8-2020

Collective Spectral Density Estimation and Clustering for Spatially-Correlated Data

Tianbo Chen
King Abdullah University of Science and Technology

Ying Sun
King Abdullah University of Science and Technology

Mehdi Maadooliat
Marquette University, mehdi.maadooliat@marquette.edu

Spatial Statistics, Vol. 38 (August 2020): 100451. DOI.

Collective Spectral Density Estimation and Clustering for Spatially-Correlated Data

Tianbo Chen
King Abdullah University of Science and Technology (KAUST), Statistics Program, Thuwal 23955-6900, Saudi Arabia
Ying Sun
King Abdullah University of Science and Technology (KAUST), Statistics Program, Thuwal 23955-6900, Saudi Arabia
Mehdi Maadooliat
Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, WI

Abstract

In this paper, we develop a method for estimating and clustering two-dimensional spectral density functions (2D-SDFs) for spatial data from multiple subregions. We use a common set of adaptive basis functions to explain the similarities among the 2D-SDFs in a low-dimensional space and estimate the basis coefficients by maximizing the Whittle likelihood with two penalties. We apply these penalties to impose the smoothness of the estimated 2D-SDFs and the spatial dependence of the spatially-correlated subregions. The proposed technique provides a

score matrix, that is comprised of the estimated coefficients associated with the common set of basis functions representing the 2D-SDFs. Instead of clustering the estimated SDFs directly, we propose to employ the score matrix for clustering purposes, taking advantage of its low-dimensional property. In a simulation study, we demonstrate that our proposed method outperforms other competing estimation procedures used for clustering. Finally, to validate the described clustering method, we apply the procedure to soil moisture data from the Mississippi basin to produce homogeneous spatial clusters. We produce animations to dynamically show the estimation procedure, including the estimated 2D-SDFs and the score matrix, which provide an intuitive illustration of the proposed method.

Keywords

Dimension reduction, Penalized whittle likelihood, Spatial data clustering, Spatial dependence, Two-dimensional spectral density functions

1. Introduction

In spatial statistics, many applications require the segmentation of a spatial region into subregions based on their similarities. Clustering methods are typically developed to address this need. For example, Ambroise et al. (1997) and Allard and Guillot (2000) presented clustering algorithms for spatial data using the EM algorithm. Sheikholeslami et al. (2000) proposed a spatial clustering approach based on wavelet transformation. Guillot et al. (2006) proposed a Bayesian multivariate spatial model to delineate homogeneous regions on the basis of categorical and quantitative measurements. Tarabalka et al. (2009) proposed a spectral–spatial classification scheme for hyperspectral images, which combines the results of a pixel-wise support vector machine classification and the segmentation map obtained by partitional clustering using majority voting.

An important challenge in clustering the spatial regions is to take into account the spatial correlation. Romary et al. (2015) proposed two clustering algorithms based on adaptations of classical algorithms to multivariate geostatistical data, and the spatial dependence is ensured by a proximity condition imposed for two clusters to merge. Fouedjio (2016) developed an agglomerative hierarchical clustering approach that takes into account the spatial dependency between observations. Fouedjio (2017b) introduced a spectral clustering approach to discover spatially contiguous and meaningful clusters in multivariate geostatistical data, in which spatial dependence plays an important role. Marchetti et al. (2018) proposed to compress the spatial data using spatial dispersion clustering, which produce contiguous spatial clusters and preserve the spatial-correlation structure of the data so that the loss of predictive information is minimal.

Note that most of the existing clustering algorithms aim at clustering spatial observations based on similarity of the mean values. Furthermore, spatial processes from real applications are often second-order nonstationary (Fouedjio, 2017a, Schmidt and Guttorp, 2020). Therefore more sophisticated methods are needed to identify stationary spatial regions with similar dependence structures, or spatial patterns. We tackle this problem via collective estimation of the spectral density functions (SDFs) that follow-up with a clustering step in the spectral domain.

Efficiency of the estimators for the SDFs are important, as the quality of the estimated SDF directly affects the clustering results. In one-dimensional (1D) cases (time series), the periodogram is a nonparametric estimation of the SDF, and the undesirable properties of the periodogram, such as roughness or inconsistency, have led to the development of many other estimators of the SDF. In order to achieve a consistent estimator, one method suggests smoothing the periodogram across frequencies. For example, Shumway and Stoffer (2016) discussed several periodogram smoothing techniques, including moving-average smoothing and tapering, and proved that the smoothed periodogram has a smaller variability than the raw periodogram. Wahba (1980) developed the optimally smoothed spline (OSS) estimator, and the smoothing parameter is selected to minimize the expected

integrated mean square error. The span selection is an important issue in periodogram smoothing. Lee (1997) used the unbiased risk estimator to produce the span selector, whereby the selector did not require strong conditions on the spectral density function. Likelihood is another common method for estimating the spectral density. Capon (1983) used the maximum-likelihood filter to produce the minimum-variance unbiased estimator of the spectral density function. Chow and Grenander (1985) proposed a sieve for the estimation of the spectral density of a Gaussian stationary stochastic process using likelihood. Whittle, 1953, Whittle, 1954b developed the now well-established Whittle likelihood for time series analysis, and this likelihood is constructed from the spectrum and periodogram. In Pawitan and O'Sullivan (1994), the spectral density function is estimated by the penalized Whittle likelihood. Besides nonparametric estimation of the spectrum, the autoregressive (AR) spectral approximation is discussed in Shumway and Stoffer (2016). Chan and Langford (1982) and Friedlander and Porat (1984) used the Yule–Walker method to estimate the spectrum.

In two-dimensional (2D) case, the 2D periodogram shares similar features to the 1D periodogram. Some examples of asymptotic theorems have been studied in Heyde and Gay (1993) and Stein (1995), and many spatial SDF estimation methods have been developed. For example, Kim and Fuentes (2000) applied tapering (data filter) to spatial data in order to reduce the bias of the periodogram. Fuentes (2002) proposed a nonstationary periodogram and some parametric approaches to estimate the spatial spectral density of a nonstationary spatial process. Fuentes (2007) proposed estimation methods for large, irregularly spaced spatial datasets using Whittle likelihood approximation. Ebeling et al. (2006) developed an efficient algorithm for adaptive kernel smoothing (AKS) of 2D data with a changeable kernel functional form.

In this paper, to cluster spatial data that share similar spectral features, we extend the methodology of collective spectral density functions estimation as proposed by Maadooliat et al. (2018) to two-dimensional case, and take the spatial dependence of the subregions into account to produce homogeneous spatial clusters. To begin, we use a framework similar to principal component analysis (PCA) to construct a low-dimensional basis expansion that explains the similar features of the 2D-SDFs. Then, we estimate the coefficients associated with the set of adaptive basis by maximizing the Whittle likelihood approximation with two penalties: one to control the smoothness of the adaptive basis functions; the other to consider the spatial dependence of the spatially-correlated subregions to provide more homogeneous spatial clusters. We call the estimated coefficients of the basis expansion as score matrix. Finally, instead of using the estimated 2D-SDFs for clustering, we propose to cluster the spatial data (2D-SDFs) based on the score matrix, which contains sufficient information on the 2D-SDFs but lives in a lower dimension.

The remainder of the paper is organized as follows. The proposed method for the 2D-SDFs estimation is introduced in Section 2, and the clustering algorithm is presented in Section 3. In Section 4, we present two simulation studies which consider two cases: with and without spatial dependence. In Section 5, we present the analysis of soil moisture data from the Mississippi basin and in Section 6, we summarize the paper.

2. Methodology

2.1. Spectral density and periodogram

In one-dimensional case, let $x_t, t=1,\ldots,l$ denote a zero-mean weakly stationary time series, and let $\gamma(h)$, denote its autocovariance function (ACF) that satisfies $\sum_{h=-\infty}^{\infty}|\gamma(h)|<\infty$, then $\gamma(h)$ has the following representation$\gamma(h)=\int_{-1/2}^{1/2}e^{2\pi i\omega h}f(\omega)d\omega, h=0,\pm1,\pm2\ldots,$where $f(\omega)$ is the spectral density function (SDF) of $x_t f(\omega)=\sum_{h=-\infty}^{\infty}\gamma(h)e^{-2\pi i\omega h}, -1/2\le\omega\le1/2.$The periodogram is a nonparametric estimate of the SDF. For a given time series x_t, the periodogram is calculated by $I(\omega_j)=|d(\omega_j)|^2$, where $d(\omega_j)$ is the discrete Fourier transform (DFT)$d(\omega_j)=l^{-1/2}\sum_{t=1}^{l}x_t e^{-2\pi i\omega_j t}, j=0,1,\ldots,l-1,$and the frequencies $\omega_j=j/l$ are called the Fourier or fundamental frequencies.

In the two-dimensional (2D) case, for a stationary spatial process z(s), s∈R2 with ACF C(s)=Cov{z(x),z(x+s)}, the 2D-SDF is defined asf(ω)=∫R2exp(−2πiω⊤s)C(s)ds,where ω=(u,v)⊤∈[−1⁄2,1⁄2]×[−1⁄2,1⁄2].

Suppose that the spatial process is observed on a regular a n1×n2 lattice D={1,…,n1}×{1,…,n2}, the 2D periodogram, In, n=n1n2, is defined asIn(ωj)=1n|∑s∈Dz(s)exp(−2πiωj⊤s)|2,j=0,…,n−1,where ωj=(uj1,vj2)⊤,j1∈{0,…,n1−1},j2∈{0,…,n2−1}.

2.2. Collective estimation

We consider m subregions that are located on a regular rectangular lattice. Let zi(s), i=1,…,m be the observations in the ith subregion and fi be the associated 2D-SDF, where s=(x,y), x=y=1,…,n1 and the size of the subregion is n=n12. We propose to estimate the spectral density functions collectively using two sets of basis functions.

We assume that the 2D log-SDFs can be represented by a linear combination of a set of linear independent common basis functions {φk(ω),k=1,…,K} due to the similar features they share. Specifically,(1)fi(ω)=exp{ui(ω)}=exp{∑k=1Kφk(ω)αik},i=1,…,m,where αik is the score. The value of K should be a small number so that the number of coefficients can be on a reasonable scale even if m is large.

The common basis functions are not prespecified and need to be determined from the data. We suppose that these common basis functions are constructed using linear combination of a rich family of basis functions, {bℓ(ω),ℓ=1,…,L} (L≫K), such that(2)φk(ω)=∑ℓ=1Lbℓ(ω)θℓk,k=1…,K.A large L ensures that the rich basis functions can represent the 2D-SDFs flexibly.

We denote the basis functions and their coefficients: φ(ω)={φ1(ω),…,φK(ω)}⊤, αi=(αi1,…,αiK)⊤, b(ω)={b1(ω),…,bL(ω)}⊤, and θk=(θ1k,…,θLk)⊤. We rewrite (1), (2) into the matrix form U=BΘA⊤, where U={u1(ω),…,um(ω)} is an n×m matrix that represents the 2D log-SDFs, Θ=(θ1,…,θK), and the score matrix A=(α1,…,αm)⊤. B={b(ω1),…,b(ωn)}⊤ is an n×L matrix that represents the rich basis functions. The choice of B is flexible. In this paper, B is the 2D B-spline basis functions matrix which is introduced in Section 2.5. We denote the unknown parameters by (Θ,A).

2.3. Whittle likelihood approach with spatial dependence

We propose to use the penalized Whittle likelihood that applies the roughness penalty (Green and Silverman, 1993) and spatial dependence penalty to estimate the unknown parameters (Θ,A):(3)−2ℓW(Θ,A)+λ1PEN1(φ)+λ2PEN2(A),whereℓW(Θ,A)=∑i=1m∑j=1nui(ωj)+Ii,n(ωj)exp{−ui(ωj)}is the Whittle likelihood approximation (Whittle, 1954a) and Ii,n is the 2D periodogram for the ith subregion. The basis roughness penalty PEN1(φ) is used to regularize the basis function to ensure that φk is smooth. Specifically,(4)PEN1(φ)=∑k=1KθkTRθk=tr{ΘTRΘ},where the penalty matrix R is introduced in Section 2.5.

We consider the spatial dependence of the spatially-correlated subregions using penalty PEN2(A). For the ith subregion, we penalize the difference between the basis coefficients of the ith subregion and the nearest subregions. Sun et al. (2016) applied a similar approach of penalizing the difference of the estimators based on the spatial locations. Let Ni be the set of the nearest neighbors of the ith subregion, with j∈Ni representing the jth subregion as one of the nearest neighbor, excluding the i=j case. Then,PEN2(A)=∑i=1m|αi−1|Ni|∑j∈Niαj|2=∑i=1mDiTDi,where Di=αi−1|Ni|∑j∈Niαj and |Ni| is the size of Ni, where |Ni|=2 if the ith subregion is at corners, |Ni| = 3 if the ith subregion is on the boundary, and |Ni|=4 if otherwise.

The penalized Whittle likelihood approximation is minimized by the Newton–Raphson algorithm. In each iteration, we update αi for i=1,…,m, and θk for k=1,…,K until the convergence. Specifically,(5)αinew=αiold−τ[∂2ℓW(Θ,A)∂αi∂αiT−λ2∂2PEN2(A)∂αi∂αiT]−1[∂ℓW(Θ,A)∂αi−λ2∂PEN2(A)∂αi]|Θ=ΘoldA=Aold=αiold−τ[ΘT∑j{b(ωj)Ii,n(ωj)exp[−ui(ωj)]b(ωj)T}Θ−λ2∑s=1m∂2DsTDs∂αi∂αiT]−1×[ΘT∑j{b(ωj)−b(ωj)Ii,n(

ωj)exp[−ui(ωj)]}−λ2∑s=1mθDsTDsθαi]|Θ=Θold,A=Aoldand(6)θknew=θkold−τ[∂2∂θk∂θkT{ℓW(Θ,A)}−λ1R]−1[∂∂θk{ℓW(Θ,A)}−λ1Rθk]|Θ=Θold,A=Aold=θkold−τ[∑i=1mαik2∑j{b(ωj)Ii,n(ωj)exp[−ui(ωj)]b(ωj)T}−λ1R]−1×[∑i=1mαik∑j{b(ωj)−b(ωj)Ii,n(ωj)exp[−ui(ωj)]}−λ1Rθk]|Θ=ΘoldA=Aoldwhere the learning rate τ is the first element in the sequence {(1⁄2)δ,δ=0,1,…}, which reduces the penalized Whittle likelihood approximation. We denote the estimator of (Θ, A) by ($\hat{\Theta}$, Â).

If we only focus on the spectral properties of the subregions where the spatial dependence is not considered, we use(7)−2ℓW(Θ,A)+λ1PEN1(φ)instead of (3), which is same as setting λ2=0 in (5). We denote the estimated coefficients from (7) as $\tilde{\Theta}$ and Ã. The comparison of the clustering results using Â and Ã is given in Sections 4 Simulation study, 5 Soil moisture data application.

2.4. Selecting the tuning parameters

We select λ1 and λ2 by minimizing the Akaike information criterion (AIC) introduced by Akaike (1974),AIC(λ1,λ2)=−2ℓW($\hat{\Theta}$,Â)+2{df(λ1)+df(λ2)}.The degrees of freedom df(λ1) and df(λ2) are defined asdf(λ1)=∑k=1Ktrace{[∂2∂θk∂θkT{ℓW(Θ,A)}−λ1R]−1[∂2∂θk∂θkT{ℓW(Θ,A)}]},anddf(λ2)=∑i=1mtrace{[∂2∂αi∂αiT{ℓW(Θ,A)}−λ2∂2PEN2(A)∂αi∂αiT]−1[∂2∂αi∂αiT{ℓW(Θ,A)}]},in which the parameters are replaced by the estimated values.

Since that it is computationally expensive to search the optimal λ1 and λ2 by training the model multiple times on sequences of λ1's and λ2's, we update them within the Newton–Raphson iterations. This method has been described by Schall (1991), Schellhase and Kauermann (2012), and Najibi et al. (2017), where in pth iteration we updateλ1(p+1)=df{λ1(p)}−(a−1)trace{Θˆ(p)TRΘˆ(p)},andλ2(p+1)=df(λ2(p))∑i=1m|αi(p)−1|Ni|∑j∈Niαj(p)|2,where a=2 provides the second-order difference penalty given in Section 2.5.

2.5. 2D basis and penalties

We choose 2D spline basis functions as B in this paper. Suppose that Bl∗ is the marginal 1D B-spline basis matrix with l basis functions of order 4 (to ensure piecewise cubic), then, in (1), B=Bl∗⊗Bl∗, where the number basis function of B is L=l2 and ⊗ is the Kronecker product.

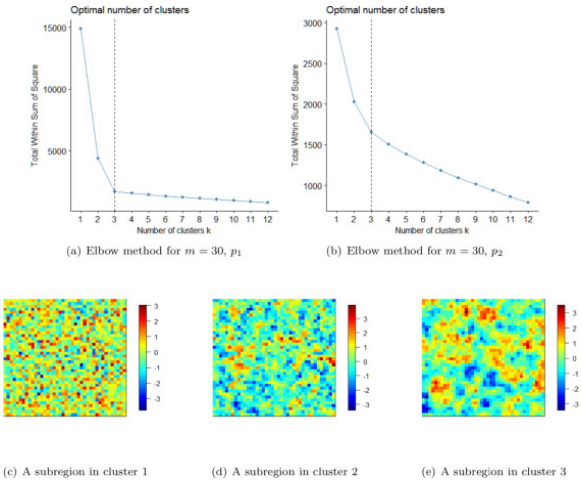
We use the spatial roughness penalty matrix R to control the roughness of common basis φk using the second-order difference penalty (Eilers and Marx, 1996) to achieve the appropriate level of smoothness. The marginal penalty matrix rl=LlTLl, whereLl=1−210…001−21⋱0⋮⋱⋱⋱⋱00⋯01−21(l−2)×l.Then, the roughness penalty matrix R in (4), (6) has the representation:R=Il⊗rl+rl⊗Il,where Il is the identity matrix.

3. Clustering algorithm

We propose to cluster spatial regions based on the estimated score matrix Â, which has the following advantages. First, Â significantly reduces the dimension from m×n, which is the dimension of the m 2D-SDFs, to m×K. Then, by using singular value decomposition (SVD), we obtain the common basis functions from the rich basis functions, and the property of SVD ensures Â contains sufficient information. Finally, by considering the spatial dependence using PEN2(A) in (3), we obtain more homogeneous spatial clusters.

A critical step in clustering real data is to identify the number of clusters, which is directly related to the choice of K. We use the elbow method (Thorndike, 1953), which is widely used in clustering analysis to choose the number of clusters. To begin, we obtain the smoothed log-periodogram estimation Usp=B(BTB)−1BTlog(I). In the elbow method, we run a hierarchical clustering method for the smoothed log-periodograms, and compute the total within-cluster sum of squares (WSS) corresponding to the number of clusters k. Then, by plotting WSS(k) against k, the optimal number of clusters K is found at the location of the elbow or turning point of the plot (see Fig. 1(a) and (b) for illustration). Alternatively, we can also use the Calinski–Harabasz index (Caliński and Harabasz, 1974) to identify the number of clusters. The Calinski–Harabasz

index ch(k)=(m−k)tr(W1)(k−1)tr(W2), where W1 is the covariance matrix between clusters and W2 is the covariance matrix within the clusters. The optimal number of clusters is chosen at K=argmaxkch(k).



(a) Elbow method for $m = 30$, p_1 (b) Elbow method for $m = 30$, p_2

(c) A subregion in cluster 1 (d) A subregion in cluster 2 (e) A subregion in cluster 3

Fig. 1. (a) and (b) The elbow method plots for the two parameter settings on two randomly selected simulation runs; (c)–(e) the generated subregions (p1, m=30) for a randomly selected simulation run.

Below is the clustering algorithm:

1. For the m subregions, we obtain the smoothed log-periodogram matrix Usp, and use the elbow method (or the Calinski–Harabasz index) based on Usp to obtain the optimal number of clusters K.
2. We apply the proposed estimation method, using K common basis functions and obtain Â.
3. We measure the importance (weights) of the columns of Â using the singular values. By denoting wk as the kth singular value and âk as the kth column of Â (k=1,…,K), we have the weighted score matrix Â∗=(â1∗,…,âK∗), where(8)âk∗=wk∑k=1Kwkâk.
4. We compute the Euclidean distance between rows of the matrix Â∗ and apply a hierarchical clustering algorithm to the distance matrix using Ward's measure as an agglomeration method (function hclust in the R package stats). Where we did not consider the spatial dependence (see Section 4.1 for example), we use Ã instead of Â, then we obtain the weighted score matrix Ã∗, and use Ã∗ for clustering. Alternative inputs for clustering include the score matrix (without weights) and the estimated 2D-SDF matrix (see the competitive estimators in Section 4.1).

4. Simulation study

In this section, we perform two simulation studies: (i) a simple case with a known number of clusters without spatial dependence consideration and the estimations are evaluated by clustering results; (ii) the subregions are located on a regular grid and the spatial dependence is considered.

We generate the spatial data from a zero-mean Gaussian process with Matérn covariance function:C(d;ν,ρ)=21−νΓ(ν)2νdρνKν2νdρ,where d is the distance, Γ is the gamma function, Kν is the modified Bessel function, ρ is the scale parameter, and ν is the smoothness parameter.

4.1. Subregions with known number of clusters and no spatial dependence

In this simulation study, we assume that there are three clusters with the same number of subregions. The scale parameters and the smoothness parameters of the Matérn covariance function that we used to generate the subregions in the three clusters are different. Specifically, we consider eight scenarios constructed by four

different number of subregions m=30,60 (to represent small numbers of subregions), and m=480,960 (to mimic large numbers of subregions); and two parameter settings for the Matérn covariance functions:

•p1: in the ith cluster, ρi=0.4×i and νi=0.4×i.

•p2: in the ith cluster, ρi=0.4×i and νi=0.4×(4−i).

Fig. 1(a) and (b) illustrate the elbow methods of the two parameter settings when m=30, where there are turning points at K=3, which is in agreement with our cluster setting. We consider three estimators from the proposed method and three competitive estimators for clustering, where the estimators are treated as input in step 4 of Section 3:

•The estimators from the proposed method (Ã∗, Ã, and estimated spectral density function matrix SDF=exp(BΘ̃Ã⊤)).

•Smoothed periodograms using the rich basis functions (SPB). We use the rich basis functions B to smooth log(I) and obtain SPB=exp{B(B⊤B)−1B⊤log(I)} as the first competitive estimator.

•Smoothed periodograms using 2D Gaussian kernel smoothing (SPK). We apply 2D Gaussian Kernel smoothing (the bandwidth is selected by generalized cross-validation) to I and obtain the second competitive estimator SPK.

•Score matrix of the separate estimations (Ãsep). For the m subregions, we maximize the Whittle likelihood separately to obtain the log-SDFs which is an n×m matrix. We use the truncated SVD of the log-SDFs to obtain the rank K approximation BΘ̃sepÃsep⊤. Then, we have the third competitive estimator Ãsep.

We first measure the performance of clustering by the adjust Rand index (ARI) introduced in Nguyen et al. (2009), which is commonly used to compare two clustering results. Note that the ARI ranges from 0 to 1, with 0 indicating that the two clusters do not agree on any pairs and 1 indicating that the clusters are exactly the same. The definition of ARI is:ARI=∑i=01∑j=01nij2−[∑ini·2+∑jn·j2]/m212[∑ini·2+∑jn·j2]−[∑ini·2+∑jn·j2]/m2.To calculate the ARI, we compute the 2 × 2 table, consisting of the following four cells:

•n11: the number of observation pairs where both observations are comembers in both clusterings.

•n10: the number of observation pairs where the observations are comembers in the one clustering but not in the other.

•n01: the number of observation pairs where the observations are comembers in the second clustering but not in the other.

•n00: the number of observation pairs where neither pair are comembers in either clustering results.

We also use the Jaccard coefficients (Jaccard, 1912), which is available in the R package clusteval, to further evaluate the clustering results.

In each simulation run, we generate the subregions for each scenario, and obtain the estimators using the proposed method (Ã∗, Ã, SDF) and the three competitive estimators (SPB, SPK, Ãsep). The clustering results of the eight scenarios and six different estimators are compared via the true clusters using ARIs and Jaccard coefficients. The associated results (mean ARIs and Jaccard coefficients based on N=100 simulation runs) are given in Table 1, in which we can see that the estimators from the proposed method (especially Ã∗) clearly outperform the other competitive estimators in the clustering task. Also, the values of the clustering indexes (ARIs and Jaccard coefficients) associated to the scenarios p1 are higher in comparing to the scenarios p2, which is reasonable since the turning point, as shown for two randomly selected simulation runs, in Fig. 1(a) is much

clearer and sharper than that in Fig. 1(b). Additionally, as m (the number of subregions) is increasing, the clustering indexes also get closer to one. We randomly pick a subregion in each cluster associated to the scenario p1, m=30 in the first simulation run and use animations to show how the algorithm updates the log-SDFs in Animation 1 of the supplementary file. We observe that the power in the low-frequency area (middle) is more dominant when the scale and smoothness parameters increase, which matches the patterns in the corresponding subregions that are shown in Fig. 1(c)–(e) for a randomly selected simulation run.

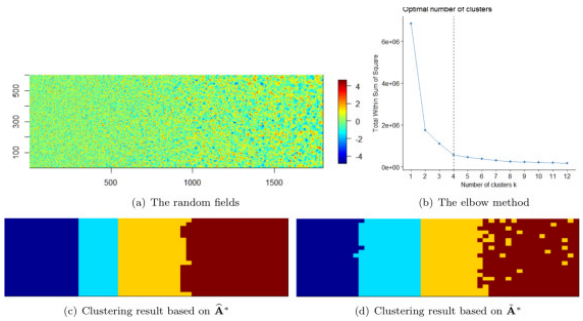
Table 1. The clustering results for Ã∗,Ã, SDF, SPB, SPK, and Ãsep using two measures of performance (ARIs and Jaccard coefficients). The results are based on 100 simulation runs and, in each simulation setup, the best performance is shown in bold. The values within parenthesis, in the first column, provide the average computational time to obtain the collective spectral densities, using a personal computer with 2.6 GHz Intel Core i7−9750H and 32 GB memory, in each simulation setup.

Scenario	**Measure**	Ã∗	Ã	**SDF**	**SPB**	**SPK**	Ãsep
p1, m=30	ARI	**1.000**	**1.000**	**1.000**	0.9844	0.9695	0.9923
(3.69 s)	Jaccard	**1.000**	**1.000**	**1.000**	0.9808	0.9619	0.9907
p1, m=60	ARI	**1.000**	**1.000**	**1.000**	0.9927	0.9742	0.9960
(6.82 s)	Jaccard	**1.000**	**1.000**	**1.000**	0.9908	0.9678	0.9949
p1, m=480	ARI	**1.000**	**1.000**	**1.000**	0.9999	0.9915	0.9998
(51.69 s)	Jaccard	**1.000**	**1.000**	**1.000**	0.9999	0.9888	0.9998
p1, m=960	ARI	**1.000**	**1.000**	**1.000**	**1.000**	0.9891	0.9994
(105.70 s)	Jaccard	**1.000**	**1.000**	**1.000**	**1.000**	0.9857	0.9992
p2, m=30	ARI	**0.9431**	0.8779	0.8943	0.5483	0.5018	0.3698
(3.65 s)	Jaccard	**0.9304**	0.8627	0.8753	0.5390	0.5030	0.4201
p2, m=60	ARI	**0.9465**	0.9265	0.9132	0.5410	0.4974	0.4125
(6.90 s)	Jaccard	**0.9331**	0.9114	0.8935	0.5345	0.5032	0.4480
p2, m=480	ARI	**0.9731**	0.9676	0.9037	0.5575	0.5029	0.4518
(50.65 s)	Jaccard	**0.9650**	0.9585	0.8845	0.5497	0.5102	0.4749
p2, m=960	ARI	**0.9688**	0.9667	0.9170	0.5721	0.5033	0.4763
(104.39 s)	Jaccard	**0.9608**	0.9582	0.8993	0.5607	0.5114	0.4888

4.2. Clustering with spatial dependence and unknown number of clusters

In this simulation, we perform a more complex case with an unknown number of clusters and the spatial dependence of the subregions is considered. The spatial region contains m=1000 (20 by 50) subregions with different Matérn covariance functions with parameters ρ's and ν's gradually increasing with the column index and the size of each subregion is 40 × 40 (n=1600). Specifically, ρcol=νcol=0.5+0.05col, where col is the column index. Fig. 2(a) shows the generated random fields, and the elbow method which indicates K=4 is shown in Fig. 2(b) for a randomly selected simulation run.

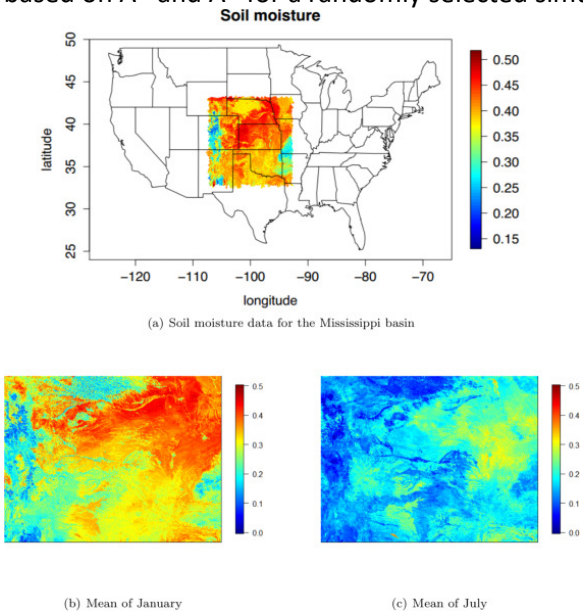
We apply our proposed method to the subregions and apply the clustering algorithm based on Â∗ given that the weighted score matrix had the best performance as outlined in Section 4.1. We also estimate Ã∗, which does not consider the spatial dependence, for comparison to show the advantage of using Â∗. Fig. 2(c) and (d) are the clustering results based on Â∗ and Ã∗. Both clustering results agree with the increasing trend in the parameters along the horizontal direction, while the proposed method provides more homogeneous clusters: clearer margins, well-separated clusters, and less isolated subregions. We use an animation to dynamically illustrate how the proposed method updating the first column of the score matrix and the corresponding clustering result in Animation 2 of the supplementary file.



(a) The random fields (b) The elbow method

(c) Clustering result based on $\hat{\mathbf{A}}^*$ (d) Clustering result based on $\tilde{\mathbf{A}}^*$

Fig. 2. (a) The random field; (b) the elbow method that indicates K = 4; (c) and (d) the spatial clustering results based on $\hat{A}*$ and $\tilde{A}*$ for a randomly selected simulation run.



(a) Soil moisture data for the Mississippi basin

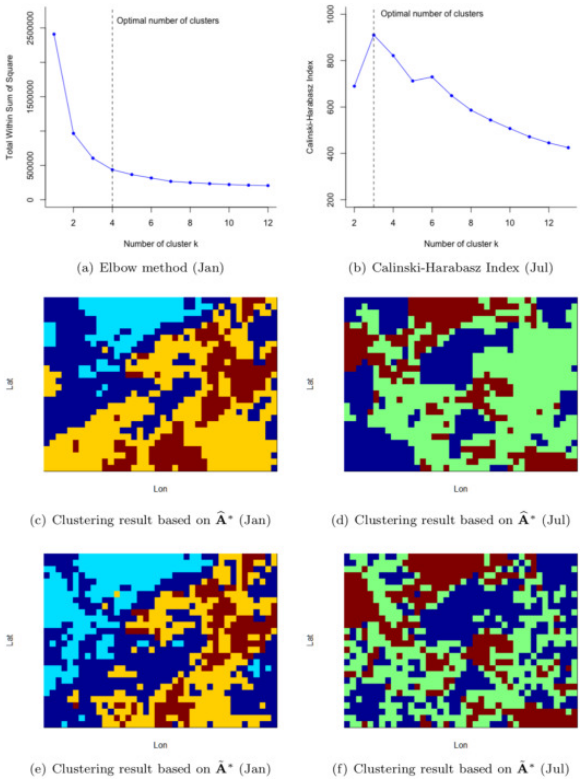
(b) Mean of January (c) Mean of July

Fig. 3. (a) The location of the soil moisture data; (b) and (c) the monthly averaged data in January and July (unit: percentage).

5. Soil moisture data application

5.1. Data description

Understanding the spatial variability, especially the spatial patterns of soil moisture is critical for many hydrological applications (Brocca et al., 2007, Brocca et al., 2012). In this application, we cluster the soil moisture data of the Mississippi basin area using the proposed method. The location (92.47°–107.72°W, 32.37°–43.44°N) of the area is shown in Fig. 3(a) (see Chaney et al. (2016) for more details). We consider the soil moisture data for January (winter) and July (summer), and we analyze them separately. For each month, we average 744 (24 × 31) hourly data and the averaged data are shown in Fig. 3(b) and (c). The size of the region is 1600 × 1120 and we divide the region into m=1120 (40 by 28) subregions with size 40 × 40 (n=1600).

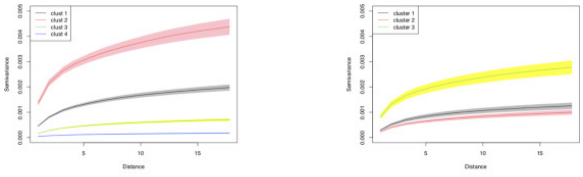


(a) Elbow method (Jan) (b) Calinski-Harabasz Index (Jul)

(c) Clustering result based on $\hat{\mathbf{A}}^*$ (Jan) (d) Clustering result based on $\hat{\mathbf{A}}^*$ (Jul)

(e) Clustering result based on $\tilde{\mathbf{A}}^*$ (Jan) (f) Clustering result based on $\tilde{\mathbf{A}}^*$ (Jul)

Fig. 4. (a) The elbow method plot (January) and (b) the Calinski–Harabasz Index (July); (c)–(f) the clustering results based on $\hat{A}*$ and $\tilde{A}*$ for the two months.



(a) Variogram in Jan (b) Variogram in Jul

Fig. 5. The averaged sample variograms in each cluster and their 95% confidence interval; left: January, right: July.

5.2. Clustering results

We apply the proposed method to the subregions of the two months, obtaining $\hat{A}*$, and apply the clustering algorithm. We also do the clustering based on $\tilde{A}*$ for purposes of comparison. Furthermore, we use the elbow method to identify the number of clusters in the month of January and the Calinski–Harabasz index for the month of July. Based on the clustering results in Fig. 4, Fig. 5, we obtain the following findings:

•For the data in January, the elbow method in Fig. 4(a) indicates that K=4. Out of the 1120 subregions, 361, 134, 437, and 188 subregions are assigned to the four clusters based on $\hat{A}*$, while 454, 237, 285, and 144 subregions are assigned to the four clusters based on $\tilde{A}*$, respectively. Fig. 4(c) and (e) present the corresponding clustering results.

•For the data in July, the Calinski–Harabasz index in Fig. 4(b) indicates that K=3. Out of the 1120 subregions, 263, 583, and 274 subregions are assigned to the three clusters based on $\hat{A}*$, while 342, 531, and 247 subregions are assigned to the three clusters based on $\tilde{A}*$, respectively. Fig. 4(d) and (f) present the corresponding clustering results.

•We observe that the clustering results based on Â∗ have more homogeneous spatial clusters: clearer margins, well-separated clusters, and less isolated subregions, which agree with the animations in Animation 3 of the supplementary file, where the estimation of the score matrices of the two months are illustrated. However, there are still some spatially non-contiguous subregions. This is due to the fact that clustering results are influenced by the spatial dependence, as well as the similarity of the spectral densities.

•For the months of January and July, in Fig. 5, we present the averaged sample variograms and the associated 95% confidence intervals of the subregions in each cluster. In Fig. 5(a), the four clusters are well-separated; while in Fig. 5(b), the black and red clusters do not have a large difference. We also estimate the parameters of the Matérn covariance function in each subregion using maximum likelihood approach. Then, we applied pairwise two-sample t-test on the estimated coefficients in each of the two clusters. In the case of January, the largest p-value is 1.680e−10 and for the month of July, the largest p-value is 0.0227, which indicates that the coefficients from each of the two clusters are significantly different.

6. Conclusion

In this paper, we developed a highly efficient collective method for 2D-SDFs estimation and clustering. A common set of adaptive basis functions spanned by a rich family of basis was used to explain the similarities among the 2D-SDFs in a lower-dimensional space. The basis coefficients were estimated by maximizing the Whittle likelihood approximation with two penalties using the Newton-type algorithm. One penalty controls the roughness of the basis functions and the other penalty takes the spatial dependence of the spatially-correlated subregions into account. The score matrix, which is the estimated coefficients associated to the basis, is a lower-dimensional representation of the 2D-SDFs which we treated as features to cluster spatial data. The two penalties provide not only smooth estimators of the 2D-SDFs but also more homogeneous spatial clusters. We produce several animations, which intuitively illustrate how the proposed method estimates the 2D-SDFs and the score matrix.

One potential limitation of this paper is that the subregions are assumed to be on a 2D regular grid. Alternatively one may use more sophisticated 2D-basis, e.g., bivariate splines over triangulations (Maadooliat et al., 2016), that works for complex geometries with unbalanced observations over irregular grid points. Another immediate extension is to introduce the collective estimation approach for multivariate spatial models.

As for the ease of use, the implementation of the proposed technique is publicly available at https://github.com/tianbochen1/NCSDE_Spatial for reproducing the results of this paper or analyzing any other spatially-correlated dataset.

Acknowledgments

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) to Ying Sun and Tianbo Chen. We would also like to thank the editor, and two referees for their constructive and thoughtful comments which helped us tremendously in improving the manuscript.

Appendix A. Supplementary data

The following is the Supplementary material related to this article.

MMC S1. Supplementary data: Animations.

References

Akaike, 1974. Akaike H. **A new look at the statistical model identification.** IEEE Trans. Automat. Control, 19 (6) (1974), pp. 716-723

Allard and Guillot, 2000. Allard, D., Guillot, G., 2000. Clustering geostatistical data. In: Proceedings of the Sixth Geostatistical Conference.

Ambroise et al., 1997. Ambroise C., Dang M., Govaert G. **Clustering of spatial data by the em algorithm.** GeoENV I—Geostatistics for Environmental Applications, Springer (1997), pp. 493-504

Brocca et al., 2007. Brocca L., Morbidelli R., Melone F., Moramarco T. **Soil moisture spatial variability in experimental areas of central Italy.** J. Hydrol., 333 (2–4) (2007), pp. 356-373

Brocca et al., 2012. Brocca L., Tullo T., Melone F., Moramarco T., Morbidelli R. **Catchment scale soil moisture spatial–temporal variability.** J. Hydrol., 422 (2012), pp. 63-75

Caliński and Harabasz, 1974. Caliński T., Harabasz J. **A dendrite method for cluster analysis.** Commun. Statist.-Theory Methods, 3 (1) (1974), pp. 1-27

Capon, 1983. Capon J. **Maximum-likelihood spectral estimation.** Nonlinear Methods Spectr. Anal. (1983), pp. 155-179

Chan and Langford, 1982. Chan Y., Langford R. **Spectral estimation via the high-order Yule-Walker equations.** IEEE Trans. Acoust. Speech Signal Process., 30 (5) (1982), pp. 689-698

Chaney et al., 2016. Chaney N.W., Metcalfe P., Wood E.F. **Hydroblocks: a field-scale resolving land surface model for application over continental extents.** Hydrol. Process., 30 (20) (2016), pp. 3543-3559

Chow and Grenander, 1985. Chow Y.-S., Grenander U. **A sieve method for the spectral density.** Ann. Statist. (1985), pp. 998-1010

Ebeling et al., 2006. Ebeling H., White D., Rangarajan F. **Asmooth: a simple and efficient algorithm for adaptive kernel smoothing of two-dimensional imaging data.** Mon. Not. R. Astron. Soc., 368 (1) (2006), pp. 65-73

Eilers and Marx, 1996. Eilers P.H., Marx B.D. **Flexible smoothing with B-splines and penalties.** Statist. Sci., 11 (2) (1996), pp. 89-102

Fouedjio, 2016. Fouedjio F. **A hierarchical clustering method for multivariate geostatistical data.** Spat. Statist., 18 (2016), pp. 333-351

Fouedjio, 2017a. Fouedjio F. **Second-order non-stationary modeling approaches for univariate geostatistical data.** Stoch. Environ. Res. Risk Assess., 31 (8) (2017), pp. 1887-1906

Fouedjio, 2017b. Fouedjio F. **A spectral clustering approach for multivariate geostatistical data.** Int. J. Data Sci. Anal., 4 (4) (2017), pp. 301-312

Friedlander and Porat, 1984. Friedlander B., Porat B. **The modified Yule-Walker method of ARMA spectral estimation.** IEEE Trans. Aerosp. Electron. Syst., AES-20 (2) (1984), pp. 158-173

Fuentes, 2002. Fuentes M. **Spectral methods for nonstationary spatial processes.** Biometrika, 89 (1) (2002), pp. 197-210

Fuentes, 2007. Fuentes M. **Approximate likelihood for large irregularly spaced spatial data.** J. Amer. Statist. Assoc., 102 (477) (2007), pp. 321-331

Green and Silverman, 1993. Green P.J., Silverman B.W. **Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach.** CRC Press (1993)

Guillot et al., 2006. Guillot G., Kan-King-Yu D., Michelin J., Huet P. **Inference of a hidden spatial tessellation from multivariate data: application to the delineation of homogeneous regions in an agricultural field.** J. R. Stat. Soc. Ser. C. Appl. Stat., 55 (3) (2006), pp. 407-430

Heyde and Gay, 1993. Heyde C., Gay R. **Smoothed periodogram asymptotics and estimation for processes and fields with possible long-range dependence.** Stochastic Process. Appl., 45 (1) (1993), pp. 169-182

Jaccard, 1912. Jaccard P. **The distribution of the flora in the alpine zone. 1.** New Phytolog., 11 (2) (1912), pp. 37-50

Kim and Fuentes, 2000. Kim, H.-J., Fuentes, M., 2000. Spectral analysis with spatial periodogram and data tapers. In: Proceedings Joint Statistical Meeting.

Lee, 1997. Lee T.C. **A simple span selector for periodogram smoothing.** Biometrika (1997), pp. 965-969

Maadooliat et al., 2018. Maadooliat M., Sun Y., Chen T. **Nonparametric collective spectral density estimation with an application to clustering the brain signals.** Stat. Med. (2018)

Maadooliat et al., 2016. Maadooliat M., Zhou L., Najibi S.M., Gao X., Huang J.Z. **Collective estimation of multiple bivariate density functions with application to angular-sampling-based protein loop modeling.** J. Amer. Statist. Assoc., 111 (513) (2016), pp. 43-56

Marchetti et al., 2018. Marchetti Y., Nguyen H., Braverman A., Cressie N. **Spatial data compression via adaptive dispersion clustering.** Comput. Statist. Data Anal., 117 (2018), pp. 138-153

Najibi et al., 2017. Najibi S.M., Maadooliat M., Zhou L., Huang J.Z., Gao X. **Protein structure classification and loop modeling using multiple ramachandran distributions.** Comput. Struct. Biotechnol. J., 15 (2017)

Nguyen et al., 2009. Nguyen, X.V., Epps, J., Bailey, J., 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th International Conference on Machine Learning (ICML-09), pp. 135.

Pawitan and O'Sullivan, 1994. Pawitan Y., O'Sullivan F. **Nonparametric spectral density estimation using penalized Whittle likelihood.** J. Amer. Statist. Assoc., 89 (426) (1994), pp. 600-610

Romary et al., 2015. Romary T., Ors F., Rivoirard J., Deraisme J. **Unsupervised classification of multivariate geostatistical data: Two algorithms.** Comput. Geosci., 85 (2015), pp. 96-103

Schall, 1991. Schall R. **Estimation in generalized linear models with random effects.** Biometrika, 78 (4) (1991), pp. 719-727

Schellhase and Kauermann, 2012. Schellhase C., Kauermann G. **Density estimation and comparison with a penalized mixture approach.** Comput. Statist., 27 (4) (2012), pp. 757-777

Schmidt and Guttorp, 2020. Schmidt A.M., Guttorp P. **Flexible spatial covariance functions.** Spat. Statist. (2020), p. 100416

Sheikholeslami et al., 2000. Sheikholeslami G., Chatterjee S., Zhang A. **Wavecluster: a wavelet-based clustering approach for spatial data in very large databases.** VLDB J.—Int. J. Very Large Data Bases, 8 (3–4) (2000), pp. 289-304

Shumway and Stoffer, 2016. Shumway R.H., Stoffer D.S. **Time Series Analysis and its Applications: with R Examples.** Springer Science & Business Media (2016)

Stein, 1995. Stein M.L. **Fixed-domain asymptotics for spatial periodograms.** J. Amer. Statist. Assoc., 90 (432) (1995), pp. 1277-1288

Sun et al., 2016. Sun Y., Wang H.J., Fuentes M. **Fused adaptive lasso for spatial and temporal quantile function estimation.** Technometrics, 58 (1) (2016), pp. 127-137

Tarabalka et al., 2009. Tarabalka Y., Benediktsson J.A., Chanussot J. **Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques.** IEEE Trans. Geosci. Remote Sens., 47 (8) (2009), pp. 2973-2987

Thorndike, 1953. Thorndike R.L. **Who belongs in the family?** Psychometrika, 18 (4) (1953), pp. 267-276

Wahba, 1980. Wahba G. **Automatic smoothing of the log periodogram.** J. Amer. Statist. Assoc., 75 (369) (1980), pp. 122-132

Whittle, 1953. Whittle P. **Estimation and information in stationary time series.** Ark. Mat., 2 (5) (1953), pp. 423-434

Whittle, 1954a. Whittle P. **On stationary processes in the plane.** Biometrika, 41 (3/4) (1954), pp. 434-449

Whittle, 1954b. Whittle P. **Some recent contributions to the theory of stationary processes.** A Study in the Analysis of Stationary Time Series, Vol. 2, Almqvist and Wiksells New Zealand (1954), pp. 196-228