Journal of Data Science 20 (4), 566–584
October 2022
DOI: 10.6339/22-JDS1077
Statistical Data Science

Supervised Spatial Regionalization using the Karhunen-Loève Expansion and Minimum Spanning Trees

Ranadeep Daw[1,*] and Christopher K. Wikle[1]
[1]*Department of Statistics, University of Missouri, Columbia, MO, USA, 65211*

Abstract

The article presents a methodology for supervised regionalization of data on a spatial domain. Defining a spatial process at multiple scales leads to the famous ecological fallacy problem. Here, we use the ecological fallacy as the basis for a minimization criterion to obtain the intended regions. The Karhunen-Loève Expansion of the spatial process maintains the relationship between the realizations from multiple resolutions. Specifically, we use the Karhunen-Loève Expansion to define the regionalization error so that the ecological fallacy is minimized. The contiguous regionalization is done using the minimum spanning tree formed from the spatial locations and the data. Then, regionalization becomes similar to pruning edges from the minimum spanning tree. The methodology is demonstrated using simulated and real data examples.

Keywords *connectivity graph; ecological fallacy; Karhunen-Loève expansion; minimum spanning tree; regionalization; spatial data*

1 Introduction

Regionalization in the geospatial literature refers to dividing a spatial domain under study into a set of geographically contiguous partitions. Conceptually, this is similar to clustering a set of spatial locations with an additional spatial contiguity constraint that only allows spatial neighbors to lie in the same partition. However, it is different from the unsupervised spatial clustering methods (e.g., DBSCAN, Ester et al., 1996) that do not consider any information from the observed random variable. In fact, regionalization usually is more similar to classification or regression methods, where the target is to partition the spatial domain based on the behavior of a spatially distributed random variable. The problem can either be discrete or continuous, depending on the measure of the domain space. For example, regionalization by joining neighboring census tracts (Spielman and Folch, 2015) is a discrete regionalization problem since the original spatial domain has a discrete measure, whereas methodologies including decision trees, random forests, and BART partition the space of continuous features into a set of compact subsets. From one perspective, the logic behind regionalization is to create homogeneous regions such that the behavior of a loss function within every partition remains similar. In this way, regionalization can help reduce the size of large spatial datasets. Moreover, the spatial boundaries between the neighboring regions are also informative about the underlying distribution of the spatial process.

One can view spatial regionalization as a bias-variance trade-off problem applied to the underlying spatial process. Real-world spatial datasets are often globally non-stationary yet contain many locally stationary patterns in smaller regions. Spatial regionalization tries to separate such

*Corresponding author. Email: ranadeepdaw@mail.missouri.edu.

Received September 1, 2022; Accepted October 30, 2022

small areas with stable spatial structures. The implicit problem of bias-variance trade-off arises from the choice of the size or the number of the regions since smaller regions may result in more accurate predictions, but simultaneously increase the uncertainties of the parameter estimation. See Lenzi et al. (2020), Fang et al. (2022) for a more detailed discussion of these issues.

Regionalization has been studied in various branches of scientific analysis, mainly in geography and the social sciences. For example, it has been applied to problems in federal policy (Gottmann, 1980; Pearson, 2007), climatology (Giorgi, 2008; Pradhan et al., 2020), public health (Ramos et al., 2020), social science (Openshaw and Rao, 1995; George et al., 1997), and data compression (Kirkley, 2022), among others. Interested readers can see Singleton and Spielman (2014), Duque et al. (2007) for surveys on spatial regionalization methodologies and applications. The methodologies can largely be divided into two parts – implicit methods and explicit methods. The implicit methods (see section 2 of Duque et al., 2007) do not use the spatial contiguity constraint to create the partitions. Instead, they create clusters based on the response variable and then try to impose the spatial compactness constraint in a second stage. Although these algorithms use location information, most methods cannot guarantee spatial contiguity. Moreover, some use the location information as a feature (e.g., Bradley et al., 2017; Chen et al., 2021), which may require one to use an additional weighting parameter determining the effect of the locations. Alternatively, explicit methods ensure the spatial contiguity of the regions. Examples of explicit methods are the `max-p` algorithm (Duque et al., 2012), minimum spanning trees (Assunção et al., 2006), and combinatorial search (Cliff and Haggett, 1970). Some other categories of regionalization include heuristic methodologies (Duque, 2004), and seeded region growing (Adams and Bischof, 1994).

Most of the aforementioned spatial regionalization methodologies do not consider the problem of the *ecological fallacy* (Robinson, 2009) that appears naturally with the spatial *change of support* (COS) problem. The ecological fallacy occurs when a process inside a common spatial boundary is studied at different resolutions and different inference is obtained simply because of aggregation. This problem is similar to Simpson's paradox, which demonstrates contrasting statistical inference drawn between an individual level and a group level study of the same data. As an example, county-level unemployment rates may show different trends from the same data aggregated to the state level. This occurs for various reasons, such as different population sizes of the group level data, high variance within groups, etc. Although most regionalization methods do not account for the ecological fallacy, Bradley et al. (2017) provided a methodology to mitigate the problem by using the ecological fallacy itself as the criterion for spatial aggregation (CAGE). However, it is known that the `k-means` algorithm used in Bradley et al. (2017) fails in creating spatially contiguous regions and only creates regions that are close in values. The authors also suggest a hierarchical agglomerative clustering (Chavent et al., 2018) and implemented it in the `R` package `rcage` (Bradley et al., 2021). However, this method suffers from the pitfalls of implicit spatial methods mentioned above. Moreover, the $\mathbb{O}(n^2 \log n)$ computational complexity of the agglomerative clustering is limiting in Bayesian applications such as used in Bradley et al. (2017). Hence, we consider further improvements in terms of both the explicit spatial neighborhood methodology and the computational complexity of the regionalization process.

We extend the two-stage method of Bradley et al. (2017) for the spatial regionalization problem. In the first stage, we mitigate the ecological fallacy issue by projecting the response variable onto the space of the optimal eigenfunctions. This is done using the Karhunen-Loève Expansion (KLE) of the underlying spatial process. KLE is an optimal feature expansion method that uses the "kernel trick" on the covariance function of the data. It uses orthonormal basis functions as the spatial features and uncorrelated random variables with zero mean and eigen-

values as the variances as basis expansion coefficients. The optimality of KLE can be established by noting that among all feature expansions with a fixed number of expansion terms, KLE has the minimum mean-square error. See Daw et al. (2022) for a review of KLE and an application in spatial modelling. Using KLE as a tool of regionalization transforms the problem onto the space of the covariance function. Thus, this ensures the same regionalization for all random variables that follow the same distributional assumptions as our observed data. Hence, KLE-based regionalization criteria are stronger than those that use only the response variable. Different from Bradley et al. (2017), in the second stage we perform spatial partitioning using minimum spanning trees (MSTs). We do this by considering the spatial data as a connected graph, where the data locations constitute the vertex set, and the edge sets only contain the location tuples that are neighbors. Based on the dissimilarity in the feature space, the edges are associated with a loss function, which ensures the existence of a unique MST. Then, the MST yields a connected graph with the minimum total cost. The spatial regions are created at the last stage by pruning edges from the MST using a heuristic algorithm. We present our model in a unified view of discrete and continuous regionalization and demonstrate the methodology on simulated and real-data examples.

The remainder of the manuscript is organized as follows. Section 2 describes the Karhunen-Loève Expansion and ecological fallacy regionalization criterion. Section 3 provides a description of the minimum spanning tree regionalization algorithm. We demonstrate the method using both simulated and real-world data in Section 4. Section 5 concludes with a summary and discussion of possible future directions.

2 KLE for Spatial Regionalization

2.1 Prerequisites

On a compact spatial domain $\mathcal{S} \subseteq \mathbb{R}^2$, consider a spatially dependent, real-valued univariate spatial process $Y(\cdot) : \mathcal{S} \to \mathbb{R}$. We assume that we have observed noisy realizations from this spatially dependent process at N locations, denoted by $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$. Unsupervised spatial regionalization only uses the distribution of $\boldsymbol{S}$ to partition $\boldsymbol{S}$ (or $\mathcal{S}$) and is not of interest in this manuscript. Rather, we also consider a corresponding observed random variable $Z(\cdot)$ and a latent process $Y(\cdot)$. The observations and the true process at locations $\boldsymbol{S}$ are denoted by $\boldsymbol{Z} = \{Z_1, \ldots, Z_N\}$ and $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$, respectively. Spatial prediction methods use these observed variables to estimate the dependence structure (or parameters in a specified spatial dependence model) in the underlying latent process $\boldsymbol{Y}$ in order to generate predictions. The dynamics of the spatial process are implicitly captured in the covariance function of the latent process. Without loss of generality, we assume that $Y(\cdot)$ is a zero-mean process with a covariance function $\mathcal{C}(\cdot, \cdot) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. In this manuscript, we do not use any stationarity assumption on the covariance function. We only require $\mathbf{Z}$ to be a continuous random variable over the real line, but it easily extends to cases with arbitrary distributions. However, this assumption can also be extended to arbitrary data types using the transformations as in generalized linear models.

The spatial regionalization problem is concerned with finding a contiguous partition $\boldsymbol{\pi} = \{P_1, \ldots, P_n\} : P_j \subset \mathcal{S}$, $P_j \cap P_k = \emptyset$ for $j \neq k$ of the spatial domain. In continuous regionalization, each P_j is a compact subset of $\mathcal{S}$ satisfying $\cup_{j=1}^{n} P_j = \mathcal{S}$, whereas discrete regionalization implies that $P_j \subset \boldsymbol{S}$ and $\cup_{j=1}^{n} P_j = \boldsymbol{S}$. The regions P_j are also known as areal units, and the data (process) over the areal units are similarly called areal data (process). We denote the observed areal data and the unobserved areal process similarly as $\boldsymbol{Z}^A = \{Z_1^A, \ldots, Z_n^A\}$ and $\boldsymbol{Y}^A = \{Y_1^A, \ldots, Y_n^A\}$.

In spatial statistics, the areal random variables $Y^A(\cdot)$ are defined by averaging the point-level random variable Y as follows (Cressie, 2015):

$$Y^A(P) = \frac{1}{\mu(P)} \int_P Y(\boldsymbol{s})\, d\mu(\boldsymbol{s}).$$

Here $\mu(\cdot)$ is the Lebesgue (counting) measure in the continuous (discrete) regionalization problem. From this definition, $Y^A(\cdot)$ is essentially the expected value of $Y(\cdot)$ over the chosen partition. Therefore, supervised regionalization problems yield an ANOVA-like decomposition of the space, meaning that $Y(\cdot)$ is constrained to be as homogeneous as possible within each spatial partition, and the between-partition discrepancies are maximized. In this way, the areal data becomes a close approximation of the original point-level spatial data. This also sets up the motivation behind many regionalization approaches. That is, the discrepancy between $Y(\cdot)$ and $Y^A(\cdot)$ should be a minimum to get an effective partition of the domain.

2.2 Ecological Fallacy as Regionalization Criterion

As noted by Bradley et al. (2017), minimizing the distance between $Y(\cdot)$ and $Y^A(\cdot)$ is not sufficient since it does not necessarily mitigate the ecological fallacy. Instead, creating areal units using the ecological fallacy itself as the loss function (i.e., their criterion for spatial aggregation) ensures that the regions do not suffer from this problem and the spatial trends (and other inference summaries) in both processes remain similar. Since the covariance function captures the spatial dependence, we use it here to create the intended loss function. In particular, the covariance function suggests features from the spatial data following the "kernel trick" for the associated reproducing kernel Hilbert space (RKHS), which represents a covariance kernel $\mathcal{C}(\cdot,\cdot)$ as inner products of feature maps: $\mathcal{C}(\boldsymbol{u},\boldsymbol{v}) = \langle \boldsymbol{\phi}(\boldsymbol{u}), \boldsymbol{\phi}(\boldsymbol{v})\rangle$. Here $\boldsymbol{\phi}(\cdot)$ is the projection of Y onto some arbitrary space of feature vectors. Although we can use any family of basis functions for the projection, it is difficult to justify the arbitrary choice of family and the number of functions. Therefore, to avoid such ambiguities, and to adhere to a more data-driven approach, we use the Karhunen-Loève Expansion (KLE, Karhunen, 1946; Loève, 1955) of the covariance function, which is well-known to be the optimal basis expansion method in the $\mathcal{L}_2$ sense.

The KLE of a univariate spatial process $Y(\cdot)$ is a consequence of Mercer's theorem. Mercer's theorem guarantees that the covariance function $\mathcal{C}(\cdot,\cdot)$ has eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$ and eigenfunctions $\{\psi_k(.) : \mathcal{S} \to \mathbb{R}\}$, which satisfy the orthonormality condition, i.e., $\int_{\boldsymbol{s}} \boldsymbol{\psi}_i(\boldsymbol{s})\boldsymbol{\psi}_j(\boldsymbol{s})\, d\boldsymbol{s} = \delta_{ij}$. Then, the KLE of a spatial process $Y(\cdot)$ is given by

$$Y(\boldsymbol{s}) = \sum_{k=1}^{\infty} \alpha_k \sqrt{\lambda_k}\psi_k(\boldsymbol{s}), \tag{1}$$

where the α_k-s are uncorrelated random variables with mean 0 and variance λ_k. As evident from the representation, the KLE satisfies the bi-orthogonality constraint, which means it uses both orthonormal basis functions and uncorrelated expansion coefficients. Using the KLE, the regionalization of Y is addressed as the following two-stage representation

$$\psi^A(P_j) = \frac{1}{\mu(P_j)} \int_{P_j} \psi(\boldsymbol{s})\, d\mu(\boldsymbol{s}), \tag{2}$$

$$Y^A(P_j) = \sum_{k=1}^{\infty} \alpha_k \sqrt{\lambda_k}\psi_k^A(P_j). \tag{3}$$

For the purpose of this manuscript, it is convenient to unify the eigenvalues and eigenvectors into a new set of scaled KL features by multiplying the eigenfunctions $\psi(\cdot)$ with the eigenvalues as $\xi_k(\cdot) = \sqrt{\lambda_k}\psi_k(\cdot)$. We call these "SKL basis functions" or "SKL expansions" (SKLE) hereafter. Areal level SKL basis functions are also defined in the same fashion as in equation (2).

2.3 Criterion for Regionalization

For regionalization, we use the following loss function Ł: $\boldsymbol{\pi} \to [0, \infty]$, defined over the partition set $\boldsymbol{\pi} = \{P_1, \ldots, P_n\}$ as follows

$$\begin{aligned} Ł(\boldsymbol{\pi}) &= \sum_{j=1}^{n} Ł(P_j), \\ Ł(P_j) &= \frac{1}{\mu(P_j)} \int_{s \in P_j} \|\boldsymbol{K} \circ Y(\boldsymbol{s}) - \boldsymbol{K} \circ Y^A(P_j)\|^2 \, d\mu(\boldsymbol{s}). \end{aligned} \tag{4}$$

To incorporate the KLE of $Y(\cdot)$, we define the operator $\boldsymbol{K}(\cdot)$ in equation (4) as the projection onto the space of the SKLE:

$$\begin{aligned} \boldsymbol{K} \circ Y &= (\xi_1, \xi_2, \ldots), \\ \boldsymbol{K} \circ Y^A &= (\xi_1^A, \xi_2^A \ldots). \end{aligned}$$

Minimizing the above loss function is the same as minimizing the ecological fallacy in the spatial regionalization $\boldsymbol{\pi}$. Equation (4) and equation (3) together ensure the decomposition of the space since the within-region discrepancies are also being minimized. The propositions 2 and 4 from Bradley et al. (2017) justify this SKLE-based loss function and connect the ecological fallacy loss in the data with the same in the eigenspace. They showed that for any measurable real-valued function f over $\mathcal{S}$, there is an equivalence between minimizing the spatial aggregation error (or the ecological fallacy loss) in f^A and maintaining the between-scale homogeneity of the underlying SKLEs. The more homogeneous the eigenfunctions are, the areal units become a better representation of the point-level data. Note that this choice of $\boldsymbol{K}$ also leads to the same regionalization for any random variable following the same distributional assumptions. This is a much stronger result than setting $\boldsymbol{K} = \boldsymbol{\mathcal{I}}$ (the identity operator) in equation (4), which is limited to the specific realization of Y. Hence, we base our regionalization algorithms on the space of the SKLE of the underlying process $Y(\cdot)$.

Remark 1. In practice, we truncate the infinite sum in the equation (1) at a large number M using the first M eigenvalue-eigenvector pairs. One can use any standard rule (e.g., 'elbow rule' or scree plots) to select the truncation parameter, which alleviates the arbitrary choice of the number of basis functions. Higher M ensures a fine-scale representation of the process but increases the computational cost significantly.

Remark 2. Following Mercer's theorem, one can show that among all possible linear expansions of $\mathcal{C}(\cdot, \cdot)$ with M expansion terms, the KLE minimizes the expected mean-square error. Therefore, KLE is the optimal feature expansion in the $\mathcal{L}_2$ sense and so justifies being the chosen feature space.

Remark 3. To compute the KLE, one can use the eigendecomposition of the empirical covariance matrix of the data if a set of replicates is available (e.g., for spatio-temporal data). However,

this is not possible when we have only one realization, more locations than replications, or when the observations are too noisy. Thus, we use a two-stage Obled-Creutin (O-C) basis calculation procedure as outlined in Bradley et al. (2017), Obled and Creutin (1986). In this approach, one starts with any choice of generating basis functions (GBF) from some commonly used family of basis functions. The corresponding O-C bases are then computed by orthonormalizing the GBFs, using the routine in Algorithm 1. Step 3 of Algorithm 2 models the observed data using these O-C bases. Then in the second step, the O-C bases are multiplied by a rotation matrix to get the Karhunen-Loève basis functions. See the routine in Algorithm 2 and Appendix B for the proof of correctness of the O-C approach.

3 Spatial Regionalization

In this section, we utilize the components of graph theory to express the spatial data as a connected graph and then partition it by removing the edges of the graph.

3.1 Spatial Data as Connected Graph

There is an equivalence between a set of geospatial locations and a connectivity graph (e.g., Dale, 2017; Anderson and Dragićević, 2020). Consider a set of spatial locations $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ with a pre-defined neighborhood map for every location. The equivalent connectivity graph $\mathcal{G}$ is defined as the tuple set $\mathcal{G} = (\boldsymbol{S}, \mathcal{V})$, where the set of the observed locations $\boldsymbol{S}$ is called the vertex set and the edge set $\mathcal{V}$ is defined as the collection of *edge*s $\{v_{jk} : (j,k) \in \{1, \ldots, N\}^2\}$, where an edge v_{jk} exists if and only if the locations $\boldsymbol{s}_j$ and $\boldsymbol{s}_k$ are spatial neighbors. Information from $Y(\cdot)$ is incorporated into $\mathcal{G}$ by associating the edges with a set of edge weights. We define a *weight* function ω_{jk} for the edge v_{jk} as $\omega_{jk} = \|\boldsymbol{K} \circ Y_j - \boldsymbol{K} \circ Y_k\|$. The choice of the absolute loss seemed to be a better choice than the squared-error distance due to the roughness of the SKLEs. We use the notation $\mathcal{G}_{\boldsymbol{\omega}} = (\boldsymbol{S}, \mathcal{V}; \boldsymbol{\omega})$ to denote a connectivity graph with edge weights $\boldsymbol{\omega}$, whereas $\mathcal{G} = (\boldsymbol{S}, \mathcal{V})$ denotes a graph without any information on edge weights.

We define a *path* from location $\boldsymbol{s}_j$ to $\boldsymbol{s}_k$ if there exists a set of edges $\{v_{ji_1}, v_{i_1i_2}, \ldots, v_{i_{t-1}i_t}, v_{i_tk}\}$ that connect the two spatial locations. A *circuit* is a path where an ordering of the edges inside the path has the same first and last location. A graph is *connected* if there is a path between any two locations. Under this representation, a (discrete) spatial partition is the same as a subset of locations connected by a path. Therefore, we can partition the connectivity graph $\mathcal{G}_{\boldsymbol{\omega}}$ into a set of connected subgraphs to create the spatial regions such that the loss function Ł in section 2 is as small as possible inside the subgraphs. Motivated by this, we will use the structure of *spanning tree* next to create the regions.

3.2 Minimum Spanning Tree

A *spanning tree* $\mathcal{T}$ of a connected graph $\mathcal{G}_{\boldsymbol{\omega}} = (\boldsymbol{S}, \mathcal{V}; \boldsymbol{\omega})$ is a connected sub-graph $\mathcal{T} = (\boldsymbol{S}, \mathcal{U})$, $\mathcal{U} \subseteq \mathcal{V}$, such that any two locations in $\boldsymbol{S}$ are connected by a unique path. In applications to spatial data, spanning trees create a unique traversal route between the locations of the data under study. If $\boldsymbol{S}$ has N locations, a spanning tree contains exactly $N - 1$ edges. Therefore, removing edges from the spanning tree leads to a discrete regionalization, where the locations inside a region are connected via the split unique path. Hence, spanning trees arise intuitively in partitioning spatial regions (Assunção et al., 2006; Teixeira et al., 2019; Luo et al., 2021).

A connected graph typically has many possible spanning trees, leading to multiple paths. Here we utilize the weights ω_{jk} to get a supervised and unique route to connect the locations. Using these edge-specific weights, the *cost* of a graph is defined as the sum of the edge weights. The *minimum spanning tree* (MST) is defined as the spanning tree that has the minimum total cost. Therefore, we use the MST of the graph $\mathcal{G}_{\boldsymbol{\omega}}$ as the unique choice of the spanning tree. Note that a MST is unique if and only if the pairwise weights are distinct real numbers. In our regionalization methodology, since the weights arise from the SKLE, they are different with probability 1 and hence lead to a unique choice of the MST. Now, we need to remove $n-1$ edges from the MST to get the discrete regionalization.

We use the well-known Kruskal algorithm (Kruskal, 1956) to create the MST from the connected graph $\mathcal{G}$. It is a greedy algorithm, i.e., the MSTs created using the Kruskal algorithm are locally optimal solutions. The Kruskal algorithm initially sorts all the weights associated with edges from low to high and adds the particular edge with the global minimum weight to the MST. Then, in each iteration, it adds one edge from the remaining edge set with the minimum weight. An edge is only rejected if the associated locations form a circuit with the spanning tree formed up to the particular iteration. The algorithm stops once all locations are added to the MST. The other famous choice for MST construction is Prim's algorithm (Prim, 1957), which uses a similar greedy routine. Prim's algorithm chooses a random location first and then finds an associated path with minimum weight. See the routines in Appendix C for the pseudocode associated with these two algorithms.

3.3 Regionalization

After finding the MST, optimal spatial regions are created by pruning the edges of the MST. The most famous method for partitioning an MST is the SKATER method (Assunção et al., 2006), which uses a combinatorial search in every iteration to find the optimal choice. For moderate-sized datasets, this approach is computationally expensive and efforts to find cheaper pruning alternatives are still an active area of research (e.g., see Xu et al., 2002; Lv et al., 2018, for a review of edge removal methodologies). The common theme of such methods is to remove the most inconsistent edges. We considered the long-edge removal approach, which creates K partitions by removing the $K-1$ edges with the largest weights from the MST (Kleinberg and Tardos, 2006, Section 4.7). However, our experiments showed that removing the highest weighted edges can lead to unbalanced-sized partitions. This occurs naturally since the space of SKLE is usually quite rough. Thus, we include a region size constraint similar to the ideas in Laszlo and Mukherjee (2005). We use the hyperparameter $min\_partition\_size$ to choose a minimum size of the partitions, i.e., an edge is only removable if the size of the two resulting subgraphs is more than $min\_partition\_size$. We proceed from the edge with the highest weight and then verify the above condition. If the condition is not satisfied, we move to the next highest weighted edge. The algorithm stops when the size of every partition satisfies the above condition. Although it can be difficult to specify the hyperparameter constraints in practice, we use the ideas from the CAGE algorithm in Bradley et al. (2017). For example, the American Community Survey CAGE example in Bradley et al. (2017) used 185 partitions and an average of 0.54% of the spatial locations clustered in the same partition. Based on this, we used $min\_partition\_size = \lceil N * \frac{0.5}{100} \rceil$ in our experiments. Although this performed well in our examples, an optimal choice of this parameter is the subject of future research.

In the examples presented here, we select the intended number of partitions as K. We use the Bayesian model (explained in Section 3.4) to sample a set of partitions following the

above edge-removal scheme. For every such sampled partition, we compute the ecological loss function from the equation (4). The optimal choice for the spatial regionalization is the one that minimizes equation (4). In this way, we choose a partition that results in the minimum amount of ecological fallacy loss while performing the spatial change-of-support.

3.4 Computation

In this section, we propose the full computational details of our approach. Our modeling approach uses basis functions and assumes prior distributions on the modeling parameters. However, it is not necessary to model the data using a Bayesian basis regression. We simply need to estimate the form of the covariance function, which is possible using any common spatial modeling approach, such as Gaussian process regression (Rasmussen, 2003), fixed rank kriging (Cressie and Johannesson, 2008), Vecchia Gaussian process approaches (Vecchia, 1988), etc. An example of estimating KLE using a Gaussian process regression is provided in Appendix D. We chose a Bayesian approach here because it allows the simple specification of different regularization approaches through the prior distribution.

Our first step is to estimate the Karhunen-Loève basis functions from the data $\mathbf{Z}$ by considering a set of mild assumptions on the data distribution. Following Algorithm 1, we need to use a set of generating basis functions (GBF) to get the KLE. Here, we have chosen the well-known 6-th order Wendland basis functions (Wendland, 1998). Wendland bases are multi-resolutional and are often a good candidate for non-stationary data. The Wendland basis function of order 6 has the following form

$$
\begin{aligned}
d(\boldsymbol{s};\boldsymbol{z},\zeta) &= \zeta^{-1}\|\boldsymbol{s}-\boldsymbol{z}\|,\\
\theta(\boldsymbol{s};\boldsymbol{z},\zeta) &= (1-d)^6\frac{(35d^2+18d+3)}{3},
\end{aligned}
$$

where $\boldsymbol{s}$ is the spatial location and $\boldsymbol{z}$ are a set of knots selected on a regular grid over the spatial domain $\mathcal{S}$. Here, ζ is a range parameter that controls the resolution of the chosen basis function. Assuming the spatial domain to be a subset of the unit square $[0,1]\times[0,1]$, we used 4 different choices of ν. Corresponding to the j-th resolution, q_j equidistant knots are placed in each dimension of $\mathcal{S}$. We considered $\boldsymbol{q}=\{5,10,15,20\}$, which leads to 25, 100, 225 and 400 basis functions. Note that model fitting can be difficult in real data applications and one needs to check the sensitivity of the hyperparameter choices. Since this is out of the scope of our manuscript, we have used a sufficiently well-performing choice of the hyperparameters to demonstrate the methodology.

Given the Wendland basis functions, Algorithm 1 is implemented to get the O-C version of the Wendland basis functions, which is equivalent to the orthonormalization of the GBFs. We denote the O-C basis function as $\phi_j(\cdot):\mathcal{S}\to\mathbb{R};\ j=1,\ldots,\tilde{M}$. In matrix form, it is denoted as $\Phi\in\mathbb{R}^{N\times\tilde{M}}$ and the j-th row corresponding to location $\boldsymbol{s}_j$ is denoted as $\boldsymbol{\phi}(\boldsymbol{s}_j)\in\mathbb{R}^{\tilde{M}}$. Using the O-C basis, we specify our choice of the model for $\mathbf{Z}$ as:

$$
\begin{aligned}
Z(\boldsymbol{s}) &= \alpha + Y(\boldsymbol{s}) + \epsilon(\boldsymbol{s}),\ \text{where} \\
Y(\boldsymbol{s}) &= \boldsymbol{\phi}^T(\boldsymbol{s})\boldsymbol{\beta}.
\end{aligned}
\tag{5}
$$

The error term $\epsilon(\boldsymbol{s})$ is assumed to follow an *iid* normal distribution with a common variance σ^2. Since basis functions are often over-specified, regularization is accomplished by using prior information $\pi(\boldsymbol{\beta},\sigma^2)$ on the model parameters. We assume $\boldsymbol{\beta}$ to be normally distributed with mean

0 and covariance Σ. We experimented with a few choices of prior for Σ, such as the conjugate multivariate normal prior, the Moran's I prior (e.g., Bradley et al., 2017), the independent ridge prior (i.e., $\Sigma = \tau^2\mathbb{I}$), and matrix half-$t$ prior (Huang and Wand, 2013). The results presented here use the matrix half-t prior. This prior has a simple interpretable hierarchical structure and performs well over arbitrary covariance matrices. Using these choices, we specify our full model below:

$$\begin{aligned} Z(\boldsymbol{s}) &\sim \mathbb{N}\big(\alpha + \boldsymbol{\phi}^T(\boldsymbol{s})\boldsymbol{\beta}, \sigma^2\big), \\ \boldsymbol{\beta} &\sim \mathbb{N}(\mathbf{0}, \Sigma), \\ \Sigma|\gamma_1, \ldots, \gamma_p &\sim \text{Inverse-Wishart}(\nu + p - 1, 2\nu\text{diag}(\gamma_1, \ldots, \gamma_p)^{-1}), \\ \gamma_j|\rho &\sim \text{Inverse-Gamma}\Big(\frac{1}{2}, \frac{1}{\rho^2}\Big). \end{aligned} \tag{6}$$

As mentioned previously, hyperparameter tuning is crucial in practical applications. Any Bayesian model-fitting diagnostic can be used to validate the choice of the models. Note that the regionalization can only perform as well as our fitted model permits. We fixed the values of the hyperparameters by experimenting with a set of pre-determined choices and checking the results on a small validation set offline. For the examples in this manuscript, we used $\nu = 2$ and $\rho = 1$ for all the examples. We use Gibbs sampling to sample from the posterior distribution of all the hyperparameters. For our purpose, the knowledge of the posterior distribution for $\big[\Sigma \mid \mathbf{Z}\big]$ is necessary. That is, we obtain samples $\Sigma_1, \ldots, \Sigma_K$ from $\big[\Sigma \mid \mathbf{Z}\big]$. Here, step 5 - 6 of Algorithm 2 gives the KLE. This is done by performing an eigendecomposition of each Σ_j as $\Sigma_j = E_j\Lambda_j E_j^T$. Then, the j-th realization of the KL eigenfunctions are obtained as $\Psi_j = \Phi E_j$ and the eigenvalues are the diagonals of Λ_j. Since we are using the SKL basis functions, they are given by $\Xi_j = \Phi E_j \Lambda_j^{1/2}$. We also calculate the edge weights here as $\omega_{jk} = \|\boldsymbol{\xi}(\boldsymbol{s}_j) - \boldsymbol{\xi}(\boldsymbol{s}_k)\|$.

Our next goal is regionalization based on the SKLE. We address both the discrete and the continuous cases simultaneously here. For the case of discrete regionalization, we evaluate the SKLEs at the data locations $\boldsymbol{S}$. To define the neighborhood, we use a distance-based method, i.e., include the set of points within given proximity in the neighborhood set. The continuous regionalization is approximated by evaluating the SKLEs at a set of gridded "pseudo locations" over the spatial domain $\mathcal{S}$. Note that since the eigenfunctions are available over the whole domain $\mathcal{S}$, it is possible to compute the SKLEs, and hence the connectivity graph $\mathcal{G}_{\boldsymbol{\omega}}$, at the pseudo locations without any complicated algebra. In our examples, we used the same amount of pseudo locations as the original data and sampled them over a rectangular grid covering the domain. We considered the neighborhood as the set of axis-wise and diagonal-wise closest grid points. Then, compact regions are formed by joining the pseudo grid points that belong to the same spatial region. Hence, together, we use the connected graph $\mathcal{G}_{\boldsymbol{\omega}} = (\mathcal{S}, \mathcal{V}; \boldsymbol{\omega})$, where $\mathcal{S}$ is either $\boldsymbol{S}$ or the set of pseudo points. Similarly, $\mathcal{V}$ and $\boldsymbol{\omega}$ correspond to the edges between either observed locations or the pseudo grid points. The rest of the algorithm is simple. In the j-th iteration, we first compute the MST $\mathcal{T}_j$ from the j-th connectivity graph $\mathcal{G}_j = (\mathcal{S}, \mathcal{V}; \boldsymbol{\omega}_j)$. Then we remove the edges from $\mathcal{T}_j$ to get the corresponding spatial regions. We implemented our method using MATLAB (MATLAB, 2018).

4 Simulation and Real Data Examples

We demonstrate our methodology with a simulated and real world example. Specifically, we use our method to create the areal units from point level data and also compute the regionalization

error over the units.

4.1 Simulation Example: Two-Dimensional Gaussian Process

Consider simulated data from a Gaussian process over the unit square with the mean function $m(\boldsymbol{s})$ and the covariance functions $c(\boldsymbol{s}_i, \boldsymbol{s}_j)$ given by the following:

$$m(\boldsymbol{s}) = 2\cos(\boldsymbol{s}) - 4\sin(\boldsymbol{s}),$$
$$c(\boldsymbol{s}_i, \boldsymbol{s}_j) = \exp\{-4\|\boldsymbol{s}_i - \boldsymbol{s}_j\|\} + 0.1\delta_{ij}.$$

Here δ_{ij} is the Dirac Delta function. We simulate from this process at $n = 2500$ randomly generated locations. To apply our method, the SKLE of the underlying spatial process was first derived by using the model from Section 3.4. In the second stage, we generated gridded pseudo locations of the same size from the dataset. The minimum partition size is chosen as 13 and the number of areal regions is fixed at 80. The MST is first computed based on the sample, then is pruned to get spatial regionalization.

Figure 1 shows the areal units obtained by our method. Note that our algorithm selects a regionalization that ensures the minimum ecological fallacy loss in equation (4). This implies that we can represent the 2,500 data points with 80 areal units, optimally minimizing the possibility of an ecological fallacy. The error map in the bottom row of Figure 1 also shows how the regionalization error is smoothed over the spatial regions (i.e., an estimate of the ecological fallacy criterion).

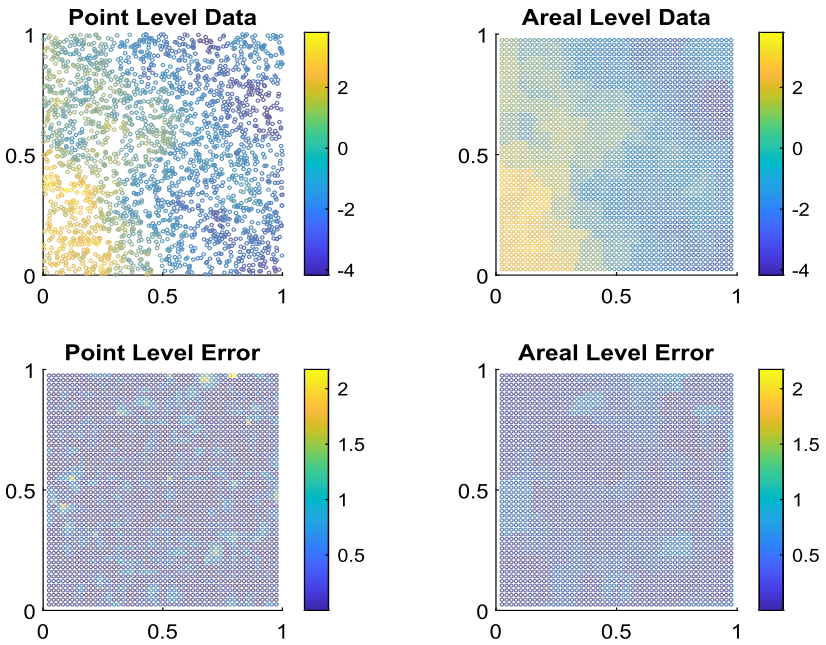

Figure 1: *Regionalization of data from a stationary Gaussian process:* Simulated data from Gaussian process is used to demonstrate the spatial regionalization methodology. The top row contains the point level (left) and areal level (right) data. The bottom row shows the point level (left) and areal level (right) aggregation errors. Note that each partition is restricted to have minimum 0.5% number of points.

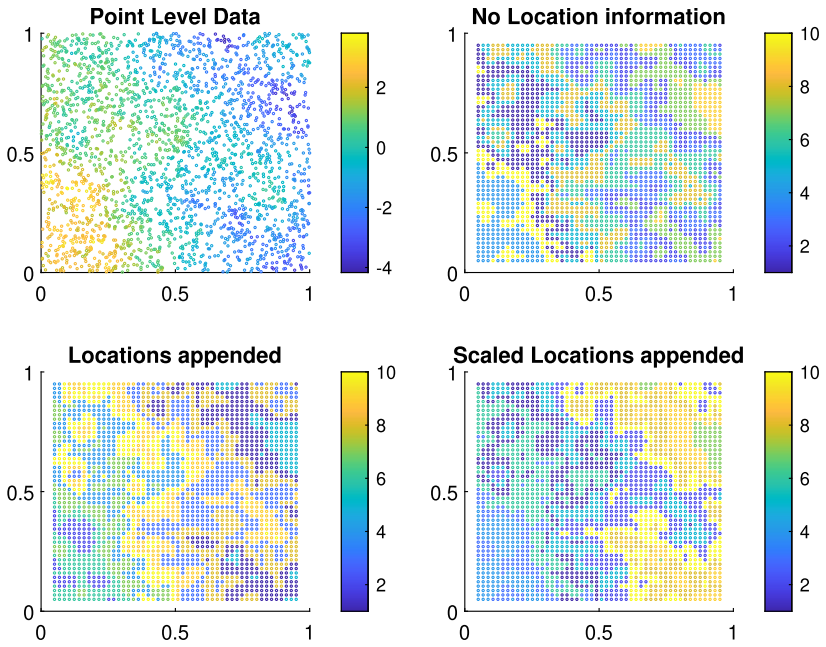

Figure 2: *An example with kmeans clustering:* We implemented the `kmeans` method in this example to find 10 spatial partitions of the Gaussian process simulated data. The left image in the top row shows the original data. The other 3 panels show the colors corresponding to cluster labels, here numbered from 1 – 10. The top right image shows the `kmeans` partitions on the sampled process as used in CAGE. In the bottom row, we append location information to the sampled process to force the soft contiguity constraint. For the bottom left image, the process is appended with locations to get the new feature $\tilde{\mathbf{Y}}_1(\boldsymbol{s}) = \big(\mathbf{Y}(\boldsymbol{s}); \boldsymbol{s}\big)$. The bottom right image scales the location information twice, i.e., $\tilde{\mathbf{Y}}_2(\boldsymbol{s}) = \big(\mathbf{Y}(\boldsymbol{s}); 2\boldsymbol{s}\big)$, to strengthen the soft contiguity constraint. It is clear that `kmeans` fails to provide contiguous partitions.

We also implemented the clustering methodology used by Bradley et al. (2017) to compare with our model for these simulated data. We have implemented both the `kmeans` and the spatial agglomerative clustering in this example, illustrating with 10 spatial partitions to facilitate interpretation. Both Figure 2 and 3 demonstrate that the `kmeans` and the 3 algorithm fail to maintain regional contiguity. Even appending the soft-clustering constraint does not guarantee the geographical contiguity. In our case, the MST-pruning-based regionalization method does not suffer from these issues.

4.2 Real Data Example: Prediction of Ocean Color

We apply our methodology to partition ocean color observations in the coastal Gulf of Alaska (Leeds et al. (2014), Wikle et al. (2013)). Ocean color is a proxy measure of the abundance of phytoplankton in the near surface ocean. The areas with more phytoplankton appear greener than the low-density regions. Hence, regionalization of ocean color provides information on regions of primary productivity in the ocean ecosystem, which is a primary component of the lower trophic levels of the ocean food chain. Many satellites, such as SeaWiFS, MODIS, and MERIS collect global and local level remotely sensed observations of ocean color. See Werdell

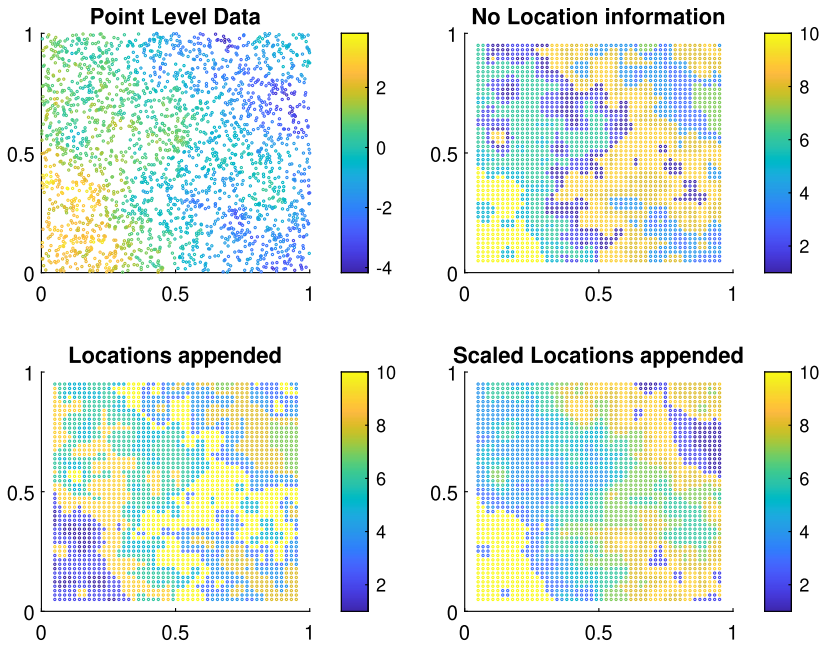

Figure 3: *An example with soft spatial constrained hierarchical clustering:* We implemented a soft hierarchical clustering method similar to that in the `R` package `ClustGeo` on the Gaussian process simulated data. The left image in the top row shows the original data. The other 3 images show colors corresponding to cluster labels, here numbered from 1 – 10. The top right image shows the partitions when no location information is appended to the process. The bottom row shows similar regionalization based on location-appended process. The hierarchical clustering is then performed on these new features to get the regions. From the best regionalization (the bottom right one), it is clear that the regions are getting more defined due to the higher scaling effect of the locations.

and McClain (2018) for more information on satellites dedicated to oceanographic data.

We consider SeaWiFS data over the coastal Gulf of Alaska for 12 May, 2000. The dataset contains observations at 4718 spatial coordinates and was analyzed previously by Daw et al. (2022) using the CAGE methodology. However, the authors used additional covariate information about ocean model (i.e., simulated ocean color output, sea surface temperature, and sea surface height). Moreover, the `kmeans` based clustering used in that paper failed to provide spatially contiguous clusters. In our case, we have not used any covariate information and have directly modeled the response surface using a basis function regression in the first stage as described above. The MST-prune-based clustering in the second stage leads to guaranteed spatially contiguous partitions in our case.

Figure 4 shows the results of using our approach on these data. We choose the minimum partition size as described in Section 3.4, which leads to the choice of a minimum partition size equal to 24. We consider 140 areal regions in the data, an order of magnitude reduction. Note the similarity between the aggregated (area-level data) and the point-level data, with similar error distribution.

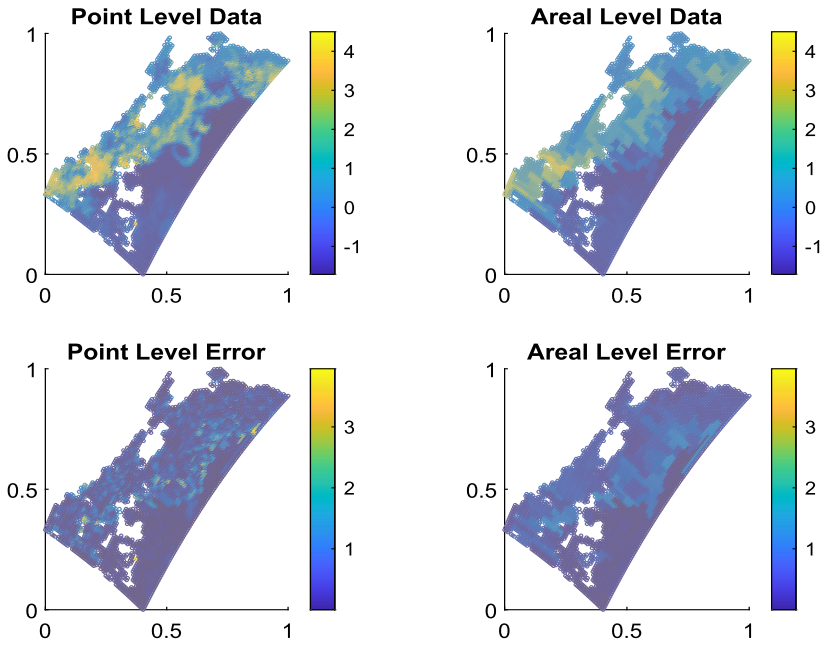

Figure 4: *Analysis of Ocean color*: We applied our methodology to regionalize the SeaWiFS ocean color data. The top row shows the point level (left) and the areal level (right) datasets; note, white areas correspond to missing data. The bottom row demonstrates the regionalization error in the two scales.

5 Discussion

In this manuscript, we developed a novel spatial regionalization methodology that seeks to minimize the ecological fallacy across spatial aggregations. Although the ecological fallacy is well-studied in the geography and ecological literature, there has been little research on using it to inform partitioning spatial domains. Many existing methodologies of special regionalization suffer from the very problem since it naturally occurs while changing the resolution of the spatial support. An exception is the CAGE approach of Bradley et al. (2017), which also minimizes the ecological fallacy, yet is not always guaranteed to give spatially contiguous regions and is costly to implement. We also consider a formal definition of the regionalization error based on the ecological fallacy between the point level and areal level data. We focus on an explicit spatial methodology, which ensures that the regions are spatially contiguous with probability 1.

The ecological fallacy is defined using the Karhunen-Loève Expansion (KLE) of the underlying spatial process. The KLE uses the "kernel trick" to derive an optimal set of features from the covariance kernel of the process. Therefore, using KLE also ensures the same regionalization for any spatial process that follows the same distributional form as the data being considered. To compute the KLE, we need to estimate the covariance function of the underlying spatial process. One can use any spatial modeling technique here to estimate the covariance. Here we used a Bayesian basis function regression model, using the O-C basis approach from Bradley et al. (2017). The O-C basis is the orthonormal version of any chosen family of generating basis functions. The KLE is computed by rotating the O-C basis functions using a matrix of eigenvectors that decorrelate the basis expansion coefficients. We use the scaled KLE by multiplying the KL eigenfunctions with the squared root of the eigenvalues to get the spatial regions in the

second stage.

Our regionalization method uses the minimum spanning tree (MST) associated with the weighted graph of the spatial data. For this, we first represent the SKLE as a connectivity graph. Spatial locations form its vertex set. An edge in the edge set represents a tuple of neighbors. The edges are associated with weights, which are the discrepancies between the SKLE of the locations. MST derives the unique subgraph that connects all the vertices and has the minimum total cost, which is the sum of the edge weights of the MST. Note that the MST also only connects the spatial neighbors. Then, we prune the MST by removing its edges to get the spatial regions. We constrain the partitions to have a minimum size and keep dividing the MST until a stopping criterion meets. We consider various choices of the minimum partition size and choose the pruned graph with the minimum value of the regionalization error. In this way, we use the optimal features (i.e., the SKLE) in regionalization that ensures both minimum ecological fallacy and spatial contiguity. It should also be noted here that there are a different set of MST partitioning methodologies that minimizes a set of different loss functions (e.g., Xu et al., 2002; Lv et al., 2018).

Despite the promise of our methodology, there are several areas for future research and application. For example, the method can easily extend to temporal, spatio-temporal, and other general stochastic processes. Furthermore, it will be interesting to see the behavior of the KLE-based regionalization for different types of observed variables. Global pruning of the MST is also a known difficulty for researchers due to the combinatorial complexity in every iteration. We also look to derive an information-based hypothesis testing procedure to facilitate a more statistically sound stopping criterion.

Another important direction would be the application to large datasets. First, large datasets would require dealing with a massive number of eigenvectors. Although low-rank analysis is arguably possible by considering the first few SKLEs, it is known to not address the small-scale structures. Another source of problem is the presence of outliers, since both the estimation KLE and the MST are highly affected by it. Therefore, very large spatial datasets are naturally difficult to handle through this approach. Future work will attempt to make this procedure more robust and computation-friendly for such applications.

Supplementary Material

The supplementary material includes the following files: (1) README: a brief explanation of all the files in the supplementary material; (2) The synthetic dataset; (3) The real-world dataset; (4) Code files; (5) Images used in the paper; (6) A miscellaneous example of KLE computation directly from covariance matrices.

A Pseudocode for O-C Basis Calculation

We provide the pseudocode for the O-C basis function in this section. O-C basis functions are useful when we have to model the data using generating basis functions (GBFs). The O-C approach simply finds the orthonormal version of our chosen GBF. Then, the KLE is computed by multiplying the O-C basis with a rotation matrix that ensures a diagonal covariance among the expansion.

Algorithm 1 OC basis function.

1: **Input** Locations $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$. GBFs $\{\theta_j(.) : \tilde{\boldsymbol{S}} \to \mathbb{R},\ j = 1, \ldots, \tilde{M}\}$ from any family $\mathcal{G}$.
2: Define the matrix $\Theta \in \mathrm{R}^{N \times \tilde{M}}$ with i, j-th element as $\theta_j(\boldsymbol{s}_i)$.
3: Define $W \in \mathrm{R}^{\tilde{M} \times \tilde{M}}$ with i, j-th element as $w_{ij} = \int_{\boldsymbol{S}} \theta_i(\boldsymbol{s})\theta_j(\boldsymbol{s})\, d\boldsymbol{s}$. Assume W to be non-negative definite.
4: Calculate the Cholesky decomposition Q, where $W^{-1} = QQ^T$.
5: Get the O-C basis function as $\Phi = \Theta Q$.
6: **Output** O-C basis functions Φ.

Algorithm 2 Estimate KLE.

1: **Input** Locations $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$. OC basis functions $\{\phi_j(.) : j = 1, \ldots, \tilde{M}\}$. Observations $\mathbf{Z} = (Z_1, \ldots, Z_N)$. Number of KL bases M.
2: Define the matrix $\Phi \in \mathbb{R}^{N \times \tilde{M}}$ with i, j-th element as $\phi_j(s_i)$.
3: Fit the linear regression model: $\mathbf{Z} = \Phi\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ under some prior information $\pi(\boldsymbol{\gamma}, \boldsymbol{\epsilon})$.
4: Define $\Sigma = \mathrm{cov}(\boldsymbol{\gamma} \mid \mathbf{Z})$.
5: Perform an eigendecomposition $\Sigma = E\Lambda E^T$.
6: Define the Karhunen-Loève basis functions $\Psi = \Phi E$ and eigenvalues Λ.
7: **Output** Return the KLE: $\{\big(\psi_j(\cdot), \lambda_j\big) : j = 1, \ldots, M\}$.

B Proof of Correctness of Algorithm 2

First, we prove that the OC basis functions from Algorithm 1 are orthonormal. Define, $F \in \mathrm{R}^{\tilde{M} \times \tilde{M}}$ to be the matrix with i, j-th element $f_{ij} = \int_{\boldsymbol{S}} \phi_i(\boldsymbol{s})\phi_j(\boldsymbol{s})\, d\boldsymbol{s}$. Then, it is straight-forward to see that $F = Q^T W Q = \mathbb{I}$. Therefore, Φ is an orthonormal basis function of the chosen GBF family $\mathcal{G}$.

Now, note that the step 2 in Algorithm 2 is same as the two-step definition: $\mathbf{Z} = \mathbf{Y} + \boldsymbol{\epsilon}$; $\mathbf{Y} = \Phi\boldsymbol{\gamma}$. From step 3, the posterior covariance matrix of $\boldsymbol{\gamma}$ is Σ. Then, using the eigendecomposition in step 5, we have the following:

$$\begin{aligned} \mathrm{cov}(\mathbf{Y}) &= \Phi\Sigma\Phi^T \\ &= \Phi E\Lambda E^T\Phi^T \\ &= \Psi\Lambda\Psi^T \end{aligned}$$

Since both Φ and E are orthonormal, Ψ is also orthonormal. Also, Λ is a diagonal matrix of non-increasing eigenvalues. Hence, (Ψ, Λ) is the eigendecomposition of the covariance kernel of $\mathbb{C} = \mathrm{cov}(\mathbf{Y})$. Thus, we can prove that $\psi_j(\cdot)$-s are the Karhunen-Loève basis functions and λ_j-s are the corresponding eigenvalues.

C Pseudocode for Finding the MST

In this section, we provide the pseudocode for the MST. We have used Kruskal's algorithm (Kruskal, 1956) in our study. Both Kruskal's and Prim's (Jarník, 1930; Prim, 1957) algorithms are greedy ways to construct the MST from a connectivity graph. Both algorithms run with a

computational cost of $\mathbb{O}(|\mathcal{S}| \log |\mathcal{V}|)$, which in our case is same as $N \log N$, since we consider $\mathbb{O}(1)$ neighbors for each spatial locations. See that Kruskal's algorithm starts with the least weighted edge and adds edges iteratively, whereas Prim's algorithm starts by choosing a location randomly and then finds the least costly edge from the edge set.

Algorithm 3 MST construction using Kruskal algorithm.

Input Weighted graph $\mathcal{G}_{\boldsymbol{\omega}} = (\boldsymbol{S}, \mathcal{V})$.
$\mathcal{U} = \emptyset$.
Sort all the edges of $\mathcal{G}_{\boldsymbol{\omega}}$ from low weight to high.
while $|\mathcal{U}| < |\boldsymbol{S}| - 1$ **do**
    Find the edge $v_{jk} \in \mathcal{V}$ such that $v_{jk} \notin \mathcal{U}$ and $\omega_{jk} = \max\{\omega_{i\ell} : v_{i\ell} \in \mathcal{V} - \mathcal{U}\}$.
    If $\mathcal{U} \cup v_{jk}$ does not create a cycle, $\mathcal{U} = \mathcal{U} \cup v_{jk}$.
end while
Output MST $\mathcal{T}_{\boldsymbol{\omega}} = (\boldsymbol{S}, \mathcal{U}; \boldsymbol{\omega})$.

Algorithm 4 MST construction using Prim's algorithm.

Input Weighted graph $\mathcal{G}_{\boldsymbol{\omega}} = (\boldsymbol{S}, \mathcal{V})$.
Randomly choose $\boldsymbol{s}_j$ from $\boldsymbol{S}$. Initialize $\bar{\mathcal{S}} = \{\boldsymbol{s}_j\}$, $\mathcal{U} = \emptyset$.
while $\bar{\mathcal{S}} \subset \boldsymbol{S}$ **do**
    Find $v_{jk} = \min : \{v_{i\ell} : v_{i\ell} \in \mathcal{V} - \mathcal{U}, \boldsymbol{s}_i \in \bar{S}\}$.
    If $\mathcal{U} \cup v_{jk}$ does not create a cycle, $\mathcal{U} = \mathcal{U} \cup v_{jk}$.
end while
Output MST $\mathcal{T}_{\boldsymbol{\omega}} = (\boldsymbol{S}, \mathcal{U}; \boldsymbol{\omega})$.

D General Methodology for KLE Estimation

In this section, we propose how to compute the KLE for any choice of model. Suppose that, under the assumptions of Section 2, one uses a spatial model of the following form:

$$\begin{aligned} Z(\boldsymbol{s}) &= Y(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \\ Y(\boldsymbol{s}) &= f(\boldsymbol{s}). \end{aligned} \tag{7}$$

Here $f(\cdot)$ can be any common spatial model, such as a Gaussian process (Rasmussen, 2003), Fixed Rank Kriging (Cressie and Johannesson, 2008), Vecchia process (Vecchia, 1988), etc. After fitting the model, we can estimate the parameters of f, which gives us an estimation of the covariance kernel as $\hat{\mathcal{C}}$. Now, using the routine in Algorithm 5, one can directly estimate the KLE from $\hat{\mathcal{C}}$.

Algorithm 5 General method for KLE.

1: **Input** Estimated covariance kernel $\hat{\mathcal{C}}(\cdot,\cdot) : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$.
2: $\{\theta_j : j = 1, \ldots, \tilde{M}\}$ from any family of GBF ϑ.
3: Calculate the matrix $\boldsymbol{B}$ with i, j-th element $b_{ij} = \int_{\mathcal{S}} \theta_i(\boldsymbol{s})\theta_j(\boldsymbol{s})\, d\boldsymbol{s}$.
4: Calculate the matrix $\boldsymbol{A}$ with i, j-th element $a_{ij} = \int_{\mathcal{S}}\int_{\mathcal{S}} \hat{\mathcal{C}}(\boldsymbol{s}_i, \boldsymbol{s}_j)\theta_i(\boldsymbol{s}_i)\theta_j(\boldsymbol{s}_j)\, d\boldsymbol{s}_i\, d\boldsymbol{s}_j$.
5: Solve the generalized eigenvalue problem $\boldsymbol{AF} = \boldsymbol{BF}\Lambda$, where Λ is the diagonal matrix of eigenvalues and $\boldsymbol{F}$ is the matrix of eigenvectors.
6: Compute the Karhunen-Loève eigenfunctions as: $\psi_j(\boldsymbol{s}) = \sum_k f_{jk}\theta_k(\boldsymbol{s})$.
7: **Output** Karhunen-Loève eigenvectors $\{\psi_j : j = 1, \ldots, M\}$ and eigenvalues λ_j.

Author Contributions

RD conceptualized the methodology, implemented it, and drafted the manuscript. CKW contributed to discussions of the methodology and helped polish the manuscript.

Financial Disclosure

None reported.

Conflict of Interest

The authors declare no potential conflict of interests.

Funding

This research was partially supported by the U.S. National Science Foundation (NSF) grant SES-1853096. The computation for this work was performed on the high performance computing infrastructure provided by the Research Computing Support Services at the University of Missouri, Columbia, MO, and is supported in part by the NSF grant CNS-1429294.

References

Adams R, Bischof L (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6): 641–647.

Anderson T, Dragićević S (2020). Complex spatial networks: Theory and geospatial applications. *Geography Compass*, 14(9): e12502.

Assunção RM, Neves MC, Câmara G, da Costa Freitas C (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7): 797–811.

Bradley JR, Wikle CK, Holan SH (2017). Regionalization of multiscale spatial processes by using a criterion for spatial aggregation error. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3): 815–832.

Bradley JR, Wikle CK, Holan SH, Holloway ST (2021). *rcage: Regionalization of Multiscale Spatial Processes*. R package version 1.1.

Chavent M, Kuentz-Simonet V, Labenne A, Saracco J (2018). Clustgeo: An R package for hierarchical clustering with spatial constraints. *Computational Statistics*, 33(4): 1799–1822.

Chen W, Castruccio S, Genton MG (2021). Assessing the risk of disruption of wind turbine operations in Saudi Arabia using bayesian spatial extremes. *Extremes*, 24(2): 267–292.

Cliff AD, Haggett P (1970). On the efficiency of alternative aggregations in region-building problems. *Environment and Planning A*, 2(3): 285–294.

Cressie N (2015). *Statistics for Spatial Data*. John Wiley & Sons.

Cressie N, Johannesson G (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 209–226.

Dale MR (2017). *Applying Graph Theory in Ecological Research*. Cambridge University Press.

Daw R, Simpson M, Wikle CK, Holan SH, Bradley JR (2022). An overview of univariate and multivariate Karhunen Loève Expansions in Statistics. *Journal of the Indian Society for Probability and Statistics*, 23: 1–42.

Duque JC (2004). *Design of Homogenous Territorial Units. A Methodological Proposal and Applications*. Universitat de Barcelona.

Duque JC, Anselin L, Rey SJ (2012). The max-p-regions problem. *Journal of Regional Science*, 52(3): 397–419.

Duque JC, Ramos R, Suriñach J (2007). Supervised regionalization methods: A survey. *International Regional Science Review*, 30(3): 195–220.

Ester M, Kriegel HP, Sander J, Xu X, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, volume 96, 226–231.

Fang K, Kifer D, Lawson K, Feng D, Shen C (2022). The data synergy effects of time-series deep learning models in hydrology. *Water Resources Research*, 58(4): e2021WR029583.

George JA, Lamar BW, Wallace CA (1997). Political district determination using large-scale network optimization. *Socio-Economic Planning Sciences*, 31(1): 11–28.

Giorgi F (2008). Regionalization of climate change information for impact assessment and adaptation. *Bulletin of the World Meteorological Organization*, 57(2): 86–92.

Gottmann J (1980). Spatial partitioning and the politician's wisdom. *International Political Science Review*, 1(4): 432–455.

Huang A, Wand MP (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8(2): 439–452.

Jarník V (1930). O jistém problému minimálním. (z dopisu panu o. borůvkovi). *Práce Moravské přírodovědecké společnosti*. 57–63.

Karhunen K (1946). *Zur Spektraltheorie Stochastischer Prozesse*. Annales Academiæ Scientiarum Fennicæ, 34.

Kirkley A (2022). Spatial regionalization as optimal data compression. *Communications Physics*, 5(1): 1–10. Nature Publishing Group.

Kleinberg J, Tardos E (2006). *Algorithm Design*. Pearson Education India.

Kruskal JB (1956). On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7(1): 48–50.

Laszlo M, Mukherjee S (2005). Minimum spanning tree partitioning algorithm for microaggregation. *IEEE Transactions on Knowledge and Data Engineering*, 17(7): 902–911.

Leeds WB, Wikle CK, Fiechter J (2014). Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. *Statistical Methodology*, 17: 126–138.

Lenzi A, Castruccio S, Rue H, Genton MG (2020). Improving Bayesian local spatial models in large datasets. *Journal of Computational and Graphical Statistics*, 30(2): 349–359.

Loève MM (1955). *Probability Theory*. Van Nostrand, Princeton, N.J.

Luo ZT, Sang H, Mallick B (2021). A bayesian contiguous partitioning method for learning clustered latent variables. *The Journal of Machine Learning Research*, 22(1): 1748–1799.

Lv X, Ma Y, He X, Huang H, Yang J (2018). Ccimst: A clustering algorithm based on minimum spanning tree and cluster centers. *Mathematical Problems in Engineering*. 2018.

MATLAB (2018). *9.7.0.1190202 (R2019b)*. The MathWorks Inc., Natick, Massachusetts.

Obled C, Creutin J (1986). Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *Journal of Applied Meteorology and Climatology*, 25(9): 1189–1204.

Openshaw S, Rao L (1995). Algorithms for reengineering 1991 census geography. *Environment and planning A*, 27(3): 425–446.

Pearson M (2007). *Us Infrastructure Finance Needs for Water and Wastewater Rural Community Assistance Partnership (RCAP)*. Community Resource Group, Washington, DC, USA.

Pradhan P, Kriewald S, Costa L, Rybski D, Benton TG, Fischer G, et al. (2020). Urban food systems: How regionalization can contribute to climate change mitigation. *Environmental Science & Technology*, 54(17): 10551–10560.

Prim RC (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, 36(6): 1389–1401.

Ramos MC, Barreto JOM, Shimizu HE, de Moraes APG, de Silva EN. (2020). Regionalization for health improvement: A systematic review. *PloS one*, 15(12): e0244078.

Rasmussen CE (2003). Gaussian processes in machine learning. In: *Summer School on Machine Learning*, 63–71. Springer.

Robinson WS (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, 38(2): 337–341.

Singleton AD, Spielman SE (2014). The past, present, and future of geodemographic research in the united states and united kingdom. *The Professional Geographer*, 66(4): 558–567.

Spielman SE, Folch DC (2015). Reducing uncertainty in the american community survey through data-driven regionalization. *PloS one*, 10(2): e0115626.

Teixeira LV, Assunção RM, Loschi RH (2019). Bayesian space-time partitioning by sampling and pruning spanning trees. *Journal of Machine Learning Research*, 20(85): 1–35.

Vecchia AV (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2): 297–312.

Wendland H (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *Journal of Approximation Theory*, 93(2): 258–272.

Werdell PJ, McClain CR (2018). *Satellite Remote Sensing: Ocean Color. Technical Report*. Elsevier.

Wikle CK, Milliff RF, Herbei R, Leeds WB (2013). Modern statistical methods in oceanography: A hierarchical perspective. *Statistical Science*, 28: 466–486.

Xu Y, Olman V, Xu D (2002). Clustering gene expression data using a graph-theoretic approach: An application of minimum spanning trees. *Bioinformatics*, 18(4): 536–545.