algorithms

Article

Shelved–Retrieved Method for Weakly Balanced Constrained Clustering Problems

Xinxiang Hou [1], Andong Qiu [2], Lu Yang [2] and Zhouwang Yang [1,*]

[1] School of Mathematical Sciences, University of Science and Technology of China, Hefei 230026, China; hxx0516@mail.ustc.edu.cn
[2] School of Data Science, University of Science and Technology of China, Hefei 230026, China; qad@mail.ustc.edu.cn (A.Q.); yl0501@mail.ustc.edu.cn (L.Y.)
* Correspondence: yangzw@ustc.edu.cn

Abstract: Clustering problems are prevalent in areas such as transport and partitioning. Owing to the demand for centralized storage and limited resources, a complex variant of this problem has emerged, also referred to as the weakly balanced constrained clustering (WBCC) problem. Clusters must satisfy constraints regarding cluster weights and connectivity. However, existing methods fail to guarantee cluster connectivity in diverse scenarios, thereby resulting in additional transportation costs. In response to the aforementioned limitations, this study introduces a shelved–retrieved method. This method embeds adjacent relationships during power diagram construction to ensure cluster connectivity. Using the shelved–retrieved method, connected clusters are generated and iteratively adjusted to determine the optimal solutions. Further, experiments are conducted on three synthetic datasets, each with three objective functions, and the results are compared to those obtained using other techniques. Our method successfully generates clusters that satisfy the constraints imposed by the WBCC problem and consistently outperforms other techniques in terms of the evaluation measures.

Keywords: weakly balanced constrained clustering; connectivity; shelved–retrieved method; centroidal power diagram

check for updates

Citation: Hou, X.; Qiu, A.; Yang, L.; Yang, Z. Shelved–Retrieved Method for Weakly Balanced Constrained Clustering Problems. *Algorithms* **2023**, *16*, 492. https://doi.org/10.3390/a16100492

Academic Editor: Yu-Chen Hu

Received: 15 September 2023
Revised: 8 October 2023
Accepted: 18 October 2023
Published: 23 October 2023

1. Introduction

Clustering is a foundational task for applications in real-world scenarios [1,2] including resource allocation [3] and site selection [4]. It typically involves partitioning a set of points into several subsets, referred to as clusters, such that the points in the same cluster are similar, while those in different clusters are dissimilar. In transportation and partitioning, it is not sufficient to merely partition points according to similarity for clustering [5,6]. Owing to limited resources and the need for efficient transport, clusters must additionally meet other specific requirements, among which we are most concerned about weight constraints and cluster connectivity [7–14].

Weight constraints originate from a prominent and challenging concern regarding resource limitations in transportation and partitioning scenarios. The weight of a point is usually associated with some problem-specified quantity, such as the size, area, or volume of the corresponding object. The cumulative weight of points within a cluster, also referred to as the cluster weight, is required to remain within a predetermined capacity range. For example, constraints of this kind are raised when the number of deliveries within the service area of an express service station must not exceed its designated capacity limits. This can be achieved by limiting the cluster weights within specific intervals.

Cluster connectivity is another type of constraint often encountered in practice. In scenarios such as farmland consolidation, some points in the area may be separated by barriers [12,13]. A route connecting two separate points is required to bypass the barriers, which can cost significantly more than connecting two points that are at the same distance but are not separated by any barrier. As such, transporting between these points directly may be impossible or unreasonable, thereby resulting in their disconnection. Generalized from disconnection between point pairs, the cluster connectivity constraint requests that a cluster cannot be split into two subsets such that any point pair between the subsets is disconnected. The inconsistency between connectivity and geometric proximity makes the cluster connectivity constraint a great challenge for the design of an algorithm.

In this paper, the similarity between points is assessed via a cost kernel, a commonly used technique that generalizes the classical Euclidean distance-based similarity. Similar to the partitioning clustering, the results should reach the goal that the costs within clusters are lower enough and the costs between the clusters are high enough according to the cost kernel. The clustering problem with cost kernel under the constraints on cluster weights and connectivity is formally known as the *weakly balanced constrained clustering* (WBCC) problem.

The clustering problem with a cost kernel is easy to solve by traditional clustering methods, but it is challenging to handle constraints on cluster weights and connectivity. Traditional clustering methods partition points into clusters using similarity metrics and optimize the compactness of clusters. The cluster weights are not relative to the reduction in the objective functions in clustering; thus, they cannot be adjusted through the optimization of the objective function. In addition, the cluster connectivity fails to be quantified by the cost kernel. Therefore, it is particularly challenging to address these constraints in clustering.

To handle the constraints on cluster weights, previous methods used power diagrams for clustering point sets [5,8,10,12,13]. They partitioned points $\boldsymbol{x} \in X$ into clusters C_i with associated sites $\boldsymbol{s}_i$ using an additively weighted distance, i.e.,

$$C_i = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{s}_i\|_2 - \alpha_i \leq \|\boldsymbol{x} - \boldsymbol{s}_j\|_2 - \alpha_j\}.$$

These methods obtain the cluster weights satisfying the constraints by the optimization of parameters α_i. Because these methods rely solely on distance metrics and optimization of parameters, they fail to automatically avoid barriers in certain areas. They cannot readily handle these constraints on cluster connectivity, so it is particularly challenging to handle the WBCC problems.

In response to those limitations, this paper introduces a shelved–retrieved method to solve the WBCC problems. The shelved–retrieved approach embeds adjacent relationships between points in the construction of power diagrams. It assigns points to the cluster to which adjacent points belong, thereby guaranteeing the connectivity of the cluster. Further, it takes advantage of power diagrams to obtain clusters satisfying the constraints in the WBCC problem by optimizing the cluster parameters and sites. The proposed method is guaranteed to produce a clustering result that satisfies the constraints on cluster weights and connectivity. Due to the versatility of the cost kernel, our method can carve out different cost functions in a variety of scenarios and obtain feasible clustering results with lower costs than existing methods. Furthermore, the clustering results generated by our method are more compact compared with other methods.

The remainder of this paper is organized as follows: Section 2 provides an overview of the existing methods for WBCC. Section 3 details the formulation of the WBCC problem and introduces the shelved–retrieved method. Section 4 presents the simulation results, and Section 5 concludes.

2. Related Work

Previous methods for WBCC can be categorized into two groups: conventional clustering on size-constrained clustering and clustering methods induced by diagrams on WBCC.

2.1. Conventional Methods on Size-Constrained Clustering

The size-constrained clustering problem [11,14,15] (characterized by assigning uniform weights) is a specialized WBCC problem. To address this problem, conventional clustering techniques have been employed to generate clusters of predetermined sizes.

The size-constrained clustering problem can be directly handled by traditional clustering methods [7,14,16]. The k-means method was modified to incorporate cluster size constraints using prior knowledge and can escape from local minima [14]. A Deterministic Annealing method [17] was used to handle clustering problems with several forms of size constraints [7]. A heuristic method [16] was incorporated into a conventional clustering approach as an extension. Additionally, matrix factorization techniques [18] were integrated into the shrinkage clustering method to identify clusters that fulfilled the size constraints. The fuzzy C-means method was used to handle the position and the shape of each cluster, and a wrapper algorithm was introduced to alleviate the cluster size insensitivity [15,19].

To reduce the complexity, other models are proposed to formulate size-constrained clustering. A Minimum Cost Flow linear network model [11] and a mixed integer programming model [20] were introduced to handle size-constrained clustering problems. These models were solved by linear programming or network simplex methods.

2.2. Clustering Methods Induced by Diagrams

Power diagrams were introduced in the clustering methods to address the general WBCC problem. In power diagrams, a geometric domain is partitioned into predefined sizes within a continuous space [21–23]. Similar to the capacity constrained partition problems, power diagrams can produce solutions to WBCC problems. The properties of power diagrams are harnessed to segment point sets into distinct clusters with specific size constraints [5,10,24]. In the clustering methods induced by diagrams, the additive weighted distances in power diagrams were introduced as the basis for classifying clusters. The clustering methods induced by diagrams used parameter tuning to adjust the cluster weights.

In clustering methods induced by diagrams, several models were proposed to formulate WBCC problems. For example, a transportation network model was constructed and resolved using network Voronoi diagrams and a pressure equalizer approach [5]. Furthermore, a quadratic optimization model was formulated to address WBCC, with its optimal solution derived from power diagrams in discrete space [9,10,12,13,24]. These models were constructed for the requirements of real-world scenarios, and they optimize different objective functions.

2.3. Analysis of Related Work

Conventional methods on size-constrained clustering and clustering methods induced by diagrams are introduced above. Table 1 summarizes the benefits and limitations of all methods. Conventional methods on size-constrained clustering have efficiently addressed size-constrained clustering problems. However, they may fail to produce the required clusters when applied to general WBCC problems in diverse scenarios. Clustering methods based on diagrams can address WBCC problems in convex cases. However, power diagrams rely exclusively on a convex partition strategy to handle this problem, thereby hindering them from ensuring cluster connectivity. Hence, this study introduces a shelved–retrieved method as an innovative approach to overcome this limitation.

Table 1. Summarizing of methods in WBCC problems.

Category	Method	Benefits	Limitations
Size-constrained	Modified k-means method [14]	Escaping from local minima	High computational complexity
	Deterministic Annealing method [7]	Fast convergence	Low accuracy
	Heuristic method [16]	Fast convergence	Low accuracy
	Shrinkage clustering method [18]	Ease of implementation	High computational complexity
	Fuzzy C-means method [15,19]	High stability	High computational complexity
	Minimum Cost Flow method [11]	Escaping from local minima	Low efficiency
	Mixed integer programming method [20]	High accuracy	Low efficiency

Table 1. *Cont.*

Category	Method	Benefits	Limitations
Diagram-induced	Network Voronoi diagrams method [5]	High accuracy	Low efficiency
	Power diagrams method in discrete space [9,10,24]	High accuracy	High computational complexity

3. Methodology

In this section, the WBCC problem and the shelved–retrieved method are formulated and introduced, respectively.

3.1. Mathematical Formulation

For a given point set $X = \{\boldsymbol{x}_1 = (x_1^1, \dots, x_1^d), \dots, \boldsymbol{x}_m = (x_m^1, \dots, x_m^d)\}$, each point is assigned a weight $\omega_j = \omega(x_j) > 0$ from a weight set $\Omega = \{\omega_1, \dots, \omega_m\}$ to represent its quantity information. This set is divided into n clusters, wherein the binary variable $\xi_{i,j}$ indicates whether the point $\boldsymbol{x}_j \in X$ belongs to cluster C_i. For instance, $\xi_{i,j} = 1$ indicates that point $\boldsymbol{x}_j$ belongs to cluster C_i. Further, the weight of cluster C_i is determined by $\omega(C_i) = \sum_{j=1}^{m} \xi_{i,j}\omega_j$, which is required to satisfy the balancing constraint $\omega(C_i) \in [\kappa_i^-, \kappa_i^+]$, wherein the minimal capacity $\kappa_i^- > 0$ and the maximal κ_i^+ constitute the set K^- and K^+, respectively.

Except for the constraints on the cluster weights, the WBCC problem requires cluster connectivity. According to the cost kernel f, a weight matrix is given in the datasets. Then, the corresponding graph $G = (V, E)$ can be generated. The node set is defined as $V = X$, and edges are added to the edge set E if the corresponding edge weight is finite in the weight matrix. Further, cluster C_i is connected if each induced subgraph $G[C_i]$ is also connected.

In this study, the WBCC problem uses the cost kernel $f(\cdot,\cdot)$ to construct the objective function. The cost kernel measures the transportation costs between points $\boldsymbol{x}_j$ and $\boldsymbol{s}_i \in C_i$ in each scenario. The decision variables in this problem formulation are clustering C_i, $i = 1, \dots, n$ and their corresponding sites $\boldsymbol{s}_i, i = 1, \dots, n$.

The mathematical formulation of the WBCC problem is as follows:

$$\begin{aligned}
\min_{C_i, \boldsymbol{s}_i \in C_i,\ i=1,2,\dots,n} \quad & \sum_{i=1}^{n} \sum_{\boldsymbol{x}_j \in C_i} f(\boldsymbol{x}_i, \boldsymbol{s}_i), \\
\text{s.t.} \quad & \sum_{j=1}^{m} \omega_j \xi_{i,j}, \in [\kappa_i^-, \kappa_i^+], \quad i = 1, 2, \dots, n, \\
& \sum_{i=1}^{n} \xi_{i,j} = 1, \quad j = 1, \dots, m, \\
& \xi_{i,j} \in \{0, 1\}, \quad i = 1, \dots, n;\ j = 1, \dots, m, \\
& G[C_i] \text{ is connected}, \quad i = 1, \dots, n.
\end{aligned} \quad . \tag{1}$$

Obviously, Model (1) has a feasible solution when all κ_i^- and κ_i^+ are set to zero and $\sum_{j=1}^{n} \omega_j$, respectively. There are some extreme situations in which Model (1) is unsolvable. To guarantee that Model (1) has a feasible solution, the dataset should satisfy Assumption 1. Then, Model (1) has a feasible solution, as proven in Theorem 1.

Assumption 1. *The graph G is connected, and K^-, K^+ satisfy the following qualities:*

$$\sum_{i=1}^{n} \kappa_i^- + n \max_{i=1,\dots,n} (\kappa_i^+ - \kappa_i^-) < \sum_{j=1}^{m} \omega_j < \sum_{i=1}^{n} \kappa_i^+ - n \max_{i=1,\dots,n} (\kappa_i^+ - \kappa_i^-), \tag{2}$$

$$\kappa_i^+ - \kappa_i^- > \max_{j=1,\dots,m} \omega_j. \tag{3}$$

Theorem 1. *Under Assumption 1, Model (1) has a feasible solution.*

Proof. $n = 2$: Graph G can be divided into two connected sub-graphs, G_1 and G_2, where $G_1 = (V_1, E_1)$, $G_2 = (V_2, E_2)$.

We assume $\omega(V_1) < \kappa_1^-$. We can select point $\boldsymbol{x}_t \in V_2$, such that $G[V_1 + \boldsymbol{x}_t]$ and $G[V_2 - \boldsymbol{x}_t]$ is connected. Due to Equality (3), we can obtain

$$\omega(V_1 + \boldsymbol{x}_t) = \omega(V_1) + \omega_t \leq \kappa_1^+.$$

If $\omega(V_1 + \boldsymbol{x}_t) < \kappa_1^-$, we repeat the above operations. If $\omega(V_1 + \boldsymbol{x}_t) \geq \kappa_1^-$, we prove this partition is a feasible solution. Due to Formula (2),

$$\omega(V_1) + \omega(V_2) > \kappa_1^- + \kappa_2^- + (\kappa_1^+ - \kappa_1^-) + \omega_t = \kappa_1^+ + \kappa_2^- + \omega_t > \kappa_2^- + \omega(V_1) + \omega_t.$$

Then, $\omega(V_2) - \omega_t \geq \kappa_2^-$. Similarly, $\omega(V_2) - \omega_t \leq \kappa_2^+$.

We assume that when $n = k$, Model (1) has a feasible solution. When $n = k + 1$, we partition the point set into two clusters V_1 and $\hat{V}_2$ with $\hat{\kappa}_1^- = \kappa_1^-, \hat{\kappa}_1^+ = \kappa_1^+, \hat{\kappa}_2^+ = \sum_{i=2}^{k+1} \kappa_i^-$, $\hat{\kappa}_2^+ = \sum_{i=2}^{k+1} \kappa_i^+$. Then, we can obtain $k + 1$ clusters by partitioning the point set $\hat{V}_2$.

By induction, Model (1) has a feasible solution under Assumption 1 for all $n \geq 2$. □

3.2. Shelved–Retrieved Method

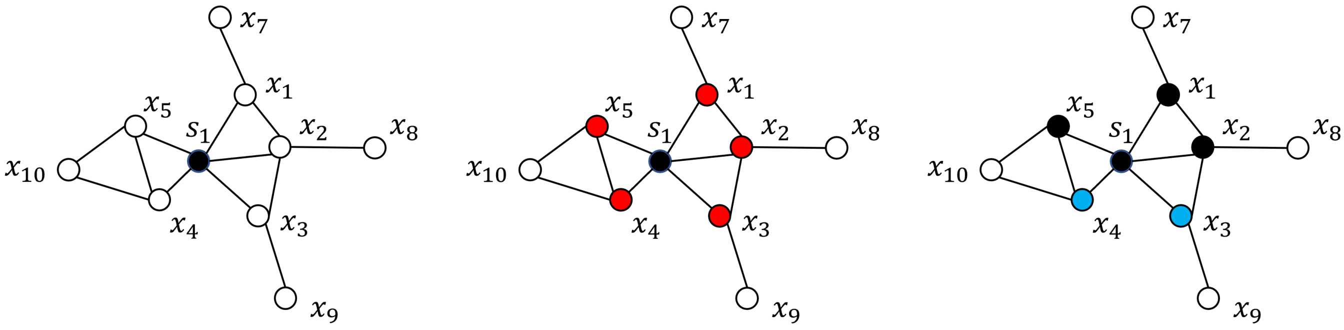
To solve the WBCC model (1), the shelved–retrieved method embeds adjacent relationships into the construction of the power diagrams. This integration is essential for the generation of connected clusters.

It is crucial that clustering results remain connected throughout the process. Then, each point should be adjacent to at least one other point within the same cluster during the clustering procedure. In other words, they can only be assigned to clusters to which adjacent points belong. Here, we specify the assignment process for the shelved–retrieved method.

We take cluster one as an example. The shelved–retrieved method randomly selects a point from the dataset to serve as the initial site for cluster one. This selected point is assigned to cluster one and colored black in Figure 1a. Further, the adjacent hollow points of black points $\boldsymbol{s}_1$, $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, $\boldsymbol{x}_4$, $\boldsymbol{x}_5$ are identified and colored red in Figure 1b. According to parameters α_i, $i = 1, \dots, n$, the shelved–retrieved method estimates whether $f(\boldsymbol{x}_j, \boldsymbol{s}_1)^2 - \alpha_1$ is smaller than $f(\boldsymbol{x}_j, \boldsymbol{s}_i)^2 - \alpha_i, i = 2, \dots, n$ at each point $\boldsymbol{x}_j$. Assuming that $f(\boldsymbol{x}_1, \boldsymbol{s}_1)^2 - \alpha_1$, $f(\boldsymbol{x}_2, \boldsymbol{s}_1)^2 - \alpha_1$, $f(\boldsymbol{x}_5, \boldsymbol{s}_1)^2 - \alpha_1$ are the smallest and $f(\boldsymbol{x}_3, \boldsymbol{s}_1)^2 - \alpha_1$, $f(\boldsymbol{x}_4, \boldsymbol{s}_1)^2 - \alpha_1$ are not, the shelved–retrieved method assigns points $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_5$ to cluster one, and we color them black in Figure 1c. Points $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ are colored blue. The shelved–retrieved method repeats the aforementioned operations until no additional adjacent hollow points are identified.

Several blue and hollow points might not have been assigned to any cluster. These blue points are adjacent to other points within the same cluster during the clustering process. They are assigned to the clusters to which their adjacent points belong.

Using blue points $\boldsymbol{x}_j$, the shelved–retrieved method identifies a set of clusters denoted by $\mathcal{A} = \{C_i : \boldsymbol{x}_k \in C_i,\ \boldsymbol{x}_k \text{ is a black point adjacent to } \boldsymbol{x}_j\}$. The minimum $d(\boldsymbol{x}_j, \boldsymbol{s}_{i^*})^2 - \alpha_{i^*}$ is within the set $\{f(\boldsymbol{x}_j, \boldsymbol{s}_i)^2 - \alpha_i : \boldsymbol{s}_i \in C_i, C_i \in \mathcal{A}\}$, while point $\boldsymbol{x}_j$ is assigned to cluster i^*. The shelved–retrieved method repeats the process until all the points are turned black, as shown in Figure 1. Here, the aforementioned mechanism for clustering points with fixed parameters is concluded in Algorithm 1.



(**a**) (**b**) (**c**)

Figure 1. The assigning process of the shelved–retrieved method: The black, hollow, red, and blue points represent those that belong to clusters, have not undergone processing yet, are eligible candidates for assignment to clusters, and are temporarily unassigned to any clusters, respectively. (**a**) Initial state; (**b**) Finding the adjacent points; (**c**) Assigning points.

Algorithm 1 Clustering by power diagrams based on connectivity

```
Require: Domain $X = \{x_1, \ldots, x_m\}$, sites $s_1, \ldots, s_n$, parameters $\alpha_1, \ldots, \alpha_n$, adjacent set $A_{x_j}, j = 1, \ldots, m$
Ensure: Clustering solution $\mathcal{C} = \{C_1, \ldots, C_n\}$
Initialize blue point set $B = \emptyset$, black point set $K = \emptyset$, red point set $R = \emptyset$, cluster $C_i = \{s_i\}$, $i = 1, \ldots, n$.
Update $R$ based on $A_{s_i}$.
while $R \neq \emptyset$ do
    for $x_j \in R$ do
        for $i = 1, \ldots, n$ do
            if $i = \arg\min_{i'=1,\ldots,n} f(x_j, s_{i'})^2 - \alpha_{i'}$ then
                Assign $x_j$ to $C_i$ and $K$, and update $R$ based on $A_{x_j}$.
            end if
        end for
        if $x_j \notin K$ then
            Assign $x_j$ to $B$.
        end if
    end for
    if $R = \emptyset$ and $K \neq X$ then
        for $x_j \in B$ do
            Calculate $i = \arg\min_{i' \in \{C_i : C_i \cap K \cap A_{x_j} \neq \emptyset\}} f(x_j, s_{i'})^2 - \alpha_{i'}$.
            Assign $x_j$ to $C_i$ and $K$, and update $R$ based on $A_{x_j}$.
        end for
    end if
end while
```

The shelved–retrieved method uses Algorithm 2 to cluster the points. In each iteration, denoted by p, parameters α_i and $e_i := \omega(C_i) - \kappa_i$ are represented by α_i^p and e_i^p, respectively. $\mathrm{sgn}(\cdot)$ denotes the sign function. The sites can be updated by the Lloyd algorithm [25] in the clustering.

$$\alpha_i^{p+1} = \begin{cases} \alpha_i^p - 0.1\mathrm{sgn}(e_i^p)\min_{j=1,\ldots,n,j\neq i} d(s_i, s_j) & \text{if } p = 1 \\ \alpha_i^p - 0.2\mathrm{sgn}(e_i^p)\min_{j=1,\ldots,n,j\neq i} d(s_i, s_j)/n & \text{if } p \geq 2 \text{ and } (\alpha_i^p - \alpha_i^{p-1})(e_i^p - e_i^{p-1}) \neq 0 \\ \alpha_i^p - \min\left\{ \left| \frac{\alpha_i^p - \alpha_i^{p-1}}{e_i^p - e_i^{p-1}} e_i^p \right|, 0.2\min_{j=1,\ldots,n,j\neq i} d(s_i, s_j)/n \right\} \mathrm{sgn}(e_i^p) & \text{otherwise} \end{cases}. \quad (4)$$

Algorithm 2 Shelved-retrieved method

```
Require: Domain $X = \{x_1, \dots, x_m\}$, weight set $\Omega = \{\omega_1, \dots, \omega_m\}$, set $K = \{\kappa_1, \dots, \kappa_n\}$, weight matrix $A_\omega$, adjacent set $A_{x_j}$ for each point $x_j$, and maximum number of iterations $M$.
Ensure: Clustering solution $\mathcal{C} = \{C_1, \dots, C_n\}$
Initialize $s_1, \dots, s_n, \alpha_1 = 0, \dots, \alpha_n = 0$.
repeat
    $p = 1$
    repeat
        Obtain the clustering $\mathcal{C}$ by the Algorithm 1.
        Update $\alpha_i^{p+1}$ by Formula (4) for each $i$.
        $p \leftarrow p + 1$
    until $p > M$ or $\mathcal{C}$ satisfies the constraints on cluster weights.
    Update $s_1, \dots, s_n$ by the Lloyd algorithm.
until $s_1, \dots, s_n$ remain the same values.
```

Using the aforementioned process, the shelved–retrieved method can produce a connected solution for Model (1), as stated in Lemma 2.

Lemma 2. *The shelved–retrieved method can produce a feasible solution for Model* (1)*.*

Proof. The shelved–retrieved method ensures that each point in a cluster is connected to the cluster site through a path, thereby resulting in connected clustering. Here, we prove that the clustering results satisfy the inequality constraints on the cluster weights.

For each cluster C_i, we can obtain point x_j that satisfies

$$j = \mathop{\arg\max}_{j: x_j \in C_i} \frac{(x_i - s_i) \cdot (s_k - s_i)}{f(s_k, s_i)}.$$

Based on the process of assigning points to the clusters, we propose the following inequality:

$$\alpha_k - \alpha_i < f(x_j, s_k)^2 - f(x_j, s_i)^2, \ \forall \ e_{ik} \in E_s.$$

Further, we transform these inequalities into Standard equality system (5).

$$\begin{aligned} \alpha'_k - \alpha''_k - (\alpha'_i - \alpha''_i) + \beta_{ik} = f(x_j, s_k)^2 - f(x_j, s_i)^2, &\qquad \forall e_{ik} \in E_s \\ \alpha'_i, \alpha''_i, \beta_{ik} \geq 0, &\qquad \forall e_{ik} \in E_s \end{aligned} \tag{5}$$

The rank of the equality system is lower than $|E_s|$, both of which are smaller than the number of variables $2n + |E_s|$. Hence, Equality system (5) has a solution, and a power diagram is built.

By fixing the other parameters α_k, $k \neq i$, the weight of cluster C_i is within the interval $[\omega(s_i), \sum_{j=1}^{m} \omega_j]$. We consider cluster weight $\omega(C_i)$ as a response variable and parameter α_i as an independent variable. A step function is defined as the mapping from the independent variable to the response variable. Parameter α_i exists such that the corresponding cluster weight $\omega(C_i)$ in the step function satisfies the inequality constraint in Model (1). Because the sites are optimized using the shelved–retrieved method, clustering $\mathcal{C}$ produced by the shelved–retrieved method is the optimal solution for Model (1). □

We offer the computational complexity analysis of Algorithm 2 as follows. In each iteration with a given red point set, Algorithm 1 handles these red points at most $O(mn\Delta(G))$ times. Algorithm 1 loops through the red point set, at most, $m - n$ times; thus, the computation burden of Algorithm 1 is $O((m - n)mn\Delta(G))$. Algorithm 2 iterates, at most, $O(k_1 k_2 (m - n)mn\Delta(G))$ times, where k_1 denotes the number of site iterations and k_2 denotes the number of parameter iterations.

4. Results

We conducted the experiments on synthetic datasets and farmland consolidation with Windows 10, Intel(R) Core(TM) i7-10700K CPU @ 3.8GHz. The models were implemented using Python 3.8.5. To compare the results, we used two metrics for evaluation: the transportation costs (objective function value) and root-mean-square standard deviation (RMSSTD) [26]. A lower RMSSTD value indicates a better performance. RMSSTD was calculated using the following formula:

$$\mathrm{RMSSTD} = \left(\sum_{i=1}^{n} \frac{\sum_{x_j \in C_i} f(\boldsymbol{x}_j, \boldsymbol{s}_i)^2}{d \sum_{i=1}^{n} (|C_i| - 1)} \right)^{\frac{1}{2}}.$$

4.1. Synthetic Datasets

Experiments were conducted on three synthetic cases with three objective functions to evaluate the validity and rationality of the shelved–retrieved method. The discrete form of the centroidal power diagram (D-CPD) method, which is specified in Appendix A, was used as a comparison method in the experiments.

For all synthetic cases, 2000 points were generated in a funnel-shaped region, and chosen to effectively represent concave conditions. Three separate cases were constructed to evaluate the effectiveness of the proposed method on datasets containing different numbers of clusters. In Case 1, three subarea capacity intervals were set as [1969.49, 2089.49], [5819.85, 5939.85], and [2182.77, 2302.77]. In Case 2, we increased the number of subareas and the capacity intervals were set as [1969.49, 2089.49], [3819.85, 3939.85], [2182.77, 2302.77], and [1819.85, 2039.85]. In Case 3, we set five subareas with capacity intervals [1969.49, 2089.49], [2819.85, 2939.85], [2182.77, 2302.77], [1819.85, 2039.85], [819.85, 1045.85].

Transportation costs in different scenarios served as objective functions. Three distinct cost kernels were constructed to quantify different transportation costs and validate the effectiveness of our method across a range of scenarios.

First, the following Euclidean function was used as the cost kernel in the experiments as expressed below:

$$\hat{f}_1(\boldsymbol{x}_j, \boldsymbol{s}_i) = \|\boldsymbol{x}_j - \boldsymbol{s}_i\|_2.$$

In addition to the Euclidean function, two additional cost kernels, $\hat{f}_2$ and $\hat{f}_3$, were used in the experiments as expressed below:

$$\hat{f}_2(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})M_1(\boldsymbol{x} - \boldsymbol{y})^T,$$

$$\hat{f}_3(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})M_2(\boldsymbol{x} - \boldsymbol{y})^T.$$

Here,

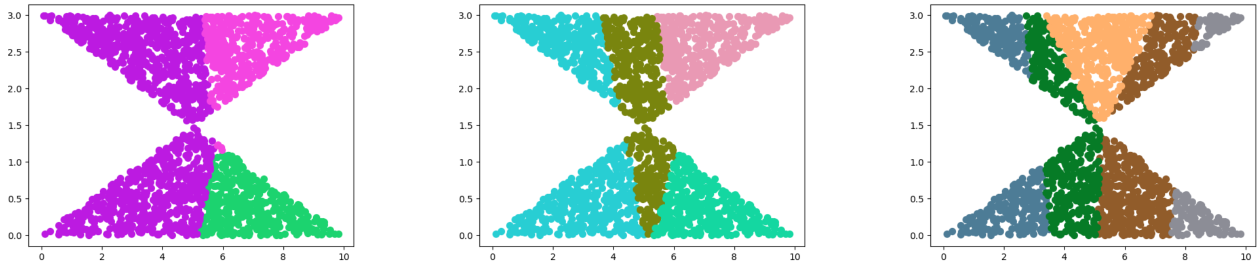
$$M_1 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$

is a positive-definite matrix, but

$$M_2 = \begin{pmatrix} 1 & 1 \\ 2 & 4 \end{pmatrix}$$

is not.

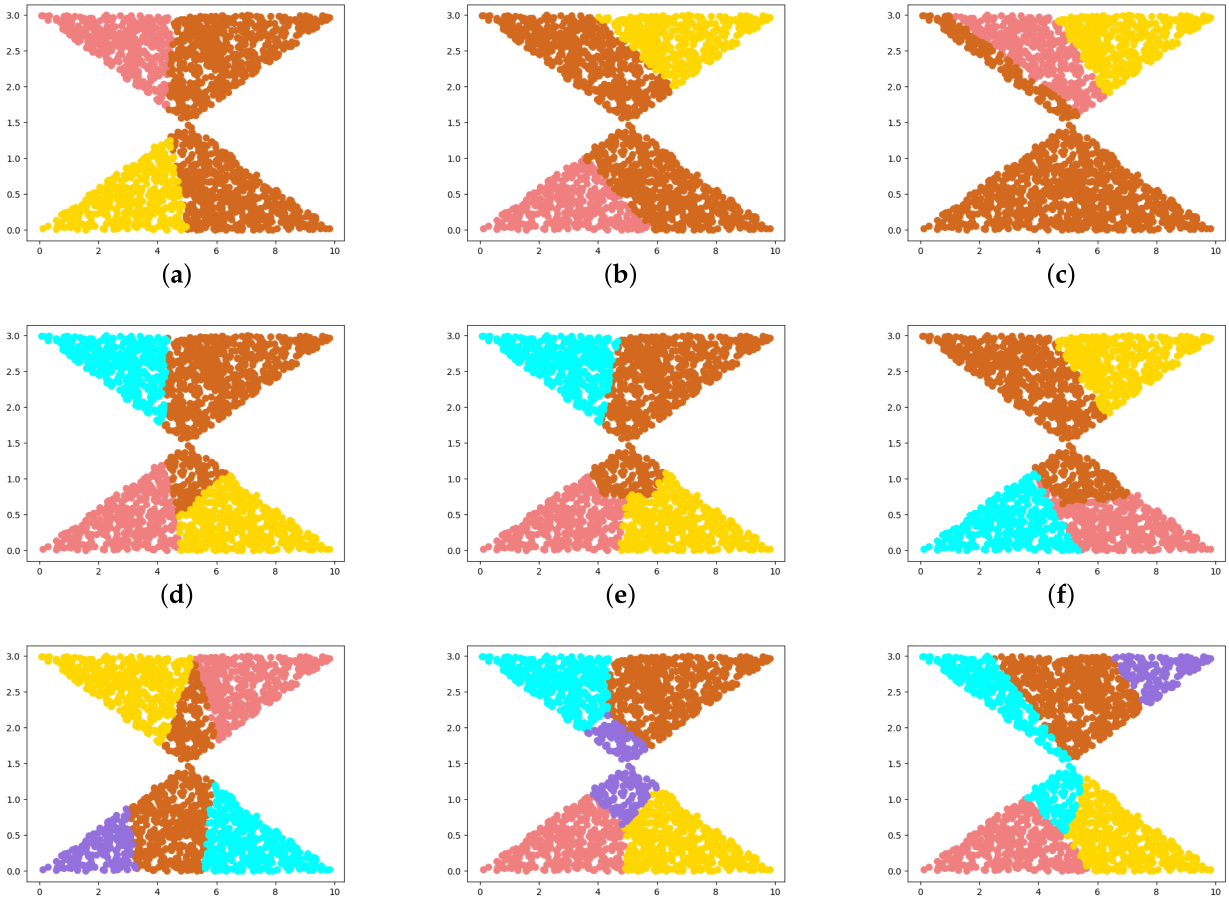
Because the D-CPD method is applicable exclusively in metric spaces, we applied it to address the WBCC problem with the cost kernel $\hat{f}_1$. The results are shown in Figure 2. These results show that the D-CPD method yields disconnected results when applied to the synthetic datasets. Consequently, it can be inferred that the D-CPD method is unsuitable for solving the WBCC problem in all scenarios.



(**a**) (**b**) (**c**)

Figure 2. Results on the three synthetic cases. In each subfigure, the color of a point represents the cluster to which it belongs. (**a**) The result in Case 1; (**b**) The result in Case 2; (**c**) The result in Case 3.

Further, the shelved–retrieved method was applied to address the WBCC problem. The results corresponding to the three distinct cost kernels are presented in Figure 3. The visual representations in Figure 3 show that our method consistently satisfies the connectivity requirements of the clusters. Hence, the shelved–retrieved method can produce connected results when applied in diverse scenarios.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

(**g**) (**h**) (**i**)

Figure 3. (**a**–**i**) Results of three synthetic cases with three cost kernels. In each subfigure, the color of a point represents the cluster to which it belongs. The results in the same row represent variations of a single case under different cost kernel functions, while those in the same column correspond to a common cost kernel. The cost kernels involved in the three columns are $\hat{f}_1(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_2$, $\hat{f}_2(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})M_1(\boldsymbol{x}-\boldsymbol{y})^T$, and $\hat{f}_3(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})M_2(\boldsymbol{x}-\boldsymbol{y})^T$, respectively.

To compare the shelved–retrieved method to the D-CPD method, we evaluated all the results using two metrics.

For each of the different cost kernels $\hat{f}_k, k = 1, 2, 3$, the transportation cost can be calculated using formula $\sum_{i=1}^{n} \sum_{x_j \in C_i} \hat{f}_k$ for each k. The results are presented in Table 2. The transportation costs obtained using the shelved–retrieved method were considerably lower than those obtained using the D-CPD method. Compared to the D-CPD method, the transportation costs of the shelved–retrieved method were reduced by an average of 22.53% for the three cases. Consequently, the shelved–retrieved method effectively reduced transportation costs.

Table 2. Transportation costs of two methods on synthetic cases. Owing to the limitation of the D-CPD method to metric spaces, the transportation costs can only be calculated using the cost kernel $\hat{f}_1$ in all the cases.

Case	Cost Kernel	Shelved–Retrieved Method	D-CPD Method	Reduction
Case 1	$\hat{f}_1$	**2614.56**	2615.82	0.0482%
	$\hat{f}_2$	3916.60	/	/
	$\hat{f}_3$	3040.00	/	/
Case 2	$\hat{f}_1$	**2615.81**	3319.67	21.20%
	$\hat{f}_2$	2635.42	/	/
	$\hat{f}_3$	2345.04	/	/
Case 3	$\hat{f}_1$	**1881.07**	3506.45	46.35%
	$\hat{f}_2$	2254.69	/	/
	$\hat{f}_3$	2069.65	/	/

Additionally, we utilized RMSSTD as a metric to measure the similarity between the clusters.

All the results were evaluated using RMSSTD, and the values are presented in Table 3. Table 3 shows that the RMSSTD of the shelved–retrieved method is considerably lower than that of the D-CPD method in each case. Consequently, the clustering results obtained using the shelved–retrieved method outperform those obtained using the D-CPD method.

Table 3. RMSSTD on synthetic cases. Since the D-CPD method can only be used in metric spaces, RMSSTD can be calculated with the objective function $\hat{f}_1$ in all the cases.

Case	Cost Kernel	Shelved–Retrieved Method	D-CPD Method	Reduction
Case 1	$\hat{f}_1$	**0.809**	0.813	0.492%
	$\hat{f}_2$	0.99	/	/
	$\hat{f}_3$	0.87	/	/
Case 2	$\hat{f}_1$	**0.74**	0.91	18.68%
	$\hat{f}_2$	0.81	/	/
	$\hat{f}_3$	0.77	/	/
Case 3	$\hat{f}_1$	**0.69**	0.94	26.60%
	$\hat{f}_2$	0.75	/	/
	$\hat{f}_3$	0.72	/	/

In synthetic cases, compared with the D-CPD method, our method can produce connected and more compact clusters, and the corresponding transportation costs are much lower. The shelved–retrieved method is more suitable for solving the WBCC problems compared to the D-CPD method.

4.2. Farmland Consolidation

Farmland consolidation is a classical scenario in the WBCC. A large number of small-sized lots cultivated by farmers are scattered over an agricultural area. In farmland consolidation, these lots are restructured into several large connected fields. The adjacent lots in a large connected field are assigned to the same farmer, and the area of lots belonging to the farmer should not change too much. These requirements in farmland consolidation can be formulated by the WBCC problem.

To verify the rationality of our method, we conduct experiments on farmland consolidation. We obtain the relative data of the agricultural area in Germany, such as the position of each lot, the barriers in the area, and the boundary of the lots. Due to the privacy of the datasets, we present the schematic map of the agricultural area in Figure 4. As Figure 4 shows, the lots are distributed over a large area, and the barriers are located in the center of the farmland.

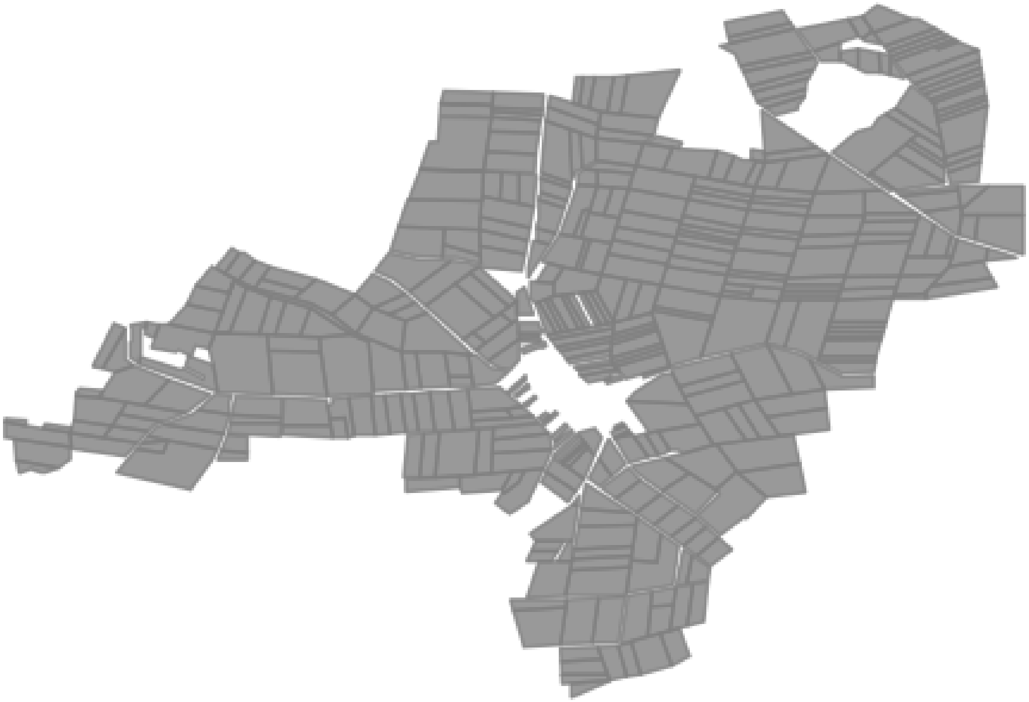


Figure 4. The schematic map of the agricultural area in German.

A total of 399 lots in Figure 4 are cultivated by seven farmers, and each farmer requires that the area should not change too much after reassignment. Each farmer, respectively, provides the lower and upper thresholds ϵ^-, ϵ^+, i.e., the maximal value of area deviation. The original farm area of farmer i is denoted as κ_i; then, the restructured farmland area is within the interval $[\kappa_i - \epsilon^-, \kappa_i + \epsilon^+]$. To harmonize expressions of formulation, we denote $\kappa_i^- = \kappa_i - \epsilon^-$ and $\kappa_i^+ = \kappa_i - \epsilon^+$.

According to the requirements in farmland consolidation, information regarding the belonging of each lot to respective farmers should be obtained. In the experiments, we set $m = 7$ and $n = 399$. Similar to the experiments on synthetic datasets, the D-CPD method is applied to handle the farmland consolidation with cost kernel $\hat{f}_1$. As presented in Figure 5, the D-CPD method produces a disconnected result in the cyan cluster. Hence, the D-CPD method is unsuitable for solving the WBCC problem in farmland consolidation.

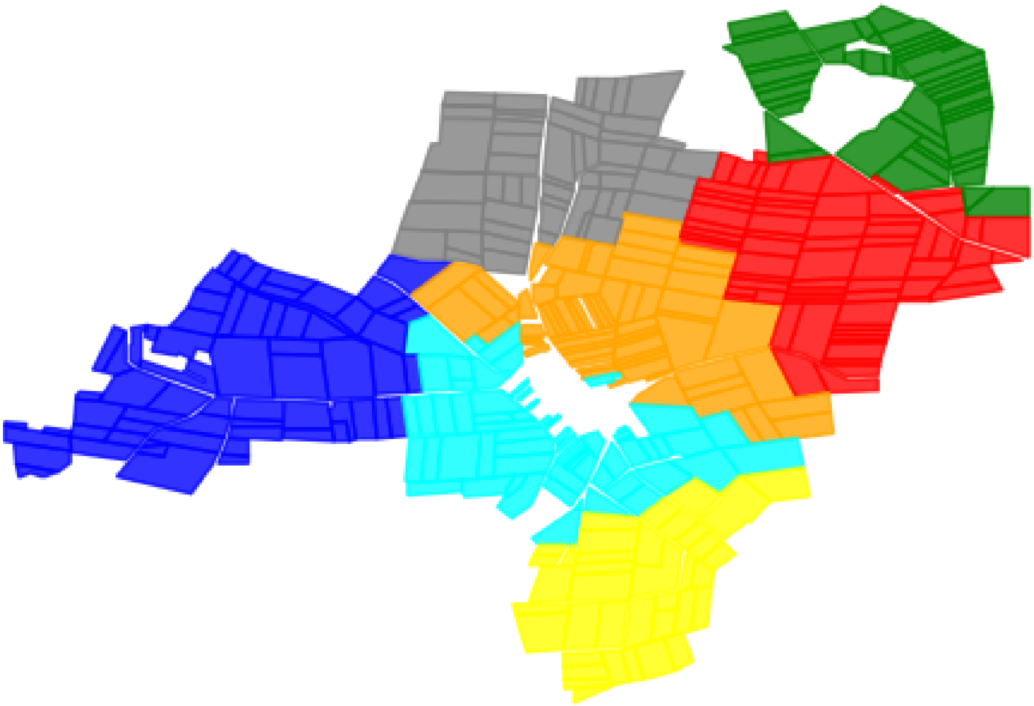
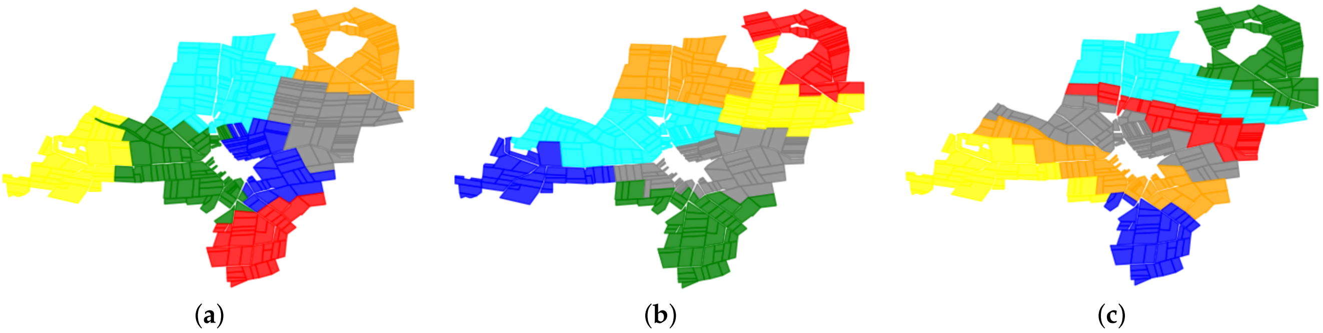


Figure 5. Results on farmland consolidation by the D-CPD method. The color of a point represents the cluster to which it belongs.

Furthermore, the shelved–retrieved method is applied to solve the farmland consolidation with $r = 18$ in the experiments. Three cost kernels $\hat{f}_1, \hat{f}_2, \hat{f}_3$ are used to measure different transportation costs in farmland consolidation. The results corresponding to the three distinct objective functions are shown in Figure 6. Compared with the result of the D-CPD method in Figure 5, the shelved–retrieved method produces connected results but the D-CPD method does not. Thus, the shelved–retrieved method can reasonably solve the WBCC problem in farmland consolidation.



(**a**) (**b**) (**c**)

Figure 6. Results on the farmland consolidation with three cost kernels. The color of a point represents the cluster to which it belongs. (**a**) The result with cost kernel $\hat{f}_1$; (**b**) The result with cost kernel $\hat{f}_2$; (**c**) The result with cost kernel $\hat{f}_3$.

We also evaluate all results by the transportation costs and RMSSTD to compare the two methods. Transportation costs are presented in Table 4. In the experiments, the transportation cost generated by the results of the shelved–retrieved method is reduced by 40.17% compared to the D-CPD method. Therefore, the shelved–retrieved method can reduce the cost of equipment transportation between lots.

Table 4. Transportation costs of two methods on farmland consolidation. Owing to the limitation of the D-CPD method to metric spaces, the transportation costs can only be calculated using the objective function $\hat{f}_1$ in all the cases.

Cost Kernel	Shelved–Retrieved Method	D-CPD Method
$\hat{f}_1$	**8239.77**	13,772.06
$\hat{f}_2$	9711.75	/
$\hat{f}_3$	10,538.33	/

The RMSSTD value of each result is presented in Table 5. Table 5 shows that the RMSSTD of the shelved–retrieved method is lower than that of the D-CPD method with a 22.67% reduction. Thus, the shelved–retrieved method outperforms other methods in farmland consolidation.

Table 5. RMSSTD of two methods on farmland consolidation. Owing to the limitation of the D-CPD method to metric spaces, the transportation costs can only be calculated using the objective function $\hat{f}_1$ in all the cases.

Cost Kernel	Shelved–Retrieved Method	D-CPD Method
$\hat{f}_1$	**3.24**	4.19
$\hat{f}_2$	3.46	/
$\hat{f}_3$	3.67	/

5. Conclusions

The WBCC problem necessitates the division of a point set into connected clusters, each with weights falling within specified intervals. To handle this problem, our study introduced the shelved–retrieved method, which incorporates adjacent relationships into power diagram construction, enabling points to be assigned to clusters based on their adjacent points. Leveraging parameters from power diagrams, this method effectively partitions the point set into connected clusters. Furthermore, it employs a specially designed loop structure to guarantee the generation of clusters that adhere to both weight and geometrical connectivity constraints.

Our experiments, which included three synthetic cases using three cost kernels, as well as their application in farmland consolidation, consistently demonstrated the effectiveness of the shelved–retrieved method. Our results consistently met the constraints of the WBCC problem, resulting in an average reduction of transportation costs by 22.53% and 40.17% for synthetic cases and farmland consolidation, respectively.

Our findings highlight the fact that the shelved–retrieved method not only addresses the WBCC problem effectively, but also ensures cluster connectivity, surpassing other techniques in terms of transportation costs and RMSSTD. This method's flexibility in quantifying different costs through cost kernels allows for substantial cost reductions in various scenarios. However, it is important to note that the shelved–retrieved method may face challenges when dealing with high-dimensional weights. Future research endeavors should aim to address the WBCC problem in such high-dimensional weight scenarios to further enhance the method's applicability and effectiveness.

Author Contributions: Conceptualization, X.H.; methodology, X.H.; software, X.H. and A.Q.; validation, X.H.; formal analysis, X.H.; investigation, X.H., L.Y., A.Q.; data curation, X.H. and L.Y.; writing—original draft preparation, X.H.; writing—review and editing, X.H., A.Q. and L.Y.; project administration, Z.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data available on request due to restrictions, e.g., privacy or ethical. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy policy.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

WBCC	Weakly Balanced Constrained Clustering
RMSSTD	Root-Mean-Square Standard Deviation

Appendix A. D-CPD Method

In the D-CPD method, parameter α can be optimized using Equation (A1).

$$\alpha_i^{p+1} = \begin{cases} \alpha_i^p - \frac{u e_i^p l_i}{\left|e_i^p\right|} & \text{if} \left|e_i^p\right| > \frac{1}{2}(\kappa_i^+ - \kappa_i^-) \\ \alpha_i^p & \text{otherwise} \end{cases}. \tag{A1}$$

The specific algorithm for D-CPD is presented in Algorithm A1.

Algorithm A1 D-CPD method

```
Require: Domain $X = \{x_1, \dots, x_m\}$, weight set $\Omega = \{\omega_1, \dots, \omega_m\}$, capacity sets $K^- = \{\kappa_1^-, \dots, \kappa_n^-\}$, $K^+ = \{\kappa_1^+, \dots, \kappa_n^+\}$.
Ensure: Clustering solution $\mathcal{C} = \{C_1, \dots, C_n\}$
Initialize the cluster $C_i = \emptyset$ for each $i = 1, \dots, n$ and the parameters $\alpha_1, \dots, \alpha_n = 0$.
Randomly select $s_1, \dots, s_n$ in X.
repeat
    $p = 1$
    repeat
        Assign the point $x_j$ to the cluster $C_{i^*}$ for each $j \in \{1, \dots, m\}$, where $i^* = \arg\min_i f(x_j, s_i)^2 - \alpha_i$.
        Update $\alpha_i^{p+1}$ by the Formula (A1) for each $i$.
        $p \leftarrow p + 1$
    until $\alpha_i^{p+1} = \alpha_i^p$ for all $i$
    Update $s_1, \dots, s_n$ by Lloyd algorithm
until $s_1, \dots, s_n$ are not changed
```

References

1. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data clustering: A review. *ACM Comput. Surv. (CSUR)* **1999**, *31*, 264–323. [CrossRef]
2. Omran, M.G.H.; Engelbrecht, A.P.; Salman, A. An overview of clustering methods. *Intell. Data Anal.* **2007**, *11*, 583–605. [CrossRef]
3. Stillwell, M.; Schanzenbach, D.; Vivien, F.; Casanova, H. Resource allocation using virtual clusters. In Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, Shanghai, China, 18–21 May 2009; pp. 260–267.
4. Fischer, D.T.; Church, R.L. Clustering and compactness in reserve site selection: An extension of the biodiversity management area selection model. *For. Sci.* **2003**, *49*, 555–565.
5. Yang, K.; Shekhar, A.H.; Oliver, D.; Shekhar, S. Capacity-constrained network-voronoi diagram. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 2919–2932. [CrossRef]
6. Chopra, S.; Rao, M.R. The partition problem. *Math. Program.* **1993**, *59*, 87–115. [CrossRef]
7. Baranwal, M.; Salapaka, S.M. Clustering with capacity and size constraints: A deterministic approach. In Proceedings of the 2017 Indian Control Conference (ICC), Guwahati, India, 4–6 January 2017; pp. 251–256.
8. Brieden, A.; Gritzmann, P.; Klemm, F. Constrained clustering via diagrams: A unified theory and its application to electoral district design. *Eur. J. Oper. Res.* **2017**, *263*, 18–34. [CrossRef]
9. Brieden, A.; Gritzmann, P. On optimal weighted balanced clusterings: Gravity bodies and power diagrams. *SIAM J. Discret. Math.* **2012**, *26*, 415–434. [CrossRef]
10. Borgwardt, S.; Brieden, A.; Gritzmann, P. Geometric clustering for the consolidation of farmland and woodland. *Math. Intell.* **2014**, *36*, 37–44. [CrossRef]
11. Bradley, P.S.; Bennett, K.P.; Demiriz, A. Constrained k-means clustering. *Microsoft Res. Redmond* **2000**, *20*. Available online: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-2000-65.pdf (accessed on 15 October 2023)

12. Brieden, A.; Gritzmann, P. A quadratic optimization model for the consolidation of farmland by means of lend-lease agreements. In *Operations Research Proceedings 2003*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 324–331.
13. Borgwardt, S.; Brieden, A.; Gritzmann, P. Constrained minimum-k-star clustering and its application to the consolidation of farmland. *Oper. Res.* **2011**, *11*, 1–17. [CrossRef]
14. Ganganath, N.; Cheng, C.; Chi, K.T. Data clustering with cluster size constraints using a modified k-means algorithm. In Proceedings of the 2014 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Shanghai, China, 13–15 October 2014; pp. 158–161.
15. Höppner, F.; Klawonn, F. Clustering with size constraints. In *Computational Intelligence Paradigms Innovative Applications*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 167–180.
16. Zhu, S.; Wang, D.; Li, T. Data clustering with size constraints. *Knowl.-Based Syst.* **2010**, *23*, 883–889. [CrossRef]
17. Rose, K. *Deterministic Annealing, Clustering, and Optimization*; California Institute of Technology: Pasadena, CA, USA, 1991.
18. Hu, C.W.; Li, H.; Qutub, A.A. Shrinkage clustering: A fast and size-constrained clustering algorithm for biomedical applications. *BMC Bioinform.* **2018**, *19*, 19. [CrossRef] [PubMed]
19. Li, J.; Horiguchi, Y.; Sawaragi, T. Cluster size-constrained fuzzy c-means with density center searching. *Int. J. Fuzzy Log. Intell. Syst.* **2020**, *20*, 346–357. [CrossRef]
20. Tang, W.; Yang, Y.; Zeng, L.; Zhan, Y. Size constrained clustering with milp formulation. *IEEE Access* **2020**, *8*, 1587–1599. [CrossRef]
21. Balzer, M. Capacity-constrained voronoi diagrams in continuous spaces. In Proceedings of the 2009 Sixth International Symposium on Voronoi Diagrams, Copenhagen, Denmark, 23–26 June 2009; pp. 79–88.
22. Xin, S.; Lévy, B.; Chen, Z.; Chu, L.; Yu, Y.; Tu, C.; Wang, W. Centroidal power diagrams with capacity constraints: Computation, applications, and extension. *ACM Trans. Graph. (TOG)* **2016**, *35*, 1–12. [CrossRef]
23. Galvao, L.C.; Novaes, A.G.; De Cursi, J.S.; Souza, J.C. A multiplicatively-weighted voronoi diagram approach to logistics districting. *Comput. Oper. Res.* **2006**, *33*, 93–114. [CrossRef]
24. Aurenhammer, F.; Hoffmann, F.; Aronov, B. Minkowski-type theorems and least-squares clustering. *Algorithmica* **1998**, *20*, 61–76. [CrossRef]
25. Lloyd, S. Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137. [CrossRef]
26. Liu, Y.; Li, Z.; Xiong, H.; Gao, X.; Wu, J. Understanding of internal clustering validation measures. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 13 December 2010; pp. 911–916.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.