mathematics

MDPI

Article

Applied Geospatial Bayesian Modeling in the Big Data Era: Challenges and Solutions

Jason S. Byers [1,†] and Jeff Gill [2,*,†]

[1] Social Science Research Institute, Duke University, Durham, NC 27708, USA
[2] Department of Government, Department of Mathematics & Statistics, Center for Data Science, American University, Washington, DC 20016, USA
* Correspondence: jgill@american.edu
† These authors contributed equally to this work.

Abstract: Two important trends in applied statistics are an increased usage of geospatial models and an increased usage of big data. Naturally, there has been overlap as analysts utilize the techniques associated with each. With geospatial methods such as kriging, the computation required becomes intensive quickly, even with datasets that would not be considered huge in other contexts. In this work we describe a solution to the computational problem of estimating Bayesian kriging models with big data, Bootstrap Random Spatial Sampling (BRSS), and first provide an analytical argument that BRSS produces consistent estimates from the Bayesian spatial model. Second, with a medium-sized dataset on fracking in West Virginia, we show that bootstrap sample effects from a full-information Bayesian model are reduced with more bootstrap samples and more observations per sample as in standard bootstrapping. Third, we offer a realistic illustration of the method by analyzing campaign donors in California with a large geocoded dataset. With this solution, scholars will not be constrained in their ability to apply theoretically relevant geospatial Bayesian models when the size of the data produces computational intractability.

Keywords: Bayesian kriging; bootstrap; spatial modeling; big data; spatial quantiles; smoothing

MSC: 62-04; 62-08

check for updates

Citation: Byers, J.S.; Gill, J. Applied Geospatial Bayesian Modeling in the Big Data Era: Challenges and Solutions. *Mathematics* **2022**, *10*, 4116. https://doi.org/10.3390/math10214116

Academic Editors: Philip D. Waggoner, Dominique Makowski and Manuel Alberto M. Ferreira

Received: 12 September 2022
Accepted: 2 November 2022
Published: 4 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In this work, we apply the well-established and geospatial method of kriging and focus on how to estimate a Bayesian kriging model when the size of the dataset poses computational problems. Previously we have demonstrated that Bayesian implementations of spatial kriging models offer important advantages in estimation power over alternatives [1,2] since stochastic simulation tools can be readily applied for marginalizing complex joint posterior distributions [3]. We also noted in these works that a fully probabilistic description of spatial quantities is both easier and more intuitive, as first show in [4]. With Bayesian and other kriging approaches, as the sample size increases, the computational complexity is proportional to the cube of the size of the input data. With small and medium datasets, therefore, kriging models can be estimated and applied to forecasting problems by a variety of programs on a standard desktop computer. However, as the size of the data expands, the estimation problem quickly becomes a major challenge.

By way of background, kriging (named in honor of a South African mining engineer, D.G. Krige) was first developed by Georges Matheron [4–6]. The main goal with kriging is, given n observations of a random field $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))'$, how do we predict $\mathbf{Y}$ at location $\mathbf{s}_0$ when it has not been observed? Kriging is a method that produces smooth estimates of unobserved data points, which is filling in predictions at new locations based on information that is available from the observable points in the dataset. The technique of ordinary kriging uses a weighted average of observable points to estimate the unobserved

points at a given location, while universal kriging (our focus here) not only uses locational information, but also covariates as predictors. Therefore, kriging is a useful technique to better understand issues that are correlated in space (e.g., disease propagation, natural resource detection, political ideology, and so on).

One important topic that is yet to be resolved is how to use the kriging technique with large datasets. Estimating a point-referenced, or kriging, model is a problem that is $\mathcal{O}(N^3)$ in complexity (meaning that the memory needs of estimating the model are a cubic function of the data size), due to the complexity of solving the inverse of the unique covariance matrix [7] (p. 780). Therefore, as the sample size increases, estimating a point-referenced model will ultimately become too computationally rigorous to complete. In addition to that, Bayesian model specifications often require estimation from large numbers of posterior samples generated from either a Monte Carlo (MC) or a Markov chain Monte Carlo (MCMC) procedure, thus exacerbating the computation time problem [8–11]. To alleviate this issue with big data, several researchers have sought to generate new statistical methodologies that serve to help researchers circumvent the issues of kriging with large datasets [12]. We describe a recent methodological approach for Bayesian kriging and apply it to big data, thus advancing Bayesian spatial modeling. The incorporation of this specialized bootstrap technique provides a subsetting algorithm to alleviate the computational changes that come from having to use the entire dataset at once thus contributing a new tool for researchers with big data and spatial models.

While Monogan and Gill [2] provide a useful application of the Bootstrap Random Spatial Sampling (BRSS) process, primarily, we are interested in illustrating the computational and statistical properties of the technique, rather than just using it to solve a single problem. Our main contribution is to attempt to build on the extant literature centered around kriging and big data by attempting to describe the statistical properties of the BRSS and describe how this method, for large spatial data, provides consistent estimates. We provide the statistical properties of the the kriging estimation procedure that we utilize for our Bayesian model, which is a hybrid Monte Carlo feasible generalized least squares (MC-FGLS) procedure described in Section 2.2.1, that allows us to adequately sample for the joint posterior of the parameters. As is explained, the kriging estimation procedure increases in computational rigor as the size of the data increases. Therefore, we describe the bootstrapping process that is utilized and illustrate the statistical consistency of the process. Largely, we are interested in providing theoretical evidence of the statistical properties of the (BRSS) process. Additionally, we attempt to apply these statistical properties through two separate big data examples.

This work proceeds in the following way. First, we describe in detail the issues associated with kriging with big data. Next, we describe our own Bayesian procedure to kriging with large data. Third, we derive analytic properties of a customized bootstrap technique. Fourth, we conduct a Monte Carlo experiment that compares two bootstrap methods to estimation over fully-sized datasets Fifth, we provide an experiment conducted with data regarding fracking in West Virginia to show the performance of this technique. Sixth, we apply full model and estimation process to a large dataset of California donors where traditional statistical approaches fail. Finally, we discuss the implications of our results for applied researchers.

2. Materials and Methods

2.1. Why Is Bayesian Kriging Difficult with Big Data?

In the current data revolution of the 21st century, very large datasets and large data problems are becoming more common. Additionally, research is focused on how the observations are related spatially, which increases the complexity of the overall data structure. Institutions such as the U.S. Census Bureau and the U.S. Centers for Disease Control and Prevention (CDC) have large datasets for which researchers struggle to provide adequate analysis based on the complex nature of spatial modeling. Many of the issues surrounding these complexities are based on the structure of the data, which generally span large

geographic definitions (i.e., a city, state, region, country, or even the globe). Traditional estimators (be they least squares, likelihood-based, or Bayesian), which assume full information from the complete dataset are not equipped to handle large spatial data structures that require the inversion of enormous covariance matrices [13].

When applying geospatial techniques like kriging models to such large datasets, the computational problems are escalated even more than regular regression specifications. The basic specification of a Bayesian linear kriging model is as follows:

$$\begin{aligned} \mathbf{Y_s} &\sim \mathcal{N}(\mathbf{X_s}\boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ \boldsymbol{\Sigma} &= \sigma^2\mathbf{H}(\phi) + \tau^2\mathbf{I} \\ H(\phi)_{ij} &= \rho(\phi; d_{ij}) \\ \rho &= \exp(-\phi^2 d_{ij}^2) \text{ (Gaussian correlation function)} \\ \pi(\boldsymbol{\beta}) &\sim \text{flat} \\ \pi(\tau^2/\sigma^2) &\sim \text{Unif}(6, 8) \\ \pi(\sigma^2) &\sim 1/\sigma^2 \\ \pi(1/\phi) &\sim \text{Unif}(0, 9000) \end{aligned} \tag{1}$$

where $H(\phi)_{ij} = \rho(\phi, \|\mathbf{s}_i - \mathbf{s}_j\|)$, $\mathbf{s}_i$ and $\mathbf{s}_j$ are two observed spatial points,and ρ is a spatially appropriate correlation function. Here, $\mathbf{Y}$ is the vector of all values of the outcome variable, which are observed at the locations $\mathbf{s}$. We assume the outcome is normally distributed with the conditional linear mean $\mathbf{X_s}\boldsymbol{\beta}$, as is the case in a standard linear model, and specify diffuse normal prior distributions with mean zero for the $\boldsymbol{\beta}$ terms (i.e., standard conjugacy). The distributions $\pi(\tau^2/\sigma^2)$, $\pi(\sigma^2)$, and $\pi(1/\phi)$ are shown to be low information prior forms from simulation and our previous work with this model. The real expansion in a kriging model, though, is that the variance-covariance matrix of the errors, $\boldsymbol{\Sigma}$, consists of two parts. The nugget variance, τ^2, is multiplied by an identity matrix and represents the homoscedastic and uncorrelated portion of the error variance. The nugget specifically represents the portion of the error variance that is independent of spatial distances among observations. The other portion of $\boldsymbol{\Sigma}$, however, consists first of the partial sill, σ^2, which represents the portion of the unexplained variance that is related to spatial distances. This term is multiplied by the correlation matrix $\mathbf{H}(\phi)$, where each entry $H(\phi)_{ij}$ is a function of the distance between observations i and j as well as the decay term ϕ, which as the name implies, is the parameter that quantifies the declining association between two locations as the distance increases. As an example, one common specification of $H(\phi)_{ij}$ is an exponential correlation function, $\exp(-\phi d_{ij})$, where d_{ij} is the distance between point i and point j. This function, as well as most of the other functions commonly used to specify $\mathbf{H}$, imply the correlation between observations diminishes with increased distance. In addition, diffuse gamma prior distributions are assigned to each of $\{\sigma^2, \phi, \tau^2\}$.

In this setup, rather than estimate the nugget effect (τ^2) directly, we estimate the relative nugget (τ^2/σ^2) or ratio of the nugget to the directly-modeled partial sill. Since a reasonably tight prior is required for numerical stability [14], we centered a uniform distribution around an approximate 7:1 empirical ratio of these two quantities. For the partial sill (σ^2), we assume a reciprocal prior, and for the range ($1/\phi$), we assume a uniform prior up to twice the maximum distance in the data. Priors on the regression coefficients are flat. The best-fitting covariance function is the Gaussian function, which will be described more in the next section, and all other model attributes are defined as before.

So why is the Bayesian model in (1) so computationally difficult with large datasets? The reason lies with the $\mathbf{H}$ matrix. Specifying $\mathbf{H}$ requires that all of the pairwise distance among observations be computed to obtain each d_{ij} term. Each distance then must be multiplied by a parameter estimate, ϕ, where the product is included in an exponential function. Whether using a hill-climbing algorithm or numerical integration with Monte Carlo or MCMC to estimate this model, all of these steps must be repeated with each new candidate value of ϕ. Plus, once all of these steps are computed, the variance-covariance

matrix must be inverted because $\boldsymbol{\Sigma}^{-1}$ is essential to efficiently estimating the model's coefficients, $\boldsymbol{\beta}$. So although distances among observations are fixed and known quantities, the fact that these distances are included in a nonlinear function alongside an unknown parameter means that these extensive computations cannot be simplified. This, in computer engineering, is labeled an $\mathcal{O}(N^3)$ problem. This means that kriging models become difficult to estimate as the number of observed points in the spatial field increases in size.

In fact, the computational time that would be involved to estimate models over large spatial datasets and to make forecasts from them would prove insurmountable for typical high-speed computing system (For a more comprehensive discussion of the the computational issues encountered when dealing with large data sets, as well as potential methods to alleviate these issues, please see: Dickinson [8], Neiswanger et al. [9], Scott et al. [10], Luengo et al. [11]). Kriging with a large number of data points requires significant time and computing power in order to produce a predictive model. Rennen [15] notes, "with current implementations and computing power, it is sometimes even impossible to fit a Kriging model using all available data" [15] (p. 546). Therefore, researchers have had to create multiple ways to circumvent these issues when utilizing kriging methods (see: [15–18]).

2.2. Procedure: Kriging with Bootstrapping

To address these computational difficulties, we build on the work of Monogan and Gill [2] and suggest that researchers utilize Bootstrap Random Spatial Sampling (BRSS) in order to estimate Bayesian kriging models over large samples. Using the BRSS method, analysts will be able to answer research questions with spatially correlated data using less computation time and power. Figure 1 offers a quick overview of this technique. Following the logic of this flowchart, assume a set of N total observations. In step (1) of the BRSS method, researchers take B random resamples of size n without replacement. In step (2), for each of the resamples, we estimate a Bayesian kriging model with n.sims iterations of the sampler. In step (3), for each fitted model, we compute the mean of all model parameters to provide a point estimate from each resample. Finally, in step (4), we collect all of the B estimates of the parameters of interest and use them to characterize the marginal posterior distribution of each parameter of interest.

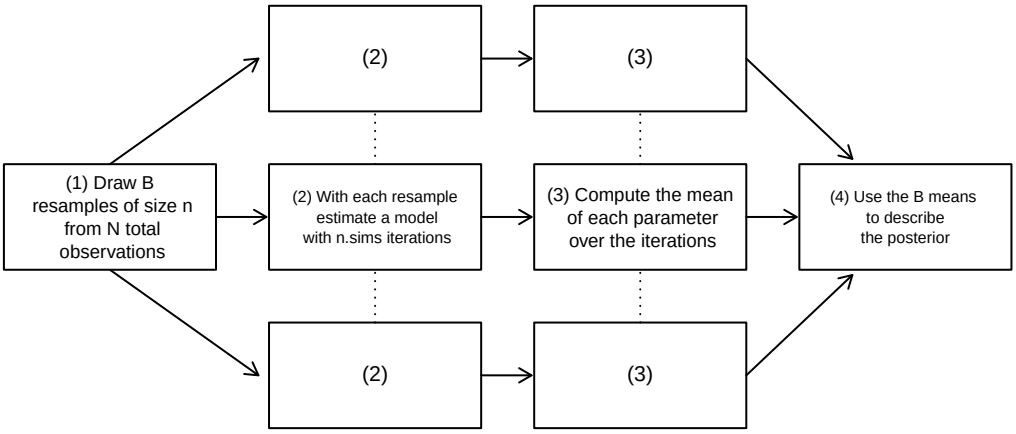


Figure 1. Flowchart Illustrating the Steps of BRSS.

2.2.1. Monte Carlo Integration of Spatial Quantities

We estimate the spatial model using the Monte Carlo-Feasible Generalized Least Squares (MC-FGLS) estimation procedure developed by Diggle and Ribeiro [19] (p. 141). The procedure is summarized as follows using the full observed dataset $(\mathbf{Y}, \mathbf{X}) = \{(\mathbf{Y}_1, \mathbf{X}_1), (\mathbf{Y}_2, \mathbf{X}_2), \ldots, (\mathbf{Y}_N, \mathbf{X}_N)\}$ (for the moment). In order to draw values

for $\frac{\tau^2}{\sigma^2}$ and $\frac{1}{\phi}$ conditional on the spatial data define first the following joint conditional posterior distribution:

$$p\left(\frac{\tau^2}{\sigma^2}, \frac{1}{\phi}\middle|\mathbf{Y}, \mathbf{X}\right) \propto \pi\left(\frac{\tau^2}{\sigma^2}\right)\pi\left(\frac{1}{\phi}\right)|V_{\tilde{\beta}}|^{\frac{1}{2}}\left|\mathbf{H}(\phi) + \left(\frac{\tau^2}{\sigma^2}\right)\mathbf{I}\right|^{-\frac{1}{2}}(\hat{\sigma}^2)^{-\frac{n}{2}}, \tag{2}$$

where $V_{\tilde{\beta}}$ is the correlation matrix of the regression coefficients estimated with FGLS using the current draw of $1/\phi$, n is the sample size, and $\hat{\sigma}^2$ is an estimate of the partial sill based on residuals drawn from initial FGLS coefficient estimates according to:

$$\tilde{\beta} = (\mathbf{X}'\mathbf{H}(\phi)^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}(\phi)^{-1}\mathbf{Y}. \tag{3}$$

Hence,

$$V_{\tilde{\beta}} = (\mathbf{X}'\mathbf{H}(\phi)^{-1}\mathbf{X})^{-1}. \tag{4}$$

So that the estimated partial sill is:

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{H}(\phi)^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \tag{5}$$

The iterative procedure then works as follows, where m is the iteration index of the algorithm and M is the total number of iterations.

1. Draw a set of initial values from the prior distributions for $\frac{\tau^2}{\sigma^2}$ and $\frac{1}{\phi}$.
2. Set the FGLS iteration counter to $m = 1$ to start the iterative part of the process.
3. At the m^{th} step using the posterior probabilities computed in Equation (2) draw a single set of sample posterior values for $\frac{\tau^2}{\sigma^2}$ and $\frac{1}{\phi}$ and label these $\frac{\tau^2}{\sigma^2}^{(m)}$ and $\frac{1}{\phi}^{(m)}$.
4. Use the sampled values of $\frac{\tau^2}{\sigma^2}^{(m)}$ and $\frac{1}{\phi}^{(m)}$ to define the conditional posterior distribution of the partial sill, σ^2:

$$\sigma^2|\mathbf{Y}, \mathbf{X}, \frac{\tau^2}{\sigma^2}^{(m)}, \frac{1}{\phi}^{(m)} \sim \chi^2_{ScI}(n, \hat{\sigma}^2) \tag{6}$$

Take a draw from this scaled inverse χ^2 distribution to determine $\sigma^{2^{(m)}}$.

5. Use the sampled values of $\frac{\tau^2}{\sigma^2}^{(m)}$, $\frac{1}{\phi}^{(m)}$, and $\sigma^{2^{(m)}}$ to estimate the regression coefficients using (3) and (4). This yields the vector of coefficients $\tilde{\boldsymbol{\beta}}$ and the covariance matrix $\sigma^2 V_{\tilde{\beta}}$. With these two terms, we define the conditional posterior distribution of the vector of regression coefficients, $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\mathbf{Y}, \mathbf{X}, \sigma^{2^{(m)}}, \frac{\tau^2}{\sigma^2}^{(m)}, \frac{1}{\phi}^{(m)} \sim \mathcal{MVN}(\tilde{\boldsymbol{\beta}}, \sigma^2 V_{\tilde{\beta}}) \tag{7}$$

Take a draw from this multivariate normal distribution to determine $\boldsymbol{\beta}^{(m)}$.

6. Update the FGLS iteration counter as $m = m + 1$.
7. Repeat steps 3–6 until $m = M$.

This hybrid MC-FGLS procedure allows us to build a sufficient sample to reflect the joint posterior for the full parameter set $(\frac{\tau^2}{\sigma^2}, \frac{1}{\phi}, \sigma^2, \boldsymbol{\beta})$. This procedure "…is direct simulation, replicated independently, rather than MCMC. Hence, issues of convergence do not arise" [20] (p. 163). While these steps provide a simulation process that is justified by the law of large numbers, it is taxing on any common high-speed cluster for data of the size that we want to work with. That is, for a more modest sized dataset this iterative process would be run once with the entire dataset to complete the analysis and report the results.

Since with very large data we cannot perform kriging estimation either through this MC-FGLS procedure or through an MCMC process over all of the data due to the demands on computing infrastructure (RAM memory, cycles, storage), we implement the Bootstrap Random Spatial Sampling procedure to estimate the full model as defined in the bootstrapping steps below:

1. Independently draw B bootstrap samples of the parameter estimate $3 + k$ length vector, $\boldsymbol{\theta} = \{\tau^2, \sigma^2, \phi, \boldsymbol{\beta}\}$, where k is the number of explanatory variables on the right-hand-side of the core model. The data of size N are drawn across iterations without replacement, meaning that these B samples will not contain overlapping data cases among them.

Predictably Step 1 becomes more complex in computation due to the kriging process described above and large sample sizes, meaning that Step 1 is efficiently executed using the following substeps:

1a. Independently draw n data cases from the full N-sized sample, without replacement since zero distances between cases are not defined in the model, producing $\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_n^*$.

1b. Take these n data cases and perform the kriging process n.sims times to get the parameter draws: $(\tau_b, \sigma_b, \phi_b, \boldsymbol{\beta}_b)$, for $b = 1, \ldots,$ n.sims. Put another way, we repeat the hybrid MC-FGLS procedure described above n.sims times in order to construct a posterior sample of the parameters over a resample.

1c. From the n.sims $\times (3 + k)$ matrix generated by the MC-FGLS process, take the means down columns to produce $\theta_b^* = \{\hat{\tau}, \hat{\sigma}, \hat{\phi}, \hat{\boldsymbol{\beta}}\}$, producing one of the $b = 1, \ldots, B$ results required in step 2 below.

This process produces an unbiased set of marginal posterior distributions as if the full sample were analyzed simultaneously based on the arguments in the previous section. The obvious advantage is the ability to parallelize the procedure after bootstrap samples are defined ($\mathbf{X}^*$). Once Step 1a through Step 1c have been completed, continue through the process by proceeding with the following steps:

2 Record the sample statistics of interest, θ_b^* for each bootstrap sample, and the mean of these statistics:

$$\bar{\boldsymbol{\theta}}^* = \frac{1}{B} \sum_{b=1}^{B} \theta_b^*$$

3 Estimate the bootstrap standard error of the statistic by:

$$\mathrm{Var}(\bar{\boldsymbol{\theta}}^*) = \frac{1}{B-1} \sum_{b=1}^{B} \left(\theta_b^* - \bar{\theta}^*\right)^2$$

These last two steps are standard bootstrap operations.

2.3. Properties of Bootstrap Random Spatial Sampling

In this section, we demonstrate that our bootstrapping algorithm introduced in the previous section produces consistent estimates of quantities of interest in the context of kriging models. While the general consistency of bootstrapping algorithms has long been established, we need to show this property for a version that performs Monte Carlo estimation of the quantities of interest at each iteration. In addition, we show that the resulting standard errors are slightly larger, reflecting the replacement of a calculation from a standard bootstrap technique with a (required) Monte Carlo step. Thus, our algorithm theoretically recovers a consistent estimate of the distribution of parameters of interest and conservative estimates of uncertainty.

2.3.1. Terminology and Consistency

We begin by deriving the simple parametric bootstrap as shown by Bickel and Freedman [21], Singh [22], Efron [23,24,25], Diaconis and Efron [26], Efron and Tbishirani [27], Hall [28], Efron and Tibshirani [29], Shao and Tu [30], and others to set standard notation and terms. Start with an iid sample of size N $(X_1, X_2, \ldots, X_N)$ and a population quantity of interest θ. The finite sample estimate of this latter quantity is defined as $\hat{\theta}_N = g(X_1, X_2, \ldots, X_N)$. Often bootstrapping is used instead of this simple approach because calculation of the standard error of this statistic is difficult or even impossible. In our case, data are sufficiently large that we can calculate neither the statistic nor its standard error with readily available high-speed computers using the Bayesian kriging model, and so we resort to a bootstrapping approach. A single bootstrap resample of size n with empirical distribution P_n is notated as $X_1^*, X_2^*, \ldots, X_n^*$ $(n < N)$ under the assumption of monotonically increasing cumulative distribution function (CDF), and it is produced by uniformly randomly sampling n values without replacement by replicate sample from the N in the full sample. Again, we sample without replacement across the full data, because we cannot have repeated values in a single sample when identifying the model. The X^* terms reflect that the values are drawn from the original X variable, but this subset is reordered based on the order in which each observation was sampled. If the original sample elements are unique then there are $N(n) = \binom{N}{n}$ unique resamples that can be generated in this fashion, a number that grows rapidly with increasing sample size N. Using this resample we can calculate the original estimate of interest, $\hat{\theta}_n^* = g(X_1^*, X_2^*, \ldots, X_n^*)$. This subsampling and estimation process is repeated B times, which is typically a large number to obtain the set of estimates $\hat{\theta}_{n,1}^*, \hat{\theta}_{n,2}^*, \ldots, \hat{\theta}_{n,B}^*$. Using these statistics under the stated assumptions, we get the full bootstrap estimates of the population statistic and its standard error by:

$$\bar{\theta}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{n,b}^* \qquad \hat{s}_{\theta_B} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_{n,b}^* - \bar{\theta}_B)^2}. \tag{8}$$

Define the CDF of the difference of the bootstrap estimate and the true statistic of interest as $F(t) = p(\sqrt{N}(\hat{\theta}^* - \hat{\theta}) \leq t)$ for any t in the support of the estimate. If $\hat{\theta}_n^*$ is an arbitrary bootstrap estimate from a single resample n corresponding to the true but unknown value $\hat{\theta}_n$ for this same resample, then the CDF of the difference is defined by

$$\hat{F}_n(t) = p(\sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) \leq t | X_1, X_2, \ldots, X_n), \tag{9}$$

which has the full bootstrap analog of:

$$\hat{F}_B(t) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}(\sqrt{n}(\hat{\theta}_b^* - \hat{\theta}_b) \leq t), \tag{10}$$

where $\mathbb{I}()$ is an indicator for the condition in the parentheses, and the b subscript identifies each bootstrap resample of equal size n. Thus $\hat{F}_B(t)$ is the bootstrap estimate of $F(t)$ under the assumption of sufficiently large B.

To be more specific (and parametric) now stipulate that $X_1, X_2, \ldots, X_N$ have population mean μ and variance σ^2. In standard situations we would simply calculate $\hat{\mu}_n^*$ from a single bootstrap sample of size n to define:

$$F_n(t) = p(\sqrt{n}(\hat{\mu}_n^* - \mu) \leq t). \tag{11}$$

Using the same logic as above we replace this quantity, which cannot be calculated directly with the resample definitional version and the full bootstrap version:

$$\hat{F}_n(t) = p(\sqrt{n}(\hat{\mu}_n^* - \hat{\mu}_n) \leq t | X_1, X_2, \ldots, X_n), \tag{12}$$

$$\hat{F}_B(t) = \frac{1}{B}\sum_{b=1}^{B} \mathbb{I}(\sqrt{n}(\hat{\mu}_{n,b}^* - \hat{\mu}_n) \leq t), \tag{13}$$

where $\hat{\mu}_{n,b}^*$ is the mean calculation of $\hat{\mu}_n$ for the bth bootstrap sample $1, 2, \ldots, B$. Therefore:

$$\bar{\mu}_B = \frac{1}{B}\sum_{b=1}^{B} \hat{\mu}_{n,b}^* \tag{14}$$

is the final bootstrap estimate of the mean. Almost sure convergence has been shown many times (see citations above) that as $B \longrightarrow \infty$:

$$\sup_t |\hat{F}_B(t) - F(t)| \xrightarrow{AS} 0, \tag{15}$$

by the strong law of large numbers. The result from Equation (15) is fundamental to our procedure because it shows that bootstrap estimators are consistent. Since it is feasible and easy to increase the number of bootstrap resamples (B) with modern computing, researchers control the rate of convergence here and are assured of validity.

In addition, standard errors are also consistent from the bootstrap procedure. Defining:

$$\hat{s}_\mu^* = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B} (\hat{\mu}_{n,b}^* - \bar{\mu}_B)^2} \tag{16}$$

gives the result that as $B \longrightarrow \infty$:

$$\frac{(\hat{s}_\mu^*)^2}{s_\mu^2} \xrightarrow{AS} 1 \tag{17}$$

as shown by Shao [31]. This discussion of bootstrapping in standard notation is used to set us up to derive the properties of the Bootstrap Random Spatial Sampling process.

2.3.2. A Modification of the Parametric Bootstrap for Big Data

One of the hallmarks of working with so-called big data is that compromises must be made in the procedures being used. Often in statistical work this comes down to a mean squared error (MSE) style tradeoff of bias versus variance in the selection of new or modified procedures to deal with large samples and in complex modeling that needs to be performed in realistic time on realistic (but large) computing systems. Here we show that the modified bootstrap procedure for our purposes retains the desired consistency properties described in the last section but pays a penalty in the resulting variance. This is actually proper and desired as the multistage process introduces a second computational level of uncertainty.

For each drawn bootstrap resample the first step is to modify the direct calculation of each $\hat{\theta}_n^* = g(X_1^*, X_2^*, \ldots, X_n^*)$ for some $g()$ (mean, median, etc.) with a Monte Carlo simulation produced step:

$$\begin{aligned} \hat{\theta}_{\text{n.sims}}^* &= g(h(X_1^*, X_2^*, \ldots, X_n^*)) \\ &= g(MC(z_1, z_2, \ldots, z_{\text{n.sims}} | X_1^*, X_2^*, \ldots, X_n^*)) \\ &= g(MC(\mathbf{Z}|\mathbf{X})) \end{aligned} \tag{18}$$

where the $\mathbf{Z}$ are quantities of interest produced by the Monte Carlo simulation process, denoted as $MC()$, and conditioned on the $\mathbf{X}$ data values. That is, the inner function introduces a Monte Carlo process. Returning to the calculation of means version of this problem, the inner function $h()$ allows us to modify (14) in the following manner:

$$\begin{aligned}\bar{\mu}_{MC} &= \frac{1}{B}\sum_{b=1}^{B}\hat{\mu}^{*}_{\text{n.sims},b} \\ &= \frac{1}{B}\sum_{b=1}^{B}\left(\frac{1}{\text{n.sims}}\sum_{k=1}^{\text{n.sims}} z_{k,b}\right) \\ &= \frac{1}{B}\frac{1}{\text{n.sims}}\sum_{b=1}^{B}\sum_{k=1}^{\text{n.sims}} MC(\mathbf{Z}|\mathbf{X})_{k,b},\end{aligned} \tag{19}$$

where the k indexes the mean operation in the Monte Carlo step and b indexes the mean in the bootstrapping step as before. From the simulation component we also get the mean from the bth simulation step:

$$\hat{\mu}^{*}_{\text{n.sims},b} = \frac{1}{\text{n.sims}}\sum_{k=1}^{\text{n.sims}} z_{k,b}. \tag{20}$$

where the Monte Carlo quantities of interest are now subscripted by k because they are summed over the number of simulations (n.sims) and by b because they are also summed over the number of bootstrap samples (B). This expression means that a very large operation $g()$ can be broken up into bootstrap samples that can be run separately in time in order not to hit memory limits, or to be run in parallel if possible. Thus a big data problem for $g()$ is a set of small data simulation problems that are manageable and scalable (see Dickinson [8], Neiswanger et al. [9], Scott et al. [10]) .

However there is a price to be paid, both in the example and in our application: the variance of the resulting estimate will be different since there are two sources of variability: one from using a mean estimate and one from the Monte Carlo process. Thus the new standard error of μ is:

$$\begin{aligned}\hat{s}_{\mu,MC} &= \left(\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\mu}^{*}_{\text{n.sims},b} - \bar{\mu}_{MC})^2\right)^{\frac{1}{2}} \\ &\quad \text{[substitute in the definition of } \hat{\mu}^{*}_{\text{n.sims},b} \text{ from (20) and } \bar{\mu}_{MC} \text{ from (19)]} \\ &= \left(\frac{1}{B-1}\sum_{b=1}^{B}\left(\frac{1}{\text{n.sims}}\sum_{k=1}^{\text{n.sims}} z_{k,b} - \frac{1}{B-1}\sum_{b=1}^{B}\left(\frac{1}{\text{n.sims}}\sum_{k=1}^{\text{n.sims}} z_{k,b}\right)\right)^2\right)^{\frac{1}{2}} \\ &\quad \text{[put constants in front for clarity]} \\ &= \left(\frac{1}{(B-1)^3}\frac{1}{\text{n.sims}^2}\sum_{b=1}^{B}\left(\sum_{k=1}^{\text{n.sims}}(B-1)z_{k,b} - \sum_{b=1}^{B}\sum_{k=1}^{\text{n.sims}} z_{k,b}\right)^2\right)^{\frac{1}{2}} \\ &\quad \text{[pull the } \textstyle\sum_{k=1}^{\text{n.sims}} \text{ out of the inner parentheses to highlight the } z_{kb} \text{ difference]} \\ &= \left(\frac{1}{(B-1)^3}\frac{1}{\text{n.sims}^2}\sum_{b=1}^{B}\left(\sum_{k=1}^{\text{n.sims}}\left((B-1)z_{k,b} - \sum_{b=1}^{B} z_{k,b}\right)\right)^2\right)^{\frac{1}{2}}.\end{aligned} \tag{21}$$

This highlights the two contributions to the variance from the BRSS procedure, both of which are affected by the magnitude form of $z_{k,b}$ and the chosen form of $h()$ (mean, median, quantiles, etc.). Of course (21) is also both a function of the chosen number of bootstrap

samples (B) and the chosen number of Monte Carlo samples (n.sims). We can also perform a similar expansion of (16) for comparative purposes:

$$\begin{aligned}
\hat{s}^*_{\mu} &= \left(\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\mu}^*_{n,b} - \bar{\mu}_B)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{1}{(B-1)^3} \sum_{b=1}^{B} \left((B-1)\hat{\mu}^*_{n,b} - \sum_{b=1}^{B} \hat{\mu}^*_{n,b} \right)^2 \right)^{\frac{1}{2}} \\
&= \sqrt{\frac{1}{(B-1)^3} \sum_{b=1}^{B} \left((B-1)\frac{1}{n}\sum_{i=1}^{n} X^*_{n,b} - \sum_{b=1}^{B} \frac{1}{n}\sum_{i=1}^{n} X^*_{n,b} \right)^2} \\
&= \left(\frac{1}{(B-1)^3} \frac{1}{n^2} \sum_{b=1}^{B} \left((B-1)\sum_{i=1}^{n} X^*_{n,b} - \sum_{b=1}^{B}\sum_{i=1}^{n} X^*_{n,b} \right)^2 \right)^{\frac{1}{2}} \\
&= \left(\frac{1}{(B-1)^3} \frac{1}{n^2} \sum_{b=1}^{B} \left(\sum_{i=1}^{n} \left((B-1)X^*_{n,b} - \sum_{b=1}^{B} X^*_{n,b} \right) \right)^2 \right)^{\frac{1}{2}}
\end{aligned} \tag{22}$$

The main difference between Equations (21) and (22), is that z is a sampled value of a test statistic that itself computed over a resample. Hence the uncertainty of the resample also goes into $\mathbf{Z}$. Recall that in Step 1a. of the tailored bootstrap we independently drew n data cases from the full N-sized sample (without replacement since zero distances between cases are not defined in the spatial model) producing $\mathbf{X}^*_1, \mathbf{X}^*_2, \ldots, \mathbf{X}^*_n$. So here we see the implications for selecting n in that step. Due to the innermost summation and the practice of having more Monte Carlo simulations than bootstrap sample sizes, in general $\hat{s}_{\mu,MC}$ will be larger than $\hat{s}^*_{\mu}$, suggesting a penalty for adding the Monte Carlo step in the Bootstrapped Random Spatial Sampling Process that has this feature. It also depends on $MC(\mathbf{Z}|\mathbf{X})$, which defines the simulation stage. Because of the simple forms of (21) and (22), we can always make the calculation for particular applications.

What does this analysis provide? First, (15) shows us that as the number of bootstrap samples (B) becomes large, our estimated distribution function of the sampling distribution is consistent. Importantly, (17) similarly shows that our bootstrapped estimate of the error variance of a statistic is also consistent. Finally, comparing (22) to (21), we can see that the standard errors from Monte Carlo estimation are similar, albeit slightly conservative, relative to the consistent standard error estimates produced from a standard bootstrapping algorithm.

3. Results

3.1. Validity Demonstration Example: Fracking in West Virginia

To illustrate the validity of our method with real data, we turn to data on fracking in West Virginia. Methodologically, this example is ideal, because the data consist of approximately 2000 observations: this is small enough that a Bayesian kriging model still can be estimated over the full dataset, yet large enough that reasonably-sized bootstrap samples can be drawn from the data. This allows us to show how closely the bootstrap results resemble results gathered from the real data.

Figure 2 illustrates our data with the locations of 2474 natural gas wells in the state (Data accessed 25 March 2019 from FracTracker: https://www.fractracker.org/map/us/west-virginia). This example is in line with the original development of kriging techniques by D.W. Krige, who sought to predict the value of minerals that could be extracted from a new borehole given information about yields from past boreholes. Fracking certainly

poses a number of risks—not only accidents to workers but also the possible induction of seismic activity. However, using natural gas to supply electricity emits far less carbon dioxide into the atmosphere than any other fossil fuel [32]. So there can be significant profits, jobs, and environmental effects from a large yield of natural gas to weigh against the risks. Ultimately, a good model that can forecast where the high yields are would help assess where the balance of environmental gains versus risks might be in favor of fracking.

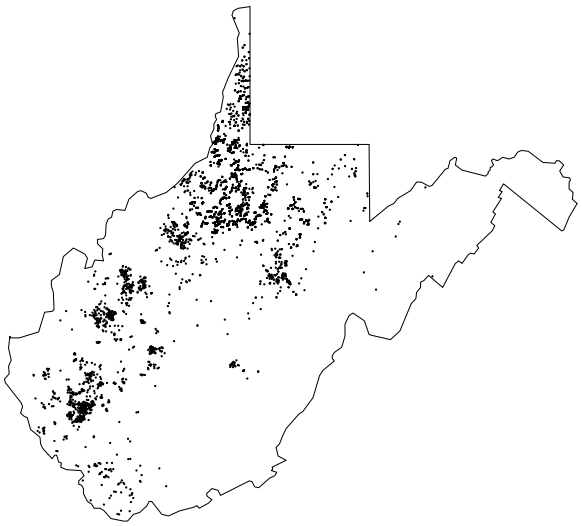


Figure 2. Locations of Oil and Gas Wells in West Virginia.

3.1.1. Fracking Model: Full Data versus Bootstrap

For each site, we observe gas production in 2012 in thousands of cubic feet (Mcf). Most wells produced little or no gas in this year in West Virginia, with a median output of 0 Mcf. However, some yielded substantial output, bringing the mean up to 83,140 Mcf. The highest output in 2012 was 2,144,000 Mcf coming from a well operated by Hall Drilling, LLC (Ellenboro, WV, USA) that was located near the northern end of the state, on the Allegheny Plateau in Harrison County. In order to demonstrate how our BRSS kriging method operates, compared to how the Bayesian kriging model with the full data set behaves, we focus on productive gas wells in 2012 and model the logarithm of natural gas production as our outcome variable.

We have two predictors pertaining to gas production. First, we consider the elevation of the well. Geological features have the potential to affect outputs, and wells in West Virginia ranged from 565 feet above sea level to 3700 feet above sea level. Meng [33] and Meng and Ashby [34] found a correlation between the elevation of the land, from sea level, and the location of natural gas well sites in Pennsylvania. Second, we consider the formation rock pressure at the surface in pounds per square inch (PSI), which ranges from 0 PSI to 7390 PSI, with an average of 842.5 PSI. The formation rock pressure at the surface is tied directly to the amount of pressure that must be used when applying the fracturing liquid to the earth's surface [35] (p. 56).

When modeling the logged gas production, we predict it with just these two covariates. Additionally, our error term follows an exponential semivariogram kriging structure. As described, this error pattern requires that, holding both predictors constant, more proximate wells are expected to bear similar yields in gas production. We estimate this model with an illustration of a model using the BRSS (without replacement), BRSS (with replacement) jittered, and with a model fitted over the full data to compare the results (For further discussion of the BRSS (without replacement) and the BRSS (with replacement) jittered, please see the supplemental materials in Appendix A.). In all three cases the prior distributions for regression parameters are diffuse normals centered at zero and the prior distributions for variance parameters are gammas with long tails.

Table 1 shows a comparison between a full-data model, a bootstrapped (without replacement) model, and a bootstrapped (with replacement) jittered model. Once the data are cleaned, we have 1949 wells in the dataset. The BRSS models were estimated using 100 bootstrap samples of 500 randomly drawn natural gas wells each, applying the Bayesian kriging model to each sample and recording the mean estimate in each iteration (All samples were estimated utilizing 1 cpu core with 10 gb of memory. The full model took approximately 15 min to execute, with the BRSS and BRSS with jitter taking 54 and 59 min, respectively.).

Recall, to obtain our estimates we take a random sample of the data as well as a random sample of the cases. With each round of the bootstrap, we create an unbiased approximation of both parameter estimates from the model and forecasts of the selected predictive observations. At the end, we pool the sample values of parameter estimates in order to summarize the marginal posterior distributions of each model parameter.

Turning to the results of Table 1, the estimates from the three models are largely similar. The estimates from the BRSS models tend to have more uncertainty than those from the model with the full data set—as our result from (22) would imply. Across all three models, both the intercept and the partial coefficient for pressure are similar in coefficient size, direction, and robustness. From either model, we would draw the conclusion that, holding elevation and location constant, on average higher pressure leads to higher expected gas yields. The only major loss we get from inference is the effect of elevation: The median value of the coefficient is comparable across the three models, which would indicate greater yields at higher elevations in West Virginia, on average an all else equal. However, in the full data model, we conclude that the probability of a positive relationship exceeds 95%, whereas with the BRSS models we cannot draw such a robust conclusion. For the variance terms of the partial sill (σ^2), the range ($1/\phi$), and the ratio of the nugget to the partial sill (τ^2/σ^2), we get similarly-scaled terms in Table 1, again with a larger spread in the credible interval for the BRSS (without replacement) model.

To truly gauge what these variance terms mean, though, the best technique is to turn to a visual representation of the semivariogram, shown in Figure 3. In this figure, the horizontal axis represents the distance between wells in meters. The vertical axis represents the semivariance for the error terms of wells separated by this distance. The semivariance has two interpretations: it is half the variance of the difference in residuals for wells separated by a given distance, or the full variance of residuals when wells at a given separation distance are pooled together. The points on the graph represent the empirically observed semivariance of the residuals when observations at an approximate distance are pooled together. The solid red line is the estimated exponential semivariogram for the full dataset, as implied by the first numerical column of Table 1. The shaded pink area is the 90% credible interval for the full-data semivariogram. The dashed blue (green) line is the estimated exponential semivariogram for the BRSS models, as implied by the third and fourth numerical columns of Table 1. The shaded light blue (green) area is the 90% credible interval for the BRSS semivariograms.

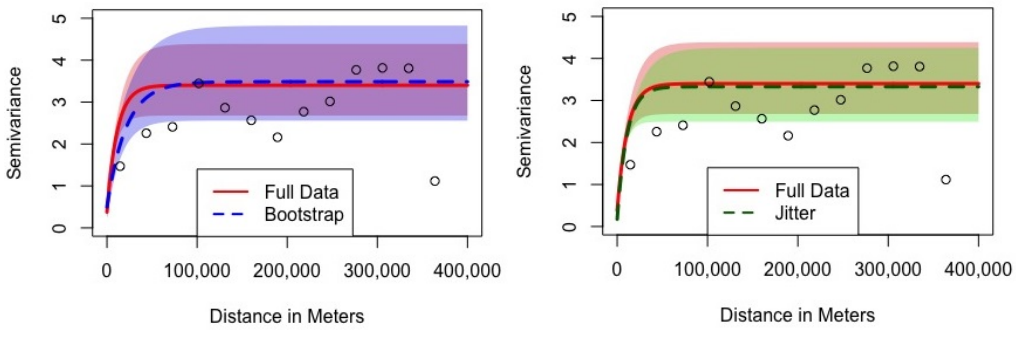


Figure 3. Rival Parametric Semivariograms for West Virginia Fracking for BRSS Without Replacement (**Left**) and BRSS (with replacement) Jittered (**Right**).

Table 1. Kriging Models of Logged Natural Gas Production in West Virginia for 2012.

	Full Data Set			BRSS Sample Size = 500			BRSS with Jitter Sample Size = 500		
Parameter	**Median**	**5th Perc.**	**95th Perc.**	**Median**	**5th Perc.**	**95th Perc.**	**Median**	**5th Perc.**	**95th Perc.**
Intercept	7.0542	5.1359	8.7443	6.9515	4.1634	9.9202	6.5613	3.5828	9.0190
Log Elevation	0.2690	0.0325	0.5235	0.2926	−0.1193	0.7113	0.3752	0.0327	0.7989
Pressure	0.0002	0.0002	0.0003	0.0003	0.0002	0.0004	0.0003	0.0002	0.0004
σ^2	3.0243	2.4428	3.7879	2.9852	2.3574	3.7473	3.1772	2.5154	3.6810
$1/\phi$	11,229.3100	8448.2760	14482.7600	21,246.9800	15,515.2600	25,635.1300	11,068.5300	6949.5260	20,982.9300
τ^2/σ^2	0.1245	0.0966	0.1586	0.1675	0.0852	0.2875	0.0570	0.0500	0.1248

Notes: $N = 1949$. BRSS estimates were computed using 100 iterations of BRSS with 500 observations per bootstrap sample. Exponential semivariogram structure.

In Figure 3, both of the estimated parametric models successfully thread through the empirical semivariogram points estimated from the raw data. The nugget of a semivariogram is the portion of the unexplained variance that cannot be attributed to geospatial factors, which is the semivariance at a distance of zero meters. In the BRSS figure (left), we can see that both models have a closely sized nugget term, with semivariances of 0.38 and 0.50, respectively. Additionally, for the BRSS with jitter (right), the semivariances are 0.38 and 0.17, respectively. The range term describes how quickly or slowly spatial association dissipates. When the semivariance is low, observations are highly spatially correlated (as is expected for physically proximate values), and when the semivariance is high, there is little correlation. With a smaller range term, the semivariance approaches its upper bound quickly and spatial association dissipates. As this figure shows, the full data model does have a smaller range term, with the solid red line reaching its asymptote a little more quickly than the blue dashed line for the BRSS model. These positions are close on the scale of the data, though. With distances as long as 400,000 m in the data, both models approach the top variance well before 100,000 m. Finally, the sill is the sum of the nugget and the partial sill, and it represents the maximum semivariance reached when a sufficient distance is reached. The sill, thus, is the asymptotic variance the semivariogram approaches, seen as the approximate values taken when the lines flatten-out. The solid red line's asymptotic variance is 3.40 for the full data model, while the blue and green dashed line's sill is 3.48 and 3.33 for the BRSS and BRSS with jitter models. Again, these are similar maxima.

It should be noted that with fracking, geological processes produce natural gas. Because similar processes are in proximate places, nearby wells are likely to produce similar outputs. This is the fundamental goal of a kriging model—unobserved processes generate similar data in similar locations. Additionally, the relationship in the semivariogram is non-linear. We would expect heavy correlation in a cluster of proximate wells, but getting out of a cluster the correlation will drop away substantially more. Hence, by 100,000 m the correlation has basically disappeared.

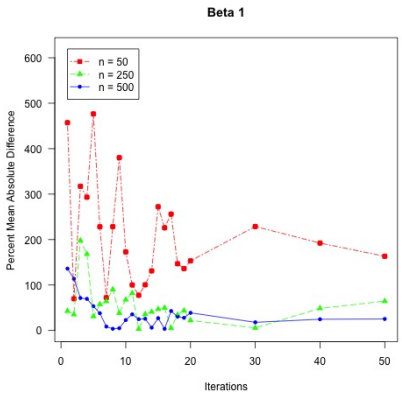
Examining these lines in Figure 3, the variance structure of the two models is nearly the same. In fact, the wider 90% credible interval for the BRSS model includes nearly all of the credible interval of the full data model, including the median-level estimate of the semivariogram. The small difference in the median estimates of the semivariograms is because these are variance components and the BRSS process adds slightly more uncertainty through the multistage generation and combination process. The median estimates of the semivariogram do cross, largely because a larger range term in the BRSS model offsets in part the larger partial sill term from that model. Again, we note that the credible intervals show that there is not a robust difference between these two models. This is an example of our point that big data statistical procedures generally impose compromises on the researcher versus analogous standard procedures on more modest sized data.

3.1.2. A Performance Experiment

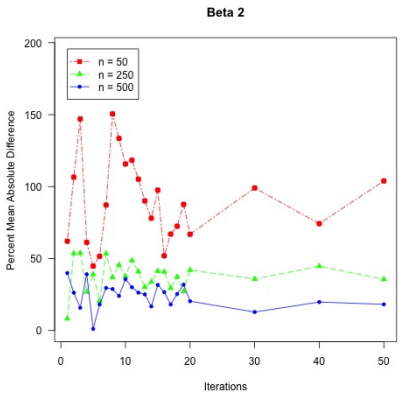
Under what circumstances does our BRSS models perform the best? To answer this, we expanded the comparison from Table 1 to consider performance when we manipulate

the number of iterations in the bootstrap as well as the resample size in each bootstrap iteration. Thus, we have two different treatments: The first treatment is to manipulate the number of iterations of the bootstrap, ranging from 1 to 50. The second treatment is to manipulate how large the resample size is per bootstrap iteration, using resample sizes of 50, 250, and 500. For each combination of treatments, we compare the results from BRSS to those of the full data model shown in Table 1. We computed the percent mean absolute deviance between the BRSS results and the full data model. Ideally, these percentages will be as small as possible, indicating that BRSS is recovering estimates closer to what would be obtained from a full data model, in cases where estimating the full data model would be feasible.

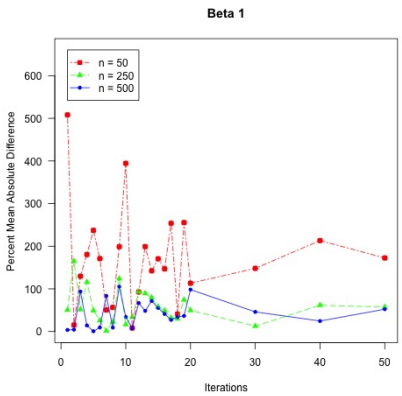
Figure 4 shows the findings from our experiment. Each of the four panels in the figure represents the two primary coefficients from the two BRSS models. In each panel, the horizontal axis represents the number of iterations used in the bootstrap. Experiments were run for all iteration counts from 1–20, as well as 30, 40, and 50. The vertical axis represents the percent mean absolute difference, or the percentage difference between the BRSS estimates and the model estimated over the full data. Each panel contains three lines: a red dot-dash line with square characters to represent the treatment of 50 observations per bootstrap resample, a green dashed line with triangle characters to represent the 250 observation treatment, and a blue solid line with dot characters to represent the 500 observation treatment.



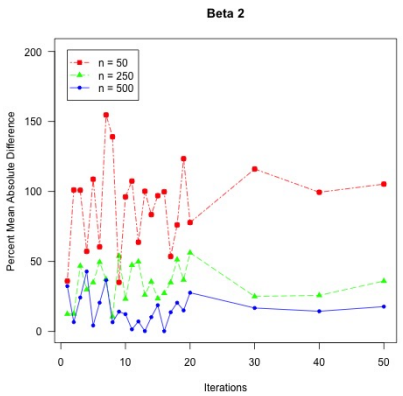
Logged Elevation (BRSS Without Replacement)



Pressure (BRSS Without Replacement)



Logged Elevation (BRSS with Jitter)



Pressure (BRSS with Jitter)

Figure 4. A Comparison of the BRSS Without Replacement and BRSS with Jitter Performance.

From this analysis of gas production in West Virginia, we have learned several important facets of the BRSS model. First, in instances in which it is computationally impossible to estimate a model over the full data, BRSS offers an estimable alternative that provides estimates that fairly approximate the results that would be obtained if a single model could be fitted over the full data. Second, BRSS does perform better with larger bootstrap sample

sizes and more iterations of the bootstrap. There is greater benefit with increasing the bootstrap resample size than the number of iterations to bringing the results in line with the full data model. Of course, the challenge that estimating a kriging model is $\mathcal{O}(N^2)$ in computational complexity remains with bootstrap samples. This means that a researcher can improve results by increasing sample size per iteration, yet cannot increase the sample size too much before the problem becomes intractable again. Having learned these features of the model, we now apply these lessons to another example in which estimating a model over all of the data is entirely impossible.

3.2. Application with Big Data: Campaign Contributions in California

The fracking data were useful to illustrate how well the BRSS method works whenever we still have few enough observations to estimate the model over all data. In this way we were able to compare the results of the model with the full data to the results when we use BRSS. We also learned that relatively large bootstrap samples and more bootstrap iterations improve the results. Now we turn to an example of data which are too big for us to estimate the full model, thereby inducing the need for us to turn to BRSS. In this way, we show the applied researcher how to implement our strategy when managing a large dataset.

This big data problem is increasingly common as campaigns and consultants scrape massive datasets on voters, donors, and consumers. To illustrate this, we analyze data that have been shared by the consulting firm ∅ptimus (a right-of-center firm that collects and integrates government and corporate produced data). Their dataset consists of over 19 million residents of California, with 475 variables about them. We specifically were interested in campaign donation behavior, so we focused on residents who had donated to a left- or right-of-center campaign from 2011–2018. This reduced the data to 199,626 donors on the left and 77,566 donors on the right. Figure 5 shows a map of the 277,192 Californians who donated enough money in this time-frame to be reported by the Federal Election Commission. Each point on the map represents the donor's location, geocoded based on his or her address.

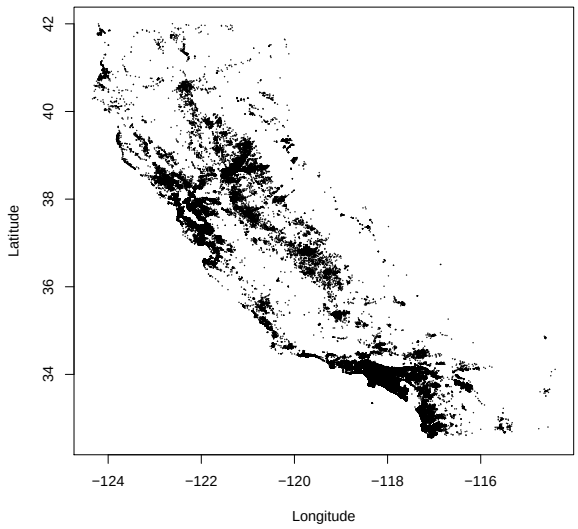


Figure 5. Donors in California.

Our Bayesian model is fashioned after [36] analysis of campaign contributions to Rick Perry's 2006 gubernatorial reelection campaign, which was not Bayesian but similar features. In that paper, they estimated an ordinary kriging model and showed that values of contributions to the Perry campaign did have a tendency to cluster together. (The implication of this finding is that certain areas could be more profitable to seek donations from.) We would like to repeat this ordinary kriging analysis, and we also want to consider whether clustering is present in the context of linear predictors in a Bayesian universal kriging model. However, a model of 77,566 right-of-center donors (much less 199,626 left-

of-center donors) cannot be readily estimated due to the computational questions involved. Thus, we have no choice but to use the BRSS algorithm to understand patterns in California political donor behavior.

Monogan and Gill [2] argue that kriging models are appropriate when analyzing political responses (of which campaign donations are a clear example) because local factors are likely to make neighbors more like-minded. Gimpel and Schuknecht note that there are compositional and contextual reasons that neighbors are politically similar [37] (pp. 2–4). Compositional reasons are that neighbors are likely to share similar economic interests, social structure, religion, and socioeconomic factors. Since it is impossible to include all of these factors in a model, we need to account for similarity among neighbors somehow, and the error term from a kriging structure is one way to do this. The contextual reasons for similarity are that proximate residents are likely to share the same information sources through media outlets and are likely to even speak with each other such that ideas and information will spread through local networks. This, too, cannot be easily modeled and calls for a spatial error term like we specify. Hence, there are several reasons that nearby citizens will be politically similar.

As a first look at the patterns in these data from ∅ptimus, we present contour maps of running local averages of logged donations levels in Figure 6. In each panel of this figure, the horizontal axis represents donors' locations in eastings, while the vertical axis represents donors' locations in northings. By rescaling longitude and latitude to eastings and northings, all distances on the graph can be evenly scaled with kilometer-based distances. The left panel shows patterns for donors whose largest contribution was to a left-of-center campaign while the right panel shows patterns for donors whose largest contribution was to a right-of-center campaign. In each panel, darker spots indicate a higher average contribution relative to other places on the panel. Contour lines also illustrate particular levels of the logged value of donations.

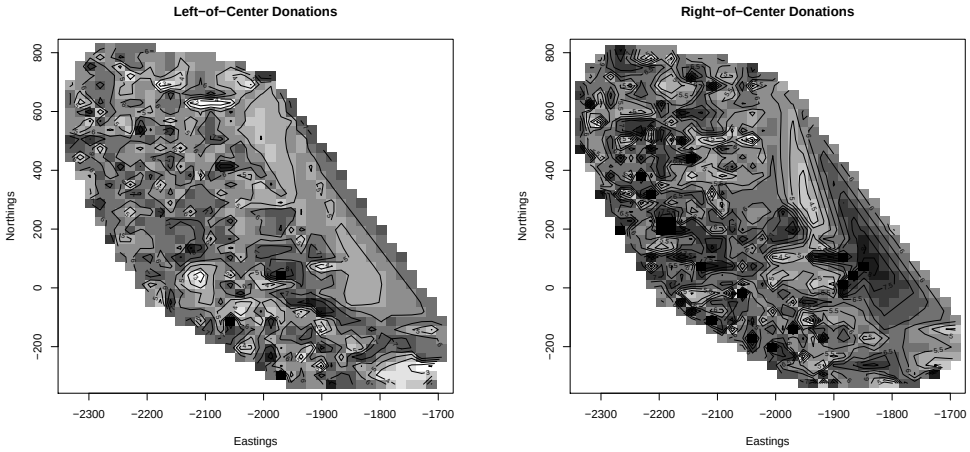


Figure 6. Contour Maps of Left- and Right-of-Center Donations.

As Figure 6 shows, local averages of contributions behave as would be expected. On the left panel of left-of-center donations, on the northern coast, there tends to be a high average, indicated by fairly consistent patterns of dark gray. The southern coast, by contrast, is more checkered. The more rural inland east is shaded lighter, indicating that relatively fewer left-of-center contributions come from this area. Meanwhile, the right-hand panel shows the patterns for right-of-center contributions. This, too, behaves as might be expected: The highest levels of right-of-center contributions tend to be scattered in a checkered pattern in coastal areas. Meanwhile, the rural inland east is a consistent cluster for right-of-center donations. In both panels, it does appear that there is noticeable clustering, which suggests that geographical space is worthwhile when considering where donations can emerge. Based on the figures, it might appear that there is endogeneity based

on the placement of donors in the state of California. Based on [2] we are confident that these concerns are alleviated.

To this end, we proceed to estimate kriging models for the larger left-of-center dataset. We have to use BRSS because the full dataset is simply too large to estimate the full model. Table 2 presents two models of the logged value of left-of-center donations. These two models are an ordinary kriging model in which levels of donations are entirely a function of a spatial error process, and a universal kriging model in which logged donations are a linear function of several predictors and the errors of the model have a spatial error process. Each row of the table represents another parameter from the model. The first three numerical columns of the table show the median and 90% credible interval of parameter estimates from the ordinary kriging model. The last three numerical columns of the table show the median and 90% credible interval of parameter estimates from the universal kriging model. Medians and percentiles are all computed from BRSS with 50 iterations and 1000 observations drawn per iteration. In this way, we follow the recommendations from the previous section to estimate the model with a large number of iterations and a resample size that is large enough for reliable results but still feasible on a computing cluster. (Utilizing 1 cpu core with 10 gb of memory, it took approximately 6 h to estimate the ordinary kriging model. Additionally, utilizing the same machine metrics, it took approximately 17 h to estimate the universal kriging model. While this is time-consuming, it is feasible, whereas a full-data model is not even feasible at present).

Table 2. Kriging Models of Logged Left-of-Center Donations.

	Ordinary Kriging			**Universal Kriging**		
Parameter	**Median**	**5th Perc.**	**95th Perc.**	**Median**	**5th Perc.**	**95th Perc.**
Intercept	5.9810	5.6986	6.3034	0.9167	−4.3045	5.5629
Vote Frequency				0.0117	−0.0277	0.0623
Logged Age				0.4534	−0.0026	0.7872
Female				−0.2561	−0.4235	−0.0542
Income (12 Cat.)				0.1030	0.0789	0.1262
Asian				0.1695	−0.1785	0.4015
Black				−0.2342	−0.6583	0.1895
Latino				−0.3553	−0.6468	−0.0557
College				0.1763	−0.0534	0.3689
Eastings				−0.0012	−0.0034	0.0007
Northings				−0.0005	−0.0015	0.0003
σ^2	1.3461	1.2053	1.4669	1.3008	1.1229	1.4787
$1/\phi$	646.2641	334.7049	752.5614	714.3906	569.0185	797.5597
τ^2/σ^2	2.2432	2.1426	2.3966	2.2555	2.1125	2.4299

Notes: $N = 199{,}626$. Estimates were computed using 50 iterations of BRSS with 1000 observations per bootstrap sample. Exponential semivariogram structure.

As Table 2 shows, the ordinary kriging model confirms substantial clustering in contributions. The partial still term (σ^2) is the portion of the variance that is attributable to spatial causes, and the posterior median of this term is approximately 1.4. Meanwhile the nugget effect (τ^2) is the portion of the variance that is independent from geographic space—individually idiosyncratic variance. We parameterize the nugget relative to the value of the partial sill, and we get a posterior median of this ratio of approximately 2.2. This tells us that, at the median parameter values, we would estimate that approximately 31% of the variance can be explained by spatial clustering (This calculation is based on the calculation $\frac{\sigma^2}{\sigma^2+\tau^2}$, which can easily be computed with the model-estimated ratio if rewritten as $\frac{1}{1+\frac{\tau^2}{\sigma^2}}$.). The range term ($1/\phi$) indicates how persistent spatial correlation is in the observations, with higher values indicating more persistence. From the range, we can mathematically deduce the effective range, which is the distance at which spatial correlation among values drops to a negligible level. The effective range of this ordinary

kriging model is approximately 1150 km (The effective range is typically defined as the distance at which spatially-based correlations drop below 0.05 [4] (p. 27). We can compute this using the formula: $-\frac{1}{\phi}\ln\left(0.05(\frac{\tau^2}{\sigma^2}+1)\right)$. For the median estimates of $1/\hat{\phi} = 630.4404$ and $\hat{\tau}^2/\hat{\sigma}^2 = 2.2294$, the effective range approximates to 1150 km.).

Does geospatial correlation in donations persist even in the context of linear predictors? The universal kriging model of the last three columns of Table 2 show us that it does. While the partial sill (σ^2) does appear smaller with a posterior median of approximately 1.3, this drop can largely be attributed to the fact that this variance is now reflective of residual variance in a model, as opposed to total variance. The ratio of the nugget to the partial sill (τ^2/σ^2) increased a slight fraction, and in the universal kriging model we still at the posterior median would estimate that approximately 31% of the unexplained variance is geospatial in nature. Hence, adding the predictive model did not effectively alter the share of spatial to idiosyncratic variance. In fact, the larger range term ($1/\phi$) implies that residual correlation in the universal kriging model is actually more persistent than observational correlations over the raw data in the ordinary kriging model. In this case, the effective range is 1304 km, which is greater than California's north-south length of 1240 km. So spatial correlation persists strongly in this case.

Meanwhile, the predictors from the universal kriging model in Table 2 offer a sense of which left-of-center donors are most likely to donate more, holding their location constant. There are robust effects in this table that, on average and ceteris paribus, women donors on average donate less, higher-income donors give more, and Latino donors donate less than non-Latino whites. Note again that the population is of people who donate to left-of-center campaigns. This model speaks less to who would choose to donate in the first place, but rather to how much individuals will contribute if they are the donating type. It is also worth noting that the estimates of these coefficients benefit from efficiency gains relative to estimation with ordinary least squares: The universal kriging model accounts for the spatial correlation of errors while estimating these coefficients, whereas OLS would ignore this important feature of the data.

As a final illustration of the variance properties from these two models, Figure 7 shows the semivariance functions implied by the two models of Table 2. In this figure, the horizontal axis represents the distance in kilometers. The vertical axis is the semivariance for observations separated by that distance. The semivariance can be interpreted in two ways: For observations separated by a certain distance (say, 1000 km) it is half of the variance of the differences between these observations, or alternatively if all of the observations are pooled it is the variance of the pooled observations. In this graph, the exponential semivariogram function for the ordinary kriging model is shown with a blue dashed line, and its 90% credible interval is shown by the shaded light blue area. The semivariogram function for the universal kriging model is shown with a red solid line, and its 90% credible interval is shown by the shaded pink area.

As Figure 7 shows, the semivariance for both models is at its lowest in short distances, which is expected because variances are lower for observations that are highly correlated. As the distance increases, the semivariance increases because there is more spread (and less correlation) among more separated observations. For any given distance, the semivariance for the universal kriging model is lower than for the ordinary kriging model. This pattern makes sense because the linear model explains some of the variance in the universal kriging model. Finally, for both models, the curve flattens-out between 1000 and 2000 km. This fits with the expected ranges of 1150 and 1304 km for the ordinary and universal kriging models, respectively. When the correlation among spatial observations becomes minimal, the variance finds an asymptotic upper bound. Also notably, there is substantial overlap between the credible intervals. Therefore adding predictors does not appear to have altered dramatically the implied variance structure.

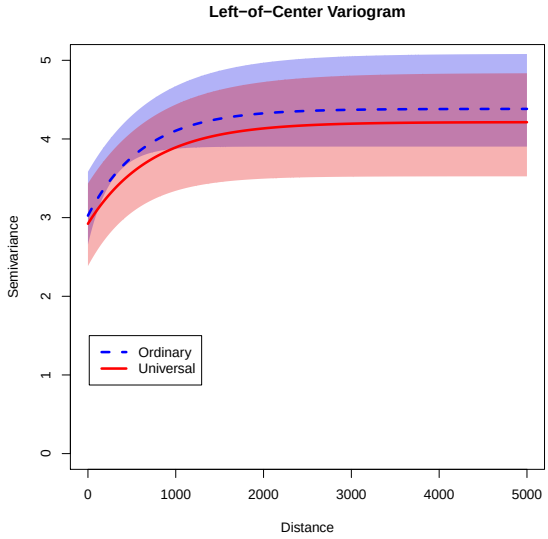


Figure 7. Rival Parametric Semivariograms for Left-of-Center Donors.

With this illustration, we show how scholars or campaign strategists can simultaneously take advantage of geospatial statistical methods while working with big datasets. In this model, we had information for 199,626 donors, which is simply too big of a dataset to estimate a kriging model with. However, using BRSS with 50 iterations and 1000 observations per sample, we were able to describe the level of clustering in the data and how predictors affect contribution levels even when holding geospatial location constant.

4. Discussion & Conclusions

Geospatial statistical models are clearly important in the big data world. There is little doubt that marketeers, political campaigns, government agencies, epidemiologists, and a wide variety of social science researchers will continue to gather and use finely-grained geographical data. But such work is uniquely challenging since they provide not only conventional covariate analysis but also relational specifications between cases, thus exploding the impact of data size. Here we address a key tradeoff that is imposed by a combination of big data and spatial model specification: the choice between realistic computational time and the ability to use the full dataset.

Specifically, the Bootstrap Random Spatial Sampling (BRSS) analytically produces consistent point estimates of parameters and slightly conservative estimates of uncertainty. Using an illustration of fracking in West Virginia, we show that this holds in practice, as our bootstrap estimates were reliable estimates to the full-data model values. Nearly all tools designed to handle big data involve tradeoffs, and while we proved consistency of the BRSS, our data analysis shows that higher uncertainty around spatial parameter estimates is the necessary price to be paid. In the example of data on California campaign donors, we do not have the luxury of comparing BRSS estimates to the full data estimates, but we know that these principles are working nonetheless.

A dataset of a few thousand spatial observations or less can be estimated with a typical desktop computer and some patience. Accordingly, a model such as our analysis of nearly 2000 natural gas wells in West Virginia took notable but reasonable time on a desktop machine. But such data size is unrealistically small for many applications in the 21st century. Our objective here is to provide a principled method for estimating geospatial models in this context in realistic computing time. Ultimately, we were able to describe the statistical properties of the BRSS method and provide multiple comparative examples of the BRSS and full model estimation results utilizing large data sets. Overall, we find that the results of the BRSS provides consistent point estimates and should allow

researchers the opportunity to explore spatial relationships within large data sets with realistic computational expectations.

Author Contributions: All authors contributed equally to this work. As for individual contributions, J.S.B. provided contributions in the following areas: Data Curation; Formal Analysis; Investigation; Resources; Software; Validation; Visualization; and the Writing of the original draft, review, and editing. J.G. provided contributions in the following areas: Conceptualization; Data Curation; Formal Analysis; Funding Acquisition; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; and the Writing of the original draft, review, and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science Foundation grants SES-1630265 and SES-1630263.

Institutional Review Board Statement: Not applicable for studies not involving humans or animals.

Informed Consent Statement: Not applicable for studies not involving humans or animals.

Data Availability Statement: Complete replication information is available at the following web address: https://doi.org/10.7910/DVN/A4A3UO (accessed on 3 November 2022).

Acknowledgments: For helpful research assistance and programming expertise, we thank Sahir Shahryar, Lynne Erika Astronomo Luz, and Le Bao. For sharing data, we thank Scott Tranter and the ∅ptimus Consulting team. This study was also supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology. Questions can be directed to Jeff Gill as corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Monte Carlo Experiment

To evaluate the performance of bootstrapping with point-referenced spatial data, we conducted a Monte Carlo experiment. In this experiment, we generated $M = 50$ spatial datasets with $N = 3000$ observations given a defined population data-generating process. For each of the M datasets, we first estimated a model over all of the data using the algorithm described in Section 2.2.1. We then modeled each dataset using two bootstrap algorithms. The first involved drawing each resample without replacement (Bootstrapping without replacement, meanwhile, typically requires a correction to retain the consistency properties that are generally associated with bootstrapping: This correction is typically based on strata-based mean deviation (see Nigam and Rao [38]). We tried two experimental treatments for bootstrapping without replacement: One which included this stratum-based correction and another which did not. Surprisingly, the corrected version performed considerably worse. Hence, we solely focus on uncorrected bootstrapping without replacement in this analysis.) The second involved drawing each resample with replacement, but jittering each observation that was a replicate of a prior observation. Whenever a jittered point is moved outside of the map (e.g., into an ocean or another country) or over a border of interest within the map (e.g., into another county or state), then the new coordinates are rejected for that observation until an appropriate in-bounds jitter is found. The jittering strategy resembles ridge regression in the sense that a slight distortion identifies the model (Such styles of solutions are common in geospatial statistics writ large. With areal data, for example, a conditional autoregressive model can only be estimated in a likelihood-based setting by introducing a ridge-style correction to the definition of the inverse of the covariance matrix [4] (pp. 81–82).). Either algorithm serves a need of kriging models that two observations cannot be observed at the exact same relocation, lest the error covariance matrix will be singular. For each bootstrap algorithm, we drew $B = 100$ resamples of size $n = 500$ observations.

Figure A1 shows the results of this experiment. In both panels of the figure, the horizontal axis presents the five parameters that we estimated: The first two are partial regression coefficients, which are a coefficient on a dummy variable predictor (β_1) and a coefficient on

a normally-distributed predictor (β_2). The last three parameters are variance terms, which are the partial sill (σ^2), the range term ($1/\phi$), and the ratio of the nugget to the partial sill (τ^2/σ^2). The vertical axis in each case is a percentage. In panel (a), this is the percentage mean absolute deviation, which is the percentage difference between the average parameter estimate over $M = 50$ trials and the true population parameter. In panel (b) this is the coverage probability of a 90% credible interval, which in other words is the percentage of time that the 90% credible interval contains the true population parameter. Ideally, we want each of these lines to be close to 90%. In each panel, the solid black line represents the results when estimating the model from the full-sized $N = 3000$ datasets, the blue dashed line represents the results from the bootstrap without replacement, and the red dotted line represents the results from the bootstrap with replacement and jittering.

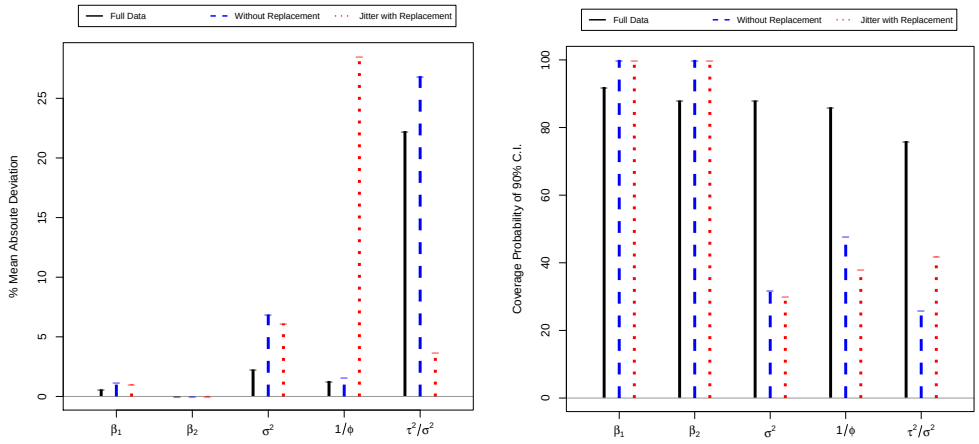


Figure A1. Results from Monte Carlo Experiment of Performance over an $N = 2000$ Dataset versus Bootstrap Methods. (**a**) % Mean Absolute Deviation, (**b**) Coverage Probability.

Starting with panel (a) of Figure A1, we see that the percent mean absolute deviation is small for the regression coefficients for all three methods, which is good across the board. Turning to the variance terms, the two bootstraps show only a little slippage for the partial sill (σ^2) relative to the full data model. Meanwhile, there is a flip-flop with the model without replacement performing better for the range term ($1/\phi$), but the jittered bootstrap performs better for the relative nugget (τ^2/σ^2). Turning to panel (b) of Figure A1, we see that the credible intervals from fully-sized datasets generally have good coverage probabilities, albeit a bit low for the relative nugget (τ^2/σ^2). We would want the true population parameters from the Monte Carlo data generating process to be included in the 90% credible interval approximately 90% of the time, and for most parameters that case is approximated. As for the bootstrap methods, both have 100% coverage of the two regression coefficients. However, there is a considerable drop-off for both methods for the three variance terms.

While the performance on separate variance terms may look concerning, it is important to recall that these three terms offset each other in many ways. That is, if one term is higher than the population value, another can compensate by being lower. To that end, we can consider all three terms in tandem by plotting the parametric semivariogram implied by the model. Figure A2 makes just such a comparison. The horizontal axis in this figure represents distance among observations, and the vertical axis represents semivariance. The wide gray line is the population semivariance, the solid black line is based on the average of the full-data treatment results, the blue dashed line is based on the results from a bootstrap without replacement, and the red dotted line is based on the bootstrap with jittering. As the results show, all four semivariograms closely resemble each other much more than the separate terms would imply.

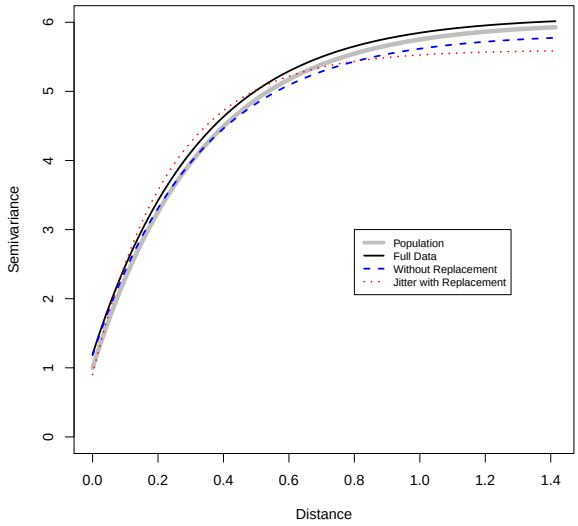


Figure A2. Parametric Semivariograms for Population Process and Three Treatments.

Altogether, a few lessons are to be learned. First, if an analyst's goal is not prediction but efficient inference regarding predictors of interest, either bootstrap will do well. For both β_1 and β_2, the average deviation from the true population parameters was small for both bootstrap methods, and the credible intervals included the true population parameters more often even than the full-data model. Hence, bootstrapping is a valid choice for inference. Second, if an analyst is particularly interested in the values of the variance terms—as he or she would be if using kriging methods to make predictions—then all three semivariograms look similar. Overall, then, both bootstrap techniques, in cases of big data, can serve as a fair approximation when estimation over full data is impossible.

References

1. Gill, J. Measuring Constituency Ideology Using Bayesian Universal Kriging. *State Politics Policy Q.* **2021**, *21*, 80–107. [CrossRef]
2. Monogan, J.E.; Gill, J. Measuring State and District Ideology with Spatial Realignment. *Political Sci. Res. Methods* **2016**, *4*, 97–121. [CrossRef]
3. Gill, J. *Bayesian Methods: A Social and Behavioral Sciences Approach*, 3rd ed.; Chapman & Hall/CRC: New York, NY, USA, 2014.
4. Banerjee, S.; Carlin, B.P.; Gelfand, A.E. *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed.; Chapman & Hall/CRC: New York, NY, USA, 2015.
5. Matheron, G. Principles of Geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266. [CrossRef]
6. Cressie, N.A.C. *Statistics for Spatial Data*, revised ed.; Wiley: New York, NY, USA, 1993.
7. van Stein, B.; Wang, H.; Kowalczyk, W.; Emmerich, M.; Bäck, T. Cluster-Based Kriging Approximation Algorithms for Complexity Reduction. *Appl. Intell.* **2020**, *50*, 778–791. [CrossRef]
8. Dickinson, J.P. Some Statistical Results in Combination of Forecasts. *J. Oper. Res. Soc.* **1973**, *24*, 253–260. [CrossRef]
9. Neiswanger, W.; Wang, C.; Xing, E. Asymptotically Exact, Embarrassingly Parallel MCMC. *arXiv* **2013**, arXiv:1311.4780.
10. Scott, S.L.; Blocker, A.W.; Bonassi, F.V.; Chipman, H.A.; George, E.I.; McCulloch, R.E. Bayes and Big Data: The Consensus Monte Carlo Algorithm. *Int. J. Manag. Sci. Eng. Manag.* **2016**, *11*, 78–88. [CrossRef]
11. Luengo, D.; Martino, L.; Elvira, V.; Bugallo, M. Efficient Linear Fusion of Partial Estimators. *Digit. Signal Process.* **2018**, *78*, 265–283. [CrossRef]
12. Bradley, J.R.; Cressie, N.; Shi, T. A Comparison of Spatial Predictors When Datasets Could be Very Large. *Stat. Surv.* **2016**, *10*, 100–131. [CrossRef]
13. Sun, Y.; Li, B.; Genton, M.G. Geostatistics for Large Datasets. In *Advances and Challenges in Space-Time Modeling of Natural Events*; Porcu, E., Montero, J.M., Schlather, M., Eds.; Springer: Berlin, Germany, 2012.
14. Hobert, J.P.; Casella, G. The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *J. Am. Stat. Assoc.* **1996**, *91*, 1461–1473. [CrossRef]
15. Rennen, G. Subset Selection From Large Datasets for Kriging Modeling. *Struct. Multidiscip. Optim.* **2009**, *38*, 545–569. [CrossRef]
16. Furrer, R.; Genton, M.G.; Nychka, D. Covariance Tapering for Interpolation of Large Spatial Datasets. *J. Comput. Graph. Stat.* **2006**, *15*, 502–523. [CrossRef]
17. Cressie, N.A.; Johannesson, G. Fixed Rank Kriging for Very Large Spatial Data Sets. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2008**, *70*, 209–226. [CrossRef]

18. Hartman, L.; Hossjer, O. Fast Kriging of Large Data Sets with Gaussian Markov Random Fields. *Comput. Stat. Data Anal.* **2008**, *52*, 2331–2349. [CrossRef]
19. Diggle, P.J.; Ribeiro, Jr., P.J. Bayesian Inference in Gaussian Model-based Geostatistics. *Geogr. Environ. Model.* **2002**, *6*, 129–146. [CrossRef]
20. Diggle, P.J.; Ribeiro, Jr., P.J. *Model-Based Geostatistics*; Springer: New York, NY, USA, 2007.
21. Bickel, P.J.; Freedman, D.A. Some Asymptotic Theory for the Bootstrap. *Ann. Stat.* **1981**, *9*, 1196–1217. [CrossRef]
22. Singh, K. On the Asymptotic Accuracy of Efron's Bootstrap. *Ann. Stat.* **1981**, *9*, 1187–1195. [CrossRef]
23. Efron, B. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. *Biometrika* **1981**, *68*, 589–599. [CrossRef]
24. Efron, B. Bootstrap Confidence Intervals: Good or Bad? *Psychol. Bull.* **1982**, *104*, 293–296. [CrossRef]
25. Efron, B. The Bootstrap and Modern Statistics. *J. Am. Stat. Assoc.* **2000**, *95*, 1293–1296. [CrossRef]
26. Diaconis, P.; Efron, B. Computer-Intensive Methods in Statistics. *Sci. Am.* **1983**, *248*, 116–131. [CrossRef]
27. Efron, B.; Tbishirani, R.J. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat. Sci.* **1986**, *1*, 54–77. [CrossRef]
28. Hall, P. *The Bootstrap and Edgworth Expansion*; Springer: New York, NY, USA, 1992.
29. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; CRC Press: Boca Raton, FL, USA, 1994.
30. Shao, J.; Tu, D. *The Jackknife and Bootstrap*; Springer: New York, NY, USA, 2012.
31. Shao, J. Bootstrap Estimation of the Asymptotic Variances of Statistical Functionals. *Ann. Inst. Stat. Math.* **1990**, *42*, 737–752. [CrossRef]
32. Alvarez, R.A.; Pacala, S.W.; Winebrake, J.J.; Chameides, W.L.; Hamburg, S.P. Greater Focus Needed on Methane Leakage from Natural Gas Infrastructure. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 6435–6440. [CrossRef]
33. Meng, Q. Modeling and Prediction of Natural Gas Fracking Pad Landscapes in the Marcellus Shale Region, USA. *Landsc. Urban Plan.* **2014**, *121*, 109–116. [CrossRef]
34. Meng, Q.; Ashby, S. Distance: A Critical Aspect for Environmental Impact Assessment of Hydraulic Fracking. *Extr. Ind. Soc.* **2014**, *1*, 124–126. [CrossRef]
35. Ground Water Protection Council and ALL Consulting. *Modern Shale Gas Development in the United States: A Primer*; United States Department of Energy: Washington, DC, USA, 2009.
36. Tam Cho, W.K.; Gimpel, J.G. Prospecting for (Campaign) Gold. *Am. J. Political Sci.* **2007**, *51*, 255–268. [CrossRef]
37. Gimpel, J.G.; Schuknecht, J.E. *Patchwork Nation: Sectionalism and Political Change in American Politics*; University of Michigan Press: Ann Arbor, MI, USA, 2003.
38. Nigam, A.K.; Rao, J.N.K. On Balanced Bootstrap for Stratified Multistage Samples. *Stat. Sin.* **1996**, *6*, 199–214.