GLS-SOD: A *G*eneralized *L*ocal *S*tatistical Approach for *S*patial *O*utlier *D*etection

Feng Chen
Department of Computer Science
Virginia Tech, USA
chenf@vt.edu

Chang-Tien Lu
Department of Computer Science
Virginia Tech, USA
ctlu@vt.edu

Arnold P. Boedihardjo
Department of Computer Science
Virginia Tech, USA
Arnold.P.Boedihardjo@vt.edu

ABSTRACT

Local based approach is a major category of methods for spatial outlier detection (*SOD*). Currently, there is a lack of systematic analysis on the statistical properties of this framework. For example, most methods assume identical and independent normal distributions (i.i.d. normal) for the calculated local differences, but no justifications for this critical assumption have been presented. The methods' detection performance on geostatistic data with linear or nonlinear trend is also not well studied. In addition, there is a lack of theoretical connections and empirical comparisons between local and global based *SOD* approaches. This paper discusses all these fundamental issues under the proposed generalized local statistical (*GLS*) framework. Furthermore, robust estimation and outlier detection methods are designed for the new *GLS* model. Extensive simulations demonstrated that the *SOD* method based on the *GLS* model significantly outperformed all existing approaches when the spatial data exhibits a linear or nonlinear trend.

Categories and Subject Descriptors

D.2.8 [**Database Management**]: Database Applications – data mining. I.5.3 [Pattern Recognition]: Outlier Detection.

General Terms

Algorithms, Theory, and Experimentation

Keywords

Spatial Outlier Detection, Spatial Gaussian Random Field.

1. INTRODUCTION

The ever-increasing volume of spatial data has greatly challenged our ability to exact useful but implicit knowledge from them. As an important branch of spatial data mining, spatial outlier detection aims to discover the objects whose non-spatial attribute values are significantly different from the values of their spatial neighbors [1]. In contrast to traditional outlier detection, spatial outlier detection must differentiate spatial and non-spatial attributes, and consider the spatial continuity and autocorrelation between nearby samples. By the first law of geography, "Everything is related to everything else, but nearby things are more related than distant things [3]."

There are two main streams for spatial outlier detection (*SOD*): local and global based approaches. Local based approach [4] first calculates the local difference (statistic) for each object, which is the difference between the non-spatial attribute of the object and the aggregated value (e.g., average) of its spatial neighbors. By assuming i.i.d. normal distributions for these local differences, the local based approach discovers outlier objects by robust estimation of model parameters, such as the aggregated values, mean, and standard deviation. Various methods have been presented by using various spatial neighborhood definitions and robust estimation techniques [5-9]. The second stream, global based, is to identify outliers using the robust estimator of a global kriging model which is the best linear unbiased estimator for geostatistical data. Particularly, Christensen et al. [10] proposed diagnostics to detect spatial outliers on the estimation of covariance function. Cerioli and Riani [11] developed a forward search procedure to identify spatial outliers for an ordinary kriging model. Militino et al. [12] further generalized the forward search method in [11] to a universal kriging model. This paper focuses on local based methods, because local based methods are simpler to understand and implement and can achieve better efficiency with minimal loss of accuracy. This will be justified by extensive simulations in Section 5.

This work is primarily motivated by the current situation where there is still no systematic study about the statistical properties of local based *SOD* methods. For example, existing works assume i.i.d. on local differences, but no justifications have ever been proposed. Also, their performance on spatial data with linear or nonlinear trends has not been well studied. There is also a lack of research on the theoretical connections and empirical comparisons between local and global based *SOD* methods. To that end, this paper provides a generalized framework for local based *SOD* methods and theoretically and empirically compares it to global based *SOD* methods. The proposed framework is casted within the statistical abstraction of a spatial Gaussian random field which is the most popular model for geostatistical data [1,2]. A major reason for its popularity is that the optimal solution based on the Gaussian random field is equivalent to a best linear unbiased estimator that imposes no particular distributional assumption.

A spatial Gaussian random field refers to a collection of dependent random variables that are associated with a set of spatial indexes, $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathbb{R}^2\}$, where D is a continuous fixed region. This family of random variables can be characterized by a joint Gaussian probability density or distribution. In real applications, only partial observations of one realization (or a partial sample of size one) are available: $\{Z(\boldsymbol{s_1}), \dots, Z(\boldsymbol{s_n})\}$. In order to make this model operational, the requirements for stationarity and isotropy, such as second-order or intrinsic stationarity, are further imposed. Imposing such an assumption reduces the number of model parameters required to be estimated. When the data is second-order stationary and isotropic, the spatial correlation structure is described by some semivariogram or covariance function, in which the correlation between two variables is dependent on their spatial distance. Statistical inferences are then performed by assuming some explicit forms of the covariance and mean functions.

Our major contributions are as follows:

- **Design of a generalized local statistical framework:** The general local statistical (*GLS*) model is a generalized statistical framework for existing local based *SOD* methods. It can effectively handle complex situations where the spatial data exhibits a global trend or non-negligible dependences between local differences.
- **Robust estimation and outlier detection methods based on the proposed *GLS* framework**: Analyze contamination issues

that cause the masking and swamping effects of outlier detection. Based on the analysis, two robust algorithms, *GLS*-backward search and *GLS*-forward search, are proposed to estimate the parameters for the *GLS* model.

- **In-depth study on the connection between different *SOD* methods**: Present theoretical foundations for existing local based *SOD* methods and discuss the crucial connections between local and global based *SOD* methods.
- **Comprehensive simulations to validate the effectiveness and efficiency of *GLS***. This is the first work that provides extensive comparisons between existing popular methods through a systematic simulation study. The results show that the proposed *GLS-SOD* approach significantly outperformed all existing methods when the spatial data exhibits a linear or nonlinear trend.

The proceeding sections are organized as follows. Section 2 provides a brief description of spatial local statistics and survey of related works. Section 3 presents the generalized local statistical model and gives a rigorous theoretical treatment of its fundamental statistical properties. Section 4 introduces several robust estimation and outlier detection methods for the *GLS* model, and analyzes the connection between different *SOD* methods. Section 5 provides the simulation and discussion. Section 6 gives the conclusion.

2. SPATIAL LOCAL STATISTICS AND RELATED WORKS

Given a set of observations $\{Z(\boldsymbol{s_1}), Z(\boldsymbol{s_2}), \ldots, Z(\boldsymbol{s_n})\}$, a local spatial statistic [4] is defined as

$$S(\boldsymbol{s}) = \left[Z(\boldsymbol{s}) - E_{\boldsymbol{s}_i \in \boldsymbol{N}(\boldsymbol{s})}\big(Z(\boldsymbol{s}_i)\big)\right], \tag{1}$$

where $\boldsymbol{G} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset \mathbb{R}^2$ is a set of spatial locations, $\boldsymbol{s} \in \boldsymbol{G}$, $Z(\boldsymbol{s}) \in \mathbb{R}$ represents the value of Z attribute at location $\boldsymbol{s}$, $\boldsymbol{N}(\boldsymbol{s})$ is the set of spatial neighbors of $\boldsymbol{s}$, and $E_{\boldsymbol{s}_i \in N(\boldsymbol{s})}\big(Z(\boldsymbol{s}_i)\big)$ represents the average attribute value for the neighbors of $\boldsymbol{s}$. It is assumed that the set of local spatial statistics $\{S(\boldsymbol{s}_1), \ldots, S(\boldsymbol{s}_n)\}$ are independently and identically normally distributed (i.i.d. normal). Then the popular Z-test [4] for detecting spatial outliers can be described as follows: Spatial statistic $Z_{S(s)} = \left|\frac{S(s)-\mu_s}{\sigma_s}\right| > \Phi^{-1}\left(\frac{\alpha}{2}\right)$, where Φ is the cumulative distribution function (*CDF*) of a standard normal distribution, α refers to significance level and is usually set to 0.05, and $\mu_{\boldsymbol{s}}$ and $\sigma_{\boldsymbol{s}}$ are the sample mean and standard deviation, respectively.

Lu et al. [5] pointed out that the Z-test is susceptible to the well-known masking and swamping effects. When multiple outliers exist in the data, the quantities $E_{\boldsymbol{s}_i \in N(\boldsymbol{s})}\big(Z(\boldsymbol{s}_i)\big)$, $\mu_{\boldsymbol{s}}$, and $\sigma_{\boldsymbol{s}}$ are biased estimates of the population means and standard deviation. As a result, some true outliers are "masked" as normal objects and some normal objects are "swamped" and misclassified as outliers. The authors proposed an iterative approach that detects outliers by multi-iterations. Each iteration identifies only one outlier and modifies its attribute value so that it will not impact the results of subsequent iterations. Later, Chen et al. [6] proposed a median based approach that uses median estimator for the quantities $E_{\boldsymbol{s}_i \in N(\boldsymbol{s})}\big(Z(\boldsymbol{s}_i)\big)$ and $\mu_{\boldsymbol{s}}$, and median absolute deviation (MAD) estimator for $\sigma_{\boldsymbol{s}}$. Hu and Sung [7] proposed an approach similar to [6], but using trimmed mean to estimate $E_{\boldsymbol{s}_i \in N(\boldsymbol{s})}\big(Z(\boldsymbol{s}_i)\big)$, instead of the median. Sun and Chawla [8] presented a spatial local outlier measure to capture the local behavior of data in their neighborhood. Shekhar et al. [9] employed a graph-based method to define spatial neighborhoods ($N(\boldsymbol{s})$) and their method is applied to a special case of transportation network.

Most existing local based methods assume that the set of local statistics $\{S(\boldsymbol{s}_1), \ldots, S(\boldsymbol{s}_n)\}$ are i.i.d. normal, but no justifications for this assumption have been proposed. As we will discuss in subsequent sections, this i.i.d. assumption is only approximately true in certain scenarios, and the dependencies between different local differences (statistics) must be considered when the spatial data exhibit linear or nonlinear trend or the selected neighborhood size for each object is small. As shown in our simulations in Section 5, the violation of i.i.d. assumption can significantly impact the accuracies of the outlier detection methods.

3. GENERALIZED LOCAL SPATIAL STATISTICS

This section first introduces some preliminary background on spatial Gaussian random field, then presents the generalized local statistical (*GLS*) model, and finally discusses the statistical properties of the *GLS* model. Table 1 summarizes the key notations used in this paper.

Table 1: Description of Major Symbols

Symbol	*Descriptions*
$\{Z(\boldsymbol{s}_i)\}_{i=1}^n$	A given set of observations, where $\boldsymbol{s}_i \in \mathbb{R}^2$ is the spatial location and $Z(\cdot)$ is the Z attribute value.
$\{\boldsymbol{x}(\boldsymbol{s}_i)\}_{i=1}^n$	$\boldsymbol{x}(\boldsymbol{s}_i)$ is a vector of covariates of $\boldsymbol{s}_i$, such as the bases of spatial coordinates of $\boldsymbol{s}_i$.
$\mathbf{Z}$	$\mathbf{Z} = [Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)]^{\mathbf{T}}$
$\mathbf{X}$	$\mathbf{X} = [\boldsymbol{x}(\boldsymbol{s}_1), \ldots, \boldsymbol{x}(\boldsymbol{s}_n)]^{\mathbf{T}}$
$\mathbf{F}$	Neighborhood weight matrix; See equation (4)
$\boldsymbol{N}(\boldsymbol{s})$	A general definition of spatial neighbors of $\boldsymbol{s}$.
$\boldsymbol{N}_K(\boldsymbol{s})$	K-nearest neighbors of $\boldsymbol{s}$**.** This paper considers $\boldsymbol{N}_K(\boldsymbol{s})$ as the specification of $\boldsymbol{N}(\boldsymbol{s})$.
K	Neighborhood size. It is the major parameter to define spatial neighbors $\big(\boldsymbol{N}_K(\boldsymbol{s})\big)$.
SOD	***S**patial **O**utlier **D**etection*
GLS	***G**eneralized **L**ocal **S**tatistics Model*
$\boldsymbol{\beta}$, σ, σ_0	The unknown parameters in the *GLS* model

3.1 Generalized Local Statistic Model (*GLS*)

Consider a spatial Gaussian random field $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathbb{R}^2\}$ with the following form:

$$Z(\boldsymbol{s}) = f(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{\beta}) + \omega(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \tag{2}$$

where D is a fixed region, $f(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{\beta})$ is the large scale trend (mean) of the process, $\omega(\boldsymbol{s})$ is the smooth-scale variation that is a Gaussian stationary process, and $\epsilon(\boldsymbol{s})$ is the white noise measurement error with variance σ_0^2.

For the large scale trend $f(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{\beta})$, $\boldsymbol{x}(\boldsymbol{s})$ is a vector of covariates, and $\boldsymbol{\beta}$ is a vector of parameters for the trend model. We assume that $\boldsymbol{x}(\boldsymbol{s})$ is a vector of the basis of spatial coordinates of $\boldsymbol{s}$, and $f(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{\beta})$ is a linear function with $f(\boldsymbol{x}(\boldsymbol{s}), \boldsymbol{\beta}) = \boldsymbol{x}(\boldsymbol{s})^T\boldsymbol{\beta}$. The nonlinear degree of the trend depends on the polynomials of the elements in $\boldsymbol{x}(\boldsymbol{s})$. For the smooth-scale variation $\omega(\boldsymbol{s})$, we assume that it is an isotropic second order stationary process, which means the covariance $\mathrm{Cov}\big(Z(\boldsymbol{s}_1), Z(\boldsymbol{s}_2)\big)$ is a function of the spatial distance between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$: $C(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$. Various distance metrics may be selected, such as L_2 (Euclidean distance), L_1 (Manhattan distance), and graph distance [10].

Given a set of observations $\{Z(\boldsymbol{s_1}), Z(\boldsymbol{s_2}), \ldots, Z(\boldsymbol{s_n})\}$ that is a partial sample of a particular realization of the spatial Gaussian random field, let $\boldsymbol{Z} = [Z(\boldsymbol{s_1}), \ldots, Z(\boldsymbol{s_n})]^T$, $\boldsymbol{\omega} = [\omega(\boldsymbol{s}_1), \ldots \omega(\boldsymbol{s}_n)]^T$, $\boldsymbol{e} = [e(\boldsymbol{s}_1), \ldots e(\boldsymbol{s}_n)]^T$, and $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^T$. Then we have

$$\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{\omega} + \boldsymbol{e} \sim N(\boldsymbol{X\beta}, \boldsymbol{\Sigma} + \sigma_0^2 \boldsymbol{I}), \tag{3}$$

where $\boldsymbol{\omega}(\boldsymbol{s}) \sim \boldsymbol{N}(\boldsymbol{0}_{n\times 1}, \boldsymbol{\Sigma}_{n\times n})$, and $\boldsymbol{e}(\boldsymbol{s}) \sim \boldsymbol{N}(\boldsymbol{0}_{n\times 1}, \sigma_0^2 \boldsymbol{I}_{n\times n})$.

The vector of local spatial statistics calculated by equation (1) can be reformulated as the matrix form

$$\mathbf{diff}(\boldsymbol{Z}) = \boldsymbol{FZ}, \tag{4}$$

where $\boldsymbol{F} \in \mathbb{R}^{n\times n}$ is a neighborhood weight matrix with $F_{ij} = 1$ when $i = j$; $F_{ij} = -\frac{1}{K}$, when $s_j \in N_K(s_i)$; and $F_{ij} = 0$ otherwise.

By equations (3) and (4), we can readily derive the generalized local statistical (*GLS*) model as

$$\mathbf{diff}(\boldsymbol{Z}) \sim \mathrm{N}(\boldsymbol{FX\beta}, \boldsymbol{F\Sigma F}^T + \sigma_0^2 \boldsymbol{FF}^T). \tag{5}$$

As shown in Section 3.2, $\boldsymbol{F\Sigma F}^T$ can be approximated by $\boldsymbol{\sigma_0^2 I}$. It follows that the *GLS* form (5) becomes asymptotically equivalent to

$$\mathbf{diff}(\boldsymbol{Z}) \sim \mathrm{N}(\boldsymbol{FX\beta}, \sigma^2 \boldsymbol{I} + \sigma_0^2 \boldsymbol{FF}^T). \tag{6}$$

As indicated in Section 3.2 Theorem 1, when the neighborhood size is relatively large with $K \geq 8$, the component $\sigma_0^2 \boldsymbol{FF}^T$ can be further approximated by $\sigma_0^2 \boldsymbol{I}$. This leads to a simpler form of *GLS* as

$$\mathbf{diff}(\boldsymbol{Z}) \sim \mathrm{N}(\boldsymbol{FX\beta}, (\sigma^2 + \sigma_0^2)\boldsymbol{I}). \tag{7}$$

This generalized local statistical model above has the unknown parameters $\boldsymbol{\beta}$, σ, and σ_0. The robust estimation of these parameters will be discussed in Section 4.

3.2 Theoretical Properties of *GLS*

This sub-section studies the properties of two major covariance components $\sigma_0^2 \boldsymbol{FF}^T$ and $\boldsymbol{F\Sigma F}^T$, and discusses the situations where they can be approximated by $\sigma_0^2 \boldsymbol{I}$ and $\sigma^2 \boldsymbol{I}$, respectively. As shown in equation (3), $\boldsymbol{\sigma_0^2 FF}^T$ and $\boldsymbol{F\Sigma F}^T$ are the covariance matrices of the random vectors $\boldsymbol{e}^* = \boldsymbol{Fe}$ and $\boldsymbol{\omega}^* = \boldsymbol{F\omega}$, respectively. We focus on the study of their correlation structures. Because they are both multivariate normally distributed, the correlation structure gives important information about the related dependence structure (e.g., in-correlation implies independence). Three related theorems are stated as follows:

Theorem 1: *The random vector $\boldsymbol{e}^*$ has two major properties*

1) *The variance $Var(e_i^*) = \frac{K+1}{K}\sigma_0^2, i = 1 \ldots n,$*
2) *The correlation $\left|\rho\left(e_i^*, e_j^*\right)\right| \leq \frac{2}{K+1}, \forall i, j$ with $i \neq j,$*

where e_i^* refers to the i-th element in the vector $\boldsymbol{e}^*$.

Proof: First, we prove property 1). Recall that $\mathrm{Var}(\boldsymbol{e}^*) = \sigma_0^{\boldsymbol{2}} \boldsymbol{FF}^T$, where $\boldsymbol{F}$ is the neighborhood weight matrix (see Section 3.1 equation (4) for the definition). For simplicity, we represent $\boldsymbol{F}$ as $[\boldsymbol{F}_1, \boldsymbol{F}_2, \ldots, \boldsymbol{F}_n]^T$ and let $\boldsymbol{F}_{ij}$ denote the j-th component of the vector $\boldsymbol{F}_i$. According to the definition of $\boldsymbol{F}$, $\boldsymbol{F}_{ii} = 1$; $\boldsymbol{F}_{ij} = -\frac{1}{K}$, if $\boldsymbol{s}_j \in \boldsymbol{N}_k(\boldsymbol{s}_i)$; otherwise, $\boldsymbol{F}_{ij} = 0$. It implies that $\mathrm{Var}(e_i^*) = \left[\sigma_0^{\boldsymbol{2}} \boldsymbol{FF}^T\right]_{ij} = \sigma_0^{\boldsymbol{2}} \boldsymbol{F_i^T F_i} = \sigma_0^{\boldsymbol{2}}\left(1 + \sum_{i=1}^{K} \frac{1}{K^2}\right) = \sigma_0^{\boldsymbol{2}}\left(1 + \frac{1}{K}\right) = \frac{1+K}{K}\sigma_0^{\boldsymbol{2}}$, $\forall i = 1, \ldots, n$. This proves property 1).

Second, we prove property 2). $\forall i, j \in \{1, \ldots, n\}$, the correlation $\rho\left(e_i^*, e_j^*\right) = \left[\sigma_0^{\boldsymbol{2}} \boldsymbol{FF}^T\right]_{ij} / \left(\frac{K+1}{K}\sigma_0^2\right) = \frac{K}{K+1}\boldsymbol{F}_i^T \boldsymbol{F}_j = \frac{K}{K+1}\sum_{t=1}^{n} \boldsymbol{F}_{it}\boldsymbol{F}_{jt} = \frac{K}{K+1}\left(\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} + \boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} + \sum_{t=1,t\neq i,j}^{n} \boldsymbol{F}_{it}\boldsymbol{F}_{jt}\right)$. The third component in this equation satisfies $\sum_{t=1,t\neq i,j}^{n} \boldsymbol{F}_{it}\boldsymbol{F}_{jt} \in \left[0, \frac{1}{K}\right]$, since $\boldsymbol{F}_{it}$ and $\boldsymbol{F}_{jt}$ can only be $-\frac{1}{K}$ or zero, and the set $\{\boldsymbol{F}_{ik}\}_{k=1,k\neq i}^{n}$ or $\{\boldsymbol{F}_{jt}\}_{t=1,t\neq i}^{n}$ has at most K elements with value $-\frac{1}{K}$. As to the components $\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji}$ and $\boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj}$, we consider four different situations:

(1) $\boldsymbol{s}_j \in \boldsymbol{N}_k(\boldsymbol{s}_i), \boldsymbol{s}_i \in \boldsymbol{N}_k(\boldsymbol{s}_j)$:

It implies that $\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} = \boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} = -\frac{1}{K}$. Then,

$\left|\rho\left(e_i^*, e_j^*\right)\right| = \frac{K}{K+1}\left|\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} + \boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} + \sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| = \frac{K}{K+1}\left|-\frac{2}{K} + \sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| \leq \frac{K}{K+1} \cdot \frac{2}{K} = \frac{2}{K+1}$.

(2) $\boldsymbol{s}_j \in N_k(\boldsymbol{s}_i), \boldsymbol{s}_i \notin N_k(\boldsymbol{s}_j)$

It implies that $\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} = 0$ and $\boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} = -\frac{1}{K}$. Then,

$\left|\rho\left(e_i^*, e_j^*\right)\right| = \frac{K}{K+1}\left|\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} + \boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} + \sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| = \frac{K}{K+1}\left|-\frac{1}{K} + \sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| \leq \frac{K}{K+1} \cdot \frac{1}{K} = \frac{1}{K+1}$.

(3) $\boldsymbol{s}_j \notin N_k(\boldsymbol{s}_i), \boldsymbol{s}_i \in N_k(\boldsymbol{s}_j)$

It implies that $\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} = -\frac{1}{K}$ and $\boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} = 0$. Then,

$\left|\rho\left(e_i^*, e_j^*\right)\right| = \frac{K}{K+1}\left|\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} + \boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} + \sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| = \frac{K}{K+1}\left|-\frac{1}{K} + \sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| \leq \frac{K}{K+1} \cdot \frac{1}{K} = \frac{1}{K+1}$.

(4) $\boldsymbol{s}_j \notin N_k(\boldsymbol{s}_i), \boldsymbol{s}_i \notin N_k(\boldsymbol{s}_j)$

It implies that $\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} = \boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} = 0$. Then,

$\left|\rho\left(e_i^*, e_j^*\right)\right| = \frac{K}{K+1}\left|\boldsymbol{F}_{ii} \cdot \boldsymbol{F}_{ji} + \boldsymbol{F}_{ij} \cdot \boldsymbol{F}_{jj} + \sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| = \frac{K}{K+1}\left|\sum_{k=1,k\neq i,j}^{n} \boldsymbol{F}_{ik}\boldsymbol{F}_{jk}\right| \leq \frac{K}{K+1} \cdot \frac{1}{K} = \frac{1}{K+1}$.

Therefore, we conclude that $\left|\rho\left(e_i^*, e_j^*\right)\right| \leq \frac{2}{K+1}, \forall i, j$ with $i \neq j$. □

Theorem 1 indicates that when the neighborhood size is relative large, the correlations between the components in $\boldsymbol{e}^*$ are very low (e.g., smaller than 0.2 when $K = 10$) and the variance of each component is very close to σ_0^2. In this case, $\boldsymbol{\sigma_0^2 FF}^T \approx \boldsymbol{\sigma_0^2 I}$. However, for a small neighborhood size, as shown in simulations (Section 5), the dependence between the components in $\boldsymbol{e}^*$ must be considered.

The next two theorems are related to the random vector $\boldsymbol{\omega}^*$. It is very difficult to analytically evaluate $\boldsymbol{\omega}^*$, because it is generated by an isotropic second order stationary process, and even when the explicit form of the covariance function is known, the statistical properties of $\boldsymbol{\omega}^*$ are still not straightforward. For this reason, several additional assumptions (constraints) need to be considered.

The following are three assumptions required for Theorem 2:

1. *If $\boldsymbol{N}_K(\boldsymbol{s}_l) \cap \boldsymbol{N}_K(\boldsymbol{s}_d) \neq \boldsymbol{\Phi}$, then, $\forall s_i, s_j, s_t \in \boldsymbol{N}_K(\boldsymbol{s}_l) \cup \boldsymbol{N}_K(\boldsymbol{s}_d)$, their between spatial distances are approximately equivalent: $\left\|\boldsymbol{s}_j - \boldsymbol{s}_i\right\| \approx \left\|\boldsymbol{s}_t - \boldsymbol{s}_i\right\| \approx \left\|\boldsymbol{s}_j - \boldsymbol{s}_t\right\|$.*
2. *If $\boldsymbol{s}_j \in \boldsymbol{N}_K(\boldsymbol{s}_i), \boldsymbol{s}_t \notin \boldsymbol{N}_K(\boldsymbol{s}_i)$, and $\boldsymbol{N}_K(\boldsymbol{s}_t) \cap \boldsymbol{N}_K(\boldsymbol{s}_i) = \boldsymbol{\Phi}$, then $\left\|\boldsymbol{s}_t - \boldsymbol{s}_i\right\| \approx \left\|\boldsymbol{s}_t - \boldsymbol{s}_j\right\|$.*

3. *The distance between any points that are k-nearest neighbors is approximately constant everywhere.*

The intuition on assumptions 1 and 2 is that, because neighbors are close to each other, they share similar between-distances, and also share similar distances to the points that are not their neighbors. The assumption 3 is valid when the spatial locations follow a uniform distribution or a grid structure. Note that, the assumption 3 holds in many applications [13]. The situations where assumptions 1 and 2 are potentially violated will be discussed in Theorem 3.

Theorem 2: *If the above assumptions 1 and 2 hold, then the random vector* $\boldsymbol{\omega}^*$ *has two major properties*

1) *The variance* $Var(\omega_i^*) \approx \frac{1+K}{K}\left(\sigma^2 - C_{x_i}\right), i = 1 \dots n$
2) *The correlation* $\rho\left(\omega_i^*, \omega_j^*\right) \approx -\frac{1}{K}$, *if* $s_j \in N_K(s_i)$ *or* $s_i \in N_K\left(s_j\right)$; *otherwise,* $\rho\left(\omega_i^*, \omega_j^*\right) \approx 0$,

where $C_{\boldsymbol{s}_i}$ *refers to the average covariance value between* $\boldsymbol{s}_i$ *and its K-nearest neighbors, and* $\sigma = C(0)$ *refers to the constant variance for each component of* $\boldsymbol{\omega}$*. Further, if the assumption 3 also holds, then the variance* $Var(\omega_i^*)$ *becomes constant everywhere.*

Proof: Let $\boldsymbol{\Sigma} = \text{Var}(\boldsymbol{\omega}), \boldsymbol{D} = \text{Var}(\boldsymbol{\omega}^*) = \boldsymbol{F\Sigma F}^T$, and $\boldsymbol{T} = \boldsymbol{F\Sigma}$. Recall that $\boldsymbol{\omega}^* = \boldsymbol{F\omega}$, where $\boldsymbol{\omega}$ is the smooth scale variation (see Section 3.1 equation (3)). The covariance component $\boldsymbol{\Sigma}_{ij} = \text{Cov}\left(\omega_i, \omega_j\right) = C\left(\left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\|\right)$, where $C(\cdot)$ is a covariance function (e.g., exponential or spherical functions) that depends on the distance $h_{ij} = \left\|\boldsymbol{s}_i - \boldsymbol{s}_j\right\|$. By the covariance function $C(\cdot)$ and the assumption 1, neighboring points must have the same covariance. For each point $\boldsymbol{s}_i$, we represent the constant covariance between $\boldsymbol{s}_i$ and its K-nearest neighbors as $C_{\boldsymbol{s}_i}$. Let $\sigma = C(0)$. The variance for each component of $\boldsymbol{\omega}$ can be calculated as: $\text{Var}(\omega_i) = \text{Cov}(\omega_i, \omega_i) = C(\|\boldsymbol{s}_i - \boldsymbol{s}_i\|) = C(0) = \sigma, \forall i = 1, \dots, n$. Then by matrix computation,

$$T_{ij} \approx \begin{cases} \sigma^2 - C_{\boldsymbol{s}_i}, & i = j; \\ \dfrac{1}{K}\left(C_{\boldsymbol{s}_i} - \sigma^2\right), & \boldsymbol{s}_j \in \boldsymbol{N}_K(\boldsymbol{s}_i) \text{ or } \boldsymbol{s}_i \in N_K\left(\boldsymbol{s}_j\right); \\ 0, & \text{Otherwise.} \end{cases}$$

Particularly, by assumption 1, if $i = j$, then $\boldsymbol{T}_{ij} = \sum_{t=1}^{n}\left[\boldsymbol{F}_{it} \cdot \boldsymbol{\Sigma}_{tj}\right] \approx \sigma^2 + K \cdot \left(-\frac{1}{K}C_{\boldsymbol{s}_i}\right) = \sigma^2 - C_{\boldsymbol{s}_i}$. If $i \neq j$ and $\boldsymbol{s}_j \in \boldsymbol{N}_K(\boldsymbol{s}_i)$ $\left(\text{or } \boldsymbol{s}_i \in N_K\left(\boldsymbol{s}_j\right)\right)$, then $\boldsymbol{T}_{ij} = \sum_k^n\left[\boldsymbol{F}_{ik} \cdot \boldsymbol{\Sigma}_{kj}\right] \approx \left[(K-1) \cdot \left(-\frac{1}{K}C_{\boldsymbol{s}_i}\right) + \left(-\frac{1}{K}\sigma^2\right)\right] + C_{\boldsymbol{s}_i} = \frac{1}{K}\left(C_{\boldsymbol{s}_i} - \sigma^2\right)$. For other cases, derived from the assumption 2, $\boldsymbol{T}_{ij} = \sum_t^n\left[\boldsymbol{F}_{it} \cdot \boldsymbol{\Sigma}_{tj}\right] = \sum_{\boldsymbol{s}_t \in \boldsymbol{N}_K(\boldsymbol{s}_i)}\left(-\frac{1}{K}C\left(\boldsymbol{s}_t - \boldsymbol{s}_j\right)\right) + C\left(\boldsymbol{s}_j - \boldsymbol{s}_i\right) \approx 0$.

As to the covariance matrix $\boldsymbol{D} = \boldsymbol{F\Sigma F^T} = \boldsymbol{TF}^T$, by matrix computation we have that

$$\boldsymbol{D}_{ij} \approx \begin{cases} \dfrac{1+K}{K}\left(\sigma^2 - C_{\boldsymbol{x}_i}\right), & i = j; \\ \dfrac{K+1}{K^2}\left(C_{\boldsymbol{x}_i} - \sigma^2\right), & \boldsymbol{s}_j \in N_K(\boldsymbol{s}_i) \text{ or } \boldsymbol{s}_i \in N_K\left(\boldsymbol{s}_j\right); \\ 0, & \text{Otherwise.} \end{cases}$$

Particularly, if $i = j$, then $\boldsymbol{D}_{ij} = \sum_t^n\left[\boldsymbol{T}_{it} \cdot [\boldsymbol{F}^T]_{tj}\right] \approx \sum_{t=1}^{K}\left(-\frac{1}{K} \cdot \frac{1}{K}\left(C_{\boldsymbol{x}_i} - \sigma^2\right)\right) + \left(\sigma^2 - C_{\boldsymbol{x}_i}\right) = \frac{1+K}{K}\left(\sigma^2 - C_{\boldsymbol{x}_i}\right)$. If $i \neq j, \boldsymbol{s}_j \in \boldsymbol{N}_K(\boldsymbol{s}_i)$, or $\boldsymbol{s}_i \in \boldsymbol{N}_K\left(\boldsymbol{s}_j\right)$, then $\boldsymbol{D}_{ij} \approx \left[\sum_{t=1}^{K-1}\left(-\frac{1}{K} \cdot \frac{1}{K}\left(C_{\boldsymbol{s}_i} - \sigma^2\right)\right) - \frac{1}{K}\left(\sigma^2 - C_{\boldsymbol{s}_i}\right)\right] + \frac{1}{K}\left(C_{\boldsymbol{x}_i} - \sigma^2\right) = \left(\frac{1}{K} + \frac{1}{K^2}\right)\left(C_{\boldsymbol{x}_i} - \sigma^2\right)$. For other cases, where $\boldsymbol{s}_j \notin \boldsymbol{N}_K(\boldsymbol{s}_i)$ and $\boldsymbol{s}_i \notin \boldsymbol{N}_K\left(\boldsymbol{s}_j\right)$, it has $\boldsymbol{D}_{ij} = \sum_t^n\left[\boldsymbol{T}_{it} \cdot [\boldsymbol{F}^T]_{tj}\right] = 0$. We prove this statement by contradiction. Assume that the value $\boldsymbol{D}_{ij}$ does not equal zero in this situation. Then there must be some $t \in \{1, \dots, n\}$ such that $\boldsymbol{T}_{it} \cdot [\boldsymbol{F}^T]_{tj} \neq 0$. This means $\boldsymbol{s}_t \in \boldsymbol{N}_k(\boldsymbol{s}_i)$ and $\boldsymbol{s}_t \in \boldsymbol{N}_K\left(\boldsymbol{s}_j\right)$. According to assumption 1, either $\boldsymbol{s}_i \in \boldsymbol{N}_K\left(\boldsymbol{s}_j\right)$ or $\boldsymbol{s}_j \in \boldsymbol{N}_K(\boldsymbol{s}_i)$ must be true, contradiction! Recall that $\boldsymbol{D} = \text{Var}(\boldsymbol{\omega}^*)$. The above results prove that $\text{Var}(\omega_i^*) = \boldsymbol{D}_{ii} \approx \frac{1+K}{K}$; $\rho\left(\omega_i^*, \omega_j^*\right) = \boldsymbol{D}_{ij}/\boldsymbol{D}_{ii} \approx -\frac{1}{K}$, if $\boldsymbol{s}_j \in N_K(\boldsymbol{s}_i)$ or $\boldsymbol{s}_i \in \boldsymbol{N}_K\left(\boldsymbol{s}_j\right)$; and $\rho\left(\omega_i^*, \omega_j^*\right) \approx 0$, in other cases. □

Theorem 2 indicates that the correlations between the components in $\boldsymbol{\omega}^*$ are mostly zero, except for neighboring points. Particularly, the correlations between neighboring points are all negative, and their major impact factor is the neighborhood size K. The greater the value of K, the less the neighbor points are correlated. However, K cannot be arbitrary large; otherwise, the assumptions made above will be violated. For example, suppose $n = 200$ and $K = 10$, then only about 5% of pairs are correlated. For these correlated components, the correlations are only close to -0.1. As shown in Figure 1, 0.1 indicates a negligible correlation.

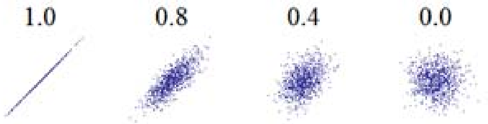


Figure 1: An example of correlation: it reflects the noise and direction of a linear relationship [13].

Theorem 2 states two approximate properties of $\boldsymbol{\omega}^*$. However, it is not directly known how these properties are impacted if assumptions 1 and 2 are violated. The next Theorem 3 will delve deeper into this issue and provide more specific analysis on $\boldsymbol{\omega}_i^*$. For Theorem 3, the following less restrictive assumptions are employed:

1. *The spatial locations* $\{s_1, \dots, s_n\}$ *follow a grid structure and* $n \leq 2500$*;*
2. *The spatial distance is defined by* L_2 *(Euclidean) distance;*
3. *The covariance function* $Cov\left(Z(s_i), Z\left(s_j\right)\right) = C(h)$*, where* $h = \left\|s_i - s_j\right\|_2$*, follows a popular spherical model;*
4. *Consider 4 or 12-nearest neighbors as spatial neighbors for each object.*

Assumptions 1 and 2 are generic properties that can be readily applied to spatial data in general [1, 2]. In many applications, the total number of spatial locations is smaller than 200. Here, we consider a much enlarged range with $n \leq 2500$, for the purpose of generality. For assumption 3, a spherical model is defined as

$$C(h; \boldsymbol{\theta}) = \begin{cases} b & if\ \ h = 0 \\ b\left(1 - \frac{3h}{2c} + \frac{1}{2}\left(\frac{h}{c}\right)^3\right) & if\ 0 < h \leq c \\ 0 & if\ h > c, \end{cases} \qquad (8)$$

where $\boldsymbol{\theta} = (b, c)^T, b \geq 0, c \geq 0$. $b = C(0; \boldsymbol{\theta})$ refers to the constant variance for each object $\boldsymbol{s}$, and $C(h; \boldsymbol{\theta})$ is a decreasing function on the distance h.

The reason for using a spherical model as opposed to exponential or Gaussian models is that the spherical model leads to closed-form analytical results. The closed-form results will provide important insights into its statistical properties. As for assumption 4, K is set to 4 or 12 due to the use of the grid structure (assumption 1). In the grid, each object has four nearest objects with the same distance r and eight next-nearest objects with the same distance $2r$, where r is the grid cell size, and so on. Hence, we can select $K = 4, 12, 24, \dots$ We select the first two values with $K = 4$ and 12, which are equivalent to defining neighborhoods with radiuses of r and $2r$, respectively.

To make the results concise, we further set $r^2 h/c^3 \approx 0$ and $r^3/c^3 \approx 0$, since r/c is usually very small (e.g., 0.1) and h ≤ c. If $h > c$, then $C(h; \boldsymbol{\theta}) = 0$ and will lead to zero covariance. These components are negligible compared to the components r/c and rh^2/c.

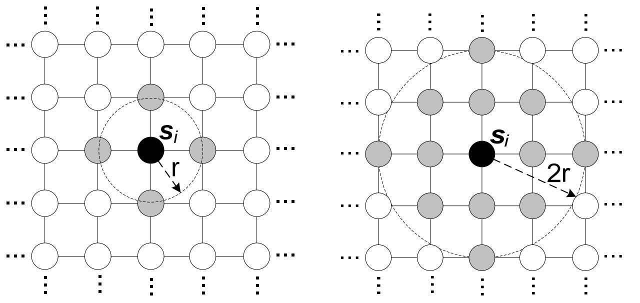
Theorem 3: *Under the above four assumptions, the random vector $\boldsymbol{\omega}^*$ has following properties on the correlation structure*

1) *If $K = 4$, then*
   a) $\rho(\omega_i^*, \omega_j^*) = 0, \quad if\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) > c + 2r,$
   b) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.4, \quad if c \leq 2r\ and\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) \leq 2r,$
   c) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.22, \ if\ c > 2r\ and\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) \leq 2r,$
   d) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.05, \ if\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) > 2r.$
2) *If $K = 12, d(\boldsymbol{s}_j, \boldsymbol{s}_i) > c + 4r$, then $\rho(\omega_i^*, \omega_j^*) = 0$*
3) *If $K = 12, c < 4r$, then*
   a) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.220, \ if\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) \leq 2r$
   b) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.110, \ if\ 2r < d(\boldsymbol{s}_j, \boldsymbol{s}_i) \leq 3r$
   c) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.050, \ if\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) > 3r$
4) *If $K = 12, c \geq 4r\ and\ row(\boldsymbol{s}_j) = row(\boldsymbol{s}_i)$ $\left(or\ col(\boldsymbol{s}_j) = col(\boldsymbol{s}_i)\right)$, then*
   a) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.4741 - \frac{0.1179 \cdot c^2/r^2}{1+c^2/(2.707 \cdot r^2)}, \ if\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) = r$
   b) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.1203, \ if\ d(\boldsymbol{s}_j, \boldsymbol{s}_i) = 2r$
   c) $|\rho(\omega_i^*, \omega_j^*)| \leq 0.1719 - \frac{0.0158 \cdot h_{ij}^2/r^2}{1+c^2/(10.5174 \cdot r^2)}$, *otherwise.*
5) *If $K = 12, c \geq 4r, row(\boldsymbol{s}_j) \neq row(\boldsymbol{s}_i),\ and\ col(\boldsymbol{s}_j) \neq col(\boldsymbol{s}_i)$,*

   then $|\rho(\omega_i^*, \omega_j^*)| \leq 0.1085 - \frac{0.0028 \cdot h_{ij}^2/r^2}{1+h_{ij}^2/(37.6723 \cdot r^2)}$,

where r refers to the grid cell size; $row(\boldsymbol{s}_i)$ and $col(\boldsymbol{s}_i)$ refer to the row and column locations of the object $\boldsymbol{s}_i$ in the grid structure; $h_{ij} = d(\boldsymbol{s}_j, \boldsymbol{s}_i)$ is the L2 (or Euclidean) distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.

Proof: The neighborhoods topologies defined by 4 and 8-nearest-neighbors rules are shown in Figure 2. The grayed objects are the spatial neighbors of the black object $\boldsymbol{s}_i$. The symbol r refers to the grid cell size.



$(a)\ K = 4$ $(b)\ K = 12$

Figure 2: The neighborhoods defined by 4 or 12-nearest-neighbors rules in gridded data, equal to those defined by radiuses r and $2r$.

Recall that $\boldsymbol{\omega}^* = \boldsymbol{F}\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is the smooth scale variation (see Section 3.1 equation (2)). Let $\boldsymbol{\Sigma} = \mathrm{Var}(\boldsymbol{\omega}), \boldsymbol{D} = \mathrm{Var}(\boldsymbol{\omega}^*) = \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^T$, and $\boldsymbol{T} = \boldsymbol{F}\boldsymbol{\Sigma}$. By assumption 3, $\boldsymbol{\Sigma}_{ij} = \mathrm{Cov}(\omega_i, \omega_j) = C(h_{ij}; \boldsymbol{\theta}) = C(h_{ij}; \boldsymbol{\theta}) - \frac{1}{K}\sum_{\boldsymbol{s}_t \in \boldsymbol{N}_K(\boldsymbol{s}_i)} C(h_{tj}; \boldsymbol{\theta})$. Given that $\boldsymbol{F}$ is a neighborhood weight matrix (see equation (4)), the component $\boldsymbol{T}_{ij} = \sum_{t=1}^{n}(\boldsymbol{F}_{it} \cdot \boldsymbol{\Sigma}_{tj})$. By the relation $\boldsymbol{D} = \boldsymbol{T}\boldsymbol{F}^T$, we have that $\boldsymbol{D}_{ij} = \boldsymbol{T}_{ij} - \frac{1}{K}\sum_{\boldsymbol{s}_t \in \boldsymbol{N}_K(\boldsymbol{s}_j)} \boldsymbol{T}_{it}$. The correlation $\rho(\omega_i^*, \omega_j^*)$ has the analytical form

$$\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta}) = \frac{\boldsymbol{D}_{ij}}{\boldsymbol{D}_{ii}} = \frac{\boldsymbol{T}_{ij} - \frac{1}{K}\sum_{\boldsymbol{s}_t \in \boldsymbol{N}_K(\boldsymbol{s}_j)} \boldsymbol{T}_{it}}{\boldsymbol{D}_{11}}, \tag{9}$$

where $\boldsymbol{D}_{ii}$ is constant and the same denominator $\boldsymbol{D}_{11}$ is used for different $\boldsymbol{D}_{ii}$. Notice that the form (9) is actually the sum of K^2 weighted spherical functions $\left(C(\cdot, \boldsymbol{\theta})\right)$. This complex form makes the function properties not well interpretable, such as the minimum value, the maximum value, and the global trend with respect to the major parameters h_{ij} and c. For this reason, we further develop a tight upper bound function of (9) that is monotone and has a simpler analytical form. The development is based on five different cases as indicated in Theorem 3. Here we focus on two representative cases, the second and the fifth cases. The upper bound functions for other cases can be proved similarly.

- **Case 1**: $K = 12$ and $d(\boldsymbol{s}_j, \boldsymbol{s}_i) > c + 4r$.

It has $C(h_{ij}; \boldsymbol{\theta}) = 0$ and $C(h_{td}; \boldsymbol{\theta}) = 0, \forall \boldsymbol{s}_t \in \boldsymbol{N}_K(\boldsymbol{s}_j) \cup \{\boldsymbol{s}_j\}, \forall s_d \in \boldsymbol{N}_K(\boldsymbol{s}_i) \cup \{\boldsymbol{s}_i\}$. It implies that $\rho(\omega_i^*, \omega_j^*) = 0$.

- **Case 5**: $K = 12, \mathrm{c} > 4r, \mathrm{row}(\boldsymbol{s}_j) \neq \mathrm{row}(\boldsymbol{s}_i)$, and $\mathrm{col}(\boldsymbol{s}_j) \neq \mathrm{col}(\boldsymbol{s}_i)$.

Based on the observations by visualization, we select a rational quadratic model $\left(f(h;\ \boldsymbol{\alpha}) = \alpha_1 + \frac{\alpha_2 h^2}{1+h^2/\alpha_3}\right)$ for the upper bounding function. The estimation of the parameters $\boldsymbol{\alpha}$ is based on the following steps:

Step 1: Let $\boldsymbol{S}_1 = \{1,2,3, \dots, 49\}$, $\boldsymbol{S}_2 = \{1,2,3, \dots, 49\}$, and $\boldsymbol{S}_3 = \{4,5,6, \dots, 15, 20, 40, 60, 80\} \subset \boldsymbol{S}_c = \{c | c \in \mathbb{R}, c \geq 4\}$.

Step 2: Solve the following optimization problem

$$\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^4} \sum_{\substack{\mathrm{row}(\boldsymbol{s}_j) - \mathrm{row}(\boldsymbol{s}_i) \in \boldsymbol{S}_1, \\ \mathrm{col}(\boldsymbol{s}_j) - \mathrm{col}(\boldsymbol{s}_i) \in \boldsymbol{S}_2, \\ c \in \boldsymbol{S}_3}} \left(f(h_{ij};\ \boldsymbol{\alpha}) - |\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|\right) \tag{10}$$

subject to

$$f(h_{ij};\ \boldsymbol{\alpha}) \geq |\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|, \quad \forall i, j, c \text{ with } \mathrm{row}(\boldsymbol{s}_j) - \mathrm{row}(\boldsymbol{s}_i) \in \boldsymbol{S}_1, \mathrm{col}\ (\boldsymbol{s}_j) - \mathrm{col}\ (\boldsymbol{s}_i) \in \boldsymbol{S}_2, \text{and } c \in \boldsymbol{S}_3,$$

where $\boldsymbol{\theta} = (b, c)$ and $b = 1$.

Step 3: $\forall (i, j) \in \boldsymbol{S}_1 \times \boldsymbol{S}_2$, solve the following optimization problem

$$\hat{c}_{ij} = \arg\min_{\mathrm{c} \in \mathbb{R}, \mathrm{c} \geq 4} \left(f(h_{ij};\ \widehat{\boldsymbol{\alpha}}) - |\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|\right), \tag{11}$$

where $\boldsymbol{\theta} = (b, c)$ and $b = 1$.

Step 4: If $\forall (i, j) \in \boldsymbol{S}_1 \times \boldsymbol{S}_2$, it satisfies the condition $f(h_{ij};\ \widehat{\boldsymbol{\alpha}}) - \rho\left(\omega_i^*, \omega_j^*; \boldsymbol{\theta} = (1, \hat{c}_{ij})\right) \geq 0$, then return $\widehat{\boldsymbol{\alpha}}$ as the estimated values of $\boldsymbol{\alpha}$ and terminate the algorithm; Otherwise, select a larger subset (e.g., $\boldsymbol{S}_3 = \{1,2,3, \dots, 100\}$) of the feasible set $\{x | x \in \mathbb{R}, x \geq 4\}$ for the parameter c, and go to step 2.

The objective of the above algorithm is to estimate a local optimal setting for $\boldsymbol{\alpha}$. Particularly, by assumption 1, the spatial locations follow a grid structure and the total number of points is smaller than 2500. It implies that the set $\boldsymbol{S}_1 \times \boldsymbol{S}_2$ includes all valid settings for the pair $\left(\text{row}(\boldsymbol{s}_j) - \text{row}(\boldsymbol{s}_i), \text{col}(\boldsymbol{s}_j) - \text{col}(\boldsymbol{s}_i)\ \right)$. The feasible set of the parameter c is $\boldsymbol{S}_c = \{c | c \in \mathbb{R}, c \geq 4\}$. At step one, we only select a representative subset ($\boldsymbol{S}_3$) of $\boldsymbol{S}_c$. The optimization problem (10) is to find a tight upper bound function based on the subset $\boldsymbol{S}_3$. Steps 2 and 3 test if the estimated parameters $\widehat{\boldsymbol{\alpha}}$ satisfy the upper bounding conditions that $f(h_{ij};\ \widehat{\boldsymbol{\alpha}}) \geq \rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})$ for every valid settings of $i, j,$ and c. If the test is passed, then we can conclude that a feasible and local optimal $\hat{\alpha}$ is obtained. Otherwise, the algorithm will start a new iteration based on an enlarged subset of $\boldsymbol{S}_c$.

The optimization problem (10) is a nonconvex problem. A local optimal solution of (10) can be obtained by numerical methods, such as interior point method [14]. The estimated parameters $\widehat{\boldsymbol{\alpha}} = (0.1085, -0.0028, 37.6723)$. A local optimal solution of $(A2)$ is acceptable for us, since our objective is to find a tight upper bound function, but not necessarily a global optimal bound.

The optimization problem (11) is also a non-convex problem. Because it is a feasibility testing procedure, a global optimal solution must be obtained. This can be achieved by exploring the special structure of (11). Particularly, first the denominator of $\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})$ is $\boldsymbol{D}_{11}$. By the equation $r^3/c^3 \approx 0$, it follows that $\boldsymbol{D}_{11} = \tau r/c$, where τ is some scalar constant. Recall that the numerator of $\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})$ is a weighted sum of 144 spherical functions. Let $S = \left\{h_{td} \mid \boldsymbol{s}_t \in \boldsymbol{N}_K(\boldsymbol{s}_j) \cup \{\boldsymbol{s}_j\}, \boldsymbol{s}_d \in \boldsymbol{N}_K(\boldsymbol{s}_i) \cup \{\boldsymbol{s}_i\}\right\}$. The set S has totally 144 components (scalars), which can be used to divide the feasible region $\boldsymbol{S}_c = \{c | c \in \mathbb{R}, c \geq 4\}$ into 145 sub-regions. It can be readily derived that, in each sub-region, the absolute value of the correlation $\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})$ has the polynomial form $\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta}) = \tau_1 + \tau_2 \cdot \frac{1}{c} + \tau_3 \cdot \frac{1}{c^2}$, where $\tau_1, \tau_2,$ and τ_3 are constant scalars depending on this sub-region. By this polynomial form, we have that $\left|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})\right|$ only has one local (global) maximum in each sub-region. By checking the maximum value in each region, we can obtain a global optimal solution for the problem (11).

- **Other Cases**:

The upper bound functions can be obtained by using similar procedures in cases 1 and 5.

The complete form of the estimated upper bound function is stated in Theorem 3. Readers are referred to Appendix for an empirical plot of the estimated bounds. □

Theorem 3 implies similar patterns as drawn by Theorem 2 although Theorem 2 provides only approximate properties. Theorem 3 is a further justification of these patterns. In the following discussions, we consider the situation with $c \geq 5$. The situation with $c < 5$ will be discussed separately. By Theorem 3, if $c \geq 5$, then $\left|\rho(\omega_i^*, \omega_j^*)\right| \leq 0.22$ when $K = 4$; and $\left|\rho(\omega_i^*, \omega_j^*)\right| \leq 0.18$ when $K = 12$. It indicates small absolute correlation values for different K values. The correlation values slightly decreases when K increases. It can also be shown that most correlations are negative and are close or equal to zero. Readers are referred to the Appendix for more detailed information about $\rho(\omega_i^*, \omega_j^*)$. All these observations are consistent with the results from Theorem 2.

We have a comparison between $\sigma_0^2\boldsymbol{FF}^T$ and $\boldsymbol{F\Sigma F}^T$**.** Consider two typical situations: $K = 4$ to represent a small neighborhood; and $K = 12$ to represent a relatively large neighborhood. If $K = 4$, then $\left|\rho(e_i^*, e_j^*)\right| \leq 0.4$ and $\left|\rho(\omega_i^*, \omega_j^*)\right| \leq 0.22$. If $\mathrm{K} = 12$, then $\left|\rho(e_i^*, e_j^*)\right| \leq 0.2$ and $\left|\rho(\omega_i^*, \omega_j^*)\right| \leq 0.18$. The impacts of these correlation values (degrees) are shown in Figure 1. Although both $\left|\rho(e_i^*, e_j^*)\right|$ and $\left|\rho(\omega_i^*, \omega_j^*)\right|$ increase when the neighborhood size K decreases, the absolute correlation $\left|\rho(e_i^*, e_j^*)\right|$ increases more drastically. Based on these results, we will approximate $\boldsymbol{F\Sigma F}^T$ by $\sigma^2\boldsymbol{I}$ for different settings of K, but will only approximate $\sigma_0^2\boldsymbol{FF}^T$ by $\sigma_0^2\boldsymbol{I}$, when K is relatively large, such as $K \geq 8$.

Theorem 3 also indicates that when c is small $(\text{e.g.}, c < 5r)$, some correlations are relatively high (e.g., $\left|\rho(\omega_i^*, \omega_j^*)\right| = 0.4$ if $K = 4, c = 1r,$ and $d(\boldsymbol{s}_j, \boldsymbol{s}_i) = r$). In this case, an important observation is that the correlation matrix of $\boldsymbol{\omega}^*$ exhibits similar structure as that of $\boldsymbol{e}^*$. Particularly, if $c < r$, these two correlation matrices become identical. In this situation, it is still reasonable to approximate the correlation matrix of $\boldsymbol{\omega}^*$ as identity or unit matrix, since the lost structure information by this approximation will be recovered while estimating the parameter σ_0 for the vector $\boldsymbol{e}^*$, because of the similar structure between the covariance matrices $\text{Var}(\boldsymbol{\omega}^*)$ and $\text{Var}(\boldsymbol{e}^*)$. For example, suppose $c < r$ and the constant variance for each component of $\boldsymbol{e}$ is σ_e^2, then we have that $\text{Var}(\boldsymbol{e}) = \boldsymbol{\Sigma} = \sigma_e^2\boldsymbol{I}$, and $\text{Var}(\boldsymbol{e}^*) = \text{Var}(\boldsymbol{Fe}) = \boldsymbol{F\Sigma F}^T = \sigma_e^2\boldsymbol{FF}^T$. By the equation (5), the true distribution model is: $\mathbf{diff}(\boldsymbol{Z}) \sim \mathbf{N}(\boldsymbol{FX\beta}, \boldsymbol{F\Sigma F}^T + \sigma_0^2\boldsymbol{FF}^T) = \mathbf{N}(\boldsymbol{FX\beta}, (\sigma_0^2 + \sigma_e^2)\boldsymbol{FF}^T)$**.** If we approximate $\boldsymbol{F\Sigma F}^T$ as $\sigma^2\boldsymbol{I}$ instead, then by the equation (6) the approximate model becomes $\mathbf{diff}(\boldsymbol{Z}) \sim \mathbf{N}(\boldsymbol{FX\beta}, \sigma^2\boldsymbol{I} + \sigma_0^2\boldsymbol{FF}^T)$. By robust parameter estimation, the approximate model can still completely recover the true distribution, ex., by setting the estimated parameters $\hat{\sigma} = 0$ and $\hat{\sigma}_0 = \sqrt{\sigma_0^2 + \sigma_e^2}$.

4. ESTIMATION AND INFERENCES

Spatial outlier detection (*SOD*) is usually coupled with a robust estimation process for the related statistical model. This section introduces ordinary estimation methods for the *GLS* model, then presents two robust estimation and outlier detection methods to reduce the masking and swamping effects, and discusses the connection between the proposed *GLS-SOD* methods with existing representative methods, such as kriging-based and Z-test *SOD* methods.

4.1 Generalized Least Squares Regression

Given a set of observations$\{Z(\boldsymbol{s_1}), Z(\boldsymbol{s_2}), \ldots, Z(\boldsymbol{s_n})\ \}$, the objective is to estimate the parameters $\boldsymbol{\beta}, \sigma,$ and σ_0 for the proposed *GLS* model. We consider mean squared error (MSE) as the score function which is the most popular error function in spatial statistics [11]. This leads to a generalized least square problem and can be formulated as:

$$\arg\min_{\beta,\sigma_0,\sigma} \left[(\boldsymbol{FZ} - \boldsymbol{FX\beta})^T\left(\sigma^2\boldsymbol{I} + \sigma_0^2\boldsymbol{FF}^T\right)^{-1}(\boldsymbol{FZ} - \boldsymbol{FX\beta})\right], \tag{12}$$

subject to $\sigma_0 +\ \sigma = 1$ and $\sigma_0, \sigma \geq 0$.

Note that we scale σ_0 and σ by a factor c with $\sigma_0^* = \sigma_0/c$ and $\sigma^* = \sigma/c$, such that $\sigma_0^* +\ \sigma^* = 1$. Without this constraint, the objective function in (12) will always be minimized by setting $\sigma_0 = \sigma = \infty$, and $\boldsymbol{\beta}$ to any value. For simplicity, we directly use the original symbols σ_0 and σ, rather than σ_0^* and σ^*. As shown in Theorem 4, the problem (12) is a convex optimization problem which can be solved efficiently by numerical optimization methods such as interior point method [14]. Note that when the neighborhood size (i.e., K) is large, the following holds: $\sigma_0^2\boldsymbol{FF}^T \approx \sigma_0^2\boldsymbol{I}$ (see Section 3.2). Then (12) reduces to a regular least squares regression problem and an explicit solution is available with $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{Z}$, and $(\sigma^2 + \sigma_0^2) = \|\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{F}\boldsymbol{Z}\|_2^2/(n-p-1)$, where p is the size of the vector $\boldsymbol{\beta}$. For the purpose of outlier detection, it is unnecessary to further derive the explicit forms of σ and σ_0.

Theorem 4: *The problem (12) is a convex optimization problem.*

Proof Sketch: Suppose λ_i and $\boldsymbol{q}_i$ are the eigenvalues and corresponding (orthonormal) eigenvectors of the matrix $\boldsymbol{F}\boldsymbol{F}^T$. It can be readily shown that the problem (12) is equivalent to

$$\arg\min_{\boldsymbol{\beta},\sigma_0,\sigma} \left[\sum_{i=1}^{n} \frac{\{(\boldsymbol{F}\boldsymbol{Z}-\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{q}_i\}^2}{\sigma^2+\sigma_0^2\lambda_i}\right],\ \text{s.t.}\ \ \sigma_0, \sigma \geq 0 \tag{13}$$

Let $f_i = \frac{\{(\boldsymbol{F}\boldsymbol{Z}-\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{q}_i\}^2}{\sigma^2+\sigma_0^2\lambda_i}$, It suffices to prove that f_i is a convex function, or equivalently $\frac{\partial^2 f_i}{\partial \boldsymbol{\theta}^2} \succcurlyeq 0$, $\boldsymbol{\theta} = [\boldsymbol{\beta}^T, \sigma^2, \sigma_0^2]^T$.

$$\frac{\partial^2 f_i}{\partial \boldsymbol{\theta}^2} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{F}\boldsymbol{q}_i(\sigma^2+\sigma_0^2\lambda_i) \\ \left(\boldsymbol{q}_i^T Z - \boldsymbol{q}_i^T\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta}\right)^T \\ \lambda_i\left(\boldsymbol{q}_i^T Z - \boldsymbol{q}_i^T\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta}\right)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{F}\boldsymbol{q}_i(\sigma^2+\sigma_0^2\lambda_i) \\ \left(\boldsymbol{q}_i^T \boldsymbol{Z} - \boldsymbol{q}_i^T\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta}\right)^T \\ \lambda_i\left(\boldsymbol{q}_i^T \boldsymbol{Z} - \boldsymbol{q}_i^T\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta}\right)^T \end{bmatrix}^T \succcurlyeq 0. \qquad \square$$

When the parameters $\boldsymbol{\beta}, \sigma,$ and σ_0 are estimated by generalized least squares, we can calculate the standard residuals and use standard statistic test procedure to identify the outliers. This method works well for sample data with small data contamination, but is susceptible to the well-known masking and swamping effects when multiple outliers exist. For the *GLS* model, the masking and swamping effects originate from two phases of the estimation process:

1) **Phase I contamination** occurs in the process of calculating local differences $\boldsymbol{F}\boldsymbol{Z}$. For example, suppose we define neighbors by the K-nearest-neighbor rule. Consider an outlier object $Z^*(\boldsymbol{s}_1) = Z(\boldsymbol{s}_1) + \zeta_1$, where $Z(\boldsymbol{s}_1)$ is the normal value but it is contaminated by a large error ζ_1 , and suppose only one of its neighbors is an outlier with $Z^*(\boldsymbol{s}) = Z(\boldsymbol{s}) + \zeta$, where ζ is the error. The local difference $\text{diff}\big(Z^*(\boldsymbol{s}_1)\big) = \left[Z(\boldsymbol{s}_1) - \frac{1}{K}\sum_{\boldsymbol{s}_i\in \boldsymbol{N}(s)}\big(Z(\boldsymbol{s}_i)\big)\right] + \zeta_1 - \frac{\zeta}{K}$. If $\zeta = K \cdot \zeta_1$, then the error is marginalized and we obtain a normal local difference for a outlier object $Z^*(\boldsymbol{s}_1)$ which will be identified as a normal object. If $Z^*(\boldsymbol{s}_1)$ is a normal object with $\zeta = 0$, then the related local difference is contaminated by the error $-\frac{\zeta}{\mathrm{K}}$. This leads to the swamping effect where the normal object $Z^*(\boldsymbol{s}_1)$ may be misclassified as an outlier. For a relatively large K (e.g., 8), it can be readily shown that Phase I contamination is more significant for a spatial sample with clusters of outliers than a spatial sample with isolated outliers. Another important observation is that the masking and swamping effects will not completely distort the ordering of true outliers. The top ranking outliers are still usually a subset of the true outliers. This observation motivates the backward algorithm presented in Section 4.3. 2) **Phase II contamination** occurs in the generalized regression process, where we regard $\boldsymbol{Z}^* = \boldsymbol{F}\boldsymbol{Z}$ as the pseudo "observed" values. The masking and swamping effects in this phase are the same effects occurred in a general least squares regression process. This is consequence of the biased estimates of the regression parameters (e.g., $\boldsymbol{\beta}$, σ , and σ_0) due to abnormal observations in $\boldsymbol{Z}^*$.

Drawbacks of existing robust estimation techniques:

Most existing robust regression techniques are designed to reduce the effect of Phase II contamination. There are two major categories of estimators [13]. The first category (also called M-estimators) is to replace the MSE function by more robust score function such as L1 norm and Huber penalty function. The second category is to estimate parameters based on a robustly selected subset of data, such as least median of square (*LMS*), least trimmed square (*LTS*), and the recently proposed forward search (*FS*) method. Unfortunately, all these robust techniques cannot be directly applied to address both Phase I and Phase II contaminations concurrently. As with the M-estimators, the application of robust penalty function (e.g., L1) will lead to a non-convex optimization problem where local optimal solution may be found. With the second type of estimators based on subset selection, the estimation results are highly sensitive to the selected objects which can detrimentally impact neighborhood quality. The next sub-section will adapt existing robust methods to the problem of concurrently handling Phase I and Phase II contaminations.

4.2 *GLS*-Backward Search Algorithm

As discussed above, the existing methods only address the Phase II contamination. The motivation for our proposed backward search algorithm is to address both Phase I and Phase II contaminations concurrently. The algorithm is described as follows:

Algorithm 1 (**Backward search algorithm**) Given a spatial data set $\{Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)\ \}$, the covariate vectors $\{\boldsymbol{x}(s_1), \ldots, \boldsymbol{x}(\boldsymbol{s}_n)\}$, the value of K for defining K-nearest neighbors, and the confidence interval $\alpha \in (0,1)$,

1. Set $\boldsymbol{S}_Z = \{Z(\boldsymbol{s_1}), \ldots, Z(\boldsymbol{s_n})\ \}, \boldsymbol{S}_x = \{\boldsymbol{x}(s_1), \ldots, \boldsymbol{x}(\boldsymbol{s}_n)\}$, and S_{output} be an empty set.

2. Estimate the parameters $\boldsymbol{\beta}, \sigma, \sigma_0$ of the *GLS* model by solving the generalized least squares regression problem (12).

3. Calculate the absolute values of standard estimated residuals
$$\boldsymbol{e} = \left[e_1, \ldots, e_{|S_Z|}\right]^T = \left|\left(\sigma^2\boldsymbol{I} + \sigma_0^2\boldsymbol{F}\boldsymbol{F}^T\right)^{-\frac{1}{2}}(\boldsymbol{F}\boldsymbol{Z} - \boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta})\right|$$

4. Set $e_m = \max\{e_i\}_{i=1}^{|\boldsymbol{S}_Z|}$.

   If $e_m \geq \Phi^{-1}(\alpha/2)$, where Φ is the *CDF* of the standard normal distribution, then update $\boldsymbol{S}_Z = \boldsymbol{S}_Z - \{Z(\boldsymbol{s}_m)\}$, $\boldsymbol{S}_x = \boldsymbol{S}_x - \{\boldsymbol{x}(\boldsymbol{s}_m)\}$, and $\boldsymbol{S}_{output} = \boldsymbol{S}_{output} + \{Z(\boldsymbol{s}_m)\}$, and go to Step 2.

   Otherwise, stop the algorithm and return $\boldsymbol{S}_{output}$ as the ordered set of candidate outliers.

In the above algorithm, the confidence interval α can be set to 0.001, 0.01, and 0.05. In step 2, we apply interior point [14] method to solve the optimization problem (12). When the neighborhood size is large, we may approximate $\sigma_0^2\boldsymbol{F}\boldsymbol{F}^T$ as $\sigma_0^2\boldsymbol{I}$. The parameters $\boldsymbol{\beta}, \sigma, \sigma_0$ can be efficiently estimated by least squares regression: $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{F}^T\boldsymbol{F}\boldsymbol{Z}$, and $(\sigma^2 + \sigma_0^2) = \|\boldsymbol{F}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{F}\boldsymbol{Z}\|_2^2/(n-p-1)$, where p is the size of the vector $\boldsymbol{\beta}$.

This backward search algorithm's design is based on the observation that top ranked outliers identified by the regular least squares method are still true outliers (in most cases) under both Phase I and II contaminations. Suppose a true outlier $\boldsymbol{s}$ is removed after the first iteration, then both Phase I and Phase II contaminations in the next iteration will be reduced. To illustrate this process, we use the same example in Section 4. Recall that an outlier object $Z^*(\boldsymbol{s})$ is decomposed into two additive components $Z^*(\boldsymbol{s}) = Z(\boldsymbol{s}) + \zeta$, where $Z(\boldsymbol{s})$ represents the normal value and ζ represents the contamination error. Suppose $\boldsymbol{s}$ is the only outlier neighbor of an object $\boldsymbol{s}_1$ that happens to be an outlier. Then the local difference $\text{diff}\big(Z^*(\boldsymbol{s}_1)\big) = \left[Z(\boldsymbol{s}_1) - \frac{1}{K}\sum_{\boldsymbol{s}_i\in \boldsymbol{N}(s)}\big(Z(\boldsymbol{s}_i)\big)\right] + \zeta_1 - \frac{\zeta}{K}$ will be marked as normal if $\zeta = K \cdot \zeta_1$. Suppose now that the true outlier $Z(\boldsymbol{s})$ is removed and the newly replaced neighbor for $\boldsymbol{s}_1$ is normal, then $\text{diff}\big(Z^*(\boldsymbol{s}_1)\big) = \Big[Z(\boldsymbol{s}_1) - \frac{1}{K}\sum_{\boldsymbol{s}_i \in \boldsymbol{N}(\boldsymbol{s})}\big(Z(\boldsymbol{s}_i)\big)\Big] + \zeta_1$. This local difference becomes an abnormal value and the masking effect is removed. Similarly, suppose $Z^*(\boldsymbol{s}_1)$ is a normal object, then its local difference is contaminated (swamped) by the error $-\frac{\zeta}{K}$, because of its outlier neighbor $Z(\boldsymbol{s})$. The removal of $\boldsymbol{s}$ will make $-\frac{\zeta}{K} = 0$ and therefore reducing the swamping effect. For Phase II contamination, the removal of $Z(\boldsymbol{s})$ leads to the removal of an abnormal difference $\text{diff}\big(Z^*(\boldsymbol{s})\big)$. The set of remaining local differences will therefore have less contamination. The center of the distribution is less attracted by outliers, and the distributional shape becomes less distorted. As a result, outliers tend to be more separated and normal objects tend to be closer together. The masking and swamping effects are therefore reduced.

4.3 *GLS*-Forward Search Algorithm

This section adapts the popular Forward Search (*FR*) algorithm [13] to the *GLS* parameters estimation problem. There are several restrictions to apply *FR* here. As discussed in Section 4.1, *FR* starts from a robustly select subset of sample, but *GLS* is a statistical model based on neighborhood aggregations. Considering only a subset of the observations $\{Z(s_1), \ldots, Z(s_n)\}$ will significantly impact the quality of the calculated local differences. To apply *FR* algorithm, we make the assumption that Phase I contamination is negligible compared to Phase II contamination. As discussed in Section 4.1, this is reasonable for the case of isolated outliers. Based on this assumption, we consider the local differences $\big\{\text{diff}\big(Z(s_1)\big), \ldots, \text{diff}\big(Z(s_n)\big)\big\}$ as pseudo "observations", and then apply *FR* algorithm to estimate the model parameters. By simulations, we also noticed that in this case there is no significant difference between applying generalized least squares regression and regular least squares regression. For the sake of efficiency, we only apply regular least squares regression to estimate the parameters β, σ, and σ_0. The *FR* algorithm is described as follows:

Algorithm 2 (**Forward Search algorithm**) Given a spatial data set $\{Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)\}$, the covariate vectors $\{\boldsymbol{x}(s_1), \ldots, \boldsymbol{x}(\boldsymbol{s}_n)\}$, and the value of *K* for defining *K*-nearest neighbors,

1. Calculate the local differences: $\mathbf{diff}(\boldsymbol{Z}) = \boldsymbol{FZ}$, and set $\boldsymbol{S}_{output}$ be an empty set.
2. Set $\boldsymbol{S} = \{\boldsymbol{s}_1, .., \boldsymbol{s}_n\}$; Set $\boldsymbol{Z}^*(\boldsymbol{S}) = [Z^*(\boldsymbol{s}_1), \ldots, Z^*(\boldsymbol{s}_n)] = \mathbf{diff}(\boldsymbol{Z})$ and $\boldsymbol{X}^*(\boldsymbol{S}) = [\boldsymbol{x}^*(\boldsymbol{s}_1), \ldots, \boldsymbol{x}^*(\boldsymbol{s}_n)] = \boldsymbol{FX}$ as the vector of pseudo "observations" and pseudo "covariates".
3. Apply least trimmed squares (LTS) [13] to find a robust subset of $\boldsymbol{S}$, defined as $\boldsymbol{S}^*$, and set $\boldsymbol{S}^*_{test} = \boldsymbol{S} - \boldsymbol{S}^*$. The size of the subset $\boldsymbol{S}^*$ is $\lfloor (n + p + 1)/2 \rfloor$ by default.
4. Estimate the parameter $\boldsymbol{\beta}$ based on $\boldsymbol{Z}^*(\boldsymbol{S}^*)$ and $\boldsymbol{X}^*(\boldsymbol{S}^*)$. Then calculate the absolute standard residuals of $\boldsymbol{S}^*_{test}$ as $\boldsymbol{e} = \sqrt{(n-p-1)}|\boldsymbol{Z}^*(\boldsymbol{S}^*_{test}) - \boldsymbol{X}^*(\boldsymbol{S}^*_{test})\boldsymbol{\beta}| / \|\boldsymbol{Z}^*(\boldsymbol{S}) - \boldsymbol{X}^*(\boldsymbol{S})\boldsymbol{\beta}\|_2$.
5. Find the minimal residual of the test set $\boldsymbol{S}^*_{test}$:

   $e_m = \min\{e_i\}_{e_i \in S^*_{test}}$.
6. Update $\boldsymbol{S}_{output} = \boldsymbol{S}_{output} + \{\boldsymbol{s}_m\}, \boldsymbol{S}^* = \boldsymbol{S}^* + \{\boldsymbol{s}_m\}, \boldsymbol{S}^*_{test} = \boldsymbol{S}^*_{test} - \{\boldsymbol{s}_m\}$. If $\boldsymbol{S}^*_{test}$ is not empty, go to step 4; otherwise, output the ordered set $\boldsymbol{S}_{output}$ and terminate the algorithm.

The proposed *FR* algorithm provides an ordering of objects based on their agreements with the *GLS* model. To identify outliers, it plots and monitors the change of the minimal residual with the increasing size of the normal set $\boldsymbol{S}^*$. A drastic drop implies that an outlier was added to $\boldsymbol{S}^*$. This plot could also help identify masked or swamped objects. Readers are referred to [13] for details. A direct method for calculating the local differences can be achieved via robust mean functions such as median and trimmed mean. However, as indicated by our simulation study, this direct approach will deteriorate the performance of *GLS*. Recall that the statistical model of *GLS*: $\mathbf{diff}(\boldsymbol{Z}) \sim \mathrm{N}(\boldsymbol{FX\beta}, \boldsymbol{F\Sigma F}^T + \sigma_0^2\boldsymbol{FF}^T)$. If we replace the left hand side $\mathbf{diff}(\boldsymbol{Z}) = \boldsymbol{FZ}$ by medians or trimmed means, the right side will remain unchanged and thus still employs the average matrix $\boldsymbol{F}$. The increased bias caused by this inconsistency is much larger than the reduction of contamination effects achieved through robust means.

4.4 Connections with Existing Methods

This section studies the connection between global (kriging) based [11, 12, 13], local spatial statistics (*LS*) based methods [4-10], and the proposed *GLS* based *SOD* approach. First, we review the first two approaches: *Kriging-SOD* and *LS-SOD*. The basic idea of *Kriging-SOD* is to first apply robust methods to estimate the parameters of a global kriging model. The method uses the estimated statistical model to predict the *Z* attribute of each sample location $\boldsymbol{s}$, denoted as $\hat{Z}(\boldsymbol{s})$, based on the *Z* values of other locations. The standardized residual $\big(\big|\hat{Z}(\boldsymbol{s}) - Z(\boldsymbol{s})\big|/\sigma_{\boldsymbol{s}}\big)$ follows a standard normal distribution, where $\sigma_{\boldsymbol{s}}$ is the estimated standard deviation. If a residual is outside the range $[-\Phi^{-1}(\alpha/2), \Phi^{-1}(\alpha/2)]$, the corresponding object is reported as an outlier, where Φ is the *CDF* and α is usually set 0.05. The *LS-SOD* approach assumes that $\mathbf{diff}(\boldsymbol{Z}) \sim \boldsymbol{N}(\mu\mathbf{1}, \sigma^2\boldsymbol{I})$. The set of components in $\mathbf{diff}(\boldsymbol{Z})$ can be regarded as an i.i.d. sample of a univariate normal distribution $N(\mu, \sigma)$. Robust techniques are designed to estimate μ and σ. The remaining steps are similar to *Kriging-SOD.*

Theorem 5: *Suppose that* $\boldsymbol{F\Sigma F}^T = \sigma^2\boldsymbol{I}$ *and the parameters of Kriging-SOD and GLS-SOD are correctly calculated by robust estimation, then Kriging-SOD and GLS-SOD are equivalent.*

Proof: For *Kriging-SOD*, we consider a universal kriging model [1], since other kriging models (e.g., ordinary kriging) are simply special cases. It suffices to prove that the standardized residuals calculated by *Kriging-SOD* and *GLS-SOD* are identical. Without loss of generality, we test the standardized residual of one particular sample point $Z(\boldsymbol{s}_n)$. Let $\boldsymbol{Z}^* = [Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_{n-1})]^T$ and $\boldsymbol{Z} = \big[\boldsymbol{Z}^{*\boldsymbol{T}}, Z(\boldsymbol{s}_n)\big]^{\mathrm{T}}$. By Section 3.1 equation (3), $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{X\beta}, \boldsymbol{D})$, where $\boldsymbol{D} = \boldsymbol{\Sigma} + \sigma_0^2\boldsymbol{I} = \begin{bmatrix} \boldsymbol{\Sigma}^* & \boldsymbol{\sigma} \\ \boldsymbol{\sigma}^T & \sigma_n^2 \end{bmatrix}$, $\text{Var}(\boldsymbol{Z}^*) = \boldsymbol{\Sigma}^*$, $\text{Cov}(Z(\boldsymbol{s}_1), \boldsymbol{Z}^*) = \boldsymbol{\sigma}$, and $\text{Var}\big(Z(\boldsymbol{s}_n)\big) = \sigma_n^2$.

Then, the standard residual by *Kriging-SOD* is

$$\text{StdRsd}_{Kriging-SOD}\big(Z(\boldsymbol{s}_n)\big) = \frac{\big[\boldsymbol{x}_n^T\boldsymbol{\beta} + \boldsymbol{\sigma}^{\mathbf{T}}\boldsymbol{\Sigma}^{*\mathbf{-1}}(\boldsymbol{Z}^* - \boldsymbol{X}^*\boldsymbol{\beta})\big]}{\sigma_n - \boldsymbol{\sigma}^{\mathbf{T}}\boldsymbol{\Sigma}^{*\mathbf{-1}}\boldsymbol{\sigma}}$$

The standard residual by LS-*SOD* is

$$\text{StdRsd}_{GLS-SOD}\Big(\text{diff}\big(Z(\boldsymbol{s}_n)\big)\Big) = \Big[(\sigma\boldsymbol{I} + \sigma_0^2\boldsymbol{FF}^T)^{-\frac{1}{2}}(\boldsymbol{FZ} - \boldsymbol{FX\beta})\Big]_n$$

The following will prove that

$$\text{StdRsd}_{Kriging-SOD}\big(Z(\boldsymbol{s}_n)\big) = \text{StdRsd}_{GLS-SOD}\Big(\text{diff}\big(Z(\boldsymbol{s}_n)\big)\Big)$$

The condition $\boldsymbol{F\Sigma F}^T = \sigma^2\boldsymbol{I}$ implies that $\sigma^2\boldsymbol{I} + \sigma_0^2\boldsymbol{FF}^T = \boldsymbol{F\Sigma F}^T + \sigma_0^2\boldsymbol{FF}^T = \boldsymbol{FDF}^T$. Then, $(\sigma\boldsymbol{I} + \sigma_0\boldsymbol{FF}^T)^{-\frac{1}{2}} = (\boldsymbol{FDF}^T)^{-\frac{1}{2}} =$

$\left(\boldsymbol{F}\boldsymbol{D}^{\frac{1}{2}}\right)^{-1} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{F}^{-1}$. It follows that $(\sigma\boldsymbol{I} + \sigma_0\boldsymbol{F}\boldsymbol{F}^T)^{-\frac{1}{2}}(\boldsymbol{FZ} - \boldsymbol{FX\beta}) = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{F}^{-1}(\boldsymbol{FZ} - \boldsymbol{FX\beta}) = \boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{Z} - \boldsymbol{X\beta})$.

Further, given that $\boldsymbol{D} = \begin{bmatrix} \boldsymbol{\Sigma}^* & \boldsymbol{\sigma} \\ \boldsymbol{\sigma}^T & \sigma_n \end{bmatrix}$, it can be readily shown that

$$\boldsymbol{D}^{-\frac{1}{2}} = \begin{bmatrix} \left[\boldsymbol{C}_1^{-1} + \boldsymbol{C}_2^{-\frac{1}{2}}\boldsymbol{\Sigma}^{*-1}\boldsymbol{\sigma}\boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{*-1}\right]^{\frac{1}{2}} & \boldsymbol{0} \\ -\boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{*-1}\boldsymbol{C}_2^{-\frac{1}{2}} & \boldsymbol{C}_2^{-\frac{1}{2}} \end{bmatrix},$$

where $\boldsymbol{C_1} = \boldsymbol{\Sigma}^{*-1} - \sigma_n\boldsymbol{\sigma}\boldsymbol{\sigma}^{\boldsymbol{T}}$ **and** $\boldsymbol{C_2} = \sigma_n - \boldsymbol{\sigma}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{*-\boldsymbol{1}}\boldsymbol{\sigma}$**.**

Then, $\left[(\sigma\boldsymbol{I} + \sigma_0\boldsymbol{F}\boldsymbol{F}^T)^{-\frac{1}{2}}(\boldsymbol{FZ} - \boldsymbol{FX\beta})\right]_n = \left[\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{Z} - \boldsymbol{X\beta})\right]_n = \left[\boldsymbol{D}^{-\frac{1}{2}}\begin{bmatrix}\boldsymbol{X}^*\boldsymbol{\beta} \\ \boldsymbol{x}_n\boldsymbol{\beta}\end{bmatrix}\right]_n = -\boldsymbol{C}_2^{-\frac{1}{2}}\boldsymbol{\sigma}^T\boldsymbol{\Sigma}^{*-1}\boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{C}_2^{-\frac{1}{2}}\boldsymbol{x}_n\boldsymbol{\beta} = \{\boldsymbol{x}_n^T\boldsymbol{\beta} + \boldsymbol{\sigma}^{\mathbf{T}}\boldsymbol{\Sigma}^{*-\mathbf{1}}(\boldsymbol{Z}^* - \boldsymbol{X}^*\boldsymbol{\beta})\}/(\sigma_n - \boldsymbol{\sigma}^{\mathbf{T}}\boldsymbol{\Sigma}^{*-\mathbf{1}}\boldsymbol{\sigma})$.

The above indicates that

$$\mathrm{StdRsd}_{Kriging-SOD}(Z(\boldsymbol{s}_n)) = \mathrm{StdRsd}_{GLS-SOD}\left(\mathrm{diff}(Z(\boldsymbol{s}_n))\right),$$

We conclude that *Kriging-SOD* and G*LS-SOD* are equivalent. □

Theorem 6. *If* $\boldsymbol{F\Sigma F}^T = \sigma^2\boldsymbol{I}$, $\sigma_0^2\boldsymbol{F}\boldsymbol{F}^T = \sigma_0^2\boldsymbol{I}$, *the parameters of GLS-SOD and LS-SOD are correctly calculated by robust estimation, and one of the following conditions is true, then GLS-SOD becomes equivalent to LS-SOD.*

(1) $\boldsymbol{Z}(\boldsymbol{s})$ *has a constant trend (mean):* $\boldsymbol{X\beta} = \mathrm{c}\mathbf{I}$, *where c is a constant value.*

(2) $\boldsymbol{Z}(\boldsymbol{s})$ *is a linear trend of spatial coordinates, and each point* $\boldsymbol{s}$ *is the geometric center (or centroid) of its neighbors.*

Proof: For either condition (1) or (2), it can be readily derived that $\boldsymbol{FX\beta} = \boldsymbol{0}$. By conditions $\boldsymbol{F\Sigma F}^T = \sigma^2\boldsymbol{I}$ and $\sigma_0^2\boldsymbol{F}\boldsymbol{F}^T = \sigma_0^2\boldsymbol{I}$, we have $\boldsymbol{FZ} \sim N(0, (\sigma^2 + \sigma_0^2)\boldsymbol{I})$ which is consistent with the i.i.d. assumption in *LS-SOD*. If we use the same robust methods to estimate the parameters, such as using median and median absolute deviation (*MAD*) to estimate the mean and standard deviation σ, then *GLS-SOD* becomes equivalent to *LS-SOD*. □

Discussion: By Theorem 6, *LS-SOD* is a special form of *GLS-SOD*. *LS-SOD* assumes $\mathrm{Var}(\mathbf{diff}(\boldsymbol{Z})) = \sigma^{\boldsymbol{2}}\boldsymbol{I}$ for some constant σ, but no justifications are presented. From this perspective, *GLS-SOD* actually provides a theoretical foundation for *LS-SOD*. Section 3.1 discusses the situations where $\mathrm{Var}(\mathbf{diff}(\boldsymbol{Z}))$ can be approximated by $(\sigma^2 + \sigma_0^2)\boldsymbol{I}$. Furthermore, under the conditions of Theorem 6, *LS-SOD* is equivalent to *GLS-SOD* and since the conditions also include "$\boldsymbol{F\Sigma F}^T = \sigma^2\boldsymbol{I}$", then by Theorem 4 we have that *GLS-SOD* is equivalent to *Kriging-SOD*. Therefore, *LS-SOD* becomes equivalent to *Kriging-SOD* in this situation. Hence, it can be seen that the proposed *GLS* framework can be parameterized to become instances of *LS-SOD* or *Kriging-SOD*. Further study on various outlier detection methods can be greatly enhanced under the lens of this unifying *GLS* framework.

As discussed in Section 3.1, $\boldsymbol{F\Sigma F}^T$ can be reasonably approximated by $\sigma^2\boldsymbol{I}$. From Theorem 5, the major difference between *Kriging-SOD* and *GLS-SOD* is for which approach the related model parameters can be estimated more accurately and efficiently. From this perspective, G*LS-SOD* is superior to *Kriging--SOD* based on three major reasons: First, G*LS-SOD* has less uncertainty than *Kriging--SOD*, since *Kriging--SOD* needs to further assume a semivariogram model. If the semivariogram model is not selected properly, the performance may be significantly impacted. Second, G*LS-SOD* is a convex optimization problem and therefore a global optimal solution exists. However, *Kriging--SOD* is a non-convex optimization problem and relies on an iteratively reweighted generalized least square (*IRWGLS*) approach [12] to determine a local solution. Finally, as shown in Section 5 simulations, *GLS-SOD* runtime performance is superior to *Kriging-SOD*.

5. SIMULATIONS

This section conducts extensive simulations to compare the performance between the proposed *GLS* based *SOD* methods and other related *SOD* methods. The experimental study follows the standard statistical approach for evaluating the performance of spatial outlier detection methods found in [11, 12, 1, 2].

5.1 Simulation Settings

Data set: The simulation data are generated based on the following statistical model:

$$Z(\boldsymbol{s}) = \boldsymbol{x}^{\boldsymbol{T}}(\boldsymbol{s})\boldsymbol{\beta} + \omega(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \qquad \text{(See Section 3.1)}$$

where $\omega(\boldsymbol{s})$ is a Gaussian random field with covariogram model $C(h; \boldsymbol{\theta})$.

We consider two popular covariogram models: spherical model and exponential model. See equation (8) in Section 3.2 for the definition of a spherical model. The exponential model is defined as

$$C(h; \boldsymbol{\theta}) = \begin{cases} b & if\ \ h = 0 \\ b\left(1 - \exp\left(-\frac{h}{c}\right)\right) & if\ 0 < h \leq c \\ 0 & if\ h > c, \end{cases} \qquad (14)$$

These two models have the same parameters b and c. Recall that b is also the constant variance for each $Z(\boldsymbol{s})$.

For the trend component $\boldsymbol{x}^{\boldsymbol{T}}(\boldsymbol{s})\boldsymbol{\beta}$, we define $\boldsymbol{x}(\boldsymbol{s}) = [1, x(\boldsymbol{s}), y(\boldsymbol{s}), x(\boldsymbol{s}) \cdot y(\boldsymbol{s}), x(\boldsymbol{s})^2, y(\boldsymbol{s})^2\ ]$, where $x(\boldsymbol{s})$ and $y(\boldsymbol{s})$ be the *X* and *Y* coordinates of the location $\boldsymbol{s}$. This implies that the trend $\boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta}$ is a polynomial of order two. The nonlinearity of the trend is decided on the regression parameters $\boldsymbol{\beta}$. For example, if $\boldsymbol{\beta} = [1,0,0,0,0,0]^{\boldsymbol{T}}$,then the trend is constant; if $\boldsymbol{\beta} = [1,1,1,0,0,0]^{\boldsymbol{T}}$, then the trend is linear trend.

For the white noise component, we employ the following standard model [1]:

$$\epsilon(\boldsymbol{s}) \sim \begin{matrix} N(0, \sigma_0^2) & \text{with probability } 1 - \alpha \\ N(0, \sigma_C^2) & \text{with probability } \alpha \end{matrix}$$

There are three related parameters σ_0, σ_C and α. σ_0^2 is the variance of a normal white noise, σ_C^2 is the variance of contaminated error that generates outliers, and α is used to control the number of outliers. Note that it is possible that the distribution $N(0, \sigma_C^2)$ will also generate some normal white noises. All true outliers must be only identified based on standard statistical test by calculating the conditional mean and standard deviation for each observation [2]. We also consider the case of clustered outliers. This can be simulated by constraining that the noises of a random cluster of $n \cdot \alpha$ points follow $N(0, \sigma_C^2)$. In the simulations, we tested several representative settings for each parameter, which are summarized in Table 2.

Table 2: Combination of Parameter settings

Variable	Settings
n	$n \in 100, 200$. Randomly generate n spatial locations $\{\boldsymbol{s}_i\}_{i=1}^n$ in the range [0,25] × [0,25].
b, c	$b = 5; c = 5,15,25$
$\boldsymbol{\beta}$	For constant trend, $\beta_1 \sim N(0,1)$ and $\beta_i = 0, i = 2, \dots, 5$; For linear trend, $\{\beta_1, \beta_2, \beta_3\} \in N(0,1)$, $\beta_i = 0$, $i = 4,5,6$; For nonlinear trend, $\{\beta_i\}_{i=1}^n \in N(0,1)$.
σ_0, σ_C	$\sigma_0^2 = 2, 10$; $\sigma_C^2 = 20$
α	$\alpha = 0.05, 0.10, 0.15.$
K	$K = 4, 8$
Covariance model	Exponential, spherical
Outlier type	Isolated, Clustered

Outlier detection methods: We compared our methods with the state of the art local and global based *SOD* methods, including *Z-test* [4], *Median Z-test* [6], *Iterative Z-test* [5], *trimmed Z-test* [7], *SLOM-test* [8], and universal kriging (*UK*) based forward search [11,12] (noted as *UK-forward*). Our proposed methods are identified as *GLS-backward-G*, *GLS-backward-R*, and *GLS-forward-R*. *GLS-backward-G* refers to the *GLS* backward algorithm using generalized least squares regression. *GLS-backward-R* refers to the *GLS* backward algorithm using regular least square regression (See section 4.2). The implementations of all existing methods are based on their published algorithm descriptions.

Performance metric: We tested the performance of all methods for every combination of parameter setting in Table 2. For each specific combination, we run the experiments six times and then calculate the mean and standard deviation of accuracy for each method. To compare the accuracies of each method, we use the standard ROC curves. We further collected accuracies of top 10, 15, and 20 ranked outlier candidates for each method, and then the counts of winners are shown in Table 3. To calculate these winning counts, we use as an example of the *GLS-backward-R* result in the top left cell of table 4: “47, 47, 45”. This column refers to the constant trend cases. If within this particular case, we only consider the true accuracy of the top 10 candidate outliers, then the *GLS-backward-R* has “won” 47 times over all combination of parameters against all other methods. A win is given to the method that exhibits the highest accuracy. Consequently, if we consider the true accuracy of the top 20 candidate outliers, then the *GLS-backward-R* has won 45 times.

All the simulations are conducted in a PC with Intel (R) Core (TM) Duo CPU, CPU 2.80 GHz, and 2.00 GB memory. The development tool is MATLAB 2008.

5.2 Detection Accuracy

We compared the outlier detection accuracies of different methods based on different combinations of parameter settings as shown in Table 2. Six representative results are displayed in Figure 4. First we considered the detection performance between local based methods. For a constant trend, our methods were competitive with existing techniques. For data sets exhibiting linear trends, our *GLS* algorithms achieved on average 10% improvement over existing local based methods. However, for data sets with nonlinear trends, our *GLS* algorithms exhibited more significant improvement (approximately 50% increase) over existing local methods. For the other combination of parameter settings in Table 2, the winning statistics for each method are displayed in Table 3. These results further justify the preceding performance results.

We also compared our *GLS* algorithms against the global based method *UK-forward*. Overall, our methods were comparable to *UK-forward*. Particularly, *GLS-backward-G* attained better accuracy than *UK-forward* on about half of the data sets. For the remaining data sets, the *GLS-backward-G* is still competitive to the *UK-forward*. Additionally, as shown in Section 5.3, the *UK-forward* incurs a significantly much higher computational cost than the *GLS* algorithms.

As discussed in section 4.1, when *K* is small, the effects of $\sigma_0^2 \boldsymbol{FF}^T$ must be considered and a generalized least regression is necessary. The theorems indicate that *GLS-backward-G* should perform better then *GLS-backward-R*, this was justified in Figure 4 c).

Table 3: Competition statistics for different combinations of parameter settings. Each cell contains three values, representing the win times for the related method based on the accuracies of top 10, 15, and 20 ranked outlier candidates for all methods.

Algorithm	Constant Trend	Linear Trend	Nonlinear Trend
GLS-backward-R	47, 47, 45	79, 72, 82	76, 81, 77
GLS-backward-G	88, 86, 89	114, 102, 120	141,144, 138
GLS-forward-R	13, 11, 14	22, 25, 27	40, 36, 47
Z-test	47, 35, 40	29, 30, 13	0, 0, 0
Iterative Z-test	35, 46, 63	16, 20, 21	0, 0, 0
Median Z-test	20, 23, 29	1, 7, 8	0, 0, 0
Trimmed Z-test	15, 23, 32	5, 13, 13	0, 0, 0
SLOM-test	0,0, 0	0, 0, 0	0, 0, 0

5.3 Computational Cost

The comparison on computational cost is shown in Figure 3. The results indicate that the time cost of *UK-forward* is much higher than other methods. Even the second slowest method *GLS-backward-G*, is still three times faster than *UK-forward*. The other local methods are approximately equal and hence much faster than *UK-forward*.

From the comparisons of both the accuracy and computational cost, it can be seen that our proposed *GLS SOD* algorithms (especially *GLS-backward-G*) is significantly more accurate than existing local based algorithms when the spatial data exhibits either a linear or nonlinear spatial trend. Our *GLS* algorithms are comparable to the global based method *UK-forward* on accuracy, but significantly faster than *UK-forward*.

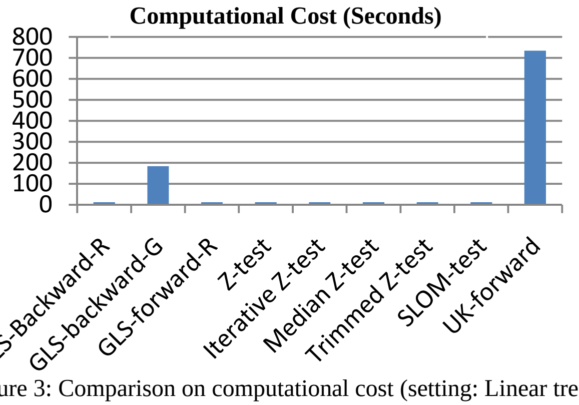
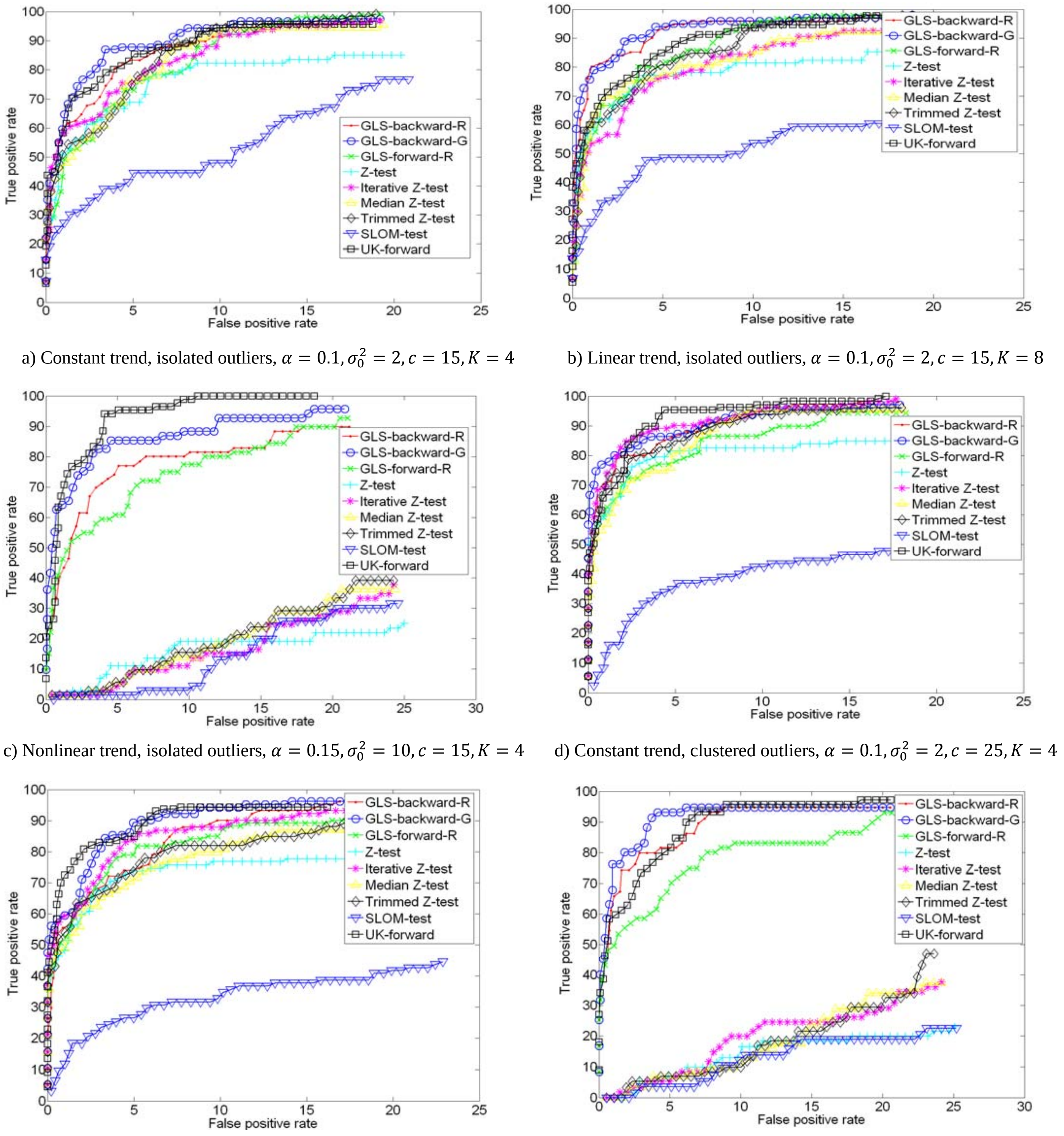


Figure 3: Comparison on computational cost (setting: Linear trend, isolated outliers, $\alpha = 0.1, \sigma_0^2 = 2, c = 15, K = 8, n = 200$)



a) Constant trend, isolated outliers, $\alpha = 0.1, \sigma_0^2 = 2, c = 15, K = 4$

b) Linear trend, isolated outliers, $\alpha = 0.1, \sigma_0^2 = 2, c = 15, K = 8$

c) Nonlinear trend, isolated outliers, $\alpha = 0.15, \sigma_0^2 = 10, c = 15, K = 4$

d) Constant trend, clustered outliers, $\alpha = 0.1, \sigma_0^2 = 2, c = 25, K = 4$

e) Linear trend, clustered outliers, $\alpha = 0.15, \sigma_0^2 = 2, c = 25, K = 8$

f) Nonlinear trend, clustered outliers, $\alpha = 0.15, \sigma_0^2 = 10, c = 5, K = 8$

Figure 4: Outlier ROC Curve Comparison (the same setting: $n = 200, b = 5, \sigma_C^2 = 20$)

6. CONCLUSTION AND FUTURE WORK

This paper presents a generalized local statistical (*GLS*) framework for existing local based methods. This generalized statistical framework not only provides theoretical foundations for local based methods, but can significantly enhance spatial outlier detection methods. This is the first paper to present the theoretical connection between local and global based *SOD* methods under the *GLS* framework. As future work we will design other algorithms to further the efficiency of the *GLS* backward and forward methods.

7. REFERENCES

[1] Cressie, N.A. 1993 Statistics for Spatial Data, Wiley.

[2] Schabenberger O. and Gotway C. A. 2005 Statistical Methods for Spatial Data Analysis. Boca Raton: Chapman and Hall–CRC, Boca Raton, Florida.

[3] Tobler, W. R. 1979 "Cellular geography," in Philosophy in Geography, 379–386, Dordrecht, Holland. Dordrecht Reidel Publishing Company.

[4] Shekhar , S., Lu, C.-T. and Zhang, P. June 2003 "A Unified Approach to Spatial Outliers Detection," Journal of GeoInformatica, Vol. 7, No.2, 139-166.

[5] Lu, C.-T., Chen, D. and Kou, Y. 2003 "Algorithms for Spatial Outlier Detection", Proceedings of the 3rd IEEE International Conference on Data Mining, (Nov. 19-22 2003), 597-600.

[6] Lu, C.-T., Chen, D. and Chen, F. 2008 "On Detecting Spatial Outliers", Journal of Geoinformatica, Vol. 12, 455-475.

[7] Hu, T. and Sung, S.Y. 2004 "A trimmed mean approach to finding spatial outliers", Journal of Intelligent Data Analysis, Vol 8, Issue 1, 79-95.

[8] Sun, P. and Chawla, S. 2004 "On Local Spatial Outliers," Proc. 4th IEEE Int'l Conf. on Data Mining, pp. 209–216.

[9] S. Shekhar, Lu, C.-T. and Zhang, P. 2001 Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In ACM SIGKDD, San Francisco, CA, USA.

[10] Christensen, R., Johnson, W. and Pearson, L.M., 1993 Covariance function diagnostics for spatial linear models. Math. Geol. 25, 145–160.

[11] Cerioli, A. and Riani, M. 1999. The ordering of spatial data and the detection of multiple outliers. J. Comput. Graphical Statist. 8, 239–258.

[12] Militino, A.F., Palacios, M.B. and Ugarte, M.D. 2006 "Outlier detection in multivariate spatial linear models", Journal of statistical planning and inference, vol. 136, 125-146.

[13] Atkinson, A.C. and Riani, M. 2000 Robust Diagnostics Regression Analysis. Springer Series in Statistics. Springer.

[14] S. Boyd and L. Vanderberghe, Convex Optimization. Cambridge Univ. Press, 2004.

8. Appendix

Theorem 3 presents an upper bound of the absolute correlation function $\left|\rho\left(\omega_i^*, \omega_j^*; \boldsymbol{\theta}\right)\right|$. The properties of this upper bound function are demonstrated in Figures 5-9, where we consider five representative cases with $\mathrm{c} = 6, 11, 15, 2, 40$, respectively. The X axis refers to the row difference between $\boldsymbol{s}_j$ and $\boldsymbol{s}_i$: $\mathrm{row}(\boldsymbol{s}_j) - \mathrm{row}(\boldsymbol{s}_i)$. The Y axis refers to the column difference between $\boldsymbol{s}_j$ and $\boldsymbol{s}_i$: $\mathrm{col}(\boldsymbol{s}_j) - \mathrm{col}(\boldsymbol{s}_i)$. The Z axis refers to the absolute correlation value. Each figure includes two surfaces. The surface with colored (yellow to red) map refers to the surface calculated by the estimated upper bound function. The surface in gray color scale refers to the surface calculated by the true correlation function (see equation (9)). These results demonstrate that the estimated upper bound function is a tight upper bound of the true absolute correlation function $\left|\rho\left(\omega_i^*, \omega_j^*; \boldsymbol{\theta}\right)\right|$.

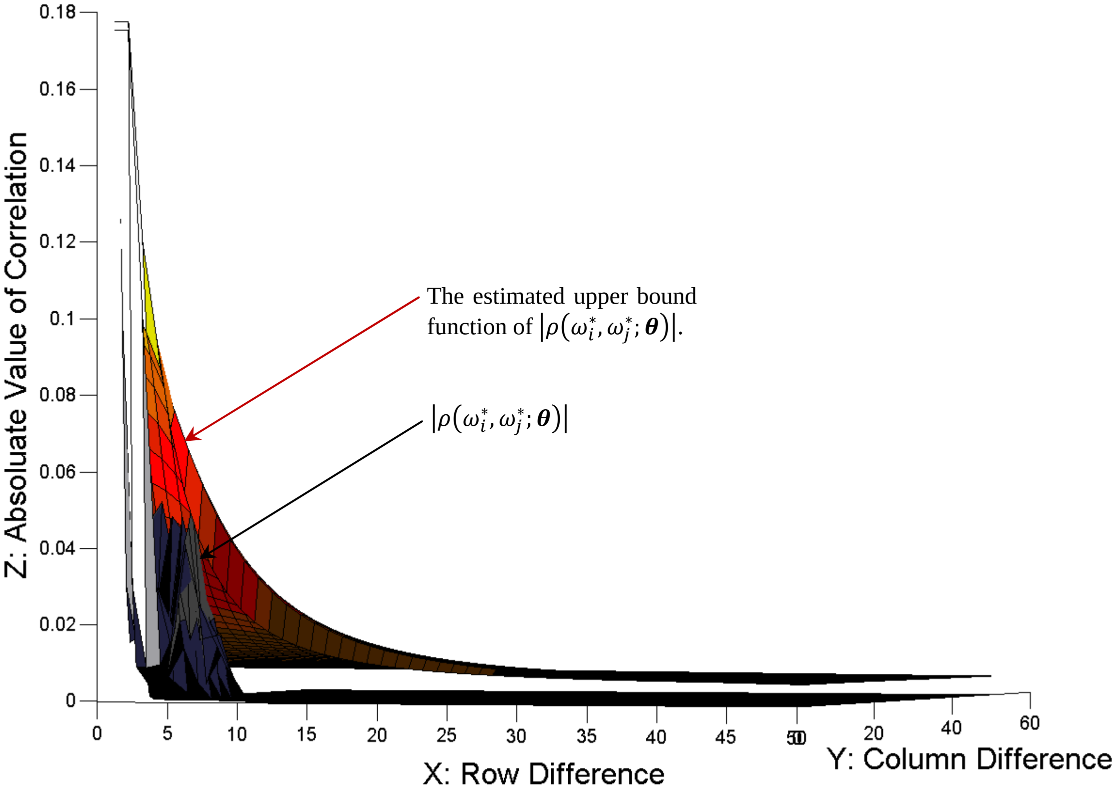


Figure 5: The comparison between the true correlation $\left|\rho\left(\omega_i^*, \omega_j^*; \boldsymbol{\theta}\right)\right|$ and the estimated bound function. Here, $K = 12, c = 6$.

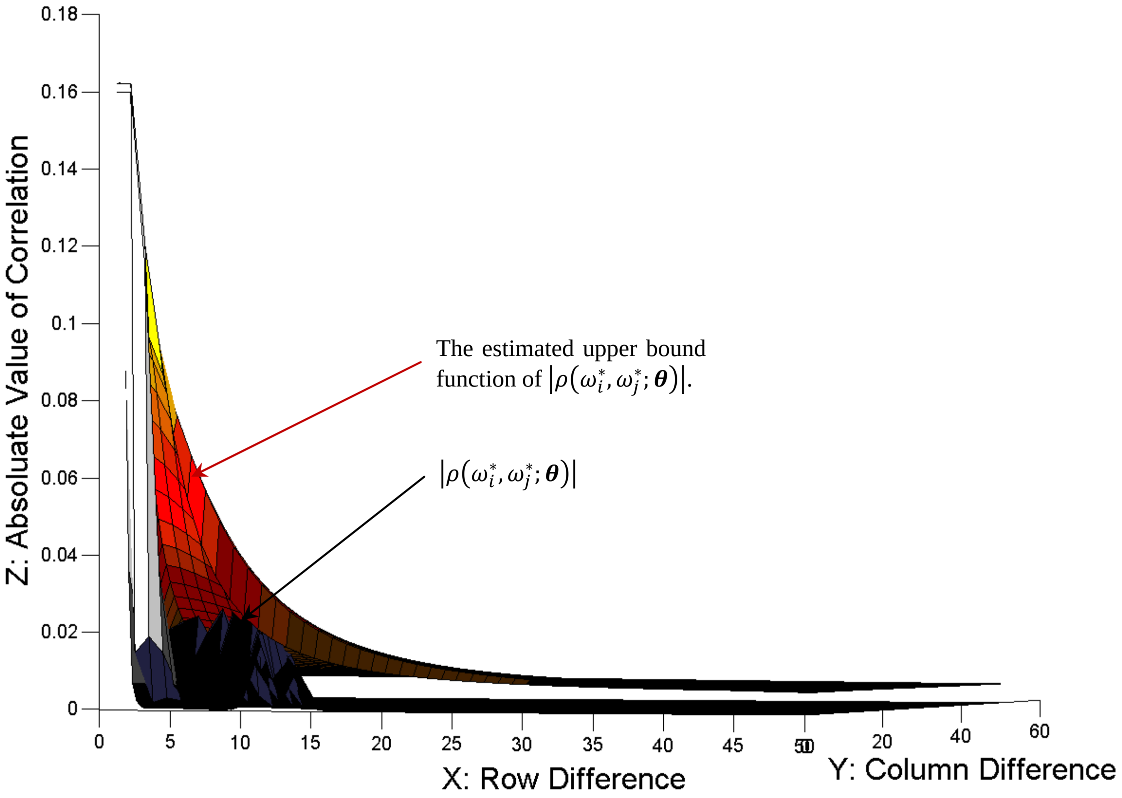


Figure 6: The comparison between the true correlation $\left|\rho\left(\omega_i^*, \omega_j^*; \boldsymbol{\theta}\right)\right|$ and the estimated bound function. Here, $K = 12, c = 11$.

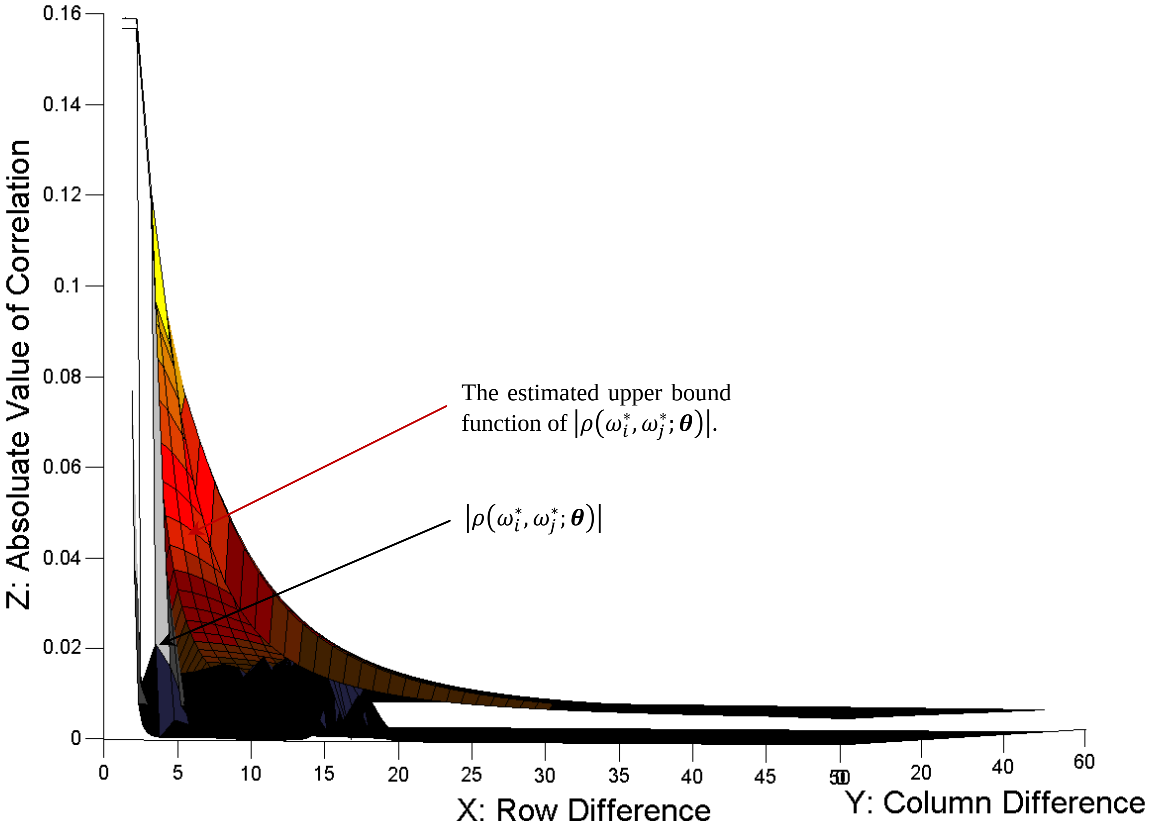
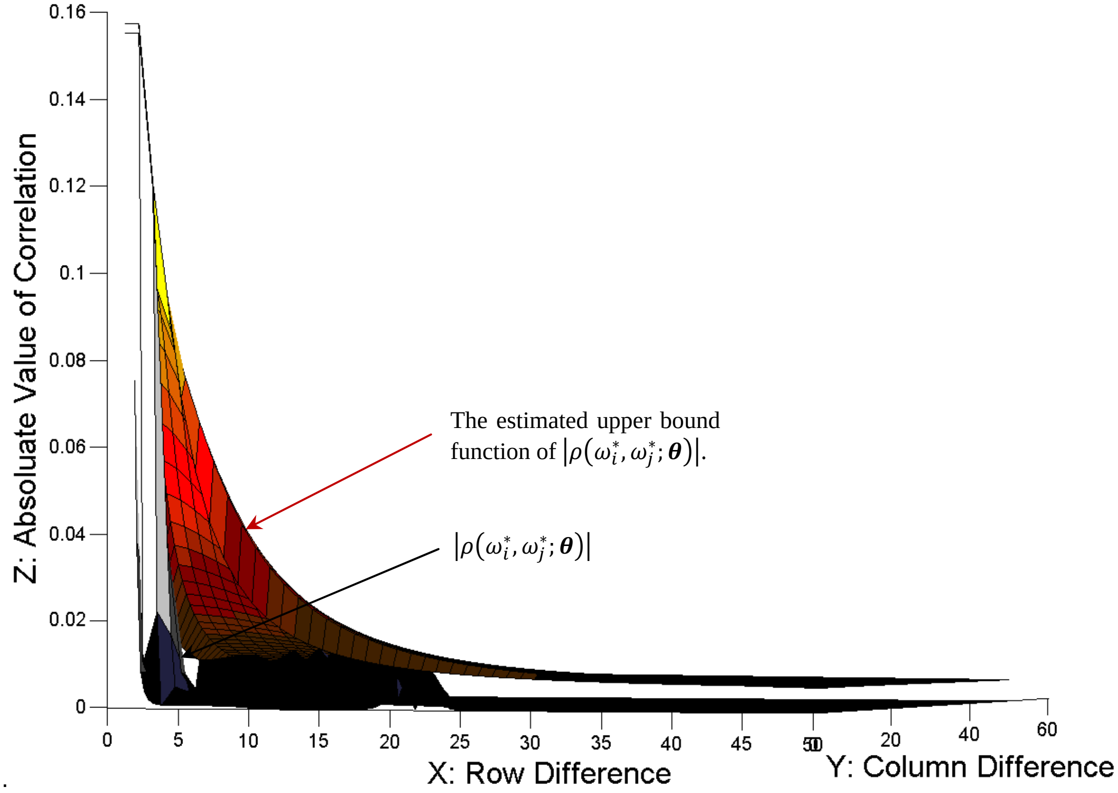


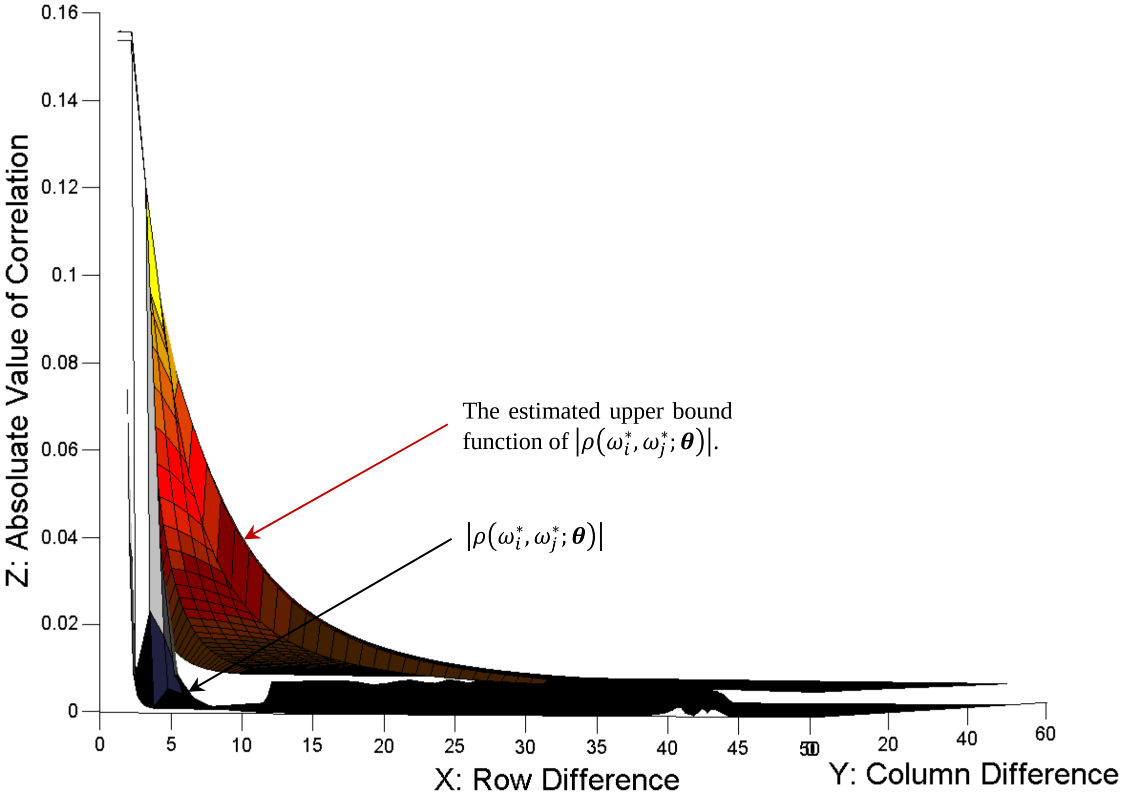
Figure 7: The comparison between the true correlation $\left|\rho\left(\omega_i^*, \omega_j^*; \boldsymbol{\theta}\right)\right|$ and the estimated bound function. Here, $K = 12, c = 15$.



The estimated upper bound function of $|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|$.

$|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|$

Figure 8: The comparison between the true correlation $|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|$ and the estimated bound function. Here, $K = 12, c = 20$.



The estimated upper bound function of $|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|$.

$|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|$

Figure 9: The comparison between the true correlation $|\rho(\omega_i^*, \omega_j^*; \boldsymbol{\theta})|$ and the estimated bound function. Here, $K = 12, c = 40$.