Constrained Spectral Clustering for Regionalization: Exploring the Trade-off between Spatial Contiguity and Landscape Homogeneity

Shuai Yuan, Pang-Ning Tan
Department of Computer Science
and Engineering
Michigan State University
East Lansing, MI-48824
Email: {yuanshu2,ptan}@msu.edu

Kendra Spence Cheruvelil
Lyman Briggs College and
Department of Fisheries and Wildlife
Michigan State University
East Lansing, MI-48824
Email: ksc@msu.edu

Sarah M. Collins, Patricia A. Soranno
Department of Fisheries and Wildlife
Michigan State University
East Lansing, MI-48824
Email: {colli636,soranno}@msu.edu

***Abstract*—A regionalization system delineates the geographical landscape into spatially contiguous, homogeneous units for landscape ecology research and applications. In this study, we investigated a quantitative approach for developing a regionalization system using constrained clustering algorithms. Unlike conventional clustering, constrained clustering uses domain constraints to help guide the clustering process towards finding a desirable solution. For region delineation, the adjacency relationship between neighboring spatial units can be provided as constraints to ensure that the resulting regions are geographically connected. However, using a large-scale terrestrial ecology data set as our case study, we showed that incorporating such constraints into existing constrained clustering algorithms is not that straightforward. First, the algorithms must carefully balance the trade-off between spatial contiguity and landscape homogeneity of the regions. Second, the effectiveness of the algorithms strongly depends on how the spatial constraints are represented and incorporated into the clustering framework. In this paper, we introduced a truncated exponential kernel to represent spatial contiguity constraints for region delineation using constrained spectral clustering. We also showed that a Hadamard product approach that combines the kernel with landscape feature similarity matrix can produce regions that are more spatially contiguous compared to other baseline algorithms.**

I. Introduction

A regionalization system delineates the geographical landscape into spatially contiguous, homogeneous units known as regions or zones. Regionalization systems are important as they provide the spatial framework used in many disciplines, including landscape ecology, environmental science, and economics, as well as for applications such as public policy and natural resources management [1], [2], [3], [4]. For example, the hierarchical system of hydrologic units described in [5] provides a standardized regionalization framework that has been widely used in water resource and land use studies [6].

McMahon et al. [7] presented two classes of approaches for region delineation. The first approach identifies regions with similar landscape characteristics from mapped data through visual pattern recognition [8], [9]. Since the region identification is performed manually, it requires considerable domain expertise to identify the primary factors that define each region. In addition, it is limited to delineating relatively small-scale regions, difficult to reproduce, and fails to document the contributions of different mapped data. Alternatively, a data-driven approach can be used to objectively identify the regions based on spatial variability of their landscape features. Multivariate clustering techniques such as k-means and hierarchical clustering [10], [11] are often employed to partition the geographical area into smaller spatial units [12]. For example, Host et al. [10] applied hierarchical k-means clustering on 20 years of monthly temperature and precipitation values across northwestern Wisconsin to identify regions with similar seasonal climatic trends. Hargrove et al. [11] performed k-means clustering using elevation, climatic, and edaphic factors to generate ecoregions for the conterminous United States. Despite their promise, one potential limitation of these existing clustering algorithms is that they do not guarantee the resulting regions will be spatially contiguous. Contiguity of the regions is a desirable criteria for applications that treat regions as individual entities representing a contiguous area of land for research, policy, and management purposes (e.g., for site-specific management in precision agriculture [13]). Therefore, alternative methods are needed that can effectively cluster similar areas based on mapped variables, but that have the added constraint of being spatially contiguous.

Constrained clustering [14] is a semi-supervised learning approach that uses the domain information provided by users to improve the clustering results. The domain information is typically provided as must-link (ML) and cannot-link (CL) constraints to be satisfied by the clustering solution. ML constraints restrict the pairs of points that must be assigned to the same cluster, whereas CL constraints specify the pairs of points that must be assigned to different clusters. Constrained clustering algorithms are designed to find a clustering solution that maximizes the within-cluster similarities and minimizes the number of violated constraints [14], [15]. Existing constrained clustering algorithms can be adapted to the region delineation problem by introducing constraints based on the proximity between the spatial units. For example, ML constraints can be created between pairs of units that are spatially adjacent to each other. These spatial constraints can then be used to guide the clustering process into finding regions that are both homogeneous and spatially contiguous.

This paper focuses on the application of constrained spectral clustering to the region delineation problem. Spectral clustering [16], [17] is a well-known clustering method that uses the eigenvector spectrum of a feature similarity matrix to find the underlying clusters of a given data set. Advantages of using spectral clustering include its flexibility in terms of incorporating diverse types of similarity functions, superiority of its clustering solution compared to the traditional k-means algorithm [18], and its well-established theoretical properties (including the consistency [19] and convergence [20] guarantees of the algorithm). Spectral clustering can also be viewed as an algorithm for solving a relaxed graph cut minimization problem [16], [17]. This fact makes it an appealing framework for constrained clustering because both the feature similarity matrix as well as the pairwise ML and CL constraints can be easily represented as an edge-weighted graph. Although there has been growing interest in developing constrained spectral clustering algorithms [21], [22], [23], [24], as will be shown in this paper, applying these algorithms to the region delineation problem is not trivial. Using a large-scale terrestrial ecology data set [25] as our case study, we showed that the existing algorithms must carefully balance the trade-off between spatial contiguity and landscape homogeneity of the regions. Otherwise, the regions produced by the existing algorithms may not be contiguous and can have arbitrary shapes and sizes. On the other hand, if the algorithms were biased toward producing only geographically connected regions, the landscape similarities within the regions might be too low.

We argued that the difficulties in applying existing constrained spectral clustering algorithms to the region delineation problem were due to the way the spatial contiguity constraints are represented and incorporated into the spectral clustering formulation. To overcome these difficulties, we introduced a new approach for representing spatial constraints in spectral clustering using truncated exponential kernels [26]. The truncated kernels can be parameterized to provide a more flexible way to specify the spatial extent to which the ML constraints are in effect, beyond just pairs of spatially adjacent units. We also proposed two algorithms, spatially-constrained spectral clustering (SSC) and binarized spatially-constrained spectral clustering (BSSC), for embedding the truncated exponential kernels into the spectral clustering formulation. Unlike previous methods, SSC and BSSC employ a Hadamard product approach to combine the truncated exponential kernel with feature similarity matrix. Our experimental results showed that the proposed algorithms produce spatially contiguous regions with higher landscape homogeneity compared to three state-of-the-art constrained clustering algorithms.

In summary, by investigating the application of constrained spectral clustering to the development of a regionalization system for landscape data, we make four important contributions:

- We demonstrated the inherent trade-off between spatial contiguity and landscape homogeneity when applying existing constrained spectral clustering algorithms to the region delineation problem.
- We proposed the truncated exponential kernels for representing spatial contiguity constraints. We showed that the flexibility in the kernels enables us to better control the trade-off between spatial contiguity and homogeneity of the resulting regions.
- We developed two algorithms, SSC and BSSC, to incorporate spatial constraints into spectral clustering formulation using the Hadamard product method with truncated exponential kernels.
- We presented the results of extensive experiments comparing the relative performance of various constrained spectral clustering algorithms and showed that the proposed algorithms are most effective in terms of producing spatially contiguous regions with homogeneous landscape features.

The remainder of this paper is organized as follows. Section II reviews previous work on the development of regionalization systems and constrained clustering algorithms. Section III formalizes the region delineation problem and presents an overview of spectral clustering. Section IV describes the different ways in which spatial constraints can be represented and augmented into the spectral clustering framework. Section V describes the application of spatially constrained spectral clustering to the region delineation problem. Section VI concludes with a summary of the results of this study.

II. Related Work

Region delineation has traditionally been studied as a spatial clustering [27] problem. Duque et al. [6] classified the existing data-driven approaches into two categories. The first category does not require explicit representation and integration of spatial constraints into the clustering procedure. Instead, the constraints are indirectly satisfied by post-processing the clusters or optimizing other related criteria. For example, Openshaw [28] applied a conventional clustering method followed by a cluster refinement step to split clusters that contained geographically disconnected patches. The second category of methods explicitly incorporates spatial constraints into the clustering algorithm [6]. Examples of such methods include adapted hierarchical clustering, exact optimization methods, and graph theory based methods. This second category also encompasses the constrained clustering methods developed n the fields of data mining and machine learning to incorporate side information from users to guide the clustering procedure. Previous studies include the development of constrained versions for K-means [15], self-organizing maps [29], and hierarchical clustering [30] algorithms.

There is also an emerging body of research that focuses on extending spectral clustering to deal with constraints [21], [22], [23], [24]. For example, Kamvar et al. [21] uses the ML and CL constraints to define the affinity matrix of the data. Shi et al. [23] proposed a constrained co-clustering method that considers both the similarity of features as well as the ML and CL constraints. All of these methods were designed to manipulate the graph Laplacian matrix using the domain constraints available. Alternatively, the constrained spectral clustering method can be designed to manipulate the feasible solution space of its optimization problem. For example, De Bie et al. [31] restricted the eigenspace to which the cluster membership vector will be projected. Wang and Davison [24] proposed a constrained spectral clustering method that considers real-valued constraint and imposed a minimum of constraints that must be satisfied in the feasible solution. None of these previous constrained spectral clustering methods were designed for the region delineation problem.

III. Preliminaries

This section formalizes region delineation as a constrained clustering problem and presents a brief overview of spectral clustering and its constrained-based method.

A. Region Delineation as Constrained Clustering Problem

Consider a spatial data set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \Re^d$ is a d-dimensional vector of landscape features associated with the geo-referenced spatial unit $\mathbf{s}_i \in \Re^2$. Let $\mathcal{R} = \{1, 2, \cdots, k\}$ denote the set of region identifiers, where k is the total number of regions, and $\mathcal{C} = \{(s_i, s_j, c_{ij})\}$ denote the set of spatial constraints. For region delineation, we consider only ML constraints, where $c_{ij} = +1$ if s_i and s_j are spatially adjacent to each other. Otherwise, $c_{ij} = 0$. The goal of region delineation is to learn a partition function $\mathcal{V}$ that maps each spatial unit s_i to its corresponding region identifier $r_i \in \mathcal{R}$ in such a way that (1) maximizes the feature similarity of the spatial units within each region and (2) minimizes the constraint violations in $\mathcal{C}$.

B. Spectral Clustering

Spectral clustering is a class of partitional clustering algorithms that relies on the eigendecomposition of feature similarity matrices to determine the cluster membership of its data points. Let $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ be a set of points to be clustered. To apply spectral clustering, we first compute an affinity (similarity) matrix $\mathbf{S}$ between every pair of data points. The affinity matrix is used to construct an undirected weighted graph $\mathcal{G} = (V, E)$, where V is the set of vertices (one for each data point) and E is the set of edges between pairs of vertices. The weight of each edge is given by the affinity between the corresponding pair of data points. The Laplacian matrix of the graph is defined as $\mathbf{L} = \mathbf{D} - \mathbf{S}$, where $\mathbf{D}$ is a diagonal matrix whose diagonal elements correspond to $\mathbf{D}_{ii} = \sum_j \mathbf{S}_{ij}$.

The spectral clustering solution can be found by solving the following constrained optimization problem [17]:

$$\arg\min_r r^T\mathbf{L}r \text{ such that } r^T\mathbf{D}r = \sum_i \mathbf{D}_{ii},\ \mathbb{1}^T\mathbf{D}r = \mathbf{0} \quad (1)$$

where $\mathbb{1}$ and $\mathbf{0}$ are vectors whose elements are all 1s and 0s, respectively. After simplification, the solution for r reduces to solving the following generalized eigenvalue problem: $\mathbf{L}r = \lambda\mathbf{D}r$. For a given number of clusters k, we extract the top k eigenvectors, which define a k-dimensional projection of the data. A standard clustering algorithm such as k-means is then applied to derive the final clusters from the k-dimensional manifold space.

C. Constrained Spectral Clustering

There are two main categories of approaches for incorporating constraints into spectral clustering algorithms. The first category encompasses methods that directly alter the graph Laplacian matrix. The simplest way to alter the matrix is by performing a weighted sum between the feature similarity matrix $\mathbf{S}$ and the adjacency matrix of the constraint graph, $\mathbf{C}$:

$$\text{Weighted sum:} \quad \mathbf{S}^{\text{total}}(\delta) = (1-\delta)\mathbf{S} + \delta\mathbf{C}, \quad (2)$$

where $\delta \in [0, 1]$ is a parameter that controls the trade-off between maximizing cluster homogeneity and preserving the ML constraints of the data. When δ approaches zero, the clustering solution is biased towards the feature similarity matrix whereas when δ approaches one, the solution is biased towards the constraint matrix. The modified graph Laplacian is given by a convex combination of the original graph Laplacian and the Laplacian induced by the constraint matrix:

$$\begin{aligned} \mathbf{D}^{\text{total}}_{ii} &= \sum_j \mathbf{S}^{\text{total}}_{ij}(\delta) \\ &= (1-\delta)\mathbf{D}_{ii} + \delta\mathbf{D}_{c,ii} \\ \mathbf{L}^{\text{total}} &= \mathbf{D}^{\text{total}} - \mathbf{S}^{\text{total}} \\ &= (1-\delta)(\mathbf{D} - \mathbf{S}) + \delta(\mathbf{D}_c - \mathbf{C}) \end{aligned} \quad (3)$$

This approach is a special case of the spectral constraint modeling (SCM) algorithm proposed by Shi et al. [23] for co-clustering problems. The altered graph Laplacian is substituted into Equation (1), which allows us to apply existing spectral clustering algorithm to identify the regions.

$$\begin{aligned} \text{SCM:} \quad & \arg\min_{r \in \mathbb{R}^N} r^T\mathbf{L}^{\text{total}}r \\ \text{s.t.} \quad & r^T\mathbf{D}^{\text{total}}r = \sum_i \mathbf{D}^{\text{total}}_{ii},\ \mathbb{1}^T\mathbf{D}^{\text{total}}r = \mathbf{0}. \end{aligned} \quad (4)$$

The second category of approaches for incorporating domain constraints is by altering the feasible solution set of the spectral clustering algorithm. For example, Wang and Davidson [24] proposed the CSP algorithm, which optimizes the following constrained optimization problem.

$$\begin{aligned} \text{CSP:} \quad & \arg\min_{r \in \mathbb{R}^N} r^T\bar{\mathbf{L}}r \\ \text{s.t.} \quad & r^T\bar{\mathbf{C}}r \geq \alpha,\ r^Tr = vol(\mathcal{G}),\ r \neq \mathbf{D}^{1/2}1, \end{aligned} \quad (5)$$

where $\bar{\mathbf{L}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$ and $\bar{\mathbf{C}} = \mathbf{D}_c^{-1/2}\mathbf{C}\mathbf{D}_c^{-1/2}$ are the normalized graph Laplacian and normalized constraint matrix, respectively. The threshold α gives a lower bound on the amount of constraints in $\mathbf{C}$ that must be satisfied by the clustering solution. Instead of setting the parameter for α, Wang and Davison [24] requires users to specify a related parameter β, which was shown to be a lower bound for α.

IV. Spatially Constrained Spectral Clustering

In this section, we describe the various ways to represent spatial contiguity constraints and to incorporate them into the spectral clustering framework.

A. Kernel Representation of Spatial Contiguity Constraints

For constrained spectral clustering, we can define a corresponding constraint graph $\mathcal{G}_C = (V, E_C)$, where V is the set of data points and E_C is the set of edges defined as follows:

$$E_{ij} = \begin{cases} 1, & (v_i, v_j) \text{ is a ML edge;} \\ -1, & (v_i, v_j) \text{ is a CL edge;} \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

For region delineation, the vertices of the constraint graph correspond to the set of spatial units to be clustered, while the ML edges correspond to pairs of spatial units that are adjacent to each other. It is also possible to define a CL edge between every pair of spatial units that are either located too far away from each other or are obstructed by certain barriers (e.g., large bodies of water) that make them unreasonable for assignment to the same region. However, since the number of CL edges tends to grow almost quadratically with increasing number of points, this severely affects the runtime of spectral clustering algorithm. Furthermore, the ML edges are often sufficient to provide guidance on how to form spatially contiguous regions. For these reasons, we consider constraint graphs that have ML edges only in this paper. Let $\mathbf{C}$ denote the adjacency matrix representation of the edge set E_C.

A constrained spectral clustering algorithm is designed to produce feasible solutions that are consistent with the constraints imposed by $\mathcal{G}_C$. Unfortunately, for region delineation, it may not be sufficient to use the adjacency information between neighboring spatial units to control the trade-off between spatial contiguity and landscape homogeneity of the regions. To improve its flexibility, we introduce a *spatially constrained kernel matrix*, $\mathbf{S}_c$. The simplest form of the kernel would be a linear kernel, which is defined as follows:

$$\text{Linear Kernel:} \qquad \mathbf{S}_c^{\text{linear}} = \mathbf{C} \tag{7}$$

More generally, we can define an exponential kernel [26] on the adjacency matrix $\mathbf{C}$ as follows:

Exponential Kernel:

$$\mathbf{S}_c^{\text{exp}} = e^{\mathbf{C}} = \mathbb{I} + \mathbf{C} + \frac{1}{2!}\mathbf{C}^2 + \frac{1}{3!}\mathbf{C}^3 + \cdots = \sum_{k=0}^{\infty} \frac{\mathbf{C}^k}{k!} \tag{8}$$

where $\mathbb{I}$ is the identity matrix. Since we consider only ML constraints, the k-th power of the adjacency matrix $\mathbf{C}$ represents the number of ML paths of length k that exist between every pair of vertices. An ML path between vertices (v_i, v_j) refers to a sequence of ML edges $e_1, e_2, \cdots, e_m$ such that the initial vertex of e_1 is v_i and the terminal vertex of e_m is v_j. It can be shown that $\mathbf{S}_c^{\text{exp}}$ is a symmetric, positive semi-definite matrix, and thus, is a valid kernel [26]. Furthermore, as the diameter of the constraint graph is finite, we also consider a truncated version of the exponential kernel:

$$\text{Truncated Exponential Kernel :} \quad \mathbf{S}_c^{\text{trunc}}(\delta) \equiv \sum_{k=0}^{\delta} \frac{\mathbf{C}^k}{k!} \tag{9}$$

where the parameter δ controls the ML neighborhood size of a vertex. The ML neighborhood specifies the set of vertices that should be in the same region as the vertex under consideration. As an example, consider the graph shown in Figure 1. When $\delta = 1$, the ML neighborhood for vertex A corresponds to its immediate neighbors, B, C, D and E. When $\delta = 2$, the ML neighborhood of vertex A is expanded to include vertices that are located within a path of length 2 or less from A, i.e., B, C, D, E, F, G, H and I. When $\delta = 3$, the ML neighborhood for vertex A includes all of the vertices in the graph. Note that each term in the summation given in Equation (8) is normalized by the path length; therefore, a vertex that is located further away from a given vertex has less influence as compared to a nearer vertex.

Finally, the truncated exponential kernel matrix can be binarized so that it can be interpreted as an adjacency matrix for an expanded constraint graph, whose ML neighborhood size is given by the parameter δ.

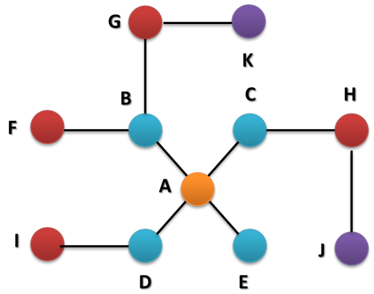


Fig. 1: An illustration of spatial contiguity constraint

Binarized Truncated Exponential Kernel :

$$\mathbf{S}_c^{\text{bin}}(\delta) \equiv \mathbf{I}\left[\sum_{k=0}^{\delta} \mathbf{C}^k > 0\right] \tag{10}$$

where $\mathbf{I}[\cdot]$ is an indicator function whose value is equal to 1 if its argument is true and 0 otherwise. Both the truncated and binarized truncated exponential kernels allow us to vary the degree to which the original constraint graph should be satisfied. As δ increases, the constraint satisfaction becomes more relaxed. Ultimately, when δ is greater than or equal to the diameter of the graph, $\mathbf{S}_c^{\text{bin}}$ reduces to a constant matrix of all 1s, which is equivalent to ignoring the spatial contiguity constraints.

B. Hadamard Product Graph Laplacian

We now describe our approach for incorporating the spatially constrained kernel matrix $\mathbf{S}_c$ (described in the previous section) into the spectral clustering formulation. Instead of using a weighted sum approach as given in Equation (2), we consider a Hadamard product approach to combine $\mathbf{S}_c$ with the feature similarity matrix $\mathbf{S}$:

$$\text{Hadamard Product:} \qquad \mathbf{S}^{\text{total}}(\delta) = \mathbf{S} \circ \mathbf{S}_c(\delta), \tag{11}$$

where the spatially constrained kernel matrix $\mathbf{S}_c(\delta)$ may correspond to the truncated exponential kernel (Equation (9)) or the binarized truncated exponential kernel (Equation (10)).

There are several advantages to using a Hadamard product approach to combine the matrices. First, unlike the weighted sum approach, it prevents spatial units that are located far away from each other from being assigned to the same cluster even though their feature similarity is high. Second, it produces a sparser kernel matrix, which is advantageous for large-scale graph analysis. Finally, it gives more flexibility to the users to specify the level of constraints that must be preserved by tuning the parameter δ, which controls the ML neighborhood size of the constraint graph.

Let $\mathbf{D}_{ii}^{\text{total}} = \sum_j [\mathbf{S} \circ \mathbf{S}_{(c)}(\delta)]_{ij}$ be elements of a diagonal matrix computed from $\mathbf{S}^{\text{total}}$. The Hadamard product graph Laplacian is given by $\mathbf{L}^{\text{total}} = \mathbf{D}^{\text{total}} - \mathbf{S} \circ \mathbf{S}_c(\delta)$. The modified graph Laplacian can be substituted into Equation (1) and solved using the generalized eigenvalue approach to identify the spatially contiguous regions.

Algorithm 1 Spatially-Constrained Spectral Clustering

```
Input:
$\mathcal{D} = \{(\mathbf{x}_1, \mathbf{s}_1), (\mathbf{x}_2, \mathbf{s}_2), ..., (\mathbf{x}_N, \mathbf{s}_N)\}$
$\mathbf{C} \in R^{N \times N}$: spatial constraint matrix.
$k$: number of clusters.
$\delta$: neighborhood size.
Output:
$\mathcal{R} = \{R_1, R_2, ..., R_k\}$ (set of regions).

1. Create similarity matrix $\mathbf{S}$ from $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$.
2. Compute the spatially constrained kernel matrix, $\mathbf{S}_c(\delta)$.
3. Compute the combined kernel $\mathbf{S}^{\text{total}}$ based on $\mathbf{S}$ and $\mathbf{S}_c$.
4. Compute $\mathbf{D}^{\text{total}}$ and $\mathbf{L}^{\text{total}}$.
5. Solve the generalized eigenvalue problem $\mathbf{L}^{\text{total}} r = \lambda \mathbf{D}^{\text{total}} r$.
   Create matrix $\mathbf{X}_r = [r_1 r_2 \cdots r_k]$ from the top-k eigenvectors.
6. $\mathcal{R} \leftarrow$ k-means($\mathbf{X}_r$, $k$)
```

C. Spatially Constrained Spectral Clustering Framework

This section summarizes our proposed spatially constrained spectral clustering approach. A high-level overview of the approach is given in Algorithm 1.

First, a feature similarity matrix is created by applying a Gaussian radial basis function kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$ to the feature set of the spatial units. The spatially constrained kernel matrix $\mathbf{S}_c$ is then computed from the constraint matrix $\mathbf{C}$, where $\mathbf{C}_{ij} = 1$ if $(\mathbf{s}_i, \mathbf{s}_j)$ is a ML edge and 0 otherwise. Note that if the truncated exponential kernel is used to represent the spatially constrained kernel matrix, we termed the approach as a spatially-constrained spectral clustering (SSC) algorithm. However, if the binarized truncated exponential kernel is used, the approach is known as a binarized spatially-constrained spectral clustering (BSSC).

Once the combined graph Laplacian, $\mathbf{L}^{\text{total}}$ is found, we extracted the first k eigenvectors as the low rank approximation of the combined kernel matrices. We then applied k-means clustering to partition the data into its respective regions. Note that the framework shown in Algorithm 1 is also applicable to the SCM and CSP algorithms, by setting their corresponding graph Laplacian, $\mathbf{L}^{\text{total}}$ and diagonal matrix, $\mathbf{D}^{\text{total}}$. Note that the computational complexity of spatially spatially constrained spectral clustering is equivalent to general spectral clustering, which is $O(N^3)$.

V. Application to Region Delineation

To evaluate the effectiveness of constrained spectral clustering for region delineation, we conducted a case study on a large-scale terrestrial ecology data set. The results of the case study are presented in this section.

A. Data set

The constrained spectral clustering methods were assessed using geospatial data from the $\text{LAGOS}_{\text{GEO}}$ [25] database. The database contains landscape characterization features measured at multiple spatial scales with a spatial extent that covers a land area spanning 17 U.S. states. The land area was divided into smaller hydrologic units (HUs), identified by their 12-Digit Hydrologic Unit Code [5]. Our goal was to develop a regionalization system for the landscape by aggregating the 20,257 HUs into coarser regions. We selected 28 terrestrial landscape variables and performed experiments on three study areas—Michigan, Iowa, and Minnesota. When the values for a landscape variable was always zero, we removed that variable before applying the clustering methods. The number of HUs to be clustered in each study region, as well as number of landscape variables for each, are summarized in Table I.

TABLE I: Summary statistics of the data set

Study Area	# HUs	# landscape variables	# PCA components	Diameter of constraint graph
Michigan	1,796	17	10	41
Iowa	1,605	19	12	43
Minnesota	2,306	19	11	57

The data set was further preprocessed before applying the constrained clustering algorithms. First, each variable was standardized to have a mean value of zero and variance of one. Since some of the landscape variables were highly correlated, we reduced the number of features by applying principal component analysis, and kept only the principal components that collectively explained at least 85% of the total variance. The principal component scores were then used to calculate a feature similarity matrix for all pairs of HUs in each study area. The ML edges for the constraint graph were determined based on whether the polygons for two HUs were adjacent to each other.

B. Baseline Methods

We compared the performance of our proposed constrained spectral clustering algorithms (SSC and BSSC) against three state-of-the-art baseline methods. The first baseline method, called SCM [23], uses the weighted sum approach (Equation (2)) to combine the feature similarity matrix $\mathbf{S}$ with the adjacency matrix $\mathbf{C}$ of the constraint graph. The algorithm has a parameter $\delta \in [0, 1]$ that controls whether the clustering should favor homogeneity or spatial contiguity of the regions. When δ approaches 0, the algorithm is biased towards maximizing the similarity of features in the regions whereas when δ approaches 1, it is biased towards producing more contiguous regions.

The second baseline method, called CSP [24], uses the spatial constraints to restrict the feasible set of the clustering solution (Equation (5)). As noted in Section III-C, the algorithm has a parameter β that gives a lower bound on the proportion of constraints that must be satisfied by the clustering solution. In addition, it was shown in [24] that $\beta < \lambda_{\max} vol(\mathcal{G})$ to ensure existence of a feasible solution. Instead of using β, we define a tuning parameter $\delta = \beta/[\lambda_{max} vol(\mathcal{G})]$ so that its upper bound is equal to 1 to be consistent with the upper bound of the tuning parameters for SCM and our proposed algorithms.

The third baseline is a spatially constrained clustering method proposed recently in the ecology literature by Miele et al. [32]. It uses a stochastic model to represent entities and interactions in a spatial ecological network. The cluster membership of each entity (spatial unit) is assumed to follow a multinomial distribution. Spatial constraints are introduced as a regularization penalty in the maximum likelihood estimation of the model parameters. We denote the model-based method as MB in the remainder of this paper.

For our experiments, we implemented the SCM, SSC, and BSSC algorithms in Matlab. For CSP and MB, we downloaded their software from the links provided by the authors[1].

C. Evaluation Metrics

We used two criteria to assess the performance of the algorithms. First, to determine whether the regions were ecologically homogeneous, we computed their within-cluster sum-of-square error (SSW), which is defined as follows [33]:

$$\text{SSW} = \sum_{i=1}^{k} \sum_{x \in C_i} dist(\mu_i, x)^2 \tag{12}$$

where μ_i is the centroid of the cluster C_i. The lower SSW is, the more homogeneous are the HUs in the regions.

The second criteria assesses the spatial contiguity of the resulting regions. We consider two metrics for this evaluation. The first metric computes the percentage of ML constraints preserved within the regions:

$$\text{PctML} = \frac{\text{\# ML edges within discovered regions}}{\text{Total \# of ML edges}} \tag{13}$$

The second metric corresponds to a relative contiguity metric proposed in the ecology literature by Wu and Murray [34]. The metric takes into consideration both the within patch contiguity (ϕ) and between patch contiguity (ν):

$$c = \frac{\phi + \nu}{\Omega} \tag{14}$$

where

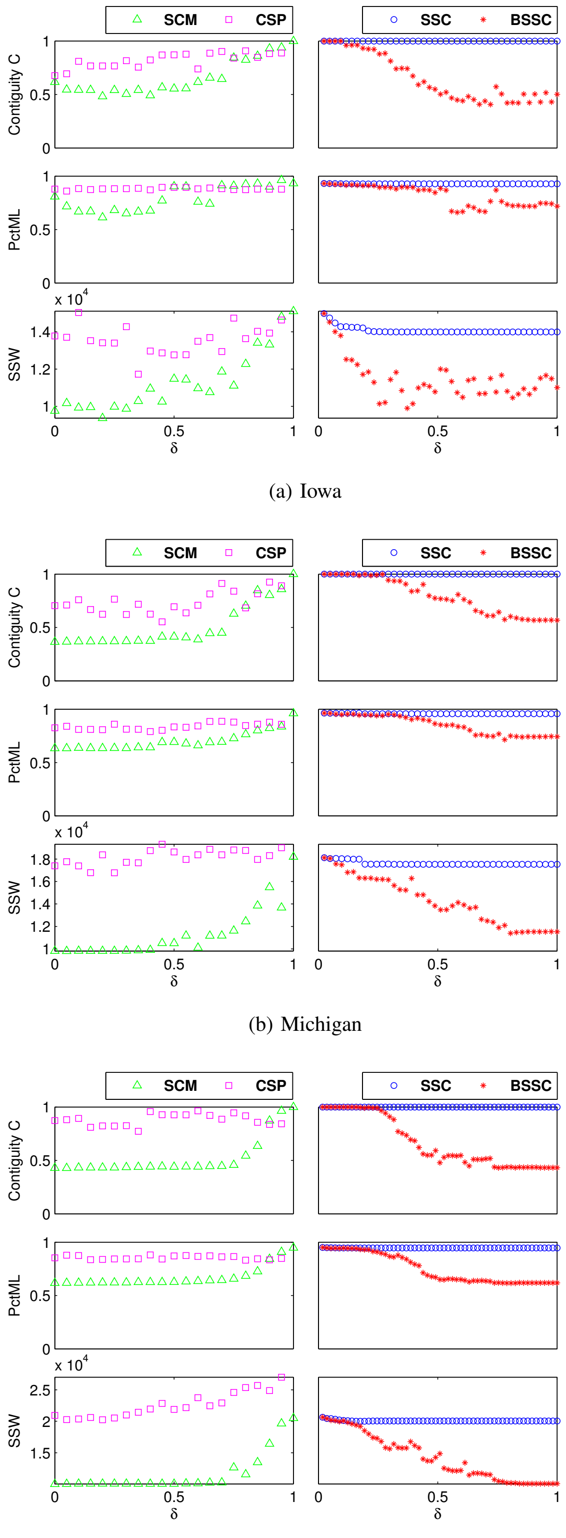
$$\begin{aligned} \phi &= \sum_{i=1}^{k} \left(\frac{N_i(N_i-1)}{2}\right), \quad \nu = \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} \left(\frac{N_i N_j}{l_{ij}^{\gamma}}\right) \\ \Omega &= \frac{(\sum_{i=1}^{k} N_i)(\sum_{i=1}^{k} N_i - 1)}{2} \end{aligned}$$

In the preceding formula, k is the number of clusters and N_i is the number of HUs assigned to the i-th cluster. l_{ij} denote the minimum spanning tree path length between clusters i and j while γ is a distance decay parameter. Since the metric is normalized by the total number of possible edges in a complete graph (Ω), it ranges between 0 and 1.

D. Results and Discussion

This section presents the results of applying various constrained clustering algorithms to the terrestrial ecology data.

1) Tradeoff between Homogeneity and Spatial Contiguity: We first analyze the trade-off between landscape homogeneity and spatial contiguity of the regions by comparing the results for four constrained spectral clustering algorithms: SCM, CSP, SSC, and BSSC. The number of clusters was set to 10. As each algorithm has a parameter δ that determines whether the clustering should be more biased towards increasing the within-cluster similarity or preserving the ML constraints, we varied the parameter and assessed their performance using the metrics described in Section V-C. The δ parameter for SSC and BSSC has been re-scaled to a range between 0 and 1 by



(a) Iowa

(b) Michigan

(c) Minnesota

Fig. 2: Comparison between various constrained spectral clustering algorithms in terms of their landscape homogeneity (SSW) and spatial contiguity (PctML and c). The horizontal axis in the plots corresponds to the parameter value δ.

[1] CSP was obtained from https://github.com/gnaixgnaw/CSP whereas MB was downloaded from http://lbbe.univ-lyon1.fr/Download-5012.html?lang=fr.

dividing the ML neighborhood size with the diameter of the constraint graph.

The results are shown in Figure 2. Observe that the contiguity score (c and PctML) for SCM increases rapidly as δ becomes closer to 1. This is because increasing δ would bias the algorithms towards preserving the spatial constraints. A similar increasing trend was also observed for CSP, especially in Iowa and Michigan, though the increase is not as sharp as SCM. In contrast, the contiguity scores would decrease for BSSC as δ increases because it allows for more pairs of spatial units to form ML edges, including pairs that are not close to each other. For SSC, the contiguity scores do not appear to change by much as δ increases. This is because the weight $1/k!$ associated with each path of length k decreases rapidly to zero as k increases. As a consequence, the ML neighborhood size for SSC grows until it reaches a maximum size by which increasing δ will not significantly alter the constraint graph. Thus, SSC is less sensitive to parameter tuning compared to BSSC. Figure 2 also shows there is generally an increasing trend in SSW for SCM and CSP as δ increases. For SSC, the SSW values do not appear to change significantly with increasing δ whereas for BSSC, the SSW curve decreases monotonically as the neighborhood size increases.

The results of this study showed that the trade-off between landscape homogeneity and spatial contiguity varies among the constrained spectral clustering algorithms. For CSP and SSC, the parameters provided by the algorithms do not allow us to achieve the full range of SSW and contiguity scores. Although these algorithms can produce regions with high contiguity scores, their SSW values were also very high. In contrast, with careful parameter tuning, SCM and BSSC can produce regions with significantly lower SSW compared to CSP and SSC. Observe that the slopes of the curves are steeper near $\delta = 1$ for SCM, which suggests that decreasing δ below 1 would lead to a dramatic reduction in the contiguity score and SSW of the regions. This makes it harder for SCM to produce regions that are both spatially contiguous and homogeneous. In contrast, the curves for the contiguity scores of BSSC are flatter near $\delta = 0$. This enables the BSSC algorithm to produce regions with homogeneous landscape features yet are still spatially contiguous.

2) Performance Comparison: In this experiment, we set the number of clusters to 10 and selected the δ parameter that gives the highest contiguity score for each constrained spectral clustering method. If there are more than one parameter values that achieve the highest contiguity score, we chose the one with lowest SSW. The Geoclust R package did not support parameter tuning, so we applied the MB algorithm using its built-in parameter values.

TABLE II: Performance comparison of the various constrained clustering algorithms on the three study regions. The number of clusters is set to 10.

Study Area	Method	PctML	c	SSW
Iowa	SCM	93.26%	1.00	15104
	CSP	87.37%	0.91	13628
	MB	89.95%	0.69	18997
	SSC	92.83%	1.00	13993
	BSSC	92.40%	1.00	14001
Michigan	SCM	96.08%	1.00	18200
	CSP	87.81%	0.92	18307
	MB	88.76%	0.65	16091
	SSC	95.69%	1.00	17534
	BSSC	94.92%	1.00	17485
Minnesota	SCM	94.78%	1.00	20506
	CSP	86.62%	0.96	23755
	MB	88.96%	0.64	20400
	SSC	94.57%	1.00	19998
	BSSC	94.12%	1.00	19594

Table II summarizes the results of our analysis. SCM, SSC, and BSSC can be tuned to produce regions that are fully contiguous ($c = 1$). The SSW for BSSC and SSC are consistently better than SCM. These results clearly showed the advantage of using a Hadamard product approach instead of a weighted sum approach to integrate spatial constraints into the feature similarity matrix. The limitation of using a weighted sum approach can be explained as follows. Since the highest contiguity score is achieved by setting $\delta = 1$, the clustering solution of SCM is equivalent to applying spectral clustering on the constraint graph only, without considering the feature similarity. If we reduce the parameter value to, say $\delta = 0.95$, its contiguity score decreases sharply (see Figure 2) while its SSW value is still worse than BSSC. The weighted sum approach has poor SSW because it significantly alters the feature similarity matrix. For example, consider the pairwise similarity values shown in the table below:

Pairs	Feature Similarity	ML Constraint	Weighted Sum	Hadamard Product
A-B	0.1	1	0.955	0.1
B-C	0.5	0	0.025	0
C-D	0.8	1	0.990	0.8

Although the A-B pair has a significantly lower similarity than C-D, the weighted sum approach inflates the similarity significantly (assuming $\delta = 0.95$) which makes it overall similarity to be comparable to C-D. In contrast, the Hadamard product approach simply zeros out the similarity of pairs that do not have ML edges without artificially inflating the similarities of pairs with ML edges.

Furthermore, since the feature similarity is computed using Gaussian radial basis function (see Section IV-C), the resulting matrix $\mathbf{S}$ for the weighted sum approach is still dense after incorporating the spatial constraints. Unless $\delta = 1$, the weighted sum approach will not prevent spatial units that are located far from each other from being placed into the same region. For example, consider the regions found by the weighted sum approach for Iowa, as shown in Figure 3. Although the regions appear to be spatially contiguous, they are not compact and have varying sizes. In fact, most of the spatial units were assigned to the same region when $\delta = 0.95$.

The contiguity scores for MB are worse than other constrained clustering methods. Nevertheless, it preserves at least 88% of the ML edges within the regions. Except for Michigan, its SSW values are also worse than other methods. In contrast, CSP has the lowest contiguity score among all the constrained spectral clustering methods. Except for Iowa, its SSW values are also among the worst. The limitation of CSP [24] is a consequence of the parameter used to control its spatial contiguity. As shown in Equation (5), the level of spatial constraints satisfied by the clustering solution depends on the

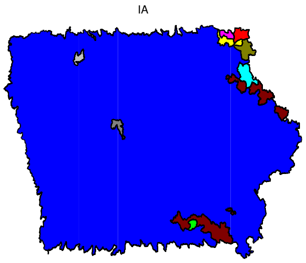


Fig. 3: The resulting regions for Iowa using the weighted sum approach (with $\delta = 0.95$).

parameter α. However, instead of directly tuning α, the authors suggested to vary another parameter, β, which was shown to be an upper bound of α. The results of our case study showed that increasing the value of β does not necessarily imply an increase in α. To illustrate this point, we randomly generated a constraint graph that has nine vertices with a randomly generated feature similarity matrix. Assuming the number of clusters is equal to 2, we ran the CSP algorithm with different parameter settings and plotted their values of α and β in Figure 9. Although this figure shows that the value of β (blue diamond) is a lower bound of α (red circle), the bound is so loose that it can not guarantee that increasing β will increase α. In fact, the figure on the right shows that α is not a monotonically increasing function of β. This is why controlling its parameter value will not always guarantee that the regions will be contiguous even when $\delta = 1$ (unlike SCM and the Hadamard product approaches).

Finally, the regionalization system generated by all the competing algorithms are shown in Figures 4 to 8. As can be seen from the figures, the regions produced by SSC and BSSC are more compact and uniform in size compared to CSP and MB. For SCM, although the regions are contiguous, their SSW values are worse than SSC and BSSC. The clustering results for SCM are also quite brittle. If δ is lowered slightly to 0.95, the regions changed significantly, as shown in Figure 3. For Iowa and Michigan, the region boundaries for SSC and BSSC are almost identical. In summary, the results in this subsection clearly shows the benefits of using BSSC to develop homogeneous and spatially contiguous regions compared to other baseline algorithms.

3) Effect of varying the number of clusters: Lastly, we varied the number of clusters k from 2 to 15 and compared the contiguity metrics as well as SSW for SCM, CSP, and BSSC. For each method, we tuned the parameter δ and plot the results with the best contiguity score in Figure 10. We observed that both BSSC and SCM can produce contiguous clusters while the CSP can not guarantee contiguity. In terms of landscape homogeneity, BSSC consistently better than the other two methods.

VI. Conclusions

This research investigated the feasibility of applying constrained spectral clustering to the region delineation problem. We compared several constrained spectral clustering methods and showed the trade-off between landscape homogeneity and spatial contiguity of their resulting regions. We also presented two algorithms, SSC and BSSC, that uses a Hadamard product approach to combine the similarity matrix of landscape features with spatial contiguity constraints. The results of our case study showed that the proposed BSSC method is most effective in terms of producing spatially contiguous regions that are homogeneous.

Acknowledgment

This work was supported by the National Science Foundation under the MacroSystems Biology Program in the Emerging Frontiers Division of the Biological Sciences Directorate through grant number EF-1065786.

References

[1] K. Cheruvelil, P. Soranno, K. Webster, and M. drivers of ecosystem state: Quantifying the spatial scale," Ecological Applications, vol. 23, pp. 1603–1618, 2013.

[2] J. A. Long, T. A. Nelson, and M. A. Wulder, "Regionalization of landscape pattern indices using multivariate cluster analysis," pp. 134–142, 2010.

[3] J. A. George, B. W. Lamar, and C. A. Wallace, "Political district determination using largescale network optimization," pp. 11–28, 1997.

[4] C. R. Margules, D. P. Faith, and L. Belbin, "An adjacency constraint in agglomerative hierarchical classifications of geographic data," Environment and Planning A, vol. 17, no. 3, pp. 397–412, 1985.

[5] P. R. Seaber, F. Kapinos, and G. L. Knapp, "Hydrologic unit maps," U.S. Geological survey water-supply papers, 1987.

[6] J. C. Duque, R. Ramos, and J. Suriach, "Supervised regionalization methods: A survey," International Regional Science Review, vol. 30, pp. 195–220, 2007.

[7] G. McMahon, S. Gregonis, S. Waltman, J. Omernik, T. Thorson, J. Freeouf, A. Rorick, and J. Keys, "Developing a spatial framework of common ecological regions for the conterminous united states," Environmental Management, vol. 28, no. 3, pp. 293–316, 2001.

[8] R. G. Bailey, "Ecoregions map of north america: Explanatory note," Miscellaneous Publication 1548, 1998.

[9] J. M. Omernik, "Ecoregions: A spatial framework for environmental management," in Biological assessment and criteria: tools for water resource planning and decision making, W. S. Davis and T. P. Simon, Eds. Boca Raton, Florida: Lewis Publishers, 1995, pp. 49–62.

[10] G. E. Host, P. L. Polzer, D. J. Mladenoff, M. A. White, and T. R. Crow, "A quantitative approach to developing regional ecosystem classifications," Ecological Applications, vol. 6, pp. 608–618, 1996.

[11] W. Hargrove and F. Hoffman, "Potential of multivariate quantitative methods for delineation and visualization of ecoregions," Environmental Management, vol. 34, pp. S39–S60, 2004.

[12] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[13] Y. Li, Z. Shi, F. Li, and H.-Y. Li, "Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land," Computers and Electronics in Agriculture, vol. 56, pp. 174–186, 2007.

[14] S. Basu, I. Davidson, and K. Wagstaff, Constrained Clustering: Advances in Algorithms, Theory, and Applications. Taylor and Francis, 2008.

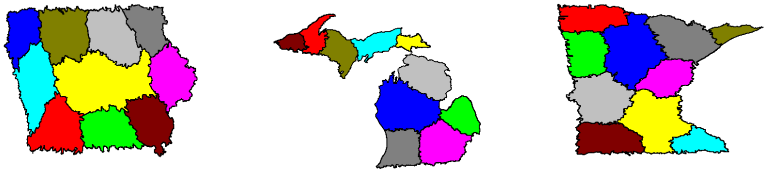
[15] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," in Proc of IEEE Int'l Conf on Machine Learning. Morgan Kaufmann, 2001, pp. 577–584.

[16] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 888–905, 1997.

[17] U. Luxburg, "A tutorial on spectral clustering," Statistics and Computing, vol. 17, no. 4, pp. 395–416, 2007.

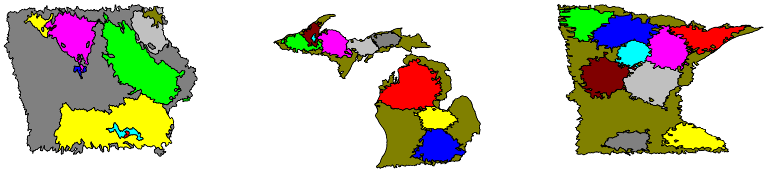
[18] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," 2001, pp. 849–856.

[19] U. Von Luxburg, M. Belkin, and O. Bousquet, "Consistency of spectral clustering," Annals of Statistics, vol. 36, no. 2, pp. 555–586, 2008.



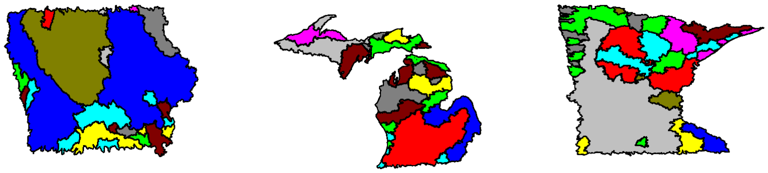
(a) Iowa (b) Michigan (c) Minnesota

Fig. 4: Regionalization system developed by the SCM algorithm (with $\delta = 1$).



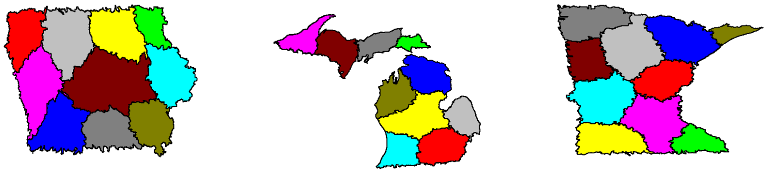
(a) Iowa (b) Michigan (c) Minnesota

Fig. 5: Regionalization system developed by the CSP algorithm.



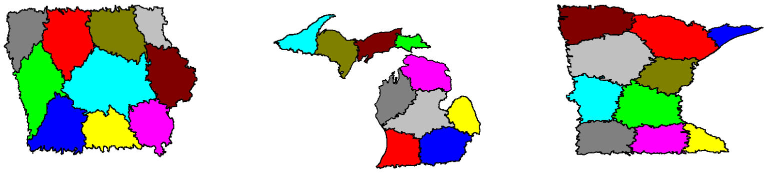
(a) Iowa (b) Michigan (c) Minnesota

Fig. 6: Regionalization system developed by the MB algorithm.



(a) Iowa (b) Michigan (c) Minnesota

Fig. 7: Regionalization system developed by the SSC algorithm.



(a) Iowa (b) Michigan (c) Minnesota

Fig. 8: Regionalization system developed by the BSSC algorithm.

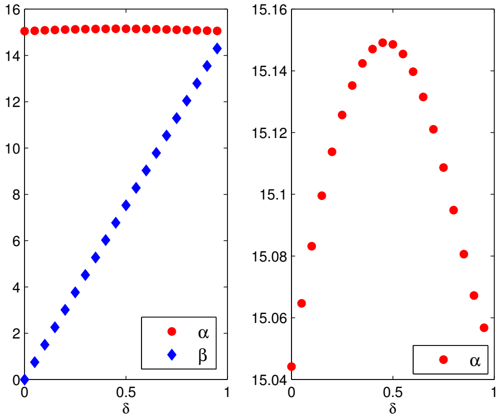


Fig. 9: The relationship between α and β parameter values for the CSP algorithm when applied to a synthetic graph data.

[20] U. Von Luxburg, O. Bousquet, and M. Belkin, “Limits of spectral clustering,” in Advances in Neural Information Processing Systems, 2004, pp. 857–864.

[21] S. D. Kamvar, D. Klein, and C. D. Manning, “Spectral learning,” in In IJCAI, 2003, pp. 561–566.

[22] D. Boley and J. Kawale, “Constrained spectral clustering using l1 regularization,” in SIAM Int’l Conference on Data Mining. SIAM, 2013, pp. 103–111.

[23] X. Shi, W. Fan, and P. S. Yu, “Efficient semi-supervised spectral co-clustering with constraints,” in Proc of IEEE Int’l Conf on Data Mining. IEEE Computer Society, 2010, pp. 1043–1048.

[24] X. Wang and I. Davidson, “Flexible constrained spectral clustering,” in 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2010, pp. 563–572.

[25] P. Soranno et al., “Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science through data reuse,” Journal of Giga Science, 2015.

[26] R. I. Kondor and J. D. Lafferty, “Diffusion kernels on graphs and other discrete input spaces,” in 19th Int’l Conference on Machine Learning. Morgan Kaufmann Publishers Inc., 2002, pp. 315–322.

[27] J. Han, M. Kamber, and A. Tung, “Spatial clustering methods in data mining: A survey,” in Geographic data mining and knowledge discovery, H. Miller and J. Han, Eds. Taylor and Francis, 2001, pp. 188–217.

[28] S. Openshaw, “A regionalisation program for large data sets,” Computer Applications, vol. 136, p. 47, 1973.

[29] F. Bacao, V. Lobo, and M. Painho, “Geo-self-organizing map (geo-som) for building and exploring homogeneous regions,” in GIScience 2004, ser. Lecture Notes in Computer Science, M. Egenhofer, H. Miller, and C. Freksa, Eds. Berlin: Springer, 2004, pp. 22–37.

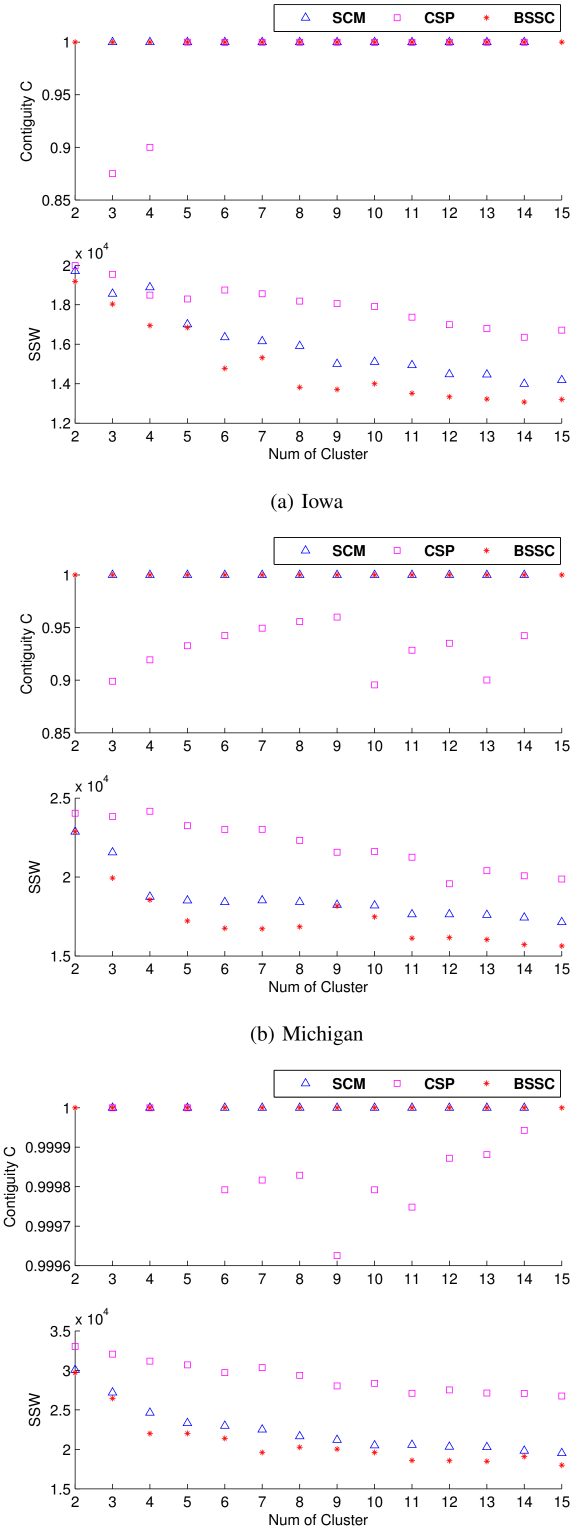
[30] I. Davidson and S. S. Ravi, “Agglomerative hierarchical clustering with constraints: Theoretical and empirical results,” in Lecture notes in computer science. Springer, 2005, pp. 59–70.

[31] T. D. Bie, J. A. K. Suykens, and B. D. Moor, “Learning from general label constraints.” in SSPR/SPR, vol. 3138. Springer, 2004, pp. 671–679.

[32] V. Miele, F. Picard, and S. Dray, “Spatially constrained clustering on ecological networks,” Methods in Ecology Evolution, vol. 5, no. 8, pp. 771–779, 2014.

[33] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to data mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc., 2005.

[34] X.Wu and A. T. Murray, “A new approach to quantifying spatial contiguity using graph theory and spatial interaction,” Journal of Geographical Information Science, vol. 22, pp. 387–407, 2008.



(a) Iowa

(b) Michigan

(c) Minnesota

Fig. 10: Comparison between SCM, CSP and BSSC with number of clusters range from 2 to 15