fitting linear mixed models when the random effect is a one-dimensional GP. Section 2, then, details the spectral approximation for intercept-only GPs in one dimension observed at equally spaced locations, and derives the simple restricted likelihood, which Section 3, then, uses to make conjectures about how data features affect parameter estimates, which simulation experiments support. Section 4 extends the approach to models with covariates. Section 5 then proposes tools for model building based on the foregoing. Section 6 extends the tools to data observed on two-dimensional regular grids, and Section 7 applies them to the forest-biomass data. The GP has been well investigated as a probability model and as an interpolator given parameter values; we focus on the GP as part of a likelihood used to estimate parameters. Finally, we discuss only finite sample inferences.

1.1. *One Dimensional Gaussian Process Fitting*

Given data $y(s)$ at location s for $s \in \{s_1, s_2, \ldots, s_M\}$, we want to fit the model

$$y(s) = x(s)\beta + w(s) + \epsilon(s) \tag{1}$$

where $w(s)$ is a stationary GP with mean 0 and isotropic covariance function $\sigma^2_s K(d;\rho)$, and $\epsilon(s)$ is normal with mean 0 and variance σ^2_e, independent between locations s and independent of $w(s)$. $K(d;\rho)$ is a correlation function; d is the distance between two locations s; ρ is an unknown range parameter; and d, ρ, and s have the same units (distance). The row p-vector $x(s)$ contains covariates including the intercept and the column p-vector β contains fixed effects.

Parameters that need to be estimated are β, σ^2_s (process variance), ρ (range), and σ^2_e (error variance). One way to fit this model is to write it as a linear mixed model

$$y = X\beta + I_M\gamma + \epsilon, \tag{2}$$

where $y = (y(s_1), y(s_2), \ldots, y(s_M))'$, X's rows are the $x(s)$, $\gamma \sim N(0, \Sigma)$ with $\Sigma = \sigma^2_s K(d;\rho)$, and $\epsilon \sim N(0, R)$ with $R = \sigma^2_e\, I_M$. Defining $V = \Sigma + R$, the unknowns in Σ and R are commonly estimated by maximizing the log restricted likelihood

$$\text{const} - 0.5(\log|V| + \log|X'V^{-1}X| + y'[V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}]y). \tag{3}$$

The restricted likelihood (3) has non-closed form terms involving the GP covariance matrix Σ, so it is a black box. The key to the desired simple form of (3) is to diagonalize V, leading to a simple matrix-free form that can be used to develop intuition about how the GP model is fit to data. (Closed form expressions for V^{-1} exist for the Ornstein–Uhlenbeck process (Finley et al. 2009, Section 2.1.1) but they are not diagonal.) To this end, we approximate the GP using orthogonal

basis functions. The result *is* an approximation but it can be used to conjecture about how the exact GP behaves, and the conjectures can be tested in simulations and used as a basis for diagnostic tools. The approximation's accuracy is not of *inherent* interest but rather only to the extent that less accuracy means poorer understanding of the exact GP and less useful tools. In this regard, we note that all diagnostics for non-normal generalized linear models and for Cox regression are based on approximations.

2. Approximating the Gaussian Process

This section develops a simple approximate form of the log restricted likelihood (3) using spectral basis functions. Section 3, then, interprets that approximate restricted likelihood as the likelihood for a particular generalized linear model and uses it to make conjectures about how features in the data, like outliers or mean-shifts, affect GP fits.

2.1. *Linear Mixed Model Representation*

An intercept-only GP model, with β the intercept, can be approximated as

$$y \approx 1_M\beta + Zu + \epsilon \tag{4}$$

where two key conditions hold: Z is an $M \times (M-1)$ matrix of basis functions that is not a function of any unknown parameters, and u is a zero-mean normal random vector with a diagonal covariance matrix, $G = \sigma_s^2 \mathrm{Diag}(b_j(\rho))$, where the $b_j(\rho)$'s are known functions of ρ. Z and $b_j(\rho)$ are chosen so $\mathrm{Cov}(Zu+\epsilon) \approx \mathrm{Cov}(y)$.

To yield the desired simplified restricted likelihood, Z must have these properties: $Z'1_M = \mathbf{0}$ and $Z'Z$ is diagonal with diagonal entries $c_1, c_2, \ldots, c_{M-1}$. For such a Z, premultiplying (4) by $(Z'Z)^{-1/2}Z'$ gives $v = \delta + \varepsilon$, where

$$\delta = (Z'Z)^{1/2}u \sim N(0, \sigma_s^2 \mathrm{Diag}\{b_j(\rho)c_j\}) \text{ for } j = 1, 2, \ldots, M-1, \text{ and}$$
$$\varepsilon = (Z'Z)^{-1/2}Z'\epsilon \sim N(0, \sigma_e^2 \mathrm{I}_{M-1}).$$

Then $v = (Z'Z)^{-1/2}Z'y$ is normal with $\mathrm{E}(v) = 0$ and diagonal covariance $\sigma_s^2 \mathrm{Diag}\{b_j(\rho)\ c_j\} + \sigma_e^2 I_{M-1}$. The distribution of the v_j's gives the log restricted likelihood for $(\sigma_s^2, \sigma_e^2, \rho)$:

$$\text{const} - \frac{1}{2}\sum_{j=1}^{M-1}\left(\log(\sigma_s^2 a_j(\rho) + \sigma_e^2) + v_j^2(\sigma_s^2 a_j(\rho) + \sigma_e^2)^{-1}\right), \tag{5}$$

where $a_j(\rho) = b_j(\rho)\ c_j$. The columns of $Z(Z'Z)^{-0.5}$ are orthonormal and the v_j^2's are the squared lengths of projections of y onto these columns. Thus, the v_j decompose the data into components corresponding to these orthonormal predictors; in the spectral approximation described in Section 2.3, these components correspond to frequencies.

2.2. *The Generalized Linear Model Form*

Given ρ, the approximate restricted likelihood (5) is identical to the likelihood arising from a gamma-errors generalized linear model with identity link, as in Hodges (2014, Ch. 15) and Henn and Hodges (2014). As such, the v_j^2 are the data, the gamma shape parameter is 1/2, $E(v_j^2) = \sigma_s^2 a_j(\rho) + \sigma_e^2$, and $\mathrm{Var}(v_j^2) = 2(\sigma_s^2 a_j(\rho) + \sigma_e^2)^2$. Thus the v_j^2's and a_j's in the approximate restricted likelihood are the keys to understanding how the GP's parameters are fit to data, giving a way to examine the model's fit that is immune to the fact that a GP can fit any y perfectly. The v_j^2's and a_j's are, in effect, the data and predictors in a regression model that provides the information about the unknowns σ_s^2, σ_e^2, and ρ.

2.3. *The Spectral Approximation*

The spectral basis is a powerful tool, widely used for correlated processes. We develop the spectral approximation of a GP following Royle and Wikle (2005) and Paciorek (2007).

Assume observations have been made at locations $s_j \in S = \{1, 2, \ldots, M\}$, $j = 1, 2, \ldots, M$, where M is a multiple of 2. Define

$$g(s_j) = \sum_{m=0}^{M-1} \varphi_m(s_j)u_m \tag{6}$$

where $u_m = a_m + ib_m$, $m = 0, 1, \ldots, M-1$, are the M spectral coefficients. The $\varphi_m(s_j) = \exp(i2\pi\omega_m s_j)$ are basis functions having frequency $\omega_m \in \{0, \frac{1}{M}, \ldots, \frac{1}{2}, -\frac{1}{2}+\frac{1}{M}, \ldots, -\frac{1}{M}\}$, $m = 0, 1, \ldots, M-1$. To apply this approximation to real-valued Gaussian processes, assume $u_0, u_1, \ldots, u_{\frac{M}{2}}$ are jointly independent; u_0 and $u_{\frac{M}{2}}$ are real valued ($b_0 = b_{\frac{M}{2}} = 0$); and $u_{\frac{M}{2}+1} = \bar{u}_{\frac{M}{2}-1}, \ldots, u_{M-1} = \bar{u}_1$. This makes $g(s_j)$ real valued:

$$g(s_j) = a_0 + 2\sum_{m=1}^{\frac{M}{2}-1}(a_m\cos(2\pi\omega_m s_j) - b_m\sin(2\pi\omega_m s_j)) + a_{\frac{M}{2}}\cos(2\pi\omega_{\frac{M}{2}}s_j), \tag{7}$$

where the a_m's and b_m's have independent mean zero Gaussian distributions with variances $\mathrm{V}(a_0) = \frac{1}{M}\sigma_s^2\phi(\omega_0;\rho)$; $\mathrm{V}(a_{\frac{M}{2}}) = \frac{1}{M}\sigma_s^2\phi(\omega_{\frac{M}{2}};\rho)$; and $\mathrm{V}(a_m) = \mathrm{V}(b_m) = \frac{1}{2M}\sigma_s^2\phi(\omega_m;\rho)$ for $m \neq 0$ or $M/2$, where $\sigma_s^2\phi(.;\rho)$ is the spectral density of the covariance function $\sigma_s^2 K(d;\rho)$. For large M, this approximate process is a Gaussian process with mean zero and covariance function close to that of the GP it approximates (Web Appendices A and E).

For data observed at locations $s_1 = 1, s_2 = 2, \ldots, s_M = M$, then, model (2) becomes

$$y = 1_M\beta + Zu + \epsilon \tag{8}$$

where $\epsilon \sim N(0, R)$, $R = \sigma_e^2 I_M$,

- β is the coefficient for the intercept, the only fixed effect,
- $u = (a_1, b_1, a_2, b_2, \ldots, a_{\frac{M}{2}-1}, b_{\frac{M}{2}-1}, a_{\frac{M}{2}})'$ is the vector of random effects, and
- Z is an $M \times (M-1)$ matrix with jth column z_j given by
  $\left(2\cos(\omega_{\frac{j+1}{2}}2\pi), 2\cos(\omega_{\frac{j+1}{2}}2\pi 2), \ldots, 2\cos(\omega_{\frac{j+1}{2}}2\pi M)\right)'$; $j \in \{1, 3, \ldots, M-3\}$,
  $\left(-2\sin(\omega_{\frac{j}{2}}2\pi), -2\sin(\omega_{\frac{j}{2}}2\pi 2), \ldots, -2\sin(\omega_{\frac{j}{2}}2\pi M)\right)'$; $j \in \{2, 4, \ldots, M-2\}$,

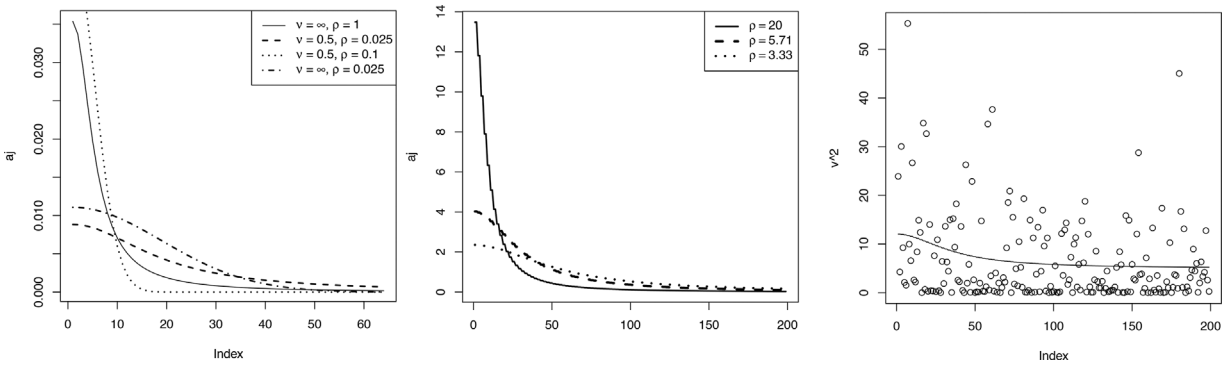

Figure 1. Left panel: $a_j(\rho)$ for Matérn $\nu = 0.5$ and Matérn $\nu = \infty$, for two values of ρ, for $j \in [1, 2, \ldots, 63, 64]$. Center panel: $a_j(\rho)$ for Matérn $\nu = 0.5$ for different values of ρ for $j \in [1, 2, \ldots, 199, 200]$. Right panel: Circles: ${v_j}^2$'s for data simulated from GP with $\sigma_s^2 = 2$, $\sigma_e^2 = 5$ and $\rho = 5$; line: $\sigma_s^2 a_j(\rho) + \sigma_e^2$.

$$\left(\cos(\omega_{\frac{j+1}{2}} 2\pi), \cos(\omega_{\frac{j+1}{2}} 2\pi 2), \ldots, \cos(\omega_{\frac{j+1}{2}} 2\pi M)\right)';$$
$$j = M - 1.$$

Finally, $u \sim N(0, G)$

$$\text{for } G = \sigma_s^2 \text{Diag}\left(\frac{1}{2M}\phi(\omega_{m(1)}; \rho), \frac{1}{2M}\phi(\omega_{m(2)}; \rho), \ldots, \frac{1}{2M}\phi(\omega_{m(M-2)}; \rho), \frac{1}{M}\phi(\omega_{m(M-1)}; \rho)\right),$$

with $m(j) = j/2$ for even j and $m(j) = (j+1)/2$ for odd j. The coefficient a_0 in (7) is not identified if an intercept is included in the model, so it has been omitted.

Z does not depend on unknowns; $Z'Z = \text{Diag}(2M, 2M, \ldots, 2M, M)$ and $Z'1_M = \mathbf{0}$, that is, Z's columns are orthogonal to each other and to the constant vector (proofs are in the Web Appendix A). The successive columns capture trends in the data corresponding to increasing frequencies, with the elements of u being the weights for these trends.

In the spectral approximation, the $a_j(\rho)$'s defined in Section 2.1 are given by

$$a_j(\rho) = \phi(\omega_{m(j)}; \rho),$$

with $m(j) = j/2$ for even j and $m(j) = (j+1)/2$ for odd j, $j = 1, 2, \ldots, M-1$. For example, for the exponential correlation function (Matérn with smoothness parameter $\nu = 0.5$),

$$K(s_i, s_j; \rho) = \exp(-\sqrt{2}|s_i - s_j|/\rho), \tag{9}$$

and $\phi(\omega; \rho)$ has the form of a Cauchy density

$$\phi(\omega; \rho) = \frac{1}{\sqrt{2}}\rho\left(1 + \frac{(\pi\rho)^2}{2}\omega^2\right)^{-1}. \tag{10}$$

For any ν, the Matérn(ν) correlation function corresponds to a particular function $a_j(\rho)$. Figure 1's left panel shows $a_j(\rho)$ for the Matérn(ν) for $\nu = 0.5$ and ∞. Given ρ, the $a_j(\rho)$ for different ν hardly differ, indicating how little information about ν the data can provide. This corroborates the well-known fact that ν is generally difficult to estimate from data.

A known aspect of the spectral approximation is that $a_j(\rho)$ is non-increasing in j and approaches zero for large j. The a_j's start higher and decline faster as ρ increases (Figure 1 center panel).

Empirically, we observe that the spectral approximation causes the correlation, as a function of the distance between observations, to decrease to zero at a faster rate than it should (Figure 2 and Web Appendix E). In one dimension, maximizing the approximate restricted likelihood compensates by making the estimate of ρ (range) larger than the estimate from the exact restricted likelihood. Also, the approximate process is periodic: $g(0) = g(2\pi)$ (Figures S1 and S2 in Web Appendix B show examples). To mitigate this, Royle and Wikle (2005) use a grid larger than the observation domain, known as padding. Paciorek (2007) pads by mapping the periodic domain $(0, 2\pi)$ to $(0, 2)$ and then mapping the observation domain onto $(0, 1)$. For our purpose, v_j must be a function of the data, so we cannot pad in this way. This is, admittedly, a weakness of the approximation but does not imply that it is useful only for analyses of periodic functions; as noted, the approximation's utility arises from its ability to provide insight, which is reasonably unimpaired, as we now argue.

3. Conjectures About Parameter Estimates

Recall from Section 2.2 that for fixed ρ, the approximate restricted likelihood (5) is identical to the likelihood from a

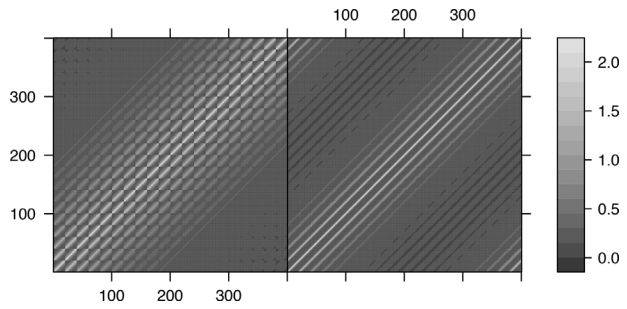

Figure 2. Left: exact 2-D exponential covariance (including the iid errors). Right: approximate covariance. $\sigma_s^2 = 2$, $\sigma_e^2 = 0.1$, $\rho = 5$. Observation domain $[1, 2, \ldots, 20] \times [1, 2, \ldots, 20]$.

gamma-errors GLM with identity link. This gives a way to generate conjectures about how parameter estimates are fit to data.

The matrix Z is the same for all GPs on a given location set so given y, the v_j^2's are also the same for all GPs; the only thing distinguishing GP models for u is their $a_j(\rho)$'s. The v_j^2's are the "data" and the parameters are fit for a model with $\mathrm{E}(v_j^2|\sigma_s^2, \rho, \sigma_e^2) = \sigma_s^2 a_j(\rho) + \sigma_e^2$ and $\mathrm{Var}(v_j^2|\sigma_s^2, \rho, \sigma_e^2) = 2\left(\sigma_s^2 a_j(\rho) + \sigma_e^2\right)^2$.

Because $a_j(\rho)$ approaches 0 for large j, $\mathrm{E}(v_j^2|\sigma_s^2, \rho, \sigma_e^2) \approx \sigma_e^2$ for large j; heuristically, v_j^2's for large j are more informative about σ_e^2 and v_j^2's for small j are more informative about σ_s^2 and ρ. Because $a_j(\rho)$ is non-increasing in j, the "data" v_j^2 have higher variance for smaller j, so the data provide more information about σ_e^2 than about σ_s^2 or ρ. For Matérn correlation functions like (10), $\sigma_s^2 a_j(\rho)$ has the form $\sigma_s^2 \rho f(\rho, \omega)$; σ_s^2 and ρ are identified in the approximate restricted likelihood (5) only by $f(\rho, \omega)$, which describes how $a_j(\rho)$ declines with j, and $\hat{\rho}$ and $\hat{\sigma}_s^2$ are chosen to fit this rate of decline to the v_j^2's. The noise in v_j^2 is a function of j so σ_s^2 and ρ are not always well identified; this and lack of consistency in joint estimates of σ_s^2 and ρ on a fixed domain are well-known problems (Ying 1991, Zhang 2004).

Figure 1's right panel shows the v_j^2's and $\sigma_s^2 a_j(\rho) + \sigma_e^2$ for a dataset simulated with exponential correlation function (9) and true $\sigma_s^2 = 2$, $\sigma_e^2 = 5$, and $\rho = 5$. As described, σ_s^2, σ_e^2, and ρ are estimated so that the $\hat{\sigma}_s^2 a_j(\hat{\rho}) + \hat{\sigma}_e^2$ fit the v_j^2 as best they can. (The estimates are in Web Appendix C, Table S1.) Thus, the GLM formulation (5) allows us to visualize how features in the data produce the parameter estimates. Sections 3.1–3.2 present and test some conjectures about how features of the data affect the parameter estimates.

To do this, data were simulated from a GP with mean 0 and correlation function (9) and normal$(0, \sigma_e^2)$ errors using each of Table 1's eight combinations of true parameter values, with observations at locations $\{1, 2, \ldots, 199, 200\}$. 100 datasets were simulated for each combination. We call these simulated datasets uncontaminated data. Parameter estimates were obtained by maximizing the exact log restricted likelihood; estimates were also obtained by maximizing the approximate log restricted likelihood. Table 1 presents averages of the estimates over these 100 datasets with Monte Carlo standard errors.

We then re-fit the GPs to two kinds of contaminated data, contaminating by:

- an outlier: the 100th observation was replaced by 18 (Section 3.1);
- a mean shift: 5 was added to the last 100 observations (Section 3.2).

Web Appendix C, also shows simulation results for a contamination in which the GP's range parameter ρ was changed halfway through the series.

3.1. *Outlier*

How does an ordinary outlier affect the estimates of the GP parameters?

We conjecture that an outlier will inflate the v_j^2's for larger j's, corresponding to high frequencies. Because $\hat{\sigma}_e^2$ is driven largely by those v_j^2's, $\hat{\sigma}_e^2$ will be inflated. The v_j^2's for smaller j's (low frequencies) will be comparatively unaffected, so the outlier will have little effect on $\hat{\sigma}_s^2$ and $\hat{\rho}$. These two effects lead to a smoother fit.

For all eight true parameter combinations, the most striking effect of the outlier contamination is in fact an inflated $\hat{\sigma}_e^2$ (Table 1), as conjectured (Figure S19a in Web Appendix C shows this for one dataset). Figure S19b in Appendix C shows that the fit is indeed smoother when the outlier is present (Appendix C, Table S1 gives the estimates). An outlier at the end of the data series has a similar effect (not shown).

3.2. *Mean Shift*

Consider the data in Figure 3's left panel, a draw from a GP with a shift in the mean halfway through the series. How does this mean shift affect the stationary GP's parameter estimates?

The contaminated data look most like the 2*nd* column of the Z matrix (Figure 3 center panel's upper right), so we conjecture that the v_j^2 arising from this column, v_2^2, will be greatly increased, which in turn will inflate $\hat{\sigma}_s^2$. The large v_2^2 will cause the v_j^2's to decline more sharply in j, so $\hat{\rho}$ will be inflated to capture that decline (recall Figure 1's center panel). The mean shift, a low frequency data feature, will not affect the v_j^2's for large j's, so $\hat{\sigma}_e^2$ should change little.

The v_j^2 arising from Z's second column, v_2^2, is indeed affected most by the mean-shift contamination (Figure 3's right panel shows this for one simulated dataset) leading, as conjectured, to inflated $\hat{\sigma}_s^2$ and $\hat{\rho}$ (Tables S1 and S2 in Web Appendix C).

4. Regressing Out Covariates

The spectral representation above is for an intercept-only GP model. If the fixed-effect design matrix X is not just a vector of ones, we propose first regressing it out as follows, and then applying Section 2's spectral representation. Let $P_X = X(X'X)^{-1}X'$ be the orthogonal projector onto X's column space. Premultiply both sides of (2) by $(I_M - P_X)$ to give

$$y^* = \gamma^* + \epsilon^*, \tag{11}$$

where $y^* = (I_M - P_X)\, y$ is the residual from a regression on X, $\gamma^* = (I_M - P_X)\, \gamma$, and $\epsilon^* = (I_M - P_X)\, \epsilon$, with $\mathrm{Cov}(\gamma^*) = (I_M - P_X)\, \Sigma\, (I_M - P_X)$ and $\mathrm{Cov}(\epsilon^*) = \sigma_e^2 (I_M - P_X)$. The likelihood arising from (11) is the restricted likelihood of the original model (2).

If $\mathrm{rank}(X)$ is small compared to M, these approximations are reasonable: $\mathrm{Cov}(\gamma^*) \approx \Sigma$, and $\mathrm{Cov}(\epsilon^*) \approx \sigma_e^2 I_M = R$, that is, ignore changes in $\mathrm{Cov}(y^*)$ induced by the residual projection, as standard linear-model diagnostics do. Priestley (1981, Ch. 7) discusses fitting stationary processes to residuals from least-squares fits. If the residuals arise from a polynomial fit, then the spectral densities estimated from y^* and y have the same asymptotic properties.

Thus, we assume the residuals γ^* after regressing on the covariates can be approximately modeled by a GP having the same covariance form as γ and the errors ϵ^* approximately modeled by the same form as ϵ, *with possibly different parameter values*. With the

Table 1

Average over 100 simulated datasets of estimates maximizing the exact and approximate restricted likelihoods; contamination by an outlier. Standard errors are in parentheses.

	Exact RL			Approximate RL		
	σ_s^2	σ_e^2	ρ	σ_s^2	σ_e^2	ρ
Actual values	2	5	5	2	5	5
Uncontaminated	2.29 (0.11)	4.75 (0.11)	6.89 (0.82)	2.39 (0.16)	4.90 (0.09)	13.51 (1.74)
Contaminated	2.44 (0.10)	5.99 (0.11)	7.25 (0.60)	2.48 (0.18)	6.04 (0.11)	12.76 (1.21)
Actual values	2	5	16.67	2	5	16.67
Uncontaminated	2.16 (0.09)	4.89 (0.07)	26.12 (4.32)	2.17 (0.09)	4.91 (0.06)	38.11 (2.53)
Contaminated	2.23 (0.11)	6.14 (0.09)	20.11 (1.84)	2.39 (0.11)	6.07 (0.08)	33.99 (2.74)
Actual values	2	0.1	5	2	0.1	5
Uncontaminated	1.94 (0.03)	0.099 (0.01)	4.97 (0.13)	1.97 (0.04)	0.232 (0.01)	10.24 (0.25)
Contaminated	1.98 (0.05)	1.36 (0.03)	5.46 (0.24)	2.03 (0.05)	1.45 (0.03)	10.94 (0.94)
Actual values	2	0.1	16.67	2	0.1	16.67
Uncontaminated	2.09 (0.07)	0.093 (0.00)	18.07 (0.84)	2.01 (0.08)	0.144 (0.00)	34.84 (1.67)
Contaminated	2.08 (0.06)	1.33 (0.02)	17.69 (1.04)	2.03 (0.08)	1.39 (0.02)	34.57 (1.76)
Actual values	10	5	5	10	5	5
Uncontaminated	10.17 (0.20)	4.99 (0.13)	5.58 (0.18)	9.98 (0.21)	5.63 (0.10)	10.87 (0.41)
Contaminated	10.02 (0.22)	6.01 (0.15)	5.59 (0.24)	10.12 (0.22)	6.70 (0.13)	10.24 (0.40)
Actual values	10	5	16.67	10	5	16.67
Uncontaminated	10.44 (0.37)	4.80 (0.06)	17.57 (0.83)	10.54 (0.43)	5.08 (0.06)	35.60 (2.05)
Contaminated	9.65 (0.29)	6.24 (0.09)	17.99 (1.01)	10.27 (0.39)	6.59 (0.08)	37.32 (2.64)
Actual values	10	0.1	5	10	0.1	5
Uncontaminated	9.70 (0.17)	0.21 (0.03)	5.26 (0.13)	9.65 (0.17)	0.76 (0.03)	10.34 (0.28)
Contaminated	10.17 (0.19)	1.36 (0.08)	5.19 (0.13)	9.55 (0.19)	2.09 (0.07)	10.12 (0.32)
Actual values	10	0.1	16.67	10	0.1	16.67
Uncontaminated	9.99 (0.34)	0.11 (0.01)	17.29 (0.67)	10.02 (0.37)	0.32 (0.01)	32.70 (1.37)
Contaminated	10.52 (0.31)	1.39 (0.05)	18.14 (0.76)	10.30 (0.35)	1.59 (0.05)	34.29 (1.29)

approximation $\mathrm{Cov}(y^*) \approx \Sigma + R$, the model becomes

$$y^* = I_M \gamma^* + \epsilon^*, \tag{12}$$

where $y^* = (I_M - P_X)\, y$, $\gamma^* = (I_M - P_X)\, \gamma \overset{approx}{\sim} \mathrm{GP}(0, \Sigma)$, $\epsilon^* = (I_M - P_X)\epsilon \overset{approx}{\sim} N(0, R)$.

It is helpful to see what this approximation does in practice. When X is a column of 1's and Z is Section 2.3's spectral basis matrix, v_j is the unshrunk projection of y onto the jth column of $W = Z(Z'Z)^{1/2}$. If we add fixed effects to X and proceed as proposed, replacing y with $y^* = (I_M - P_X)y$ but keeping the same Z, then the unshrunk projections of y^* onto the columns of W, that is, the $W'y^*$, are

$$W'(I_M - P_X)y = (Z'Z)^{1/2}Z'y - (Z'Z)^{1/2}Z'P_X y = v - (Z'Z)^{1/2}Z'P_X y \tag{13}$$

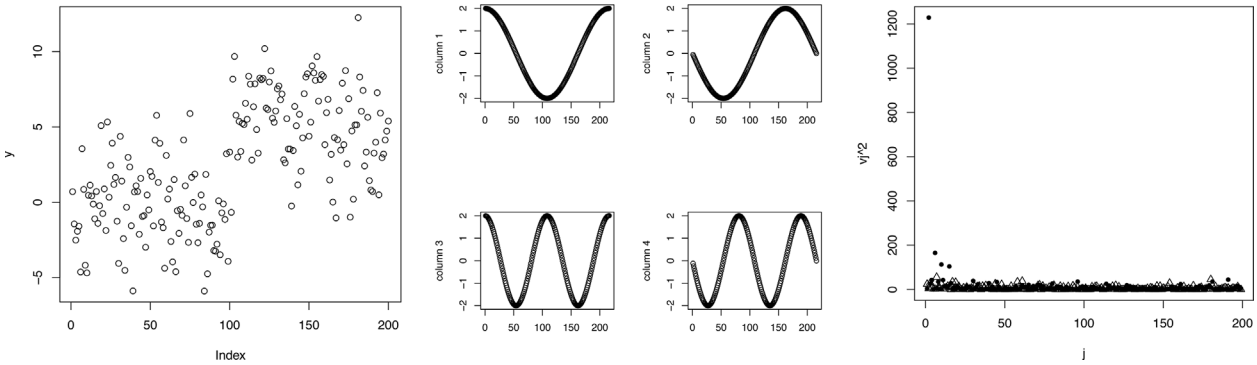

Figure 3. Left panel: Data simulated from GP with $\sigma_s^2 = 2$, $\sigma_e^2 = 5$, and $\rho = 5$ with mean shift from 0 to 5 midway. Center panel: First four columns of Z, the spectral basis matrix, on the domain [1, 2,..., 199, 200]. Right panel: Dots: ${v_j}^2$'s from uncontaminated simulated data; triangles: ${v_j}^2$'s from data with mean shift.

$$= v - [P_X Z(Z'Z)^{1/2}]'y. \quad (14)$$

Each v_j is reduced by an amount depending on how much of it "goes away" when y is projected onto the orthogonal complement of X's column space, as in (13), or by an amount determined by the projection of Z's jth column onto X, as in (14). The approximation $\text{Cov}(y^*) \approx \Sigma + R$ reduces y's projection onto Z's jth column to an extent depending on how X (collectively) is correlated with Z's jth column. Note that in computing the exact restricted likelihood, all of y's variation in X's column space is attributed to X; the approximation retains this key feature.

For another view, consider the Kullback–Leibler distance between the densities $N(0, \Sigma + R)$ and $N(0, P\Sigma P + R)$,

$$0.5 \log[\det(P\Sigma P+R)/\det(\Sigma+R)] - M + \text{trace}[(P\Sigma P+R)^{-1}(\Sigma+R)],$$

where P is an $M \times M$ projector of rank $M - p$. If $(P\Sigma P + R)^{-1}(\Sigma + R)$ is approximately the identity, this distance is small, as is the case here.

Figures S3–S18 in Web Appendix B show $\text{Cov}(y^*)$ and $\text{Cov}(y)$ for various M, σ_s^2/σ_e^2, ρ, correlation functions, and X. These figures show that the structure of $\text{Cov}(y^*)$ is (for our purposes) satisfactorily approximated by the functional form of $\text{Cov}(y)$.

5. A Small Toolkit for Assessing Goodness of Fit and Considering Covariates

Conventional residuals can highlight a few extreme outliers in data space but cannot identify lack of fit, especially in the GP part of the model. Indeed, any model of this type can be made to fit any y arbitrarily well by setting ρ small and σ_s^2 large. This section shows how to use the tools from earlier sections to avoid this problem, in particular highlighting lack of fit in the GP part of the model. When covariates are available, we present tools for considering which covariates to add. If no covariates are available, the tools identify properties of potential covariates, to aid in seeking them.

When covariates (potential fixed effects) are available, a modeler must make a choice: either let the fitting machinery interpret strong low-frequency data features as evidence of stationary GP errors with large σ_s^2 and ρ, or attribute those features to covariates to the extent possible. Whatever your view on this matter, it is essential to know whether such features are present so a well-informed choice can be made. A plot with v_j^2 on the vertical axis and j on the horizontal axis (henceforth "the v_j^2 plot") shows such prominent low-frequency data features as large v_j^2 for small j. If, in the same plot, some v_j^2 are large for large j, that is evidence of outliers (high frequency trends). Generally, a large v_j^2 suggests a missing covariate with high power at the frequency corresponding to j. A polynomial or sinusoidal curve of that frequency could be added and this may be defensible in some cases, for example, a linear term parallel to a coordinate axis or an annual cycle. However, adding substantively meaningful covariates is generally more satisfactory.

The v_j^2 plot is visually dominated by low frequency j; outlying y_i or high frequency trends in y will be spread out among high-frequency v_j^2 and may not be visible in the v_j^2 plot. If potential covariates are available, however, a modest adaptation of the familiar added variable plot can be used to examine covariates irrespective of their prominent frequencies. We now describe added variable plots in the observation and spectral (frequency) domains.

5.1. *Added Variable Plots*

In an ordinary linear model, the added variable plot for a candidate predictor C is drawn as follows (Cook and Weisberg 1982, p. 44; Atkinson, 1985, Section 5.2):

(1) Compute residuals from regressing the outcome y on all predictors except C.
(2) Compute residuals from regressing C on all predictors other than C.
(3) Plot the residuals from steps 1 and 2 on the vertical and horizontal axes, respectively.
(4) Fit a regression through the origin to the plotted data; this estimates the coefficient of C if it were included in the model.

This usual added variable plot assumes errors are independent with constant variance. A linear mixed model with a GP random effect has neither property; for this model, we describe how to adapt added variable plots in both the observation and spectral domains. Both plots estimate the same slope for the covariate C.

5.1.1. Observation domain. Consider adding C to give the model $y = X\beta + C\alpha + \gamma + \epsilon$, where X contains predictors already in the model including the intercept, γ is a GP-distributed random effect, and ϵ is iid normal errors. Pre-multiply both sides of the model equation by $\hat{V}^{-0.5}$, where $\hat{\cdot}$ denotes estimates from fitting the model without $C\alpha$. Then pre-multiply both sides of the model equation by $\hat{P} = I - \hat{V}^{-0.5}X(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-0.5}$ to give $\hat{P}\hat{V}^{-0.5}y = \hat{P}\hat{V}^{-0.5}C\alpha + \hat{P}\hat{V}^{-0.5}(u + \epsilon)$. The added variable plot shows $\hat{P}\hat{V}^{-0.5}y$ vs $\hat{P}\hat{V}^{-0.5}C$.

5.1.2. Spectral domain. For the same model equation, pre-multiply both sides by $(I - P_X)$, then pre-multiply by $(Z'Z)^{-0.5}Z'$, then pre-multiply by $\hat{D} = \text{Diag}(1/\sqrt{\hat{\sigma}_s^2 a(\hat{\rho}) + \hat{\sigma}_e^2})$, to give $\hat{D}v^* = \hat{D}v_C^*\alpha + \hat{D}(Z'Z)^{-0.5}Z'(\gamma + \epsilon)$, where $v^* = (Z'Z)^{-0.5}Z'(I - P_X)y$ are the v_j from the residuals y^* and $v_C^* = (Z'Z)^{-0.5}Z'(I - P_X)C$ are v_j from the residuals for C. This added variable plot shows $\hat{D}v^*$ vs $\hat{D}v_C^*$.

A common method for assessing the effect of a predictor is to fit the model with and without the predictors and compare the fits; the model is fit twice, which may be inefficient depending on the cost of a model fit. In contrast, added variable plots do not require re-fitting the model for each predictor and also show the effect of individual observations.

In the example below and the Web Supplement, we demonstrate the model building tools described here, that is, the v_j^2 plot and the added variable plots. (Web Appendix G uses a simulated example; Appendices H, I, and J use Finley et al's (2008) forest biomass data.) These examples show how a low-frequency covariate can have a weak signal in the observation-domain added variable plot but a strong signal in the spectral-domain added variable plot; similarly, the two

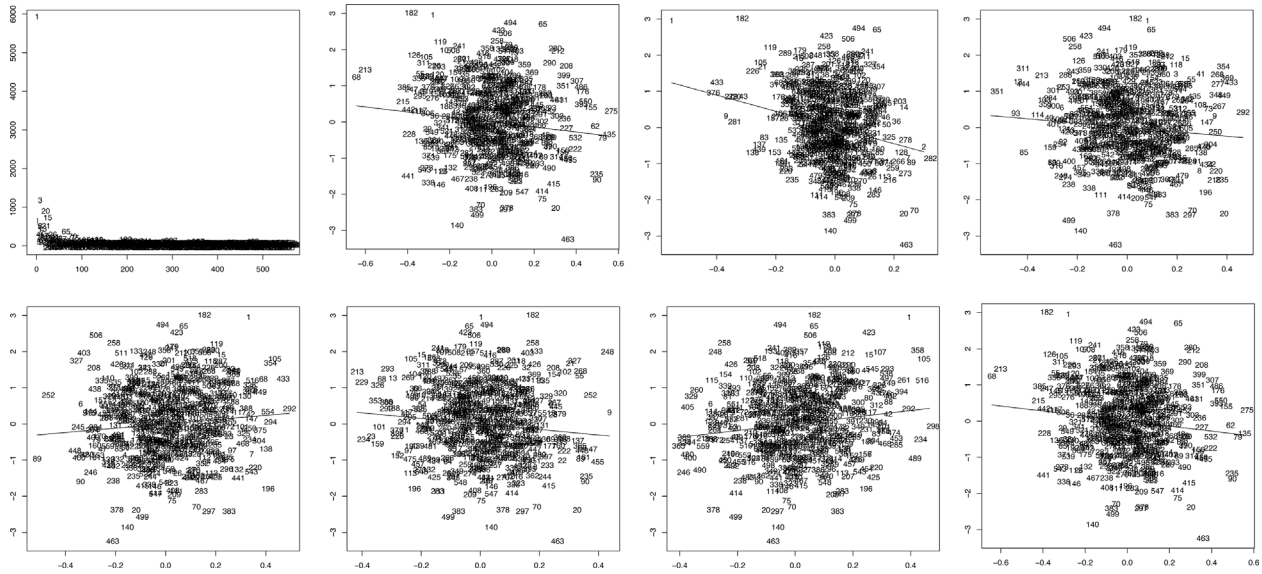

Figure 4. After intercept-only fit; leftmost figure in top line is v_j^2 vs j. All other figures are spectral-domain added variable plots; Top line second-left to right: Elevation, Slope, SpringTC2; Bottom line left to right: SpringTC3, SummerTC1, SummerTC3, FallTC2.

added variable plots may have signals of differing strength for a covariate with power mainly in high frequencies.

6. Gaussian Process on Two Dimensions

Section 2 described the spectral basis and GLM interpretation of the approximate restricted likelihood for observations at locations in one dimension. Spatial data are commonly observed at two-dimensional locations; this section outlines derivation of the 2-dimensional (2-D) approximate restricted likelihood; Web Appendix D gives details.

For observations on an equally spaced $M_1 \times M_2$ grid, the restricted likelihood has the same form as (3) and as in the 1-D case, we approximate the intercept-only GP using spectral basis functions. If the model includes fixed effects X, we proceed as in Section 4's 1-D case, that is, assume (approximately) that the residuals follow a linear mixed model with a GP-distributed random effect, then approximate this GP using the spectral basis. For the 2-D model, the approximate log restricted likelihood has the matrix-free form

$$\mathrm{ALR}(\sigma_s^2, \sigma_e^2, \rho) = \mathrm{const} - \frac{1}{2} \sum_{j=1}^{M_1 M_2 - 1} \left(\log(\sigma_s^2 a_j(\rho) + \sigma_e^2) + v_j^2 (\sigma_s^2 a_j(\rho) + \sigma_e^2)^{-1} \right); \quad (15)$$

the $a_j(\rho)$ are sorted to be non-increasing in j; it is easy to prove this order is invariant to ρ.

We can now ask for the 2-D case how GP fits respond to features in the data. As in the 1-D case, by construction sin/cos column pairs of the spectral basis matrix Z decompose the data into frequency components. Thus a high frequency feature, for example, an outlier, falls in the space spanned by columns of Z with large j and inflates v_j^2 for larger j, which in turn inflates $\hat{\sigma}_e^2$. A low frequency feature in y, for example, a linear trend, inflates v_j^2 for smaller j, which in turn inflates $\hat{\rho}$ and $\hat{\sigma}_s^2$.

For data observed on a regular grid, we have made two approximations: approximating $(I_{M_1 M_2} - P_X)\ \Sigma\ (I_{M_1 M_2} - P_X) + \sigma_e^2 (I_{M_1 M_2} - P_X)$ by $\Sigma + R$, and the spectral approximation. If the observation locations are not on a regular grid, we suggest first smoothing the data onto such a grid, as follows.

6.1. *Smoothing Observed Data onto a Regular Grid*

Construct a rectangular uniformly spaced grid on the observation domain; the grid size must be a multiple of 2 in each dimension. Label the grid locations $\{1, 2, \ldots, M_1\} \times \{1, 2, \ldots, M_2\}$ and re-scale the actual observation domain to $[1, M_1] \times [1, M_2]$. To minimize space in the grid with no observations, the map of observation locations may need to be rotated to make it more nearly rectangular, and M_1/M_2 should be close to the aspect ratio of the observation locations. Then for each location on the grid, we suggest constructing an artificial datum using inverse distance weighting (IDW) (Shepard 1968, Zimmerman 1999): if the data are $(y_1, y_2, \ldots, y_n)'$ at 2-D locations $(s_1, s_2, \ldots, s_n)'$, the value at grid location ϱ is

$$\frac{\sum_{k=1}^n t_k\ y_k}{\sum_{k=1}^n t_k}, \qquad \text{where} \qquad t_k = \frac{1}{d(\varrho, s_k)^\lambda}, \quad (16)$$

where d is Euclidean distance and λ is a tuning constant.

In Section 7's example, we chose M_1, M_2, and λ so σ_s^2, σ_e^2, and ρ would have estimates as similar as possible from the artificial and actual data. Other choices of M_1, M_2, and λ would give results qualitatively similar to those we present. Web Appendix K describes a simulation experiment showing the consequences of different M_1, M_2, and λ. Broadly, IDW largely preserves y's low-frequency trends while sacrificing

Table 2
Slopes of spectral-domain added variable plots, after the intercept-only fit

Candidate covariate	Slope	p-value	j with top 5 Cook's dist
Elevation	−3.17	10^{-10}	1,182,9,181,434
Slope	−2.24	10^{-9}	1,182,463,70,65
SpringTC2	−0.60	0.03	20,499,354,268,369
SpringTC3	0.59	0.02	1,506,196,403,463
SummerTC1	−0.77	0.007	248,463,20,268,327
SummerTC3	0.92	0.0004	1,258,378,358,248
FallTC2	−0.69	0.004	182,463,1,212,280

power at high frequencies, less so for larger λ. Thus $\hat{\sigma}_\epsilon^2$ is smaller for the artificial data than for the actual data, $\hat{\rho}$ is inflated slightly, and $\hat{\sigma}_s^2$ is affected but not in a systematic way. Added variable plots in the observation domain preserve y's high-frequency information; spectral-domain plots lose some of it. IDW's main virtue for our purpose is computing speed; an ideal method, if one exists, would minimally affect power at all frequencies.

7. Application to Data

We illustrate use of the tools with the 2002 forest inventory data analyzed by Finley et al. (2008) and included with the R package spBayes (BEF.dat; Finley et al. 2007). The outcome y is red maple total basal area (RM_02BAREA × BAREA02_TOT) and potential predictors are ELEV, SLOPE, SPR_02_TC2, SPR_02_TC3, SUM_02_TC1, SUM_02_TC3, and FALL_02_TC2, all measured at 437 locations. The data were smoothed to a grid using IDW with grid size 28 × 20, with $\lambda = 7$ for y and $\lambda = 9$ for the predictors. We fit an intercept-only model in which the GP had the exponential covariance function, that is, Matérn with $\nu = 0.5$, then used the tools to examine the fit and consider adding predictors. Section 7.1 considers the spectral-transformed data and spectral-domain added variable plots for selecting predictors, Section 7.2 considers observation-domain added variable plots and Section 7.3 shows fits and tests for added variables using the exact restricted likelihood and the original data. Here, we show just the first of a sequence of model-building steps; Web Appendices H, I, and J give all of the steps. We present this only to illustrate how the tools could be used; we make no claim that these steps are optimal.

Table 3
Slopes of observation-domain added variable plots, after the intercept-only fit

Candidate covariate	Slope	p-value
Elevation	−2.02	0.07
Slope	−1.45	0.004
SpringTC2	−0.28	0.42
SpringTC3	0.96	0.002
SummerTC1	−0.98	0.002
SummerTC3	1.25	10^{-5}
FallTC2	−0.83	0.004

7.1. *Model-Building in the Spectral Domain*

The questions are: does the intercept-only model suffice to explain variation in y or should covariates be added and if so, which ones? A stationary GP model is flexible enough that the fit to y is never bad but we presume that apparent deviations from stationarity are better modeled using covariates than with the GP.

Consider Figure 4's leftmost figure in the top panel, the v_j^2 plot from the intercept-only fit; the plotting symbol is j. The point $j = 1$ corresponds to a cubic or linear north-south trend (the spectral basis has no linear-like component), so the huge v_1^2 indicates a north-south trend. Adding a covariate could remove this trend. Figure 4 shows spectral-domain added variable plots for the candidate covariates, each of which (after pre-smoothing to the grid) has been standardized by subtracting its average and dividing by its standard deviation, so the covariates' slopes are comparable. Table 2 suggests that adding covariates will improve the fit. Note that for elevation and slope, Figures 4's top panel second and third figures from left, the signal for adding the covariate is concentrated in few transformed observations with extreme values on the horizontal axis, including the lowest frequencies $j = 1$ and 2, while the signal for the other covariates is more diffuse in j.

The natural impulse is to add the covariate with the largest slope, which is Elevation. However, because v_1^2, a north-south trend, is the most prominent v_j^2, we also want the added covariate to explain some of this trend, the more the better. To this end, we calculated Cook's distance, describing the influence of each point on the regression slope; we want $j = 1$ to have a large Cook's distance. Among the candidate covariates, Elevation has the smallest p-value for its added variable plot's slope and $j = 1$ has the largest Cook's distance. Therefore, we add Elevation to the model. (Slope has results similar to Elevation so we could have added Slope instead. Had we done so, then applying the same considerations in the next step would lead to adding Elevation next.)

Web Appendices H, I, and J show further model-building steps. Again, we make no claim this is an optimal sequence of steps but merely intend to illustrate how the tools can be used.

7.2. *Added Variable Plots in the Observation Domain*

Added variable plots in both domains estimate the same slope (apart from effects of the approximations used for the spectral-domain plot) but they emphasize different aspects of the outcome y and the candidate predictors. Table 3 shows slopes and p-values for observation-domain added variable

Table 4
Exact restricted likelihood maximized for the raw data: estimated coefficient and Wald-test p-value for each candidate covariate, added individually (one at a time) to the model, and resulting estimates of σ_s^2, σ_e^2, and ρ.

	Coefficient		Variance parameters		
Candidate covariate	slope	p-value	$\hat{\sigma}_s^2$	$\hat{\sigma}_e^2$	$\hat{\rho}$
Intercept-only	–	–	29.62	16.20	5.96
Elevation	−2.52	$<10^{-12}$	21.96	13.82	2.85
Slope	−1.63	10^{-6}	20.31	16.11	3.97
SpringTC2	−0.28	0.16	29.69	16.35	6.13
SpringTC3	0.99	10^{-5}	26.80	17.15	6.93
SummerTC1	−1.03	10^{-5}	30.98	17.54	8.88
SummerTC3	1.25	10^{-7}	26.91	16.19	6.14
FallTC2	−0.87	0.0001	26.50	17.33	6.98

plots, comparable to Table 2. The most striking differences are for Elevation and Slope: their strong low-frequency components are emphasized by the spectral-domain added variable plot but de-emphasized by the observation-domain plot. Thus, although in the observation-domain plots these two predictors have the largest slopes in absolute value (Table 3), they have much larger p-values than in the spectral-domain plots, reflecting lower power in the observation domain. Among the other predictors, Spring TC3, Summer TC1, and Summer TC3 have somewhat larger slopes and somewhat smaller p-values in the observation domain, reflecting their relative strength in high frequencies.

7.3. *Fits Using the Raw Data and Exact Restricted Likelihood*

Table 4 shows estimates of coefficients of the candidate covariates using the original data y and maximizing the exact restricted likelihood to estimate σ_s^2, ρ, and σ_e^2. These estimates are almost exactly equal to the slopes of the observation-domain added variable plots (Section 7.2), as might be expected given that the latter plot involves less approximation. The p-values in Table 4 are from Wald tests that treat estimates of σ_s^2, ρ, and σ_e^2 as if they are known to be true, as is typical in non-Bayesian analyses, which ignores variability accounted for in the p-values for both added variable plots.

Table 4 also shows $\hat{\sigma}_s^2$, $\hat{\sigma}_e^2$, and $\hat{\rho}$ for the intercept-only model and for models adding each candidate covariate. Elevation and Slope, which have strong low-frequency components and low p-values in the spectral-domain added variable plots, have the biggest impact on $\hat{\sigma}_s^2$ and $\hat{\rho}$. Entering either moves substantial variation out of the GP part of the fit and into the fixed effects, leaving the GP part of the fit with less variation (σ_s^2) and spatial correlation range (ρ). The other candidate covariates have much smaller effects on estimates of the GP parameters. Elevation is the only candidate covariate with a noteworthy impact on $\hat{\sigma}_e^2$. Web Appendices H, I, and J show analogous estimates for later steps in the model-building process.

8. Discussion

This article is a step toward tools for doing data-analysis with linear mixed models with a random effect distributed as a stationary Gaussian process, specifically tools for model-building and understanding model fits. We used the spectral approximation for stationary isotropic GPs on regularly spaced grids to give a linear mixed model in which the random effect has a design matrix with orthogonal columns not depending on any unknowns and a diagonal covariance matrix. The resulting approximate restricted likelihood is formally identical to the likelihood from a GLM with gamma errors and identity link. The transformed observations v_j^2 and the functions $a_j(\rho)$—the spectral density of the GP's covariance—are the keys to understanding the restricted likelihood as a function of the unknowns. The spectral approximation could also contribute to understanding spatial confounding; this is secondary to the present purpose but is discussed in Web Appendix L.

The approximate restricted likelihood fits the v_j^2's to $\sigma_s^2 a_j(\rho) + \sigma_e^2$, so a prominent v_j^2, especially for small j, suggests a missing covariate with substantial power at the corresponding frequency. If covariates are available, added variable plots in the observation and spectral domains can be used to examine the support in the data for adding each potential covariate.

The spectral representation requires data observed on an regular grid; we suggested a way to apply these methods to data observed at irregularly spaced locations but other approaches are possible. We view the present work as a beginning, not a definitive approach. Finally, we used a dataset to sketch how to use the model building tools developed here. A future publication will include a more fully worked example.

9. Supplementary Materials

Web Appendices A, E, F, referenced in Section 2; Web Appendix B, referenced in Sections 2 and 4; Web Appendices C, D, G, K, L, referenced in Section 3, Section 5, Section 6, Section 8, respectively, and Web Appendices H, I, J, referenced in Sections 6 and 7 are available with this article at the *Biometrics* website on Wiley Online Library. R code implementing the spectral approximation and constructing added variable plots for 2-D data is available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors would like to thank the editor and two anonymous referees for several constructive comments regarding

the article. The work of Professor Banerjee was supported by NIH/NIEHS R01-ES027027, from NSF DMS-1513654 and NSF IIS-1562303.

References

Atkinson, A. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* New York: Oxford University Press.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: CRC Press.

Chiles, J. P. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty* (Vol. 497). New York, NY: John Wiley & Sons.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

Cressie, N. (2015). *Statistics for Spatial Data*. New York, NY: John Wiley & Sons.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software* **19**, 1–24.

Finley, A. O., Banerjee, S., Ek, A. R., and McRoberts, R. E. (2008). Bayesian multivariate process modeling for prediction of forest attributes. *Journal of Agricultural, Biological, and Environmental Statistics* **13**, 60–83.

Finley, A. O., Banerjee, S., Waldmann, P., and Ericsson, T. (2009). Hierarchical spatial modeling of additive and dominance genetic variance for large spatial trial datasets. *Biometrics* **65**, 441–451.

Fuentes, M. (2006). Testing for separability of spatial-temporal covariance functions. *Journal of Statistical Planning and Inference* **136**, 447–466.

Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2014). Do we need non-stationarity in spatial models? *arXiv preprint,* arXiv: 1409.0743.

Henn, L. and Hodges, J. S. (2014). Multiple local maxima in restricted likelihoods and posterior distributions for mixed linear models. *International Statistical Review* **82**, 90–105.

Hodges, J. S. (2014). *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models Using Random Effects*. Boca Raton, FL: Chapman and Hall, CRC Press.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially correlated errors can mess up the fixed effect you love. *The American Statistician* **64**, 325–334.

Paciorek, C. J. (2007). Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *Journal of Statistical Software* **19**, 1–38.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science* **25**, 107–125.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly spaced data. In *Proceedings of the 1968 23rd ACM National Conference*, 517–524. New York, NY.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association* **106**, 6–20.

Royle, J. A. and Wikle, C. K. (2005). Efficient statistical mapping of avian count data. *Ecological and Environmental Statistics* **12**, 225–243.

Wikle, C. (2002). Apatial Modeling of Count Data: A case study in modelling breeding bird survey data on large spatial domains. In A Lawson, D Denison (eds.), *Spatial Cluster Modelling*, Boca Raton, FL: Chapman & Hall, 199–209.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis* **36**, 280–296.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association* **99**, 250–261.

Zimmerman, D., Pavlik, C., Ruggles, A., and Armstrong, M. P. (1999). An experimental comparison of ordinary and universal kriging and inverse distance weighting. *Mathematical Geology* **31**, 375–390.

Received April 2016. *Revised September* 2017.
Accepted November 2017.