Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utch20

Nonstationary Gaussian Process Models Using Spatial Hierarchical Clustering from Finite Differences

Matthew J. Heaton, William F. Christensen & Maria A. Terres

To cite this article: Matthew J. Heaton, William F. Christensen & Maria A. Terres (2017) Nonstationary Gaussian Process Models Using Spatial Hierarchical Clustering from Finite Differences, Technometrics, 59:1, 93-101, DOI: 10.1080/00401706.2015.1102763

To link to this article: https://doi.org/10.1080/00401706.2015.1102763

Supplementary materials for this article are available online. Please go to http://www.tandfonline.com/r/TECH

Nonstationary Gaussian Process Models Using Spatial Hierarchical Clustering from Finite Differences

Matthew J. HEATON and William F. CHRISTENSEN

Department of Statistics
Brigham Young University
Provo, UT 84602
(*mheaton@stat.byu.edu*; *william@stat.byu.edu*)

Maria A. TERRES

Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203
(*materres@ncsu.edu*)

Modern digital data production methods, such as computer simulation and remote sensing, have vastly increased the size and complexity of data collected over spatial domains. Analysis of these large spatial datasets for scientific inquiry is typically carried out using the Gaussian process. However, nonstationary behavior and computational requirements for large spatial datasets can prohibit efficient implementation of Gaussian process models. To perform computationally feasible inference for large spatial data, we consider partitioning a spatial region into disjoint sets using hierarchical clustering of observations and finite differences as a measure of dissimilarity. Intuitively, directions with large finite differences indicate directions of rapid increase or decrease and are, therefore, appropriate for partitioning the spatial region. Spatial contiguity of the resulting clusters is enforced by only clustering Voronoi neighbors. Following spatial clustering, we propose a nonstationary Gaussian process model across the clusters, which allows the computational burden of model fitting to be distributed across multiple cores and nodes. The methodology is primarily motivated and illustrated by an application to the validation of digital temperature data over the city of Houston as well as simulated datasets. Supplementary materials for this article are available online.

KEY WORDS: Geostatistics; Large datasets; Parallel computing; Spatial partition.

1. RESEARCH MOTIVATION AND CONTRIBUTIONS

Spatial data collected by means of remote sensing, computer model simulation, or field measurement are continually increasing in size and complexity. This increase in data complexity presents a challenge for researchers who seek to use such data to answer scientific questions. Take, as a motivating example, the radiative temperature (in degrees Celsius) across the city of Houston displayed in the first row of Figure 1. Panel (a) displays temperature measurements on June 8, 2011 according to the MODIS satellite while Panel (b) displays a corresponding temperature simulation from the High Resolution Land Data Assimilation System (HRLDAS)—a high-resolution computer model specialized to simulate temperatures over urban environments (see Hu et al. 2014, for more information on the HRLDAS). The data in Figure 1 are collected on a 125×125 grid (14,309 land observations and $125^2 - 14{,}309 = 1316$ uncollected observations over water). Note that, due to cloud cover, the MODIS satellite is only able to collect temperature measurements at $9168/14{,}309 \approx 64\%$ of the land locations.

The primary scientific goals related to the radiative temperature data are to (i) jointly use the MODIS and HRLDAS data to construct an accurate, high-resolution temperature map over Houston and (ii) use the MODIS data to validate the accuracy of the HRLDAS simulation. Similar to Berrocal, Gelfand, and Holland (2010), a simple method to achieve these goals would be to regress the MODIS data on the HRLDAS data. However, the associated residuals from this regression mapped in Figure 1(c) reveal a few issues with this approach. Notably, the residuals in Figure 1(c) are spatially correlated so that appropriate uncertainty quantification of the relationship between MODIS and HRLDAS would require that this correlation be modeled. While the Gaussian process (GP) is the most common approach to model such spatial correlation, the size of the dataset makes fitting a GP computationally burdensome due to the required matrix inversions and determinant calculations (Sun, Li, and Genton 2011). Also, due to the urban heat island effect (Stone, Hess, and Frumkin 2010), the relationship between MODIS and HRLDAS varies across the spatial domain and subsequently leads to nonstationary behavior of the spatial process.

To facilitate use of such data for scientific inquiry, the purpose of this article is to develop general and computationally feasible methodology for modeling large, nonstationary spatial datasets such as the temperature data discussed above. While various methods have been proposed to model large spatial data (e.g.,

TECHNOMETRICS, FEBRUARY 2017, VOL. 59, NO. 1
DOI: 10.1080/00401706.2015.1102763
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/tech*.

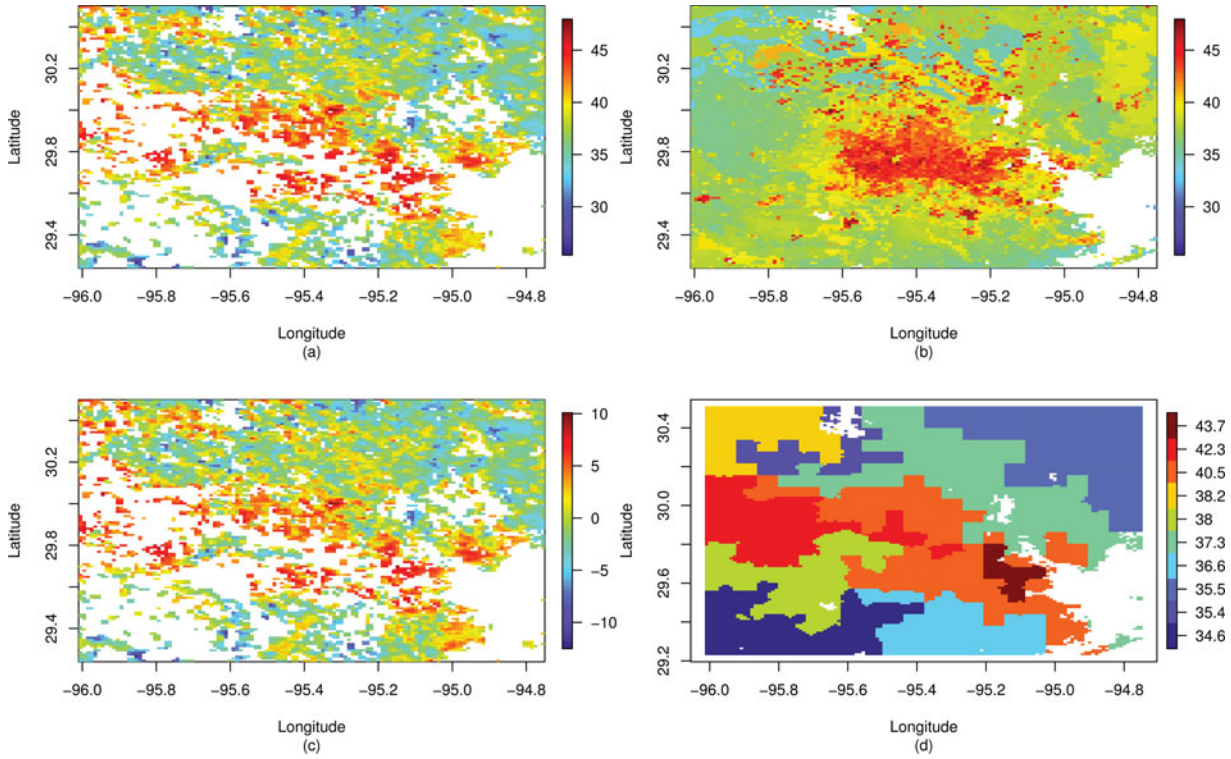

Figure 1. Houston radiative temperature data according to (a) MODIS satellite measurements and (b) HRLDAS simulation. White areas in panel (b) denote water. (c) Residuals of regressing MODIS data on HRLDAS output and (d) estimated spatial clusters. The color bar in (d) indicates the mean radiative temperature of the associated cluster.

Banerjee et al. 2008; Cressie and Johannesson 2008; Sang and Huang 2012; Eidsvik et al. 2014; Nychka et al. 2015), we take a fundamentally different approach by partitioning the spatial region into distinct regions using spatial hierarchical clustering algorithms. By partitioning the spatial domain, independent spatial models can be fit to each region thereby inducing nonstationary behavior in the global spatial process. Furthermore, the computational demand to fit each model is reduced by assuming conditional independence (i.e., independent conditional on the parameters) across spatial regions.

Within the geological science community, several approaches have been developed for clustering spatially correlated observations. Examples of applications include the domaining of potential mining locations (Romary et al. 2012) and the clustering of soil survey data (Oliver and Webster 1989). The earliest approaches by Oliver and Webster (1989) and Bourgault, Marcotte, and Legendre (1992) were based on weighting the distance between samples using the semivariogram for the data, yielding clusters that tended to be more spatially grouped, although not necessarily spatially contiguous. Romary et al. (2012) proposed the use of a clustering algorithm that constrains the clusters to be spatially contiguous. This is accomplished by first computing a Delaunay triangulation graph using the sample locations, then using standard hierarchical clustering methods that have been adapted to allow cluster mergers only when clusters contain connections between samples.

More recent model-based approaches developed within the statistics community partition spatial regions using cluster centroids (Knorr-Held and Rasser 2000; Han, Kamber, and Tung 2001; Kim, Mallick, and Holmes 2005), binary splitting of the spatial axes (Konomi, Sang, and Mallick 2014), or mixture modeling (Neelon, Gelfand, and Miranda 2014). However, each of these methods result in circular (or rectangular) clusters that are not always ideal for spatial problems because physical landmarks (e.g., rivers, mountain ranges, urban sprawl) do not necessarily have these shapes (see the urban sprawl in Figure 1(b), e.g.). Rather, we seek to develop methods that allow more general cluster shapes such as those shown in Figure 1(d).

Our approach follows closely with the clustering techniques for areal data proposed by Anderson, Lee, and Dean (2014) who clustered areal units based on the change in the observed surface relative to each areal neighbor. In contrast, we develop both agglomerative and divisive hierarchical clustering algorithms for point-referenced spatial data (rather than areal data) and we rely on the spatial finite difference as a measure of dissimilarity between observations. In this way, similar to a wombling analysis, clusters are divided along directions of a large spatial gradient (Banerjee and Gelfand 2006) rather than perpendicular to any axis or nearness to a cluster center. Our approach also shares a motivating philosophy with the contribution of Romary et al. (2012) but our approach differs in two ways. First and most importantly, our approach is motivated not by denoting regions with "high" and "low" values (although our approach generally succeeds in identifying such regions), but rather by identifying regions of local stationarity. That is, we use spatial finite differences in defining dissimilarities between points so that the clustering algorithm is able to partition nonstationary random fields along boundaries with large derivatives. In this sense, our approach is fundamentally different from existing methods for the clustering of geostatistical data. A second difference with

Romary et al. (2012) is that the Delaunay-based approach requires ad hoc pruning to avoid spurious connections whereas the Voronoi-based approach we propose is fully automated.

Previous model-based clustering approaches of, for example, Kim, Mallick, and Holmes (2005) or Konomi, Sang, and Mallick (2014), treat the number of clusters as unknown and seek to estimate them using information in the data. This approach, however, can be computationally demanding of itself, and typically relies on reversible-algorithms across model spaces (Green 1995). We seek to avoid these computational complexities by clustering observations prior to model fitting. Furthermore, hierarchical clustering allows us to efficiently explore cluster configurations without undergoing the seemingly insurmountable task of exploring all possible partitions of data points into clusters.

In Section 2, we propose a clustered Gaussian process (cGP) model for large spatial datasets based on a partition of the spatial region. Section 3 discusses the algorithms to perform spatial clustering and exploratory techniques to determine an appropriate number of clusters. Section 4 uses the developed techniques to predict and validate the Houston radiative temperature dataset described above. Section 5 evaluates the performance of the clustering algorithms and associated cGP model on simulated data. Section 6 concludes and outlines directions for future research.

2. A SPATIALLY CLUSTERED GAUSSIAN PROCESS MODEL

Let $y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_N) \in \mathbb{R}$ be random variables observed at the spatial locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ in a spatial domain $\mathcal{S} \subseteq \mathbb{R}^2$. Assume that the spatial locations $\{\boldsymbol{s}_i\}_{i=1}^N$ are partitioned into K distinct clusters (regions) $\mathcal{S}_1, \ldots, \mathcal{S}_K$ such that $\cup_{k=1}^K \mathcal{S}_k = \mathcal{S}$ and $\mathcal{S}_{k_1} \cap \mathcal{S}_{k_2} = \emptyset$ for all $k_1 \neq k_2$ according to one of the algorithms described below in Section 3. Let $\boldsymbol{y}_k = \{y(\boldsymbol{s}_i) : \boldsymbol{s}_i \in \mathcal{S}_k\}$ denote the vector of observations that belong to region $\mathcal{S}_k$ and n_k denote the number of observations in region k with $N = \sum_{k=1}^K n_k$.

We define a spatially clustered GP (cGP) model by assuming conditional independence between the regions so that, for $k = 1, \ldots, K$,

$$\boldsymbol{y}_k \overset{\text{ind}}{\sim} \mathcal{N}\left(\alpha_k \mathbf{1}_{n_k} + \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{H}_k\boldsymbol{\eta}, \sigma_k^2\boldsymbol{R}_k(\boldsymbol{\phi}_k) + \tau_k^2\boldsymbol{I}_{n_k}\right), \quad (1)$$

where α_k is a region-specific intercept, $\boldsymbol{X}_k$ is an $n_k \times P_k$ matrix of covariates with associated coefficients $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kP_k})'$, $\boldsymbol{H}_k$ is a matrix of basis functions with global coefficients $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_G)'$, $\sigma_k^2\boldsymbol{R}_k(\boldsymbol{\phi}_k)$ is a spatial covariance matrix that depends upon region-specific parameters $\boldsymbol{\phi}_k$, and τ_k^2 is the nugget variance for region k. If $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_K')'$ is the vector of all observations, then (1) implies,

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{H}\boldsymbol{\eta}, \boldsymbol{\Sigma} + \boldsymbol{T}\right), \quad (2)$$

where $\boldsymbol{\alpha} = (\alpha_1\mathbf{1}_{n_1}', \ldots, \alpha_K\mathbf{1}_{n_K}')'$, $\boldsymbol{X}$ is block-diagonal with matrices $\boldsymbol{X}_k$ on the main diagonal, $\boldsymbol{\beta}^\star = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_K')'$, $\boldsymbol{H}$ is the $N \times B$ matrix of stacked $\{\boldsymbol{H}_k\}$ matrices, and $\boldsymbol{\Sigma}$ and $\boldsymbol{T}$ are block-diagonal with main diagonal matrices $\sigma_k^2\boldsymbol{R}_k(\boldsymbol{\phi}_k)$ and $\tau_k^2\boldsymbol{I}_{n_k}$, respectively.

For many applications, the intercepts $\alpha_1, \ldots, \alpha_K$ and coefficients $\{\boldsymbol{\beta}_k\}_{k=1}^K$ will be assumed constant across regions such that $\alpha_1 = \cdots = \alpha_K = \alpha$, $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K = \boldsymbol{\beta}$ with common length $P_1 = \cdots = P_K = P$. Under this assumption, $\boldsymbol{\beta}$ would represent the global effect of covariates on the response, $\boldsymbol{X}$ in (2) is redefined to be of dimension $N \times P$ obtained by stacking the matrices $\{\boldsymbol{X}_k\}$ and we replace $\boldsymbol{\beta}^\star$ by $\boldsymbol{\beta}$ in (2). We do note, however, that assuming common $\boldsymbol{\beta}$ coefficients across regions will not be preferred in many instances. That is, the clustering algorithms discussed below are intended to partition $\mathcal{S}$ into dissimilar regions. Assuming globally constant coefficients ($\boldsymbol{\beta}$) would then neglect this dissimilarity by assuming covariates have a spatially constant effect.

For the model in (2), if $\boldsymbol{\eta} = \mathbf{0}$ then the spatial surface will have potentially large discontinuities at the region boundaries. The global set of basis functions $\boldsymbol{H}_k$ and the globally common $\boldsymbol{\eta}$ serve to smooth the predicted spatial surface across these boundaries. To see this, note that if $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$, then the marginal variance for $\boldsymbol{y}$ is

$$\boldsymbol{\Sigma}_y = \boldsymbol{H}\boldsymbol{\Sigma}_\eta\boldsymbol{H}' + \boldsymbol{\Sigma} + \boldsymbol{T} \quad (3)$$

so that the covariance between $\boldsymbol{y}_{k_1}$ and $\boldsymbol{y}_{k_2}$ is $\boldsymbol{H}_{k_1}\boldsymbol{\Sigma}_\eta\boldsymbol{H}_{k_2}' \neq \mathbf{0}$ thereby enforcing correlation across spatial regions. The matrix $\boldsymbol{H}$ can be any basis function matrix but we recommend those basis functions commonly used in fixed rank kriging (Cressie and Johannesson 2008; Kang and Cressie 2011), predictive processes (Banerjee et al. 2008; Finley et al. 2009), or LatticeKrig (Nychka et al. 2015) as these are specifically designed for spatial data. We note that in predictive processes that the $\boldsymbol{H}$ matrix is unknown but parameterized using a few parameters.

In contrast to $\boldsymbol{\eta}$, which is designed to smooth the spatial surface *across* $\mathcal{S}$, the region-specific spatial covariance matrices $\sigma_k^2\boldsymbol{R}_k(\boldsymbol{\phi}_k)$ in (1) serve to smooth the spatial surface *within* $\mathcal{S}_k$. The correlation matrix $\boldsymbol{R}_k(\boldsymbol{\phi}_k)$ is general but will most commonly be constructed using covariance functions such as the Matérn where $\boldsymbol{\phi}_k$ will represent decay and smoothness parameters. In this way, the cGP model is similar to the full-scale approximation (FSA) of Sang and Huang (2012) where $\boldsymbol{\eta}$ controls large-scale spatial correlations and $\sigma_k^2\boldsymbol{R}_k(\boldsymbol{\phi}_k)$ is meant to capture smaller-scale correlations. The cGP model, however, assumes independence across spatial clusters whereas Sang and Huang (2012) assumed all observations are in the same cluster and use a tapered covariance function to enforce sparsity in the covariance matrix. We do note that Sang, Jun, and Huang (2011) presented an alternative to FSA that uses an arbitrary spatial partitioning scheme but, here, we seek to partition the spatial region based on properties of the data rather than for convenience.

Computationally, the cGP model presents a number of benefits. First, inverting the marginal variance $\boldsymbol{\Sigma}_y$ in (3) is most efficiently done using the Sherman–Morrison–Woodbury (SMW) formula. That is, calculation of $\boldsymbol{\Sigma}_y^{-1}$ requires inverting $\boldsymbol{\Sigma}_\eta$ and $\boldsymbol{\Sigma} + \boldsymbol{T}$. First, $\boldsymbol{\Sigma}_\eta$ is of dimension $G \times G$ where G is chosen by the modeler. Hence, if $G \ll N$ then calculating $\boldsymbol{\Sigma}_\eta^{-1}$ is not computationally troublesome. Likewise, the matrix $\boldsymbol{\Sigma} + \boldsymbol{T}$ is block diagonal; thus, calculating $(\boldsymbol{\Sigma} + \boldsymbol{T})^{-1}$ involves inverting K matrices of rank $n_1, \ldots, n_K$ where each $n_k \ll N$. A second computational benefit of the cGP model is the ability to trivially parallelize computations. In (1), conditional on $\boldsymbol{\eta}$, the parameters α_k, $\boldsymbol{\beta}_k$, σ_k^2, τ_k^2, and $\boldsymbol{\phi}_k$ can be estimated independently for each region. This is particularly useful when adopting the Bayesian paradigm because Markov chain Monte Carlo updates of α_k, $\boldsymbol{\beta}_k$, σ_k^2, τ_k^2, and $\boldsymbol{\phi}_k$ are done in parallel.

Under the cGP model, the predictive distribution of $y(\boldsymbol{s}_0)$ conditional on $\boldsymbol{y}$ at the unobserved location $\boldsymbol{s}_0 \in \mathcal{S}_k$ is

$$\begin{aligned} y(\boldsymbol{s}_0) \mid \boldsymbol{y} &\sim \mathcal{N}(\mu_{s_0|y_k}, \sigma^2_{s_0|y_k}) \\ \mu_{s_0|y_k} &= \mu_k(\boldsymbol{s}_0) + \left[\sigma_k^2 \boldsymbol{R}'_k(y(\boldsymbol{s}_0), \boldsymbol{y}_k \mid \boldsymbol{\phi}_k)\right] \\ &\quad \times \left(\sigma_k^2 \boldsymbol{R}_k(\boldsymbol{\phi}_k) + \tau_k^2 \boldsymbol{I}_{n_k}\right)^{-1} [\boldsymbol{y}_k - \boldsymbol{\mu}_k] \\ \sigma^2_{s_0|y_k} &= (\sigma_k^2 + \tau_k^2) - \left[\sigma_k^2 \boldsymbol{R}'_k(y(\boldsymbol{s}_0), \boldsymbol{y}_k \mid \boldsymbol{\phi}_k)\right] \\ &\quad \times \left(\sigma_k^2 \boldsymbol{R}_k(\boldsymbol{\phi}_k) + \tau_k^2 \boldsymbol{I}_{n_k}\right)^{-1} \left[\sigma_k^2 \boldsymbol{R}_k(y(\boldsymbol{s}_0), \boldsymbol{y}_k \mid \boldsymbol{\phi}_k)\right], \end{aligned} \tag{4}$$

where $\mu_k(\boldsymbol{s}) = \alpha_k + \boldsymbol{x}'(\boldsymbol{s}_0)\boldsymbol{\beta}_k + \boldsymbol{h}'(\boldsymbol{s}_0)\boldsymbol{\eta}$, $\boldsymbol{\mu}_k = \mathbb{E}(\boldsymbol{y}_k) = \alpha_k \mathbf{1}_{n_k} + \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{H}_k\boldsymbol{\eta}$, $\boldsymbol{x}'(\boldsymbol{s})$ and $\boldsymbol{h}(\boldsymbol{s})$ are the covariates and basis functions measured at location $\boldsymbol{s}$, respectively, and $\boldsymbol{R}'_k(y(\boldsymbol{s}_0), \boldsymbol{y}_k \mid \boldsymbol{\phi}_k)$ is the vector of correlations between $y(\boldsymbol{s}_0)$ and $\boldsymbol{y}_k$. From the predictive distribution in (4), we note an advantage of the clustered approach is that predictions are made within each cluster and, hence, can be done in parallel to facilitate computational speed.

3. SPATIAL CLUSTERING

We propose to use hierarchical clustering algorithms (divisive and agglomerative) to determine an appropriate a priori partition of the spatial region $\mathcal{S}$ for use within the above cGP model. That is, we seek to a priori divide observations into groups based on a measure of dissimilarity. We argue that using a priori partitioning in this manner is appropriate because (i) it avoids complex and computationally demanding statistical procedures for modeling the cluster membership and (ii) it avoids defining partitions based on binary splitting of the spatial axes.

3.1 Cluster Dissimilarity and Spatial Contiguity Constraints

To split $y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)$ into K clusters, we define the "dissimilarity" between $y(\boldsymbol{s}_i)$ and $y(\boldsymbol{s}_j)$ to be

$$d_{ij} = d(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)) = \frac{|y(\boldsymbol{s}_j) - y(\boldsymbol{s}_i)|}{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|}, \tag{5}$$

where $\|\boldsymbol{s}_j - \boldsymbol{s}_i\|$ is the Euclidean distance between the points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. The dissimilarity metric d_{ij} in (5) is motivated by the spatial finite differences (an estimate of the directional derivative) described in Banerjee, Gelfand, and Sirmans (2003) and Banerjee and Gelfand (2006). By using the finite difference as a measure of dissimilarity, observations will tend to cluster based on the change in the spatial surface. That is, the dissimilarity d_{ij} will be large when the spatial surface $y(\boldsymbol{s})$ changes rapidly in the $\boldsymbol{s}_j - \boldsymbol{s}_i$ direction from $\boldsymbol{s}_i$ leading to $y(\boldsymbol{s}_i)$ and $y(\boldsymbol{s}_i)$ being assigned to different clusters. Using (5), cluster boundaries will be placed along directions of a large derivative. These large observed rates of change are natural regional boundaries because they represent areas where assumptions of isotropy and stationarity may not hold.

As indicated by the dissimilarity defined in (5), we recommend that the clustering algorithm be conducted on the observations $\{y(\boldsymbol{s}_i)\}_i$. That said, in many settings the primary interest is to estimate spatial random effects that explain the behavior of the response after controlling for available covariates $\boldsymbol{x}(\boldsymbol{s}_i)$, as in (1). Following this line of thinking, one could envision redefining the dissimilarity instead based on the residuals from a regression with nonspatial error (i.e., residuals from the model $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$). The dissimilarity would replace $y(\boldsymbol{s}_i)$ with $r(\boldsymbol{s}_i) = y(\boldsymbol{s}_i) - \widehat{\alpha} - \boldsymbol{x}'(\boldsymbol{s}_i)\widehat{\boldsymbol{\beta}}$, where $\widehat{\alpha}$ and $\widehat{\boldsymbol{\beta}}$ are maximum likelihood estimates (or Bayesian posterior means in a nonspatial model). However, in our analyses when we compared these two approaches for the application in Section 4 and the simulation studies in Section 5 we found that clustering based on the observations performed better in every case (results based on clustering of the residuals are omitted for brevity). This is not to say that clustering residuals is never appropriate, but in all of the cases we investigated clustering based on the observations was preferable.

Traditional hierarchical clustering based on d_{ij} will not produce spatially contiguous clusters and will, hence, forfeit the ability to borrow information from spatial neighbors to perform predictions. Instead, to enforce spatial contiguity of clusters, we choose to cluster observations only if they are Voronoi neighbors. That is, we first calculate a Voronoi tessellation of the points $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$ and define $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ to be neighbors if they share a border. Let $\boldsymbol{s}_i \overset{v}{\sim} \boldsymbol{s}_j$ denote that $\boldsymbol{s}_i$ is a Voronoi neighbor of $\boldsymbol{s}_j$.

While d_{ij} denotes the dissimilarity between two observations, we also need to define the dissimilarity between two clusters. This is analogous to choosing the type of linkage in traditional clustering methods. Let $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$ denote the set of locations that belong to clusters k_1 and k_2. Here, we consider the following four options as measures of dissimilarity between $\mathcal{C}_{k_1}$ and $\mathcal{C}_{k_2}$,

$$d(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \min_{i,j}\{d_{ij} : \boldsymbol{s}_i \in \mathcal{C}_{k_1}, \boldsymbol{s}_j \in \mathcal{C}_{k_2}, \boldsymbol{s}_i \overset{v}{\sim} \boldsymbol{s}_j\} \tag{6}$$

$$d(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \max_{i,j}\{d_{ij} : \boldsymbol{s}_i \in \mathcal{C}_{k_1}, \boldsymbol{s}_j \in \mathcal{C}_{k_2}, \boldsymbol{s}_i \overset{v}{\sim} \boldsymbol{s}_j\} \tag{7}$$

$$d(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \operatorname*{avg}_{i,j}\{d_{ij} : \boldsymbol{s}_i \in \mathcal{C}_{k_1}, \boldsymbol{s}_j \in \mathcal{C}_{k_2}, \boldsymbol{s}_i \overset{v}{\sim} \boldsymbol{s}_j\}, \tag{8}$$

and

$$d(\mathcal{C}_{k_1}, \mathcal{C}_{k_2}) = \left[\frac{n_{\mathcal{C}_{k_1}} n_{\mathcal{C}_{k_2}}}{n_{\mathcal{C}_{k_1}} + n_{\mathcal{C}_{k_2}}} (\overline{y}_{\mathcal{C}_{k_1}} - \overline{y}_{\mathcal{C}_{k_2}})^2\right] \frac{1}{\overline{E}}, \tag{9}$$

where $\overline{y}_{\mathcal{C}_k}$ is the average of observations in cluster k and $\overline{E}$ is the average Euclidean distance between points $\boldsymbol{s}_i \in \mathcal{C}_{k_1}$ and $\boldsymbol{s}_j \in \mathcal{C}_{k_2}$, where $\boldsymbol{s}_i \overset{v}{\sim} \boldsymbol{s}_j$. The linkage specifications in (6)–(8) are analogous to "single," "complete," and "average" linkage in traditional agglomerative clustering while (9) is analogous to Ward's clustering method (see Rencher and Christensen 2012, chap. 15) but appropriately accounts for the spatial distance over which the change occurs.

3.2 Clustering Algorithms

We first consider an agglomerative clustering approach where each observation starts as its own cluster and then are linked together based on the smallest d_{ij}. Table 1 describes an agglomerative clustering algorithm. We emphasize that in the algorithm of Table 1, two clusters k_1 and k_2 are agglomerated if and only

Table 1. Agglomerative spatial clustering algorithm

1:	Initialize $\mathcal{C}_k = \boldsymbol{s}_k$ for $k = 1, \ldots, N$ to be the set of locations that belong to cluster k.
2:	For $k = 0, \ldots, N - K + 1$,
3:	(a) Find cluster indices k_1 and k_2 that correspond to the minimum dissimilarity $d(\mathcal{C}_{k_1}, \mathcal{C}_{k_2})$ subject to the constraint that there exists $\boldsymbol{s}_i \in \mathcal{C}_{k_1}$ and $\boldsymbol{s}_j \in \mathcal{C}_{k_2}$ where $\boldsymbol{s}_i \overset{v}{\sim} \boldsymbol{s}_j$.
4:	(b) If $k_1 < k_2$, redefine $\mathcal{C}_{k_1} = \mathcal{C}_{k_1} \cup \mathcal{C}_{k_2}$ and set $\mathcal{C}_{k_2} = \emptyset$; otherwise, redefine $\mathcal{C}_{k_2}$ similarly and set $\mathcal{C}_{k_1} = \emptyset$.
5:	End k-for loop.

if they have minimum dissimilarity among all cluster pairs *and* there exists an $\boldsymbol{s}_i \in \mathcal{C}_{k_1}$ and $\boldsymbol{s}_j \in \mathcal{C}_{k_2}$ that are Voronoi neighbors. This ensures that the resulting clusters will be spatially contiguous.

Similar to any hierarchical clustering algorithm, we emphasize that the choice of dissimilarity metric will change the resulting clusters. In our experience, we found that using Ward's dissimilarity divided observations into clusters more evenly, whereas the other dissimilarity metrics in (6)–(8) tended to produce unrealistic cluster sizes. Specifically, using the single or average linkage often resulted in clusters of unit size for the simulation studies in Section 5. As mentioned earlier, unit clusters are undesirable because they assign unique intercepts and slopes to individual observations, and when the number of clusters is specified a priori, it forces all of the other observations into a small number of clusters. The large size of these other clusters may be problematic because the cluster-specific parameters will become more computationally challenging to estimate.

Rather than agglomerating similar observations, clustering can be done using divisive clustering wherein all the observations start in the same cluster and are subsequently divided into smaller clusters. An example algorithm for implementing a divisive approach is given in Table 2.

For the divisive approach, when splitting the cluster $k^\star$ the order in which each $\boldsymbol{s}_h \in \mathcal{C}_{k^\star}$ is assigned to either the new or old cluster is important. If, for example, during Step 5(c) of Table 2 we wish to assign the location $\boldsymbol{s}_h$ to either the new or old cluster but $\boldsymbol{s}_h \overset{v}{\not\sim} \boldsymbol{s}_i$ and $\boldsymbol{s}_h \overset{v}{\not\sim} \boldsymbol{s}_j$ then d_{hi} and d_{hj} will not be defined by the spatial contiguity constraint. Hence, for step 5 (c), we first assign cluster memberships to each $\boldsymbol{s}_h \in \mathcal{C}_{k^\star} \setminus \{\boldsymbol{s}_i, \boldsymbol{s}_j\}$ where $\boldsymbol{s}_h \overset{v}{\sim} \boldsymbol{s}_i$ and $\boldsymbol{s}_h \overset{v}{\sim} \boldsymbol{s}_j$ thereby defining a spatial boundary.

Both agglomerative and divisive clustering have drawbacks. For agglomerative clustering, if the number of observations is large then the number of required iterations may be computationally prohibitive. On the other hand, divisive clustering, while computationally cheap, is overly sensitive to the sampling locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N$. That is, when $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ is small then d_{ij} will be large even if $|y(\boldsymbol{s}_i) - y(\boldsymbol{s}_j)|$ is small, which, consequently, leads to an unnecessary declustering of observations. To overcome both of these problems, we propose to perform clustering after aggregating observations to a lattice. Details are as follows.

Let $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_L^\star$ define a lattice over the spatial region $\mathcal{S}$, $\mathcal{N}_\ell = \{\boldsymbol{s}_i : \|\boldsymbol{s}_i - \boldsymbol{s}_\ell^\star\| < \|\boldsymbol{s}_i - \boldsymbol{s}_{\ell'}^\star\| \text{ for all } \ell' \neq \ell\}$ (i.e., $\mathcal{N}_\ell$ is the subset of all observed locations whose closest lattice point is $\boldsymbol{s}_\ell^\star$) and, finally, let $\overline{y}(\boldsymbol{s}_\ell^\star) = |\mathcal{N}_\ell|^{-1} \sum_{\boldsymbol{s}_i \in \mathcal{N}_\ell} y(\boldsymbol{s}_i)$ be the average of the observed values in $\mathcal{N}_\ell$. We propose to cluster $\{\overline{y}(\boldsymbol{s}_1^\star), \ldots, \overline{y}(\boldsymbol{s}_L^\star)\}$ using the above algorithms rather than $\{y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n)\}$ for two reasons. First, because $\boldsymbol{s}_1^\star, \ldots, \boldsymbol{s}_L^\star$ constitutes a lattice, the points are equally spaced which, subsequently, prevents sampling designs from creating an unstable dissimilarity metric in (5). And, second, by choosing $L \ll N$, the dimension of the problem is reduced and agglomerative algorithms are then computationally feasible.

In practice, the cluster assignment of $y(\boldsymbol{s}_i)$ is the same as the cluster assignment of its nearest lattice point. Hence, first aggregating to a lattice allows L, the number of lattice points, to be used as another adjustment (besides linkage) to control the clustering. That is, adjusting L will result in different cluster assignments of the observations. We also note that by first aggregating to a lattice, $\overline{y}(\boldsymbol{s}_\ell^\star)$ may be undefined if $\mathcal{N}_\ell$ is empty (i.e., no observations are "closest" to $\boldsymbol{s}_\ell^\star$). In these cases, those points where $\mathcal{N}_\ell$ is empty are removed and clustering still proceeds as above.

3.3 Determining the Number of Clusters/Regions

The cGP model in Section 2 and the clustering algorithms in Section 3.2 assume that the number of clusters, K, is known a priori. Spatial partitioning methods by Kim, Mallick, and

Table 2. Divisive spatial clustering algorithm

1:	Initialize $\mathcal{C}_1 = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ to be set of all locations that belong to cluster 1.
2:	For $k = 1, \ldots, K - 1$,
3:	(a) Find cluster index $k^\star \leq k$ and location indices i and j that maximize the dissimilarity d_{ij} subject to the constraints that $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathcal{C}_{k^\star}$ and $\boldsymbol{s}_i \overset{v}{\sim} \boldsymbol{s}_j$.
4:	(b) Define $\mathcal{T}_1 = \boldsymbol{s}_i$, $\mathcal{T}_2 = \boldsymbol{s}_j$.
5:	(c) Agglomerate each $\boldsymbol{s}_h \in \mathcal{C}_{k^\star} \setminus \{\boldsymbol{s}_i, \boldsymbol{s}_j\}$ to $\mathcal{T}_1$ or $\mathcal{T}_2$ using single, complete, average or Ward linkage.
6:	(d) Redefine $\mathcal{C}_{k^\star} = \mathcal{T}_1$ and set $\mathcal{C}_{k+1} = \mathcal{T}_2$.
7:	End k for-loop.

Holmes (2005) and Konomi, Sang, and Mallick (2014) seek to estimate K but doing so requires additional computation for an already computationally demanding problem. Hence, to avoid this computation, we propose to use exploratory techniques to determine K a priori.

The optimal choice of K will balance predictive accuracy and model complexity. Ideally, cross-validation would be used to compare predictive accuracy for various values of K but this comes at a high computational cost and may not be feasible for large samples. An alternative to cross-validation is to use Akaike's information criterion (AIC), the Schwarz-Bayesian criterion (BIC), adjusted-R^2 or the like from a model that includes the covariates ($\boldsymbol{X}_k$), basis functions ($\boldsymbol{H}_k$), and cluster assignment as predictors. For example, one can first cluster observations using $K = 2$ and successively increase K until the increase in AIC is negligible (the so-called "elbow" method). Because we are using clustering techniques, clustering diagnostics, such as the GAP statistic of Tibshirani, Walther, and Hastie (2001) or the GUD statistic of Kou (2014), could also be used to choose K. We illustrate the proposed methodology using a variation on the elbow method for the simulation studies in Section 5 and the GAP statistic is illustrated on the temperature example in Section 4.

The choice of K may depend on subjective factors such as computational resources, a priori information, or sample size. Much of the motivation behind the development of cGPs herein lies in distributing the computational load of fitting a spatial model across nodes. For large sample sizes, for example, larger values of K might need to be used simply to make the computation feasible. Likewise, smaller values of K may be required due to lack of computational resources (fewer available nodes).

4. HOUSTON REMOTE SENSING APPLICATION

We apply the above methods to analyze the Houston radiative temperature dataset introduced in Section 1. The goals of this analysis are to (i) use both the MODIS and HRLDAS data to construct a complete map of the radiative temperature over the city of Houston and (ii) validate the HRLDAS data. To accomplish these goals, we fit the cGP model presented in Equation (1) where $\boldsymbol{X}_k$ is the vector of HRLDAS simulated temperatures (hence, $\boldsymbol{\beta}_k$ in (1) is a scalar β_k). Under this model, the α_k and β_k parameters represent regional biases in the HRLDAS model. That is, α_k would represent an additive bias whereas β_k represents a multiplicative bias. Ideally, $\alpha_k = 0$ and $\beta_k = 1$ for all k, in which case the HRLDAS simulation mimics the MODIS data and is an unbiased digital reconstruction of radiative temperature. Otherwise, the HRLDAS model is a biased reconstruction of radiative temperature.

To determine the clusters for the cGP model, we cluster using the agglomerative clustering algorithm after aggregating to a 20^2 lattice (this clustering algorithm is seen to perform the best in the simulation studies in Section 5). We fit $K = 2, 3, \ldots, 20$ clusters and, using the GAP statistic of Tibshirani, Walther, and Hastie (2001), found that $K = 10$ was optimal. We also used the elbow method based on adjusted-R^2 and found that the results were similar. The cluster assignments when $K = 10$ are shown in Figure 1(d) where the color bar indicates the average MODIS temperature within each cluster. Comparing Figure 1(a) with Figure 1(d) notice that the clusters successfully separate regions of higher and lower temperatures. In this application, the clusters are, themselves, of interest in that the clusters likely show urban heat island effects wherein the temperature in the city is higher than in more rural areas.

For the cGP model, we define $\boldsymbol{H}$ using three resolutions of bisquare basis functions on 4×4, 10×10, and 15×15 regular grids over the spatial domain. Following Cressie and Johannesson (2008), we take the tapering distance for the bisquare function to be 1.5 times the intergrid point distance. Because the β_k parameters are of interest, we orthogonalized the $\boldsymbol{H}$ matrix to avoid spatial confounding (Hodges and Reich 2010). As suggested by Kang and Cressie (2011), we use the Givens prior for $\boldsymbol{\Sigma}_\eta$. The correlation matrices $\{\boldsymbol{R}_k\}$ are given by the Matérn correlation function with smoothness 0.5 and decay parameter ϕ_k (which reduces to the exponential correlation function $\exp\{-\phi d\}$ where d is the distance between locations). Rather than determining a prior for ϕ_k, we find it more intuitive to place a prior on the spatial range d_0 due to the one-to-one relationship between ϕ_k and a spatial range (for a fixed smoothness). We define the spatial range d_0 to be the distance at which the spatial correlation decays to 0.05 and assume that d_0 follows a uniform distribution with upper and lower bounds of 1% and 99% of the maximum spatial distance. We a priori assume $\{\alpha_k\}_k$ are independent Gaussian random variables with mean $\alpha^\star$ and variance σ_α^2. Likewise, we assume $\{\beta_k\}_k$ are iid $\mathcal{N}(\beta^\star, \sigma_\beta^2)$ random variables. We subsequently use Jefferys' priors for $\alpha^\star$, $\beta^\star$, σ_α^2, and σ_β^2. Finally, vague, inverse gamma priors are assumed for σ_k^2 and τ_k^2.

We fit the above cGP model by simulating 5000 draws from the posterior distribution after a burn-in of 5000 draws. Subsequent analysis of trace plots suggested that convergence had been reached. Furthermore, the Monte Carlo standard errors were estimated to be less than 0.01 for all parameters (Jones et al. 2006).

For comparison, we also fit a cGP model where $\boldsymbol{\eta} = \boldsymbol{0}$ (no spatial smoothing across regions), which we will denote as cGP_0, a global fixed rank (FR) model wherein we let $\boldsymbol{y} \sim \mathcal{N}(\alpha + \boldsymbol{X}\beta + \boldsymbol{H}\boldsymbol{\eta}, \sigma^2\boldsymbol{I})$ and a full-scale approximation (FSA) model where we assume $\boldsymbol{y} \sim \mathcal{N}(\alpha + \boldsymbol{X}\beta + \boldsymbol{H}\boldsymbol{\eta}, \boldsymbol{\Sigma})$ where, following Sang, Jun, and Huang (2011), Σ is a block-diagonal matrix. The spatial blocks for the FSA model were formed by dividing the spatial region into nine equal areas using latitude cut-points of -95.58 and -95.17 and longitude cut-points of 29.66 and 30.07. The FR and FSA models are computationally reasonable alternatives to fitting a full GP model and allows us to compare an unclustered spatial model (the FR model) and a conveniently clustered model (the FSA model) with the finite difference-based cluster models described herein. We note that fitting a full GP model in this example would require us to work with a $14,309 \times 14,309$ dimensional matrix that is not computationally feasible. However, the simulation studies in the supplementary material directly compare the cGP model to a full GP model.

Figure 2 displays the posterior predictive mean of radiative temperature across the city of Houston while Figure 3 displays

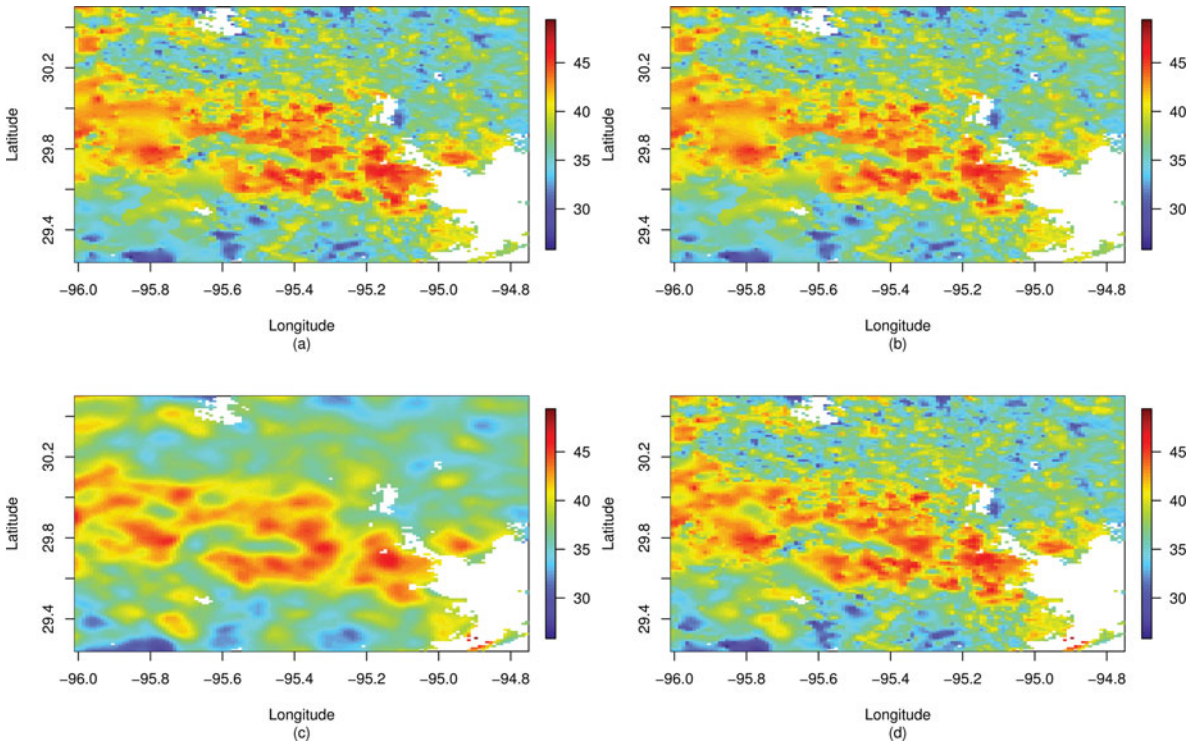

Figure 2. Posterior predictive mean of interpolated radiative temperature across Houston according to (a) the cGP model, (b) the cGP_0 model, (c) the FR model, and (d) the FSA model.

the log-standard deviation of the posterior predictive distribution. The posterior predictive means of all four models give reasonable results but, comparing panels (a), (b), and (d) to (c), the FR model seems to smooth the temperature surface more than the cGP, cGP_0 models and the FSA model. This smoothing is to be expected as Stein (2014) notes that low rank approximations have a tendency to produce overly smooth predictions.

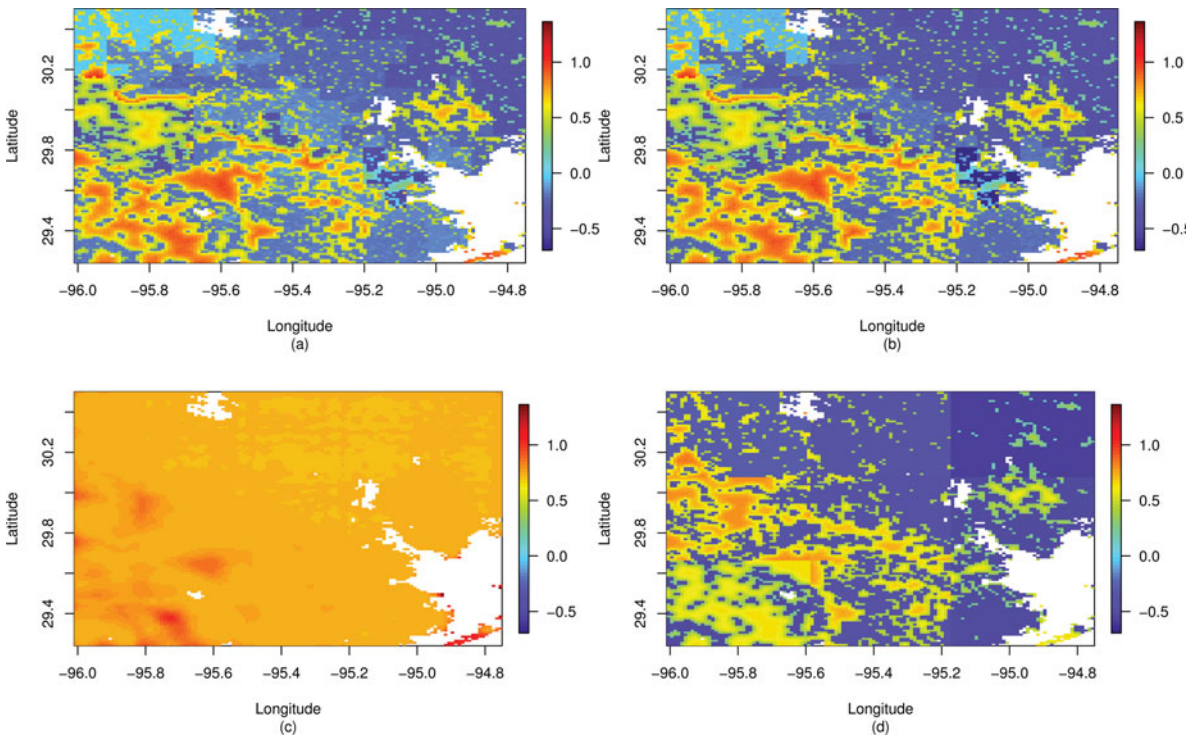

Figure 3. The log standard deviation of the posterior predictive distribution of radiative temperature across Houston according to (a) the cGP model, (b) the cGP_0 model, (c) the FR model, and (d) the FSA model.

Table 3. Predictive diagnostics of the cGP, cGP_0, FR, and FSA models on a hold-out sample in the Houston study

Model	Bias	RMSE	Coverage	Width
cGP	−0.07	1.34	0.95	6.03
cGP_0	−0.07	1.34	0.95	6.11
FR	−0.02	2.05	0.95	10.40
FSA	−0.07	1.71	0.95	7.34

According to Figure 3, the models also differ in their estimate of the posterior predictive standard deviation. Noticeably, the FR model has a near-spatially constant predictive standard deviation while the cGP and FSA models rely heavily on the density of observations to decrease the predictive standard deviation (a distinct advantage for the cGP, cGP_0, and FSA models). Contrasting the cGP and cGP_0 models to the FSA model in Figure 3(d), we see that the clusters are apparent in both the cGP and FSA models. That is, artificial boundaries in the standard deviation surfaces appear along the cluster borders. We note, further, that there are regions where the cGP and FSA uncertainties do not match. Particularly, in the bottom left corner of the spatial region the cGP model has higher uncertainty than the FSA model. However, in the area centered at, approximately, −95.85 degrees longitude and 29.9 degrees latitude, the cGP and cGP_0 models have lower uncertainty than the FSA model.

Comparing the models further, Table 3 assesses the predictive ability of each model using a hold-out sample of 500 observations. Specifically, Table 3 displays the bias and root mean square error (RMSE) of the posterior predictive mean as well as the coverage and width of 95% predictive intervals on the hold-out sample. The cGP and cGP_0 models have similar practical performance for these data implying that the cGP model does not need the global spatial component to adequately capture the regionally heterogenous temperature process. Encouragingly, both the cGP and cGP_0 models outperform the FR and FSA models in terms of RMSE and the predictive interval width (but only slightly in the FSA case) while having a slightly higher bias (which is likely negligible). The similar performance between the cGP and FSA models is, somewhat, to be expected because, as discussed in Section 2, the cGP model is a special case of the FSA model, albeit with a different partition of the spatial region.

Finally, Figure 4 displays the posterior means for α_k and β_k across all clusters. Recall that $\alpha_k = 0$ and $\beta_k = 1$ suggests that HRLDAS is an unbiased reconstruction of the MODIS data. Noticeably, the HRLDAS is a negatively biased reconstruction of the MODIS temperature field in that the estimates of α_k are all above zero with posterior probabilities of 1. The β_k estimates were all significantly different from 1, confirming the bias in the HRLDAS data. Most regions had positive estimated β_k values, indicating a bias in the magnitude of the temperatures, but there was also one region where β_k had a negative value. Comparing to the observed temperatures in Figure 1, the positive β_k correspond to higher temperatures while negative estimates of β_k correspond to regions of cooler temperatures. Since the values of β_k were closer to 1 for warmer regions, this suggests that the HRLDAS is more effective at simulating warmer urban temperatures than cooler rural temperatures.

5. SIMULATION STUDIES

Three simulation studies were performed to evaluate (i) the performance of the individual clustering algorithms, (ii) the performance of each algorithm in selecting an appropriate number of clusters, and (iii) the predictive and inference performance of each algorithm. Details of these simulation studies are available in the supplementary materials but we summarize the main conclusions here. First, the agglomerative algorithm on the raw data typically performs the best in terms of the adjusted rand index (ARI; a measure of cluster agreement with the true clusters (see Hubert and Arabie 1985) and adjusted-R^2. However, in spite of vast reductions in computation time, the agglomerative and divisive algorithms using 20 lattice points performed very well (the agglomerative algorithm, however, is preferred). And, second, in terms of inference, simulation results lead to the intuitive conclusion that correctly estimating regionally specific parameters hinges upon the ability to identify the underlying "true" clusters. However, in terms of prediction, the spatially clustered Gaussian process model in Section 2 predicts well even when the estimated clusters are different than the "true" clusters.

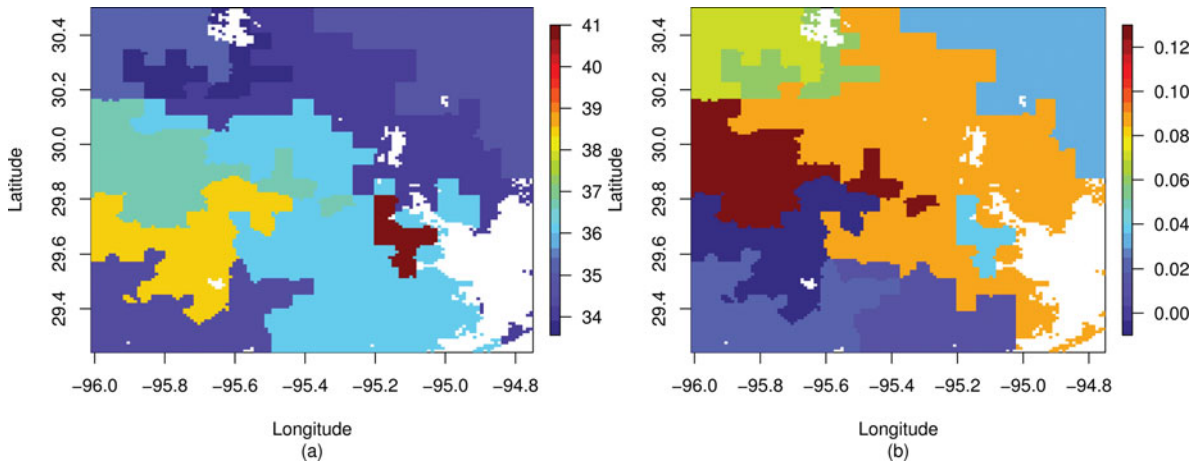

Figure 4. Posterior means of (a) α_k and (b) β_k for the cGP model.

6. DISCUSSION

The spatial clustering algorithms and associated cGP model developed herein have proven useful for finding spatial clusters as well as performing spatial prediction and inference. By using hierarchical clustering based on spatial finite differences (an estimate of the directional derivative), the spatial cluster boundaries follow lines of steep change in the spatial surface and not artificial circular or rectangular boundaries induced by the model specification. In addition, the inherent reduction in computational demands for this method makes it an attractive choice for approximating a full GP model in the context of large spatial datasets.

An inherent potential shortcoming of the proposed methods is the reliance on a priori clustering. That is, herein the clusters are determined prior to model fitting and fixed throughout the analysis. This shortcoming is the price paid for enormous gains in computational efficiency. While this is advantageous in terms of model fitting, the uncertainty associated with the clustering is not quantified in estimates of the model parameters. One possible approach to account for the cluster uncertainty is to condition the observed temperature surface and cluster assignment on the derivative process. Alternatively, Dirichlet process clustering may also provide a way to directly model the cluster membership. Details on how to account for uncertainty in the clustering is left for future work.

The algorithms and cGP detailed herein are for purely spatial data. However, future extensions to multivariate and spatio-temporal data would be of interest. We imagine that extensions to spatio-temporal data could rely on information regarding gradients across space and time. Furthermore, extensions to the multivariate setting would employ traditional clustering techniques for nonspatial multivariate data.

SUPPLEMENTARY MATERIALS

The supplementary materials contain an illustrative example of agglomerative clustering and simulation studies.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number DMS-1417856. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

[Received September 2014. Revised September 2015.]

REFERENCES

Anderson, C., Lee, D., and Dean, N. (2014), "Identifying Clusters in Bayesian Disease Mapping," *Biostatistics*, 15, 457–469. [94]

Banerjee, S., and Gelfand, A. E. (2006), "Bayesian Wombling: Curvilinear Gradient Assessment Under Spatial Process Models," *Journal of the American Statistical Association*, 101, 1487–1501. [94,96]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Datasets," *Journal of the Royal Statistical Society,* Series B, 70, 825–848. [94,95]

Banerjee, S., Gelfand, A. E., and Sirmans, C. F. (2003), "Directional Rates of Change Under Spatial Process Models," *Journal of the American Statistical Association*, 98, 946–954. [96]

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010), "A Spatio-Temporal Downscaler for Output From Numerical Models," *Journal of Agricultural, Biological and Environmental Statistics*, 15, 176–197. [93]

Bourgault, G., Marcotte, D., and Legendre, P. (1992), "The Multivariable Covariogram as a Spatial Weighting Function in Classification Methods," *Mathematical Geology*, 24, 463–478. [94]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 70, 209–226. [94,95,98]

Eidsvik, J., Shaby, B., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and Prediction in Spatial Models With Block Composite Likelihoods," *Journal of Computational and Graphical Statistics*, 23, 295–315. [94]

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009), "Improving the Performance of Predictive Process Modeling for Large Datasets," *Computational Statistics and Data Analysis*, 53, 2873–2884. [95]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [95]

Han, J., Kamber, M., and Tung, A. K. H. (2001), "Spatial Clustering Methods in Data Mining: A Survey," in *Geographic Data Mining and Knowledge Discovery*, eds. H. Miller, and J. Han, Boca Raton, FL: Taylor and Francis, pp. 1–26. [94]

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. [98]

Hu, L., Brunsell, N. A., Monaghan, A. J., Barlage, M., and Wilhelmi, O. V. (2014), "How Can We Use Modis Land Surface Temperature to Validate Long-Term Urban Model Simulations?" *Journal of Geophysical Research*, 119, 3185–3201. [93]

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [100]

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006), "Fixed-Width Output Analysis for Markov Chain Monte Carlo," *Journal of the American Statistical Association*, 101, 1537–1547. [98]

Kang, E. L., and Cressie, N. (2011), "Bayesian Inference for the Spatial Random Effects Model," *Journal of the American Statistical Association*, 106, 972–983. [95,98]

Kim, H. M., Mallick, B. K., and Holmes, C. C. (2005), "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes," *Journal of the American Statistical Association*, 100, 653–668. [94,95,98]

Knorr-Held, L., and Rasser, G. (2000), "Bayesian Detection of Clusters and Discontinuities in Disease Maps," *Biometrics*, 56, 13–21. [94]

Konomi, B. A., Sang, H., and Mallick, B. K. (2014), "Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations," *Journal of Computational and Graphical Statistics*, 23, 802–829. [94,95,98]

Kou, J. (2014), "Estimating the Number of Clusters via the GUD Statistic," *Journal of Computational and Graphical Statistics*, 23, 403–417. [98]

Neelon, B., Gelfand, A. E., and Miranda, M. L. (2014), "A Multivariate Spatial Mixture Model for Areal Data: Examining Regional Differences in Standardized Test Scores," *Journal of the Royal Statistical Society,* Series C, 63, 737–761. [94]

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), "A Multi-Resolution Gaussian Process Model for the Analysis of Large Spatial Data Sets," *Journal of Computational and Graphical Statistics*, 24, 579–599. [94,95]

Oliver, M., and Webster, R. (1989), "A Geostatistical Basis for Spatial Weighting in Multivariate Classification," *Mathematical Geology*, 21, 15–35. [94]

Rencher, A. C., and Christensen, W. F. (2012), *Method of Multivariate Analysis* (3rd ed.), Hoboken, NJ: Wiley. [96]

Romary, T., Rivoirard, J., Deraisme, J., Quinones, C., and Freulon, X. (2012), "Domaining by Clustering Multivariate Geostatistical Data," in *Geostatistics Oslo*, eds. P. Abrahamsen, R. Hauge, and O. Kolbjornsen, Amsterdam, the Netherlands: Springer, pp. 455–466. [94]

Sang, H., and Huang, J. Z. (2012), "A Full Scale Approximation of Covariance Functions for Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 74, 111–132. [94,95]

Sang, H., Jun, M., and Huang, J. Z. (2011), "Covariance Approximation for Large Multivariate Spatial Data Sets With an Application to Multiple Climate Model Errors," *Annals of Applied Statistics*, 5, 2519–2548. [95,98]

Stein, M. L. (2014), "Limitations on Low Rank Approximation for Covariance Matrices of Spatial Data," *Spatial Statistics*, 8, 1–19. [99]

Stone, B., Hess, J. J., and Frumkin, H. (2010), "Urban Form and Extreme Heat Events: Are Sprawling Cities More Vulnerable to Climate Change Than Compact Cities?" *Environmental Health Perspectives*, 118, 1425–1428. [93]

Sun, Y., Li, B., and Genton, M. (2011), "Geostatistics for Large Datasets," in *Advances and Challenges in Space-time Modelling of Natural Events*, eds. J. M. Montero, E. Porcu, and M. Schlather, New York: Springer, pp. 55–77. [93]

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the Number of Clusters in a Dataset via the Gap Statistic," *Journal of the Royal Statistical Society,* Series B, 63, 411–423. [98]