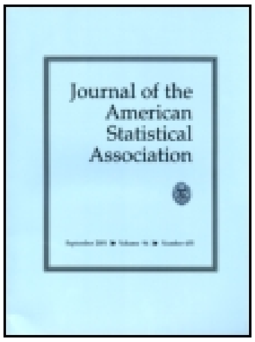

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets

Furong Li & Huiyan Sang

To cite this article: Furong Li & Huiyan Sang (2018): Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets, Journal of the American Statistical Association, DOI: 10.1080/01621459.2018.1529595

To link to this article: https://doi.org/10.1080/01621459.2018.1529595

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 00, NO. 0, 1–13: Applications and Case Studies
https://doi.org/10.1080/01621459.2019.1529595

Taylor & Francis
Taylor & Francis Group

Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets

Furong Li[a] and Huiyan Sang[b]

[a]School of Mathematical Sciences, Ocean University of China, Qingdao, China; [b]Department of Statistics, Texas A&M University, College Station, TX

Abstract

Spatial regression models have been widely used to describe the relationship between a response variable and some explanatory variables over a region of interest, taking into account the spatial dependence of the observations. In many applications, relationships between response variables and covariates are expected to exhibit complex spatial patterns. We propose a new approach, referred to as spatially clustered coefficient (SCC) regression, to detect spatially clustered patterns in the regression coefficients. It incorporates spatial neighborhood information through a carefully constructed regularization to automatically detect change points in space and to achieve computational scalability. Our numerical studies suggest that SCC works very effectively, capturing not only clustered coefficients, but also smoothly varying coefficients because of its strong local adaptivity. This flexibility allows researchers to explore various spatial structures in regression coefficients. We also establish theoretical properties of SCC. We use SCC to explore the relationship between the temperature and salinity of sea water in the Atlantic basin; this can provide important insights about the evolution of individual water masses and the pathway and strength of meridional overturning circulation in oceanography. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received 7 June 2017
Revised 31 August 2018

KEYWORDS
Fused lasso; Minimum spanning tree; Penalized least squares; Spatially clustered coefficient; Varying coefficient regression.

1. Introduction

Numerous problems today in the environmental, earth, and biological sciences involve large amounts of spatial data that are obtained from remote sensors, satellite images, scientific climate computer models, and so forth. In many such applications, a main problem of interest is investigation of the relationship between a response variable and a set of explanatory variables over a region of interest, taking into account the spatial dependence of the observations. Spatial regression models such as Gaussian process regression (Cressie 1993) or spatial generalized linear regression models (Diggle, Tawn, and Moyeed 1998) have been widely adopted for this problem; these account for spatial dependence by adding a spatial random effect to the (generalized) linear regression models or, equivalently, by assuming a spatially varying intercept that absorbs a spatial random effect. The effects of explanatory variables in such models are often assumed to be constant across the entire region. However, for data that are collected from a large region, relationships between response variables and covariates may exhibit complex spatially dynamic patterns that cannot be captured by constant regression coefficients. In particular, relationships among spatial variables may abruptly change across the boundaries of adjacent clusters but stay relatively homogeneous within clusters.

The need to detect such clusters arises in many biological, ecological, agricultural, environmental, and real estate applications, to name a few. Detecting these clusters allows straightforward interpretations of local associations between response variables and covariates. For example, climate/environmental conditions on the ground often exhibit abrupt changes across certain topological boundaries. Housing prices per square footage can differ substantially on opposite sides of a street. Neuroscientists are interested in finding clusters of human brain subregions that react to certain stimuli. Relationships among underground geophysical properties as well as atmospheric properties are also expected to change abruptly because both the underground and the atmosphere consist of complex and heterogeneous multiple layers.

Our specific motivating problem comes from an important scientific question in geoscience. Geophysical fluids (i.e., air and sea water) consist of distinct fluid masses (Talley 2011). Within each fluid mass, the physical and chemical properties are relatively homogeneous, but they change rapidly across the narrow boundaries (termed *fronts* in geoscience) between adjacent fluid masses. This phenomenon is a result of the nonlinear nature of geophysical fluid dynamics and is ubiquitous in the atmosphere and the ocean (Vallis 2006). The relationships between different characteristics of fluids are likely to change abruptly across fronts. One notable instance is the relationship between the temperature and the salinity of sea water (referred to hereafter as the T-S relationship). In oceanography, temperature and salinity are two important features of water masses that strongly affect ocean currents (Talley 2011). Knowledge of the spatial distribution of the T-S relationship in the ocean can provide important information about the evolution of individual water masses. Such information can be further used to monitor the

CONTACT Huiyan Sang huiyan@stat.tamu.edu Department of Statistics, Texas A&M University, College Station, TX.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

These materials were reviewed for reproducibility.

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

pathway and strength of the meridional overturning circulation (MOC), which plays a key role in the global climate system.

In the context of spatial statistics, various models have been developed to capture spatially varying regression coefficients. geographically weighted regression (GWR) (Fotheringham, Brunsdon, and Charlton 2003) and spatially varying coefficient (SVC) models (Gelfand et al. 2003) are two popular methods of this type. GWR extends the ordinary least-square regression by fitting a local regression model at each observation. Assuming a linear model with $\mathbf{y}$ denoting the response variable and $\mathbf{X}$ the design matrix, the regression coefficient at the ith location is estimated from $\widehat{\boldsymbol{\beta}}_i = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_i\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_i\mathbf{y}$, where $\mathbf{W}_i$ is a diagonal weight matrix defined by a kernel function of distance to point i. Such weighting schemes make the GWR method more efficient to deal with smoothly varying than clustered regression coefficients. In Gelfand et al. (2003), spatially varying regression coefficients are modeled as a multivariate spatial Gaussian process. It fits into the Bayesian paradigm with posterior inference attainable for all model parameters and thus provides a richer inferential framework. Such an advantage comes at the cost of a large computational burden due to its requirement of the Metropolis algorithm. Recent advances in dealing with large spatial datasets such as the predictive process (Banerjee et al. 2008) and the Gaussian Markov random field approximation (Lindgren, Rue, and Lindström 2011) could be utilized in the SVC models to alleviate its computational burden. Although the SVC models offer flexibility in modeling spatially varying coefficients through the choice of covariance/cross-covariance functions, to our knowledge, there are very limited existing covariance/cross-covariance functions appropriate to capture arbitrary clustered patterns.

We seek to develop a spatial modeling approach that is directly applicable for detecting spatially contiguous clusters in regression coefficients. One potential solution would be motivated by the studies on the homogeneity pursuit of regression coefficients in high-dimensional data analysis. In these studies (e.g., Tibshirani et al. 2005; Ke, Fan, and Wu 2015), pairwise coefficient differences are penalized to encourage homogeneity among coefficients. A key ingredient is the appropriate selection of sets of pairs on which to impose penalties. When a complete order of regression coefficients is available, such as in time series problems, we could construct fusion penalties (Tibshirani et al. 2005) on successive differences between coefficients in order to encourage similarity in adjacent coefficients. Then, the problem can be transformed to lasso regularization (Tibshirani 1996), an optimization problem for which efficient algorithms are available. However, this strategy is not applicable to spatial data, as they do not have a natural order. For cases where there is no prior information on the order of coefficients, several recent studies propose establishing a coefficient order based on preliminary coefficient estimates (Ke, Fan, and Wu 2015; Tang and Song 2016). Clearly, such methods require adequate sample replicates to obtain reliable preliminary coefficient estimates for ordering. But in spatial statistics, it is common to have spatial observations from only one snapshot or spatio-temporal observations with a strong dependence in time. Independent replicates are unavailable in both of these cases.

In this article, we propose a new spatial regression modeling approach, called spatially clustered coefficient (SCC) regression, to estimate regression coefficients when there are spatial patterns, especially clustered patterns, in the relationship between a response variable and explanatory variables. SCC imposes penalties on the difference between regression coefficients at any two locations connected in an edge set. To address the challenge of selecting an edge set that incorporates spatial information while maintaining computational efficiency, we propose using a minimum spanning tree (MST). An MST is a subgraph that connects all vertices of an undirected graph with no cycles and with minimum total edge weights. For a spatial problem, an MST can be constructed efficiently based on spatial locations with distances between locations serving as the edge weights. It compactly represents a spatial topology of observed points: two locations that are connected by an MST tend to be close in space. Therefore, the penalties on the edges of an MST encourage spatial homogeneity of the coefficients at proximate locations. Moreover, such a choice of the edge set facilitates computation: after a linear reparameterization, the estimation problem is reduced to the usual lasso-type optimization, which has a highly scalable algorithm suitable for large datasets.

SCC has several other advantages. It allows the investigation of different clustered patterns in different regression coefficients. The number of clusters estimated from SCC is completely data driven. Furthermore, although designed for clustered coefficients, our numerical studies show that the SCC model has strong local adaptivity and can also successfully capture a highly spatially variable pattern.

The rest of the article is organized as follows. Section 2 details the SCC model and discusses its theoretical properties. In Section 3, we present simulation studies to illustrate the performance of SCC. An application of the method is shown in Section 4 using the aforementioned temperature and salinity data in the Atlantic basin. Section 5 summarizes the major conclusions of this study followed by discussion. Related proof, more simulation results, and codes are provided in the supplementary materials.

2. Methodology

2.1. SCC Model

Suppose a set of spatial data $\{(\mathbf{x}(s_i), y(s_i)), i = 1, \ldots, n\}$ is observed at locations $s_1, \ldots, s_n \in \mathbb{R}^2$, where the response variable $y(s_i)$ is assumed to be spatially correlated and $\mathbf{x}(s_i) = (x_1(s_i), \ldots, x_p(s_i))^{\mathrm{T}}$ is the p-dimensional vector of explanatory variables for the observation located at s_i. Consider the standard linear regression, $y(s_i) = \sum_{k=1}^{p} x_k(s_i)\beta_k + \epsilon(s_i)$, where $\beta_k, k = 1, 2, \ldots, p$, are the regression coefficients and $\epsilon(s_i)$ are independently identically distributed random noises with mean 0 and variance σ^2. The intercept can be accommodated by including 1 as an entry of $\mathbf{x(s_i)}$. Without loss of generality, we assume that the explanatory variables are standardized to have mean 0 and unit variance. The extension of the linear regression model to allow spatially varying regression coefficients is straightforward,

$$y(s_i) = \sum_{k=1}^{p} x_k(s_i)\beta_k(s_i) + \epsilon(s_i). \tag{1}$$

In many spatial datasets, it is very common to observe only one or a limited number of replicates at each location, making the model (1) ill-posed if without any assumptions on $\beta_k(s_i)$. For example, with only one spatial realization $\{(\mathbf{x}(s_i), y(s_i)), i = 1, \dots, n\}$ at n observed locations, the regression model (1) can be written in the matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Here, we stack all the coefficients into a vector $\boldsymbol{\beta} = (\beta_1(s_1), \dots, \beta_1(s_n), \dots, \beta_p(s_1), \dots, \beta_p(s_n))^{\mathrm{T}}$, and the design matrix $\mathbf{X} = [\mathrm{diag}(\mathbf{x}_1), \dots, \mathrm{diag}(\mathbf{x}_p)]$ is an $n \times np$ matrix with $\mathbf{x}_k = (x_k(s_1), \dots, x_k(s_n))^{\mathrm{T}}$. Clearly, this regression problem needs to be regularized since there are more variables than observations. For spatial problems, it is expected that association between a response variable and explanatory variables at nearby locations is highly homogeneous. This motivates us to assign a regularization function for $\boldsymbol{\beta}$ reflecting such spatial homogeneity patterns.

Specifically, we propose to estimate $\boldsymbol{\beta}$ by minimizing the following objective function

$$\frac{1}{n}\sum_{i=1}^{n}\{y(s_i) - \sum_{k=1}^{p} x_k(s_i)\beta_k(s_i)\}^2 + \sum_{k=1}^{p}\sum_{(i,j)\in\mathbb{E}} P_\lambda(\beta_k(s_i) - \beta_k(s_j)). \qquad (2)$$

Here, $\mathbb{E}$ is the edge set of a graph consisting of n vertices, where each vertex corresponds to one observed spatial location. The term P_λ is a penalty function to encourage homogeneity between two regression coefficients if their corresponding locations s_i and s_j are connected by an edge in $\mathbb{E}$. λ is a tuning parameter determining the strength of penalization. The selections of the penalty function P_λ, the edge set $\mathbb{E}$ and the tuning parameter λ are the three ingredients of the model (2). Below, we discuss strategies to choose them.

2.1.1. Selection of the Penalty Function P_λ

There are various forms of penalty functions encouraging sparsity in the literature of variable selection. The simplest and perhaps the most widely adopted one is the lasso (Tibshirani 1996) that employs L_1-penalty of the form

$$P_\lambda(t) = \lambda|t|. \qquad (3)$$

As the penalty in (3) is a convex function, efficient convex optimization algorithms can be readily applied. In this case, the L_1 penalty $P_\lambda(\beta_k(s_i) - \beta_k(s_j))$ enforces sparsity of the difference in two edge-connected coefficients. This allows the estimation of regression coefficients with a spatially piecewise constant (i.e., clustered pattern) if edge sets are selected appropriately to incorporate spatial information. The nonzero elements of $|\beta_k(s_i) - \beta_k(s_j)|$ correspond to boundary points, whereas any two edge-connected coefficients with zero difference belong to the same cluster. Naturally, spatial clusters for each explanatory variable can be automatically detected. However, as lasso assigns large penalties to large values of t, it tends to underestimate t, in our case, the difference in two regression coefficients, when its true value is large. To remedy this flaw, various penalty functions have been proposed, including adaptive lasso (Zou 2006), smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), minimax concave penalty (MCP) (Zhang 2010), and reciprocal L_1-regularization (rlasso) (Song and Liang 2015). Adaptive lasso assigns larger weights to the terms with small values in the L_1 penalty. SCAD and MCP adopt some concave functions that converge to constants as the penalized term becomes large. The rlasso uses a class of penalty functions that are decreasing in $(0, \infty)$ with a discontinuity at 0 and converging to infinity when the penalized term approaches zero. These forms have smaller estimation errors compared to lasso, which, however, is at the expense of considerable increase in computational cost. In practice, penalty functions are often selected by weighing a trade-off between statistical efficiency and computational complexity for specific problems.

It should be noted that the reduction of computational burden is critically important in spatial analysis as large datasets have become common in diverse fields such as geoscience, ecology, and econometrics. In this study, we mainly focus on the lasso penalty to demonstrate the power of SCC for its computational simplicity. We remark that SCC can adopt other forms of penalty functions which may further improve its performance.

2.1.2. Edge Selections Based on Minimum Spanning Tree

The edge set $\mathbb{E}$ is the key ingredient in the SCC model since it reflects the prior assumption about the structure of regression coefficients. However, as mentioned in Section 1, unlike temporal data, spatial data do not have a natural order, which makes it challenging to construct the set $\mathbb{E}$.

We note that in many spatial problems, regression coefficients at proximate locations are likely to be similar due to their homogeneous properties within a certain subregion. It is therefore desirable to construct $\mathbb{E}$ such that only proximate coordinate pairs are included in order to reflect spatial homogeneity among coefficients.

A common choice of an edge set $\mathbb{E}$ that satisfies this criterion is the set consisting of neighboring coordinate pairs, that is, $\mathbb{E} = \{(s_i, s_j) : i = 1, \dots, n; s_j \in \mathbb{N}_{s_i}\}$, where $\mathbb{N}_{s_i}$ is the set of neighbors of s_i. The neighboring set $\mathbb{N}_{s_i}$ can be defined in many different ways, for example, the k nearest neighbors of s_i or neighbors within a certain radius. Although such selections of $\mathbb{E}$ seem natural, they suffer from two evident deficiencies. First, $\mathbb{E}$ defined as above does not necessarily connect all the points in irregular spatial locations, resulting in isolated points. In such a case, (2) will not reduce to a constant regression coefficient model when $\lambda \to \infty$. Second, and more problematically, the penalties on the pairwise differences based on this choice of $\mathbb{E}$ include many redundant terms, and the computation is very challenging when solving the optimization problem for a graph with large numbers of nodes and edges (see Tang and Song 2016). Specifically, given an edge set $\mathbb{E}$, the corresponding model can be formulated as a generalized lasso problem as follows:

$$\frac{1}{n}\sum_{i=1}^{n}\{y(s_i) - \sum_{k=1}^{p} x_k(s_i)\beta_k(s_i)\}^2 + \lambda\sum_{k=1}^{p}\|\mathbf{H}\boldsymbol{\beta}_k\|_1, \qquad (4)$$

where $\mathbf{H}$ is an $m \times n$ matrix constructed from the edge set $\mathbb{E}$ with m edges. For an edge connecting two locations s_i and s_j, we represent the penalty term $|\beta_k(s_i) - \beta_k(s_j)|$ as $|\mathbf{H}_m\boldsymbol{\beta}_k|$, where $\mathbf{H}_m$ is a row vector of $\mathbf{H}$ and contains only two nonzero elements, 1 at the ith index and -1 at the jth.

For most graphs constructed by k-nearest neighbors or neighbors within a certain radius such that all points are connected, the number of edges is greater than the number of nodes. Path-following algorithms (Shen and Huang 2010; Arnold and Tibshirani 2016) and the alternating direction method of multipliers (ADMM) (Boyd 2011; Zhu 2017) have been developed to solve the generalized lasso problem for this case. Nevertheless, for an arbitrary $\mathbf{H}$ with $m > n$ for large m and n, these algorithms are computationally very costly. Indeed, the computational complexities of these algorithms are typically at least $O(n^3)$ for the problem we are considering here, which is far slower than solving a lasso problem.

The previous discussion suggests that for large-scale problems, an appropriate choice for $\mathbb{E}$ that balances model accuracy and computational efficiency should only include coordinate pairs close to each other, should lead to connectivity of all data points, and should have no redundant pairs. One choice of $\mathbb{E}$ that satisfies all three of these criteria is the edge set of an MST. Given an undirected graph $G = (\mathbb{V}, \mathbb{E}_0)$ with a weight function $d(e)$ that assigns a weight to each edge e in an edge set $\mathbb{E}_0$, an MST is defined as the subgraph $T = (\mathbb{V}, \mathbb{E}), \mathbb{E} \subseteq \mathbb{E}_0$ that connects all vertices without any cycles and minimizes $\sum_{e \in \mathbb{E}} d(e)$. It is known that an MST has $|\mathbb{V}|$ vertices and $|\mathbb{V}| - 1$ edges.

MSTs were originally motivated by applications in the optimal design of networks, such as computer networks, telecommunications networks, and transportation networks. Another important application of MSTs is clustering (Grygorash, Zhou, and Jorgensen 2006), where the MST for a given point set and distance measure is first constructed, and then some edges are removed from the MST according to certain criteria to form clusters of points. In particular, MSTs have a close connection with the single-linkage agglomerative clustering algorithm. The clusters obtained from removing the edges in the MST whose weights are greater than a given threshold are the same as those obtained from applying the same threshold as a cut-off distance for a single-linkage dendrogram. Indeed, this is the result of a unique property of MSTs, referred to as the *cut property*. It is known that for any cut of a given connected graph G, the minimum-weight edge that crosses the cut is in the MST for G.

Based on this property, efficient algorithms for constructing MSTs have been developed (Section 2.2 provides the computation details), with a computational complexity close to $O(n\log n)$, where n is the number of vertices in a 2D Euclidean space. In addition to their computational advantages, MST-based clustering algorithms are also known to be effective in detecting clusters with irregular boundaries, due to the fact that they do not rely on the assumption of a spherically shaped cluster structure for the underlying data.

In our application, taking a given set of n spatial locations as vertices and the Euclidean distances between locations as the edge weights, we can construct an MST, $T = (\mathbb{V}, \mathbb{E})$, consisting of $n - 1$ edges with all n locations connected. By the construction, T provides a connected, acyclic graph that compactly represents the spatial topology of the n observed points, and is thus an appropriate choice satisfying the aforementioned desired properties for edge set selections. In this case, $\mathbf{H}$ is a full row rank matrix encoding the $n - 1$ penalty terms on the coefficients $\boldsymbol{\beta}_k$.

An illustration of an MST based on 1000 spatial locations is provided in Figure 1(a). By penalizing the coefficients at two locations connected in the edge set of the MST, the estimated coefficients form different clusters (see Figure 1(b)). Within each cluster, the vertices (spatial locations) remain connected and the corresponding coefficients take the same value, while across different clusters, the connection is cut and the coefficients take different values.

Finally, we remark that for data observed on a regular lattice, all spanning trees are also MSTs, as all edges have the same weight. To ensure that edges are selected such that the MST properly represents the spatial topology, we can consider a so-called *random* MST (Braunstein et al. 2007) that is formed by assigning random weights to the edges of a lattice graph using a uniform distribution from 0 to 1. As the weights are drawn from a continuous distribution, the probability of having a tie between two edge weights becomes zero, and the resulting MST is unique (Dobrin and Duxbury 2001). See Figure 1(c) for an example of the MST on a regular lattice.

2.1.3. Selection of Tuning Parameter λ

The tuning parameter λ controls the number of clusters in regression coefficients in our context. When $\lambda \rightarrow \infty$, the model (2) yields a constant regression coefficient; when $\lambda = 0$, it reduces to the ordinary least square with all different coefficients across the region. With an appropriate choice of λ, the penalized least-square model (2) produces

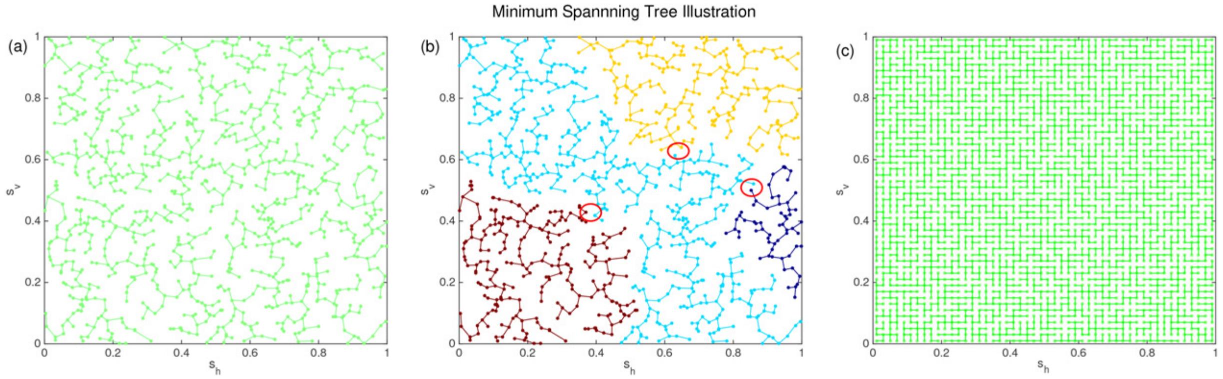

Figure 1. A schematic diagram for the MST constructed for (a) 1000 spatial irregular locations and (c) 50 × 50 regular grids. The points of different colors in (b) belong to different clusters with the cut edges marked by red circles.

clustered regression coefficients. In practice, the optimal λ can be determined via some data-dependent model selection criteria, such as generalized cross-validation (Golub, Heath, and Wahba 1979), Akaike information criterion (AIC), Bayesian information criterion (BIC) (Schwarz et al. 1978), and extended Bayesian information criterion (EBIC) (Chen and Chen 2008, 2012).

2.2. Computation

The model fitting of SCC involves two major computation steps: construction of the MST based on the spatial locations of the dataset and regularized optimization in (2) for the given MST.

In the first step, we use Prim's algorithm developed from the cut property of the MST. The computational complexity of this algorithm is $O(m + n\log n)$ on a graph with n vertices and m edges ($m = n(n-1)/2$ for a complete graph). For Euclidean MSTs, that is, MSTs using Euclidean distance as the edge weight function, computation can be further reduced by using Delaunay triangulation (March, Ram, and Gray 2010). Given n points in two-dimensional space, a Delaunay triangulation forms a connected graph. Prim's algorithm is then applied to the Delaunay triangulation to find the corresponding MST. Note that the construction of a Delaunay triangulation requires $O(n\log n)$ time and $O(n)$ storage, and the number of edges in a Delaunay triangulation is $O(n)$. The final algorithm that combines Prim's algorithm and the Delaunay triangulation method takes $O(n\log n)$ time and $O(n)$ space. We remark that for data on a sphere, we can use a so-called geodesic MST, in which the great circle distance is used as the distance metric, that is, the edge weight function, between any two locations (nodes). Dolan, Weiss, and Smith (1991) have developed algorithms for computing the Delaunay triangulation and geodesic MST for a graph on a sphere, with the computational complexity remaining at $O(n\log n)$.

Once the MST is constructed, the resulting penalties no longer contain redundant terms and hence can be easily transformed into a lasso, or lasso-type problem after suitable reparameterization. Define new parameters $\boldsymbol{\theta}_k, k = 1, \ldots, p$ as

$$\boldsymbol{\theta}_k = \begin{pmatrix} \mathbf{H} \\ \frac{1}{n}\mathbf{1}^{\mathrm{T}} \end{pmatrix} \boldsymbol{\beta}_k = \widetilde{\mathbf{H}}\boldsymbol{\beta}_k.$$

The new design matrix can be written as $\widetilde{\mathbf{X}} = [\mathrm{diag}(\mathbf{x}_1)\widetilde{\mathbf{H}}^{-1}, \ldots, \mathrm{diag}(\mathbf{x}_p)\widetilde{\mathbf{H}}^{-1}]$. Note that $\widetilde{\mathbf{H}}$ is an $n \times n$ invertible matrix since $\mathbf{H}$ has full row rank, and thus there is a one-to-one transformation between $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$. Then, the SCC model in (4) can be rewritten as

$$\frac{1}{n}\|\mathbf{y} - \widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{\ell \in B} |\theta_\ell|, \tag{5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_p^{\mathrm{T}})^{\mathrm{T}}$, a vector of size np, and B represents the index set $B = \{\ell : \mathrm{mod}(\ell, n) \neq 0, \text{for } \ell = 1, \cdots, np\}$ excluding every nth element. Henceforth, we will denote $\sum_{\ell \in B} |\theta_\ell|$ as $\|\boldsymbol{\theta}_B\|_1$ for neatness.

Therefore, the solution to the SCC model (2) with a lasso penalty can be obtained by solving the lasso problem in (5) with respect to the parameters $\boldsymbol{\theta}$. Estimators for $\boldsymbol{\beta}$ are then given by $\widehat{\boldsymbol{\beta}}_k = \widetilde{\mathbf{H}}^{-1}\widehat{\boldsymbol{\theta}}_k$, for $k = 1, \ldots, p$. Many efficient lasso solvers, such as the LARS (Efron et al. 2004), coordinate decent (Friedman, Hastie, and Tibshirani 2010), and stochastic Frank-Wolfe (Frandi et al. 2016) algorithms, can be readily applied to the SCC model for large datasets. For example, using the stochastic Frank-Wolfe algorithm, the lasso optimization with a size of up to one million penalties can be handled within a minute.

2.3. Theoretical Properties

In this subsection, we consider theoretical results concerning the behavior of SCC. We provide error bounds for its estimation and prediction, as well as its performance in detecting spatially clustered patterns of unknown regression coefficients. As there is a one-to-one transformation between $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$, we present theorem in terms of $\boldsymbol{\theta}_k$.

Assumptions 1. (a) There is a positive constant C_1 so that $n^{-1}\sum_{i=1}^{n} \widetilde{X}_{i,\ell}^2 \leq C_1$ for any $n > 0$ and $\ell \in \{1, \ldots, np\}$.

(b) There is a positive constant Φ so that for any vector $\mathbf{u} \in \mathbb{R}^{np}$ satisfying $\|\mathbf{u}_{A^c}\|_1 \leq 3\|\mathbf{u}_A\|_1$ where $A = \{\ell : \theta_\ell \neq 0, \ell \in B\} \cup B^c$ and $|A|$ denotes its cardinality, we have

$$\frac{1}{n}\mathbf{u}^{\mathrm{T}}(\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}})\mathbf{u} \geq \Phi||\mathbf{u}||_2^2. \tag{6}$$

Assumption 1(a) means the random variables $V_\ell = n^{-1}\sum_{i=1}^{n} \widetilde{X}_{i,\ell}\varepsilon_i$ to be sub-Gaussian for any $\ell \in \{1, \ldots, np\}$. Assumption 1(b) is known as the restricted eigenvalue condition (Tibshirani, Wainwright, and Hastie 2015) .

Theorem 1. Suppose that Assumption 1 holds. If $\lambda_n\sqrt{n/\log(n)} \geq 4\sqrt{(1 + C_2)2C_1^2\sigma^2}$ where C_2 is a positive constant for any $n > 0$, the following inequalities hold with probability tending to unity as $n \to \infty$

$$\frac{1}{n}||\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}||_2^2 \leq \frac{9\lambda_n^2|A|}{4\Phi}, \tag{7}$$

$$\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \leq \frac{3\lambda_n\sqrt{|A|}}{2\Phi}. \tag{8}$$

The detailed proof for (7) and (8) is provided in the supplementary materials. Assuming that $|A|$ grows with a rate as $o(n/\log n)$, the right-hand sides of (7) and (8) decrease asymptotically to zero as $n \to \infty$. We note that such an assumption for $|A|$ is generally satisfied under infilling domain asymptotics in which case the number of clusters is often fixed.

Finally, we demonstrate in Corollary 1 how the estimation error bound (8) can be used to guide the detection of clusters in practice. Define $\theta_M = \min(|\theta_\ell|, \ell \in S)$, where $S = \{\ell : \theta_\ell \neq 0, \ell \in B\}$. We have the following corollary based on Theorem 1:

Corollary 1. Assuming that $|A|$ grows with a rate as $o(n/\log n)$ and the conditions in Theorem 1 hold, the following statement holds with probability tending to unity as $n \to \infty$:

$$|\widehat{\theta}_\ell| < \delta \iff \ell \notin S$$

for any $\ell \in B$ and any $0 < \delta \leq \theta_M$.

In many applications, the possible range of θ_M could be known a priori, that is, $\theta_M^l \leq \theta_M \leq \theta_M^u$. According to Corollary 1, $\delta = \theta_M^l$ can be used as a threshold for $\widehat{\theta}_\ell$ to detect clusters.

3. Simulation Studies

In this section, we present two simulation studies to illustrate the robust performance of the SCC method under two different scenarios. The true regression coefficients in Study 1 are designed to have clustered patterns. In practice, we may not know whether the true regression coefficients are clustered or smoothly varying. To examine the behavior of SCC in capturing spatially highly variable patterns even under unfavorable scenarios, we design Study 2 in which the true regression coefficients are generated from a spatial Gaussian process.

In both studies, we randomly generate 1000 spatial locations in the square domain $[0, 1] \times [0, 1]$. Then, the response at each location is generated according to

$$y(s_i) = \beta_1(s_i)x_1(s_i) + \beta_2(s_i)x_2(s_i) + \beta_3(s_i) + \epsilon(s_i), \qquad (9)$$

where $\epsilon(s_i) \overset{\text{iid}}{\sim} N(0, \sigma^2)$. We set σ to be 0.1.

Numerical data analyses in previous work often generated values of predictors from a white-noise process (e.g., Finley 2011; Wheeler and Calder 2007). However, in geoscience studies, many variables serving as predictors of a regression model have evident spatial structures (Talley 2011). Therefore, we generate covariates from spatial processes to mimic real situations in the numerical studies. Let $\{z_1(s_i)\}$ and $\{z_2(s_i)\}$ denote the two independent realizations of a spatial Gaussian process with mean zero and a covariance matrix defined from an isotropic exponential function: $\mathrm{Cov}\{z_k(s_i), z_k(s_j)\} = \exp(-\|s_i - s_j\|/\phi), k = 1, 2$, where ϕ is the range parameter. We generate two covariates $x_1(s)$ and $x_2(s)$ by linearly transforming $z_1(s)$ and $z_2(s)$. Specifically, we set $x_1(s_i) = z_1(s_i)$ and $x_2(s_i) = rz_1(s_i) + \sqrt{1 - r^2}z_2(s_i)$, allowing dependence between the two spatially varying predictors. In the following analysis, we set $r = 0.75$, corresponding to moderate collinearity, and we consider three range parameters $\phi = 0.1, 0.3, 1$ corresponding to weak, moderate, and strong spatial correlations. For each value of ϕ, we run 100 simulations to examine the behavior of SCC in parameter estimation.

As indicated in Section 1, GWR and SVC are two popular existing spatially varying coefficient models. Previous studies suggest that SVC typically produces results that are comparable to GWR (Finley 2011) since it can be viewed as a model-based version of GWR (LeSage 2004). Therefore, we compare the results for SCC with those of GWR. To quantify the performance of each method in estimation, we consider the mean-squared error of estimation (MSE_β), defined as

$$\mathrm{MSE}_\beta = \frac{1}{np}\sum_{i=1}^{n}\sum_{k=1}^{p}(\beta_k(s_i) - \widehat{\beta}_k(s_i))^2.$$

For SCC, the tuning parameter λ is chosen using BIC. Our numerical experiments (see the supplementary materials) indicate that the performances of AIC and BIC are comparable in terms of MSE_β while EBIC performs much worse. Indeed, previous studies (Foygel and Drton 2010; Song and Liang 2015) suggest that EBIC tends to produce oversparsity in the penalized term (the difference of regression coefficients in our model). For the GWR method, we employ an exponential kernel function with the optimal bandwidth chosen by the cross-validation method.

We use the package *glmnet* to solve the lasso optimization and the package *gwr* to implement GWR (both packages are available in R and Matlab), and we use the Matlab function *graphminspantree* and the R function *mst* in the *igraph* package to find the MST. We use the R function *admm.genlasso* in the *penreg* package to solve the generalized lasso problem for comparisons. The computations were performed on a Mac Pro with a 3.0GHz eight-core processor and 64GB of memory.

3.1. Study 1: Clustered Coefficients

The true regression coefficients in this study are set to be spatially clustered. As shown in the top panel of Figure 2, each of the coefficients $\beta_1(s_i)$, $\beta_2(s_i)$, and $\beta_3(s_i)$ is assigned a distinct cluster pattern to reveal the ability of SCC in detecting various patterns.

The first part of the simulation study compares the performance of the GWR and SCC methods using an MST as the edge set. The coefficients estimated from the GWR and SCC methods in one simulation are plotted in Figure 2(d)–(i). It can be seen that the spatial patterns of the coefficients derived by SCC, as shown in the bottom panel of Figure 2, are highly consistent with the true regression coefficients shown in the top panel. SCC successfully captures the cluster structure in the regression coefficients and detects the abrupt changes across the boundaries of adjacent clusters. In contrast, the coefficients estimated from GWR do not exhibit a clear cluster structure. Specifically, GWR produces poor estimations of regression coefficients both within clusters and near their boundaries.

We further examine the performance of SCC in terms of parameter estimation. Table 1 compares MSE_β for GWR and SCC under three different settings of spatial correlation for the covariates. For coefficient estimation, SCC clearly outperforms GWR, with considerably smaller values of MSE_β in all three settings. As the spatial correlation in the covariates becomes stronger, the performance of GWR degrades substantially, whereas the SCC estimates are relatively more stable. For instance, the mean MSE_β for the GWR estimates over 100 simulations is 0.36 when spatial correlation among the covariates is weak (ϕ=0.1) but increases to 3.73 in the case of strong spatial correlation (ϕ=1). In contrast, the mean MSE_β for the SCC estimates changes by less than a factor of three. Therefore, SCC provides more robust inferences for the regression coefficients, especially in the presence of strong spatial correlation in the covariates.

We end the first part of Study 1 by examining the performance of SCC in recovering clusters of coefficients. Table 2 lists the individual Rand index for the SCC estimates, $\widehat{\beta}_1(s_i)$, $\widehat{\beta}_2(s_i)$, and $\widehat{\beta}_3(s_i)$ averaged over 100 simulations. The values of the Rand index range from 0.72 to 0.85. In particular, the Rand index

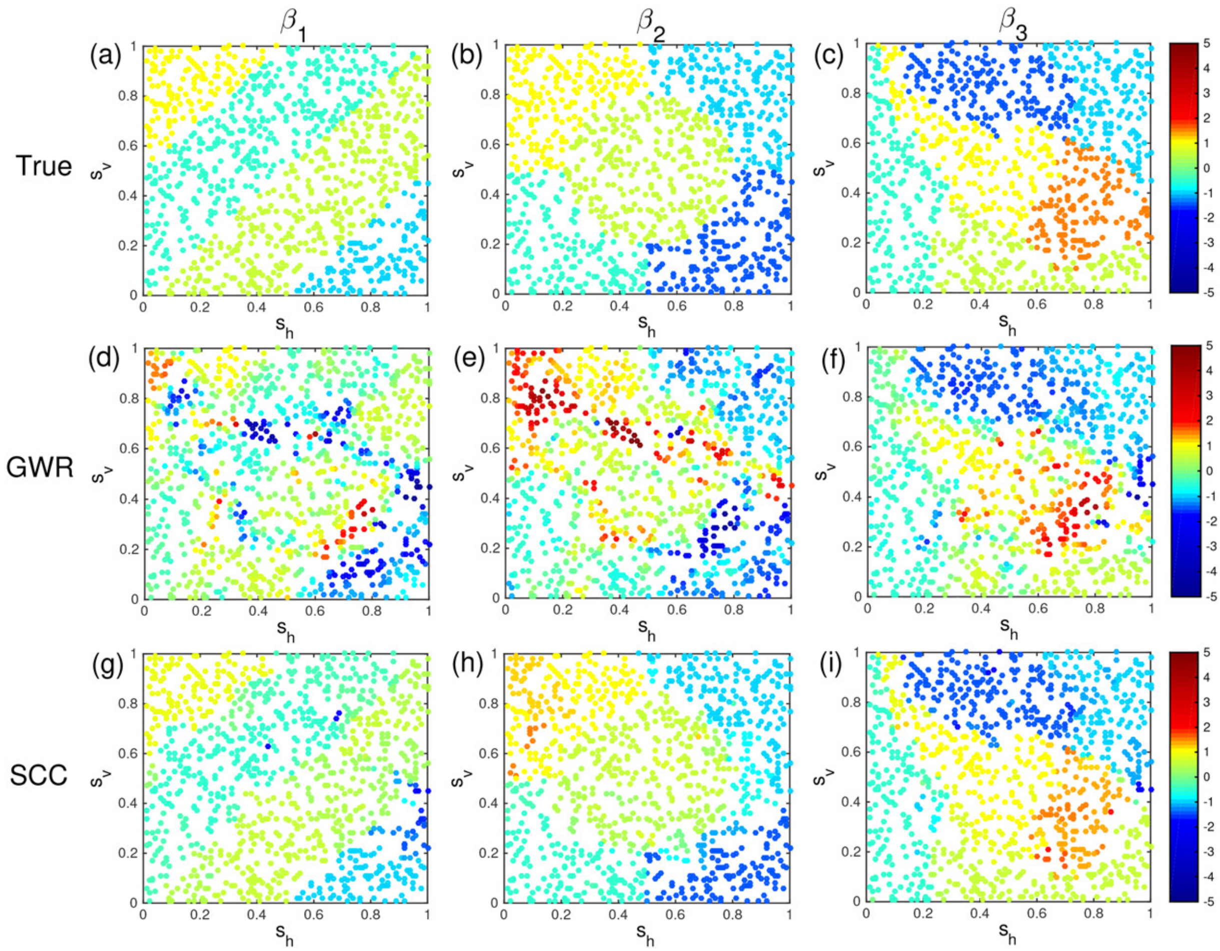

Figure 2. Study 1: spatial structures of (a)–(c) true coefficients β_1, β_2, and β_3; the estimated coefficient surfaces from (d)–(f) GWR; and (g)–(i) SCC in one simulation with the spatial range parameter $\phi = 0.3$ for predictors.

Table 1. Summary of Study 1 (clustered coefficients): the mean MSE$_\beta$ for the SCC and GWR methods over 100 simulations, under various spatial correlations for predictors.

Spatial correlation	MSE$_\beta$	
	GWR	SCC
Weak	0.36	0.09
Moderate	1.07	0.12
Strong	3.73	0.22

Table 2. Rand index for the SCC method in Study 1.

Rand Index	β_1	β_2	β_3
Weak	0.72	0.82	0.85
Moderate	0.72	0.82	0.85
Strong	0.73	0.83	0.85

does not degenerate as the spatial correlation among predictors becomes stronger, suggesting that SCC is robust for detecting clusters in regression coefficients.

For the second part of the study, our goal is to examine the performance of SCC using different choices of edge sets, including the MST (denoted as "SCC-MST"), the radius-based nearest-neighbor graph (denoted as "SCC-RNN")—using 0.05 as the radius since it is the smallest threshold that guarantees that each node has at least one neighbor, and the four-nearest-neighbor graph (denoted as "SCC-KNN"). Figure 3 shows the computation time and the boxplots of MSE$_\beta$ for each method for the case where the spatial correlation among the covariates is weak. Looking first at computation time, SCC-MST took an average of 1.4 sec to compute solutions for 200 values of λ. In comparison, GWR took an average of 7.5 sec, SCC-RNN an average of 972.0 sec, and SCC-KNN an average of 961.9 sec to compute 200 solutions for the same dataset.

All of the SCC-based methods achieve much smaller MSE$_\beta$ values than GWR, with SCC-MST outperforming GWR in terms of both computation time and parameter estimation accuracy. Among the SCC methods using different edge sets, there is clearly a trade-off between statistical efficiency and computational efficiency. Both SCC-RNN and SCC-KNN produced a more accurate parameter estimation than SCC-MST, with about a 50% reduction in MSE$_\beta$, which is not surprising since RNN and KNN graphs are denser than MSTs and hence, allow more edges to be checked to detect changes among nodes. However, as mentioned above, the improvement in statistical efficiency from using denser graphs comes with a substantially greater computational burden; for example, the computation time for SCC-KNN with four neighbors was nearly 700 times greater than that of SCC-MST. These results suggest that for a small-to-medium size dataset—say, when the number

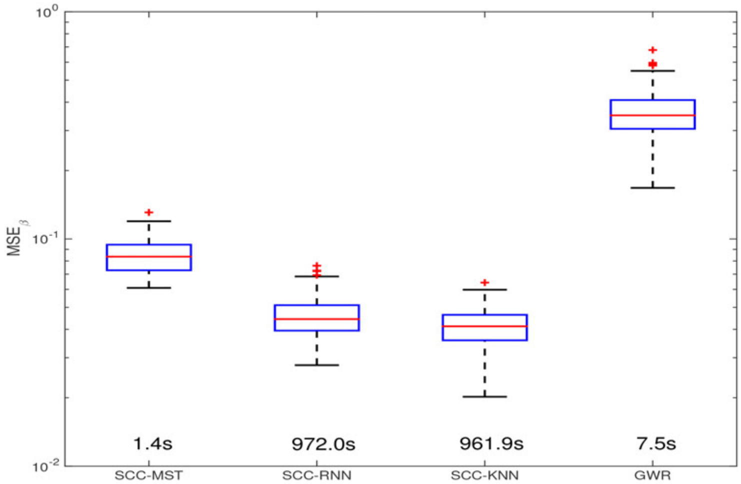

Figure 3. Study 1: the boxplots of MSE_β for SCC-MST, SCC-RNN, SCC-KNN, and GWR based on 100 simulated datasets. Computation times are reported above each method's label.

of the nodes is less than a few thousand—a denser graph such as KNN or RNN can be used to implement the SCC method to achieve more accurate parameter estimation. But for a larger dataset size, such as the real data example with 10,000 locations shown in Section 4, SCC-MST has a strong computational advantage, as it becomes computationally too expensive to implement SCC-RNN or SCC-KNN.

3.2. *Study 2: Smoothly Varying Coefficients*

The design of this study is similar to Study 1, except that the regression coefficients are independently generated from a Gaussian spatial process. Here, all the coefficient processes have a zero mean and an anisotropic exponential covariance function:

$$\mathrm{cov}(\beta_k(s_i), \beta_k(s_j)) \\ = \sigma_\beta^2 \exp\left(- \sqrt{\frac{(s_{h,i} - s_{h,j})^2}{\phi_{h,k}^2} + \frac{(s_{v,i} - s_{v,j})^2}{\phi_{v,k}^2}} \right), k = 1, 2, 3,$$

where $(\phi_{h,k}, \phi_{v,k})$ is the anisotropic range parameter and σ_β^2 is the variance parameter. σ_β^2 is fixed at 4, and (ϕ_h, ϕ_v) is set to be $(3, 1)$ for $\beta_1(s_i)$, $(1, 3)$ for $\beta_2(s_i)$, and $(2, 2)$ for $\beta_3(s_i)$.

The estimated coefficients from GWR and SCC in one simulation are displayed in Figure 4. The spatial pattern of coefficients derived by SCC agrees reasonably well with that of the true model. The estimates from GWR are, however, quite noisy, with artificially large coefficient values in some parts of the domain. This is partly because the isotropic kernel function used for GWR is too restrictive when fitting an anisotropic spatial field. In contrast, an advantage of SCC is its strong local adaptivity. The use of a local pairwise penalty function allows the fitting of a spatial field that is homogeneous in one direction but highly varying in the other. Comparisons of MSE_β for the two methods further confirm the superiority of SCC in estimating smoothly varying coefficients. For instance, in the presence of strong spatial correlation in the predictors, MSE_β for SCC is only 1/6 of that for GWR (Table 3).

3.3. *Summary of Simulation Results*

In both cases, SCC is capable of capturing the spatial pattern in coefficients, and it outperforms GWR with a considerably smaller estimation error. It should be noted that Study 1 and Study 2 can be treated as two distinct scenarios of spatial patterns of coefficients. The former is consistent with the assumption underpinning the SCC method, while the latter favors the setting of the GWR method. Simulations for a hybrid scenario with $\beta_1(s_i)$ and $\beta_2(s_i)$ clustered but $\beta_3(s_i)$ smoothly varying are included in the supplementary materials, which also reveals the superiority of SCC over GWR. The results of simulation studies show that even under misspecified models, SCC is capable of producing reasonable estimates, illustrating its robustness and strong adaptivity.

4. Real Data Analysis

4.1. *Dataset*

In this section, we use the SCC method for the detection of the Antarctic intermediate waterway (AAIW) (Talley 2011) by investigating the temperature–salinity (T-S) relationship in the Atlantic Ocean, along with a comparative analysis using the GWR method. The AAIW, formed and subducted at the high midlatitudes of the Southern Ocean, is moved northward by the upper limb of the Atlantic meridional overturning circulation (AMOC), which has a major influence on Earth's climate system and acts as an effective sink of anthropogenic CO_2 (Sabine et al. 2004). Knowledge of the extent of the AAIW can help in inferring the strength of the AMOC (Came et al. 2008; Xie, Marcantonio, and Schmidt 2012; Oppo and Curry 2012). The AAIW is characterized by a negative T-S relationship, which is different from other water masses, so that the AAIW can be identified from the T-S relationship.

We obtained temperature and salinity records from the World Atlas 2013, version 2 (WOA 13 V2) (Locarnini et al. 2013; Zweng et al. 2013), which is archived at the National Oceanographic Data Center (*https://www.nodc.noaa.gov/OC5/woa13/*). To facilitate analysis, we take a meridional segment

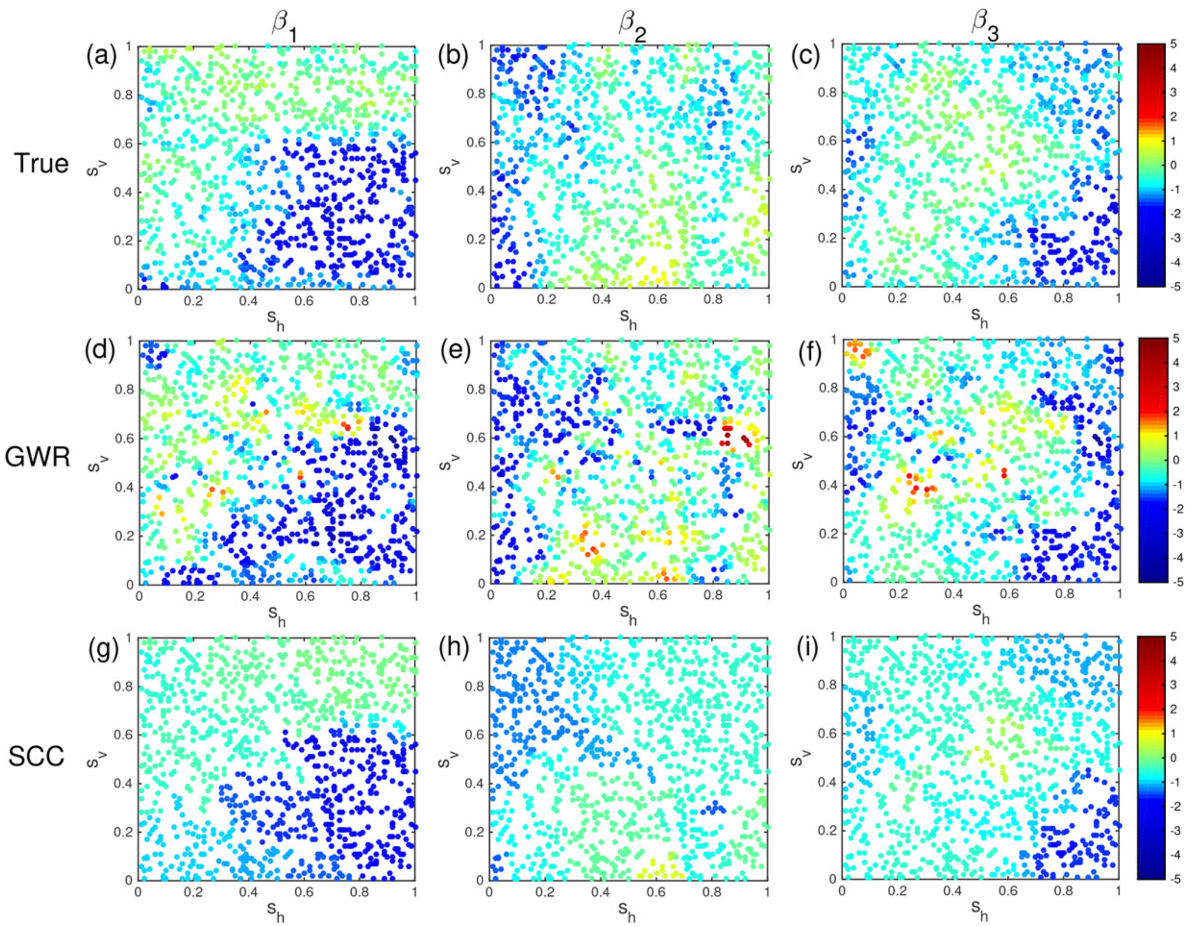

Figure 4. Study 2: spatial structures of (a)–(c) true coefficient β_1, β_2, and β_3; the estimated coefficient surfaces from (d)–(f) GWR; and (g)–(i) SCC in one simulation with the spatial range parameter $\phi = 0.3$ for predictors. Note that the color scale is nonlinear to accommodate outliers in the GWR estimation.

Table 3. Summary of Study 2 (smoothly varying coefficients): the mean MSE$_\beta$ for the SCC and GWR methods over 100 simulations, under various spatial correlations for predictors.

Spatial correlation	MSE$_\beta$	
	GWR	SCC
Weak	0.18	0.17
Moderate	0.58	0.22
Strong	2.04	0.34

of temperature and salinity in the Atlantic basin along 25°W between 60°S and 60°N (Figure 5), a standard segment widely used in oceanographic studies because it is highly representative of the spatial variations of oceanic variables. We use measurements of temperature and salinity at 10,000 locations in total, with the density of points decreasing vertically due to the fact that oceanic variables typically change much more rapidly in the upper ocean than in the abyss. Figure 5(a,b) displays the spatial distributions of temperature and salinity, respectively, along the 25°W segment.

This dataset has three notable features. First, the temperature and salinity are not randomly distributed but have well-organized spatial structures. Specifically, the temperature is generally higher at lower latitudes and in the upper ocean as a result of solar radiation. The spatial distribution of salinity is somewhat more complicated. Near the sea surface, the salinity values peak around 30°S and 30°N due to the low precipitation rates at these latitudes. In addition, there is a pronounced low-salinity tongue originating from the sea surface around 50°S to 60°S and extending northward and downward. This low-salinity tongue corresponds to the AAIW. The AAIW's encounter with a high-salinity water mass centered at 30°S leads to a strong salinity front.

The second notable feature is that the temperature and salinity are highly anisotropic. The temperature and salinity gradients in the vertical direction are several orders of magnitude larger than those in the horizontal direction. The anisotropy is essentially a result of the ocean's geometry. It has a width of around 20,000 km but a thickness of about 4 km. To account for the geometry of the ocean, oceanic studies typically adopt the nondimensional coordinates $(s_h, s_v) = (s_h^0/L, s_v^0/H)$, where s_h (s_v) is the nondimensional horizontal (vertical) coordinate transformed from the original coordinate s_h^0 (s_v^0), and L (H) is the horizontal (vertical) length of the ocean (Vallis 2006). In the nondimensional coordinates, the horizontal and vertical gradients are at the same order of magnitude, largely eliminating the anisotropy. In this study, we follow the convention of oceanic studies and adopt this scaling technique.

The third notable feature is that the distributions of temperature and salinity appear to be nonstationary. As mentioned above, there is a strong salinity gradient around the front formed by the AAIW and the high-salinity water mass centered at 30°S.

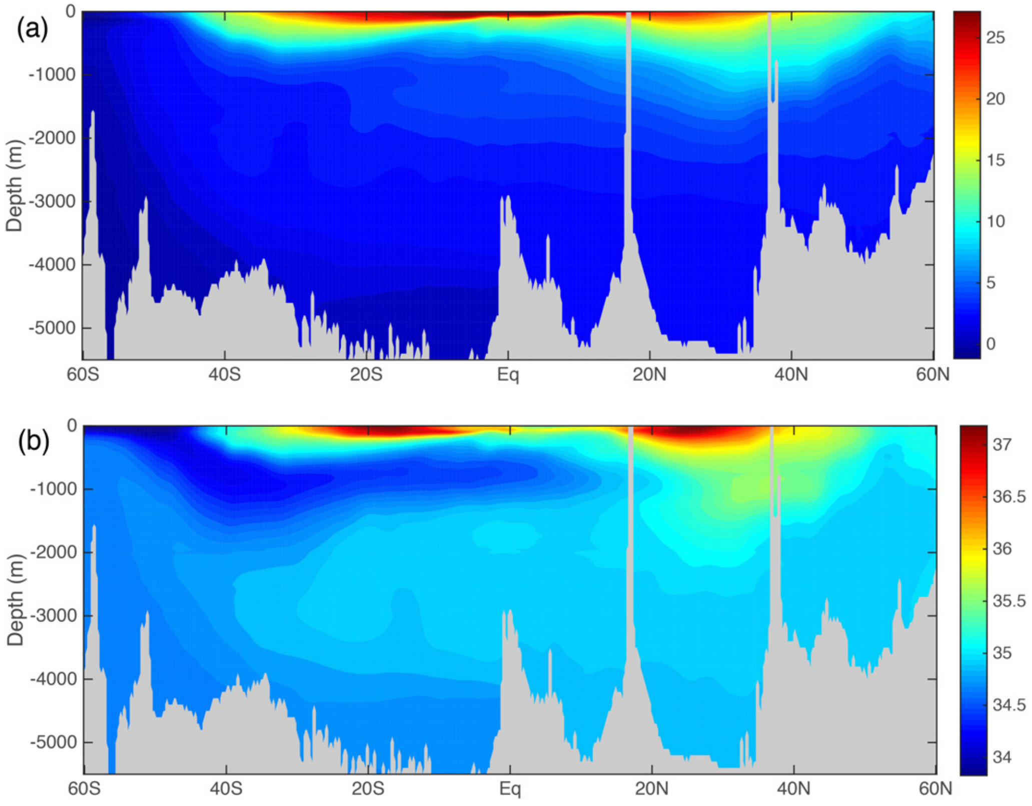

Figure 5. Spatial distributions of (a) temperature in °C and (b) salinity in PSU along the meridional segment 25°W.

In addition, there is also a strong temperature gradient near the sea surface around 40°S and 40°N. These temperature fronts are maintained by the energetic eastward ocean currents through the thermal wind relation. Furthermore, the gradients of both temperature and salinity are generally stronger in the upper ocean than in the abyss. This is because the turbulent mixing, a process that homogenizes the fluid properties, dominates the evolution of temperature and salinity in the abyss (Vallis 2006).

4.2. Analysis Results

To detect the spatial structure of the T-S relationship, we construct the regression model as follows:

$$\mathcal{S}(s_i) = \beta_1(s_i)\mathcal{T}(s_i) + \beta_0(s_i) + \epsilon(s_i).$$

where the response variable $\mathcal{S}(s_i)$ denotes salinity at location $(s_{h,i}, s_{v,i})$, $\mathcal{T}(s_i)$ denotes temperature, the regression coefficient $\beta_1(s_i)$ measures the T-S relationship of interest, and $\beta_0(s_i)$ is the intercept.

The coefficient surface of $\beta_1(s_i)$ estimated from SCC is shown in Figure 6(a). As discussed above, the boundary of the AAIW can be identified as the contour of $\beta_1 = 0$. Its encompassing region covers the well-recognized generation site of the AAIW and the low-salinity tongue that is believed to be associated with the AAIW (Talley 2011). To investigate the change rate of the T-S relationship around the boundary, we compute the magnitude of the spatial difference quotient of the T-S relationship (Simmonds 2012) derived by SCC defined as follows:

$$D(s_i) = \sqrt{\begin{array}{l}\frac{(\beta_1(s_i)-\beta_1(s_{i_1}))^2}{d_1^2 \sin^2\gamma} + \frac{(\beta_1(s_i)-\beta_1(s_{i_2}))^2}{d_2^2 \sin^2\gamma} \\ -2\frac{(\beta_1(s_i)-\beta_1(s_{i_1}))(\beta_1(s_i)-\beta_1(s_{i_2}))\cos\gamma}{d_1 d_2 \sin^2\gamma}\end{array}},$$

where s_{i_1} and s_{i_2} are the two nearest points of s_i, γ is the angle between vectors $(s_{h,i_1} - s_{h,i}, s_{v,i_1} - s_{v,i})$ and $(s_{h,i_2} - s_{h,i}, s_{v,i_2} - s_{v,i})$, and d_1 (d_2) is the distance between s_{i_1} (s_{i_2}) and s_i. The value of $D(s_i)$ exhibits evident enhancement around the identified boundary of the AAIW (Figure 7), indicating a rapid change of the T-S relationship across the boundary. This feature is consistent with the geophysical fluid dynamics, as different water masses are formed at different sites and characterized by distinct T-S relationships (Talley 2011), giving us confidence regarding the good performance of SCC.

The GWR estimates, on the other hand, is quite noisy with occasional outliers (Figure 6(b)). We note that the noisy estimates similarly occurred in the simulation studies (see Section 3), which strongly suggests that such noise in estimation is perhaps not due to an actual feature of this dataset but probably due to the GWR method's deficiency in cases with spatially correlated explanatory variables. Indeed, the noise, especially in the abyss, is not consistent with the fluid dynamics, since there is no dynamical process that can lead to changes of the T-S relationship over such a short distance in the abyssal ocean (Talley 2011). According to the above heuristic interpretation of the results with regard to fluid dynamics, the SCC method seems to produce a more reasonable estimate for the T-S relationship than the GWR method.

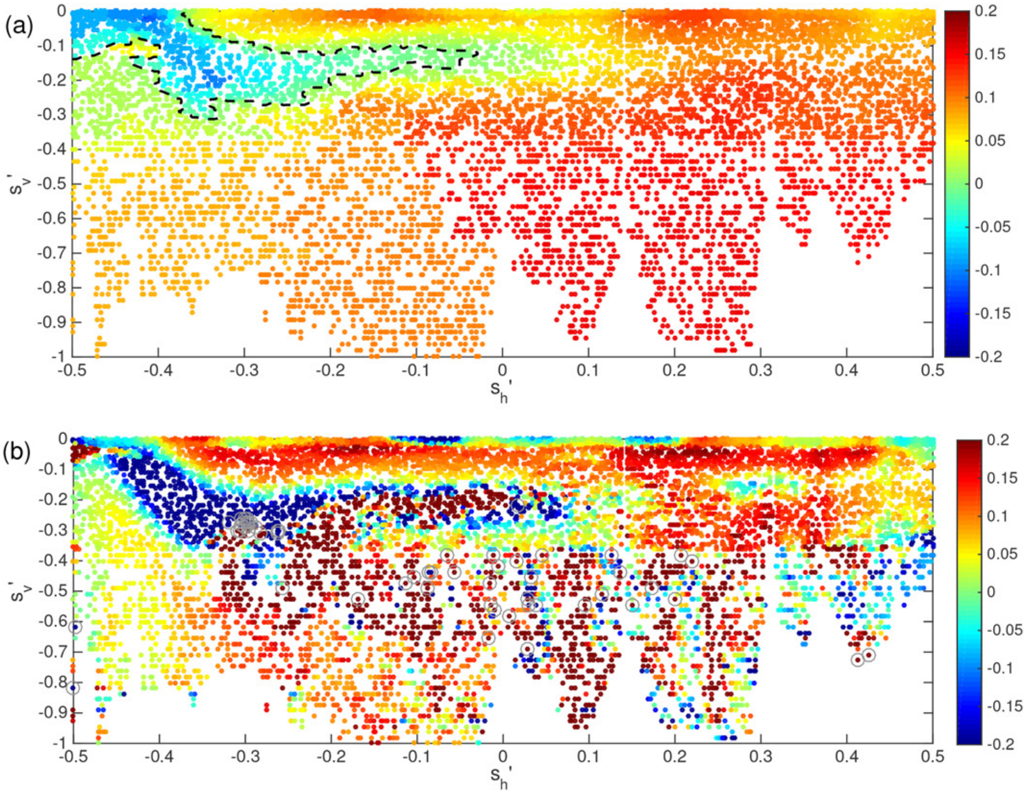

Figure 6. The T-S relationship β_1 estimated from (a) SCC and (b) GWR. The black dashed line in (a) is the contour of $\beta_1 = 0$. Note that the colorbar in (b) is saturated. The largest positive and negative values of the GWR estimates are 17.02 and -12.34, respectively. Data points with $|\beta_1| > 2$ are marked by gray circles.

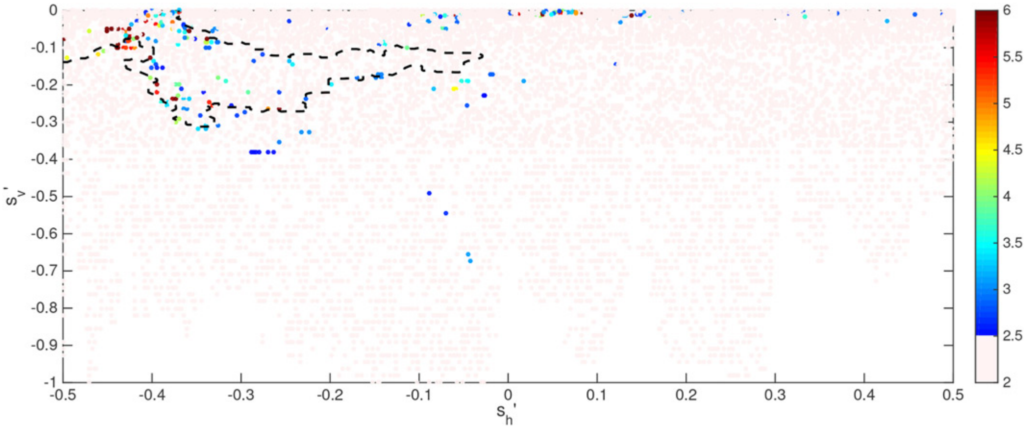

Figure 7. The magnitude of spatial difference quotient of T-S relationship and the contour of $\beta_1 = 0$ (black dashed). Note that the values less than 2.5 are masked by light pink.

5. Conclusions and Discussion

This article presents a new spatial regression approach, called the SCC method, to capture spatial patterns, especially clustered patterns in regression coefficients. SCC accommodates spatial dependence through structured coefficients and employs penalized least squares for estimation; the penalty function is carefully specified to encourage similarity in coefficients between locations connected by an MST. The estimation algorithms of SCC are easy to implement and allow highly scalable computation, as the penalized optimization can be transformed into a lasso (or lasso-type) problem. Although SCC is designed to detect a cluster structure in coefficients, our numerical studies show that it also works reasonably well in capturing a wide range of spatial patterns. Thus, SCC is a robust and flexible tool that will allow researchers to explore spatial patterns in regression coefficients without any priori information. We apply the SCC method to the analysis of a large temperature and salinity dataset for the Atlantic Ocean. The estimated regression coefficients reveal a spatially clustered pattern in the relationship between temperature and salinity. Our statistical findings are consistent with the interpretations from ocean dynamics.

SCC could be potentially improved from several aspects. As an MST is determined by spatial locations and may not be fully compatible with cluster structures of regression coefficients, it tends to miscluster especially for small sample sizes. A potential remedy suggested by an anonymous reviewer is to employ a two-step approach, using the SCC method to obtain an initial estimate of β, and then obtaining a refined SCC estimator using a new edge set $\mathbb{E}$ constructed based on both the initial estimate of β and spatial information. Alternatively, clusters could be refined by thresholding the estimated coefficients according to some criteria. Such a thresholding idea may also be applied to GWR and SVC to create clusters. Furthermore, the random errors in (1) are currently assumed to be independent, and we rely on the spatially varying intercept to capture some of the spatial dependence that is unexplained by the covariates. In future work, we may consider a model with a spatially dependent random error.

Finally, as pointed by an anonymous reviewer regarding the theoretical side, more attention needs to be paid to the validity of condition (6) for the design matrix. This restricted eigenvalue condition is a commonly used regularity condition on the design matrix for lasso regressions. However, its verification can be very challenging in practice (Dobriban and Fan 2016). For many spatial problems, covariates may show strong spatial dependence within each covariate as well as cross-dependence among covariates, so that both rows and columns of the design matrix become dependent. In this case, the distribution of the design matrix depends not only on the number of locations but also on their point pattern distributions in space, among a few other determining factors. Although L_1-methods still work very well in practice in spatial settings, as evidenced by our numerical examples and several other previous studies (Zhu, Huang, and Reyes 2010; Sun, Wang, and Fuentes 2016), further theoretical investigations are needed for this random and dependent design. One could assume a stationary multivariate Gaussian process for the covariates and follow a similar idea to that in Basu et al. (2015) for a time series context in order to establish conditions using multivariate spectral properties. We are currently investigating these topics.

Funding

This work was supported by NSF [grant nos. DMS-1343155, DMS-1622433, and DMS-1737885] and NIH [grant no. R01-CA19776101].

Supplementary Materials

Data and Codes: Data and codes to reproduce the numerical results are posted on Github page (*https://github.com/furong-tamu/Supplementary-files-for-SCC*).

Supplementary Numerical Results: Auxiliary numerical simulation results: (a) influence of the selection criteria for the tuning parameter λ on the MSE_β derived from SCC, (b) comparisons of the MSE_β derived from GWR using different bandwidths, (c) performance of SCC on regular grids, and (d) performance of SCC for a hybrid scenario with clustered coefficients and a smoothly varying intercept (or equivalently, a spatially random effect).

Supplementary Proof: Proof of Theorem 1.

References

Arnold, T. B., and Tibshirani, R. J. (2016), "Efficient Implementations of the Generalized Lasso Dual Path Algorithm," *Journal of Computational and Graphical Statistics*, 25, 1–27. [4]

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian Predictive Process Models for Large Spatial Data Sets," *Journal of the Royal Statistical Society,* Series B, 70, 825–848. [2]

Basu, S., and Michailidis, G. (2015), "Regularized Estimation in Sparse High-dimensional Time Series Models," *Annals of Statistics 43*, 1535–1567. [12]

Boyd, S. (2011). "Alternating Direction Method of Multipliers," in *Talk at NIPS Workshop on Optimization and Machine Learning.* [4]

Braunstein, L. A., Wu, Z., Chen, Y., Buldyrev, S. V., Kalisky, T., Sreenivasan, S., Cohen, R., López, E., Havlin, S., and Stanley, H. E. (2007), "Optimal Path and Minimal Spanning Trees in Random Weighted Networks," *International Journal of Bifurcation and Chaos*, 17, 2215–2255. [4]

Came, R. E., Oppo, D. W., Curry, W. B., and Lynch-Stieglitz, J. (2008), "Deglacial Variability in the Surface Return Flow of the Atlantic Meridional Overturning Circulation," *Paleoceanography*, 23(1), PA1217. [8]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection With Large Model Spaces," *Biometrika*, 95, 759–771. [5]

——— (2012), "Extended Bic for Small-n-Large-p Sparse GLM," *Statistica Sinica*, 22, 555–574. [5]

Cressie, N. (1993), "*Statistics for Spatial Data*. New York: Wiley. [1]

Diggle, P. J., Tawn, J., and Moyeed, R. (1998), "Model-Based Geostatistics," *Journal of the Royal Statistical Society,* Series C, 47(3), 299–350. [1]

Dobriban, E., and Fan, J. (2016), "Regularity Properties for Sparse Regression," *Communications in Mathematics and Statistics*, 4, 1–19. [12]

Dobrin, R., and Duxbury, P. (2001), "Minimum Spanning Trees on Random Networks," *Physical Review Letters*, 86, 5076–5079. [4]

Dolan, J., Weiss, R., and Smith, J. M. (1991), "Minimal Length Tree Networks on the Unit Sphere," *Annals of Operations Research*, 33, 501–535. [5]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, *32*, 407–499. [5]

Fan, J., and Li, R.(2001), "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [3]

Finley, A. O. (2011), "Comparing Spatially-Varying Coefficients Models for Analysis of Ecological Data with Non-Stationary and Anisotropic Residual Dependence," *Methods in Ecology and Evolution*, 2, 143–154. [6]

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2003), *Geographically Weighted Regression*, Chichester: Wiley. [2]

Foygel, R., and Drton, M. (2010), "Extended Bayesian Information Criteria for Gaussian Graphical Models," in *Advances in Neural Information Processing Systems*, eds. J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Vancouver, British Columbia, Canada: Curran Associates, Inc, pp. 604–612. [6]

Frandi, E., R. Ñanculef, S. Lodi, C. Sartori, and J. A. Suykens (2016), "Fast and Scalable Lasso Via Stochastic Frank–Wolfe Methods With a Convergence Guarantee," *Machine Learning*, 104, 195–221. [5]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models Via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [5]

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003), "Spatial Modeling With Spatially Varying Coefficient Processes," *Journal of the American Statistical Association*, 98, 387–396. [2]

Golub, G. H., Heath, M., and Wahba, G. (1979), "Generalized Cross-validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223. [5]

Grygorash, O., Zhou, Y., and Jorgensen, Z.(2006), "Minimum Spanning Tree Based Clustering Algorithms," in *Tools with Artificial Intelligence, 18th IEEE International Conference on 2006 (ICTAI'06)*, pp. 73–81. Arlington, VA: IEEE. [4]

Ke, Z. T., Fan, J., and Wu, Y. (2015), "Homogeneity Pursuit," *Journal of the American Statistical Association*, 110, 175–194. [2]

LeSage, J. P. (2004). "A Family of Geographically Weighted Regression Models," in *Advances in Spatial Econometrics*, eds. J. G. M. Raymond and S. J. Rey, pp. 241–264. [6]

Lindgren, F., Rue, H., and Lindström, J. (2011), "An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach," *Journal of the Royal Statistical Society,* Series B, 73, 423–498. [2]

Locarnini, R., Mishonov, A., Antonov, J., Boyer, T., Garcia, H., Baranova, O., Zweng, M., Paver, C., Reagan, J., Johnson, D., et al. (2013), *World ocean atlas 2013, volume 1: Temperature*. Boulder, CO: NOAA Atlas NESDIS. [8]

March, W. B., Ram, P., and Gray, A. G. (2010), "Fast Euclidean Minimum Spanning Tree: Algorithm, Analysis, and Applications," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 603–612. Washington, DC: ACM. [5]

Oppo, D., and Curry, W. (2012), "Deep Atlantic Circulation During the Last Glacial Maximum and Deglaciation," *Nature Education Knowledge*, 3, 1. [8]

Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., Wanninkhof, R., Wong, C., Wallace, D. W., Tilbrook, B., et al. (2004), "The Oceanic Sink for Anthropogenic CO_2," *Science*, 305, 367–371. [8]

Schwarz, G. et al. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464. [5]

Shen, X., and Huang, H.-C. (2010), "Grouping Pursuit Through a Regularization Solution Surface," *Journal of the American Statistical Association*, 105, 727–739. [4]

Simmonds, J. G. (2012), *A Brief on Tensor Analysis*, New York: Springer Science & Business Media. [10]

Song, Q., and Liang, F. (2015), "High-dimensional Variable Selection With Reciprocal ℓ_1-Regularization," *Journal of the American Statistical Association*, 110, 1607–1620. [3,6]

Sun, Y., Wang, H. J., and Fuentes, M. (2016), "Fused Adaptive Lasso for Spatial and Temporal Quantile Function Estimation," *Technometrics*, 58, 127–137. [12]

Talley, L. D. (2011), *Descriptive Physical Oceanography: An Introduction*, London: Academic Press. [1,6,8,10]

Tang, L., and Song, P. X. (2016), "Fused Lasso Approach in Regression Coefficients Clustering–Learning Parameter Heterogeneity in Data Integration," *Journal of Machine Learning Research*, 17, 1–23. [2,3]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [2,3]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness Via the Fused Lasso," *Journal of the Royal Statistical Society*, Series B, 67, 91–108. [2]

Tibshirani, R., Wainwright, M., and Hastie, T. (2015), *Statistical Learning With Sparsity: The Lasso and Generalizations*, Boca Raton, FL: Chapman and Hall/CRC. [5]

Vallis, G. K. (2006), *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-scale Circulation*. Cambridge: Cambridge University Press. [1,9,10]

Wheeler, D. C., and Calder, C. A. (2007), "An Assessment of Coefficient Accuracy in Linear Regression Models With Spatially Varying Coefficients," *Journal of Geographical Systems*, 9, 145–166. [6]

Xie, R. C., Marcantonio, F., and Schmidt, M. W. (2012), "Deglacial Variability of Antarctic Intermediate Water Penetration Into the North Atlantic From Authigenic Neodymium Isotope Ratios," *Paleoceanography*, 27, PA3221. [8]

Zhang, C. -H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *Annals of Statistics*, 38, 894–942. [3]

Zhu, J., Huang, H.-C., and Reyes, P. E. (2010), "On Selection of Spatial Linear Models for Lattice Data," *Journal of the Royal Statistical Society,* Series B, 72, 389–402. [12]

Zhu, Y. (2017), "An Augmented ADMM Algorithm With Application to the Generalized Lasso Problem," *Journal of Computational and Graphical Statistics*, 26, 195–204. [4]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [3]

Zweng, M., Reagan, J., Antonov, J., Locarnini, R., Mishonov, A., Boyer, T., Garcia, H., Baranova, O., Johnson, D., Seidov, D., and M. M. Biddle. (2013), *World Ocean Atlas 2013*, *Vol. 2: Salinity*, Boulder, CO: NOAA Atlas NESDIS Berlin, Heidelberg: Springer. [8]