Algorithm for Two-Phase Facility Planning via Balanced Clustering and Integer Programming

Larkin Liu

University of Toronto

Abstract

We present a solution for a two-phase facility planning scenario where in the first phase, there is some flexibility in determining where the locations of facilities (or sources) should fall. And in the second phase, new waypoints (or sinks) are added, but the location of the facilities are static. This solution applies the use of balanced clustering - using a modified K-Means approach, ensuring the cardinality of each group to be equal. Subsequently, it is followed by an integer programming solution, to solve the Hitchcock Transportation Problem. We show that the final solution can justifiably approximate the near optimal solution, and be a successful guide for facility planning in this specific scenario.

Keywords: Balanced Clustering, Integer Programming, Facility Planning

1. Introduction

The optimization of facilities - also referred to as sources in the network flow formulation - in modern supply chains play a critical role in the profitability and market competitiveness of any modern business. Such facilities can represent the function of a distribution center for goods or supplies; in our work we refer to these as depots. Such distribution centers can supply the stores or service centers in the surrounding location. In the network flow formulation these are referred to as sinks, in our work we refer to them as waypoints.

Therefore, it is essential to ensure that in the path planning of any supply chain network is properly optimized to reduce transportation distance. We wish to find an appropriate solution for a unique problem that involves the positioning and assignment of depots and waypoints in a realistic situation. We are presented with a two-phase optimization process, in the first phase, the Positioning Phase denoted as Phase I, we must determine where facilities must fall on an unconstrained \( \mathbb{R}^2 \) plane, to obtain a fixed positioning of depots. Subsequently we are presented with the Assignment Phase denoted as Phase II, where an arbitrary set of waypoints are placed in \( \mathbb{R}^2 \) representing new locations of waypoints.

For the new and preexisting waypoints the distance from itself to each depot is known, and thus is represented by a set of coordinates are transformed into a symmetric distance matrix, or cost-matrix, \( M \). This cost-matrix is based the Euclidean distance from any waypoint to any depot. From this we must assign the optimal waypoint to depot assignment such that the total travel distance is minimized.

The situation stipulates that certain locations, or waypoints, must be assigned to respective facilities, or depots. Phase I presents an optimization problem where the location
of the depot lies unconstrained in $\mathbb{R}^2$, thus we option to use a clustering algorithm to solve iteratively, specifically the K-Means algorithm.

The K-Means algorithm can also be interpreted as the Expectation Maximization (EM) algorithm (Dempster et al., 1977), without the use of a probability distribution. Specifically the assignment step involves assigning each waypoint to its nearest neighbour depot. The result provides a series of Voronoi cells (Voronoi, 1908) in $\mathbb{R}^2$, where each depot forms the centroid of each Voronoi cell. Yet, we understand that the major limitation of K-Means for our use-case is that it does not ensure that the number of waypoints belonging to each Voronoi cell is equal. For this, we must apply a modified K-Means implementation authored by (Schubert et al., 2015), ensuring that each cluster of waypoints contains the same number, and the centroid of each cluster is the location of the depot. We refer to this as K-Means Clustering with Equal Size Constraints, or balanced clustering.

In Phase II, the situation is fundamentally different. There exists no freedom in the movement of the depots in $\mathbb{R}^2$, and thus the problem is purely an assignment problem, which waypoints to assign to with depots. All waypoints, including the preexisting ones from Phase I as well as newly added ones from Phase II undergo a reassignment.

The solution to this type of problem has been investigated for example in (Hakimi, 1964) offering a reformulation of the Weber Problem (Weber, 1909) for facilities, where each facility in $\mathbb{R}^2$ can function as both a depot or waypoint. (Hakimi, 1964) proves that the optimal solution must intersect with the set of facilities. In other words, the optimal depots locations correspond to facility locations. However, in our application there is a distinction in the type of facility, that is depots and waypoints are not interchangeable. Thus we must model the problem as an Hitchcock Transportation Problem (Hitchcock, 1941), or HTP. The HTP is also another extension of the Weber Problem where, the placement of each depot and waypoint are fixed, non-interchangeable, and a solution must be provided to assign each waypoint to a depot, such that some distance function is reduced.

In the business setting, the introduction of new waypoints can represent a decision to add more location to a store branch, as determined external sources, such as executive management, at a later time. The idea is that a full-disclosure of where the waypoints location are, and how many are to be planned, is not given ahead of time. For example, an initial deployment of store branches followed by a second wave of store openings. Where in the initial phase of optimal locations must be planned for the supply depots in terms of minimizing travel distance. Therefore, there is uncertainty regarding where or how many of the waypoints can be placed. We only have control over the placement of the depots in the first phase, and assignment of waypoints to depots in the second phase.

1.1 Experimental Outline

The data content we select to model consists of approximately 25,000 world wide locations of Starbucks chain store locations. The store location’s longitude and latitude are provided. We assume that all distances from any location on the 2D world map to be the Euclidean distance, with no geographical, topographical, land, or ocean boundary which can affect the Euclidean distance. We also assume, that from the initial placement of waypoints in Phase I, there exists no correlation or deterministic relationship to the subsequent waypoint placements in Phase II.
We begin by selection at random percentage of waypoints, represented by $\gamma = 0.05$, for Phase I. It is evident that the algorithm presented in Phase II will solve for the majority of the total waypoint to depot assignment. Subsequently, we introduce the rest of the waypoints, $1 - \gamma$, for the assignment phase. In each of the respective phases, we present a specific solution to the different scenarios. In Phase I we apply the balanced clustering algorithm (Schubert et al., 2015), and in Phase II we present a solution to the HTP. Both to be illustrated with further detail in the later sections.

2. Problem Statement

The number of depots, $K$, are fixed and equal for both Phase I and Phase II. The waypoints are given in two phases, we let $N_I$ represent the number of waypoints in the Phase I and $N_{II}$ for Phase II respectively. In our notation, $N$ flexibly represents the number of waypoints in either Phase I or Phase II. Most importantly, the number of waypoints assigned to each depot must be equal, outlined in Eq. 1. In our specific problem, we pose a situation where only a subset of the entire set waypoints are initially know.

$$n_k = N/K$$  \hspace{1cm} (1)

In the Phase I, or the Positioning Phase some waypoints are provided, however, the depot locations lie unconstrained in $\mathbb{R}^2$, and can be modified accordingly. We can infer that the distance function is not discrete, rather can be modeled as a continuous function. In this experiment we use the simple Euclidean distance in $\mathbb{R}^2$, between two 2-dimensional points. The earliest notion of such an unconstrained facility planning problem can be illustrated in the Weber Problem (Weber, 1909). Its objective is to find a single depot by minimizing the Euclidean distance between all waypoints and its nearest depot. This problem can be solved using various geometric methods, such as some illustrated in (Fekete et al., 2003). The big drawback using such geometric methods are that they are not easily extendable beyond $\mathbb{R}^2$, and is not easily programmable into computers. For this reason, we propose to use computational methods, such as balanced clustering, to solve Phase I.

After we present Phase II, or the Assignment Phase, where additional waypoints are added to the set of waypoints, denoted as $\mathbb{W}$. At this stage the location of the depots are fixed, and denoted as $\mathbb{D}$. Due to this, the distance function can be represented as a discrete distance matrix, $\mathbb{M}$. The mapping of depot to waypoint can be altered, in the Assignment Phase, which is a clear statement of the HTP. The HTP can be solved using the Simplex Algorithm, and alternatively using the Hungarian Algorithm, also known as Kuhn’s Combinatorial Algorithm (Ford and Fulkerson, 1956). To note further, Aardal (1998) also investigates the capacitated facility location problem using a polyhedral approach, solved specifically using a cutting-planes approach. However, the implementation of this procedure is beyond the scope of this research due to complexity.

For the Assignment Phase, we define a standard assignment problem to be solved using integer programming. We select the Integer Programming approach over the Hungarian Algorithm formulation because the integer programming is a more generalized formulation and does not require redundancy in $\mathbb{M}$.* In the proposed problem, we seek to find a

*See Appendix A for details.
| Notation | Definition |
|----------|------------|
| $\mathbb{D}$ | Set of all potential depots. |
| $\tilde{\mathbb{D}}$ | Set of all depots after Phase I. |
| $d$ | Location of an individual depot. |
| $d^*$ | Optimal location of individual depot for Phase I. |
| $d'$ | Best alternate depot assignment for a waypoint $w$ corresponding to $\Delta$. |
| $M$ | Distance matrix representing distance from all possible waypoints to depots. |
| $N$ | Total number of waypoints in either Phase I or Phase II. |
| $N_I$ | Total number of waypoints in Phase I. |
| $N_{II}$ | Total number of waypoints in Phase II. |
| $n_k$ | Number of waypoints assigned to each depot. |
| $K$ | Number of depots. |
| $\mathbb{W}$ | Set of all possible waypoints. |
| $\tilde{\mathbb{W}}$ | Set of new waypoints introduced in Phase II. |
| $\hat{\mathbb{W}}$ | Set of all waypoints in Phase I and II. |
| $\mathbb{W}$ | Sorted $\hat{\mathbb{W}}$ based on Euclidean distance difference between $w$ to $d$ versus to $d'$. |
| $\mathbb{W}_d$ | Subset of $\mathbb{W}$ corresponding to all $w$ assigned to $d$. |
| $\mathbb{W}_{d'}$ | Subset of $\mathbb{W}$ corresponding to all $w$ assigned to $d'$. |
| $w$ | An individual waypoint. |
| $\Delta$ | The assignment plan, defined as the set of all tuples $(w, d)$. |
| $\Delta^*$ | The optimal assignment plan, as defined in Eq. (6). |
| $\Delta_d$ | Subset of $\Delta$ representing the set of waypoints only assigned to depot $d$. |
| $\Delta'$ | Assignment plan corresponding to the assignment of all waypoints in $\Delta$ to its second closest alternate depot. |
| $\Gamma$ | Set of all possible distances from all waypoints to all depots. |
| $\psi(w, d)$ | Distance function of waypoint to depot. |
| $\Psi(\Delta)$ | Total cost of the assignment plan in terms of Euclidean distance. |
| $1_{\Delta}$ | $\begin{cases} 1, & \text{if } w \text{ is assigned to } d. \\ 0, & \text{otherwise} \end{cases}$ |
Algorithm for Two-Phase Facility Planning

placement of depots, such that the total distance metric from all waypoints to its assigned depot is minimized. We first must define a distance function relating each element the set of waypoints, \( w \in \mathbb{W} \), to each element in the set of depots, \( d \in \mathbb{D} \). This mapping we define as \( \psi \) in Eq. (2),

\[
\psi : (\mathbb{W}, \mathbb{D}) \rightarrow \Gamma \in \mathbb{R}
\]  

(2)

Where the \( \Gamma \) is the set of all distances from from any \( w \) to any \( d \), which we can also write as \( \psi(w, d) \). Thus the distance function \( \psi \) is a surjection from the world of n-dimensional co-ordinate tuples to \( \mathbb{R} \). We define this as simply the Euclidean distance between 2-dimensional points in our example. Though it can be extended to the n-dimensional case with the same reasoning. In our use case, since the goal is to minimize the sum of \( \Gamma \). We define \( \Delta \), as the assignment plan from each \( w \) to \( d \) obeying the constraints specified in Eq. (1). \( \Delta \) can be represented as a set of tuples \((w, d)\) representing a collection of selected pairings of waypoints to depots. Given this definition our goal is to find a mapping, or assignment, from the set of \( \mathbb{W} \) to \( \mathbb{D} \) that will be most optimal in terms of minimizing \( \psi(w, d) \) \( \forall \Gamma \). We illustrate this relationship where \( \mathbb{W} \times \mathbb{D} \) contains all the possible combinations of \( w \) and \( d \), defined in Eq. (3).

\[
\Delta \subset \mathbb{W} \times \mathbb{D} \tag{3}
\]

s.t. \( \mathbb{D} \in \mathbb{R}^2 \) \tag{4}

Furthermore, we define \( 1_\Delta \) as a matrix of indicator variables, stored in a matrix indicating if waypoint \( w \) was assigned to depot \( d \).

\[
\Delta = \begin{bmatrix}
1_\Delta(w, d) & \cdots & 1_\Delta(w, d) \\
\vdots & \ddots & \vdots \\
1_\Delta(w, d) & \cdots & 1_\Delta(w, d)
\end{bmatrix}
\]  

(5)

Our motivation is to find the optimal assignment, denoted as \( \Delta^* \), that will minimize the Euclidean distance function. This is illustrated in Eq. (6), where \( 1_\Delta(w, d) \in (0, 1) \) is an indicator function denoting whether or not \( w \) was assigned to \( d \) under assignment plan \( \Delta \).

\[
\Delta^* = \arg\min_\Delta \sum_{(w, d) \in \mathbb{W} \times \mathbb{D}} 1_\Delta(w, d) \psi(w, d)
\]

(6)

\[
= \arg\min_\Delta \Psi(\Delta)
\]  

(7)

3. Solution

Given a two-phase problem we provide a two-phase algorithm as a solution - outlined in Section (3.3). First we utilize balanced clustering, and subsequently, an integer programming

\(^\dagger\)As defined by the arithmetic sum of all rows and columns of \( \Gamma \)
solution to the assignment problem is applied. In Phase I, the balanced clustering algorithm adapts a modified K-Means algorithm to determine the placement of depots freely in \( \mathbb{R}^2 \), this is known as balanced clustering. Subsequently in Phase II, we apply an integer programming solution to the assignment problem, when the depot location have been already determined in Phase I. We provide further details, and an outline, in the Section (3.1) and (3.2).

### 3.1 Balanced Clustering

First described in (MacQueen, 1967), the K-Means algorithm presents a geometric interpretation of the classification problem. The algorithm assigns a set of observations into K unconstrained centroids via an iterative algorithm. Consequently we apply this type of algorithm to minimize the total Euclidean distance from each waypoint to its assigned depot. However, in our scenario, we must consider the constraint that \( n_k \) must be equal for all groups. For this solution we implement a variation of the k-means algorithms, referred to as the same-size K-means algorithm developed by (Schubert et al., 2015), presented in Algorithm (1).

To satisfy the constraint that \( n_k \) must be equal for all groups, we constrain the number of waypoints assigned to any \( d \in \mathbb{D} \) to be equal. Let \( K = |\mathbb{D}| \) denote the cardinality of the set of depots, that is the number of depots. We also constrain the cardinality of each of the assignment subset for any \( d \), as denoted by \( \Delta_d \), must be equal to the cardinality of all other assignment subsets, represented as \( n_k \). We illustrate the simple K-Means algorithm which involves first assigning each waypoint to its respective depot, which is exactly the closest depot to each waypoint, as illustrated in Eq. (8). Subsequently, we re-estimate new depot location, by taking the arithmetic mean of all waypoints assigned to \( d \) under \( \Delta \), as denoted as \( \Delta_d \), as illustrated in Eq. (9). In K-Means, this alternation between Eq. (8) and Eq. (9) begins initially with a random initialization of candidate depot locations, and ends when either the maximum number of iterations is reached, or when the reduction of \( \Psi(\Delta) \) from iteration to iteration is static or below a certain threshold.

\[
\Delta(w, d) = \left\{ w : |w - d^*| \leq |w - d|, \forall d \right\}
\]

\[
d^* = \frac{1}{|\Delta_d|} \sum_{w} w \mathbb{1}_{\Delta}(w, d)
\]

Where \( \Delta_d \) represents the set of waypoints assigned to depot \( d \), and \( d^* \) is the proposed optimal depot location. Nevertheless, it is evident that Eq. (8) and Eq. (9) alone does not satisfy the constraint that the cardinality of each \( d \in \mathbb{D} \) to be equal, as illustrated by Eq. (10). Thus to accomplish satisfying Eq. (10), we must apply the algorithm from (Schubert et al., 2015), and presented in Algorithm (1).

\[
|\Delta_d| = \frac{|\mathbb{W}|}{|\mathbb{D}|}, \forall d \in \mathbb{D}
\]
Algorithm 1 Same-size K-Means Algorithm (Schubert and Zimek, 2019)

1: Using K-Means, via Eq. (8) and Eq. (9) propose a set of candidate depot locations.
2: Compute $\psi(w,d), \forall \mathbb{W} \times \mathbb{D}$.
3: Sort $\mathbb{W}$ based on the difference under $\Delta$ and the best possible alternate assignment, denoted by $\Delta'$, producing ordered set, denoted by $\mathbb{W}$.

for $w \in \mathbb{W}$ do:
   for $d \in \mathbb{D}$ do:
      Initialize $\mathbb{W}_d'$ as all $w \in \Delta_d'$
      while $|\mathbb{W}_d'| > 0$ do:
         for $w$ in $\Delta_d$ do
            if Swapping $w$ with $w'$ from $d$ to $d'$ reduces $\Psi(\Delta')$ then:
               Assign $w$ to $d'$ and $w'$ to $d$.
               Remove $w$ from $\mathbb{W}_d$.
            end if
            if Reassigning $w$ to $d'$ does not violate Eq. (10) then:
               Assign $w$ to $d'$.
               Remove $w$ from $\mathbb{W}_d$.
            end if
            if Maximum iterations reached then
               Terminate algorithm.
            end if
         end for
      end while
   end for
end for

3.2 The Assignment Problem

After the depot locations have been determined in Phase I, as outlined previously in Section 3.1, we can formulate a tractable solution for the Assignment Phase. This can be constructed as an Hitchcock Transportation Problem. Eq. (13) illustrates the objective function and constraints for such an optimization problem. In our formulation, we assign to each waypoint, $w$ to a fixed depot, $d$, where $\mathbb{D}$ no longer lies unconstrained in $\mathbb{R}^2$. $\mathbb{D}$ is a fixed set, which we denote as $\tilde{\mathbb{D}}$. Provided $\tilde{\mathbb{D}}$, we introduce a set of new fixed waypoints $\tilde{\mathbb{W}}$, where we must assign each $w \in \tilde{\mathbb{W}}$ to a specific depot in $\tilde{\mathbb{D}}$. Alternatively, Malinen and Fr"anti (2014) suggests that using the Hungarian Algorithm, also called Munkres algorithm, (Kuhn, 1955) is capable to solve up to approximately 1000 waypoints on regular computers for this specific problem. However, in order to formulate the problem using the Hungarian Algorithm, it is necessary to repeat the depot locations on the cost matrix - see Appendix A. This is inefficient, and we opt for a direct Integer Programming (IP) solution which we will present in this paper. We present a solution where we optimally assign each waypoint to depot such that $\Psi(\Delta)$ is minimized.
\[
\min_{\Delta} \sum_{\Gamma} 1_{\Delta}(w,d)\psi(w,d) \quad (11)
\]

s.t. \[
\sum_{\hat{D}} 1_{\Delta}(w,d) = \left| \hat{W} \right|, \forall w \in \hat{W} \quad (12)
\]
\[
\sum_{\hat{D}} 1_{\Delta}(w,d) = 1, \forall d \in \tilde{D} \quad (13)
\]

In Eq. (11) we illustrate the objective function that must be minimized in our IP, and subsequently Eq. (12) and (13) we illustrate the constraints on the IP. In our experiment, previously assigned waypoints assigned in Phase I may be reassigned to another depot, however, the depot locations \(\tilde{D}\) are fixed. We use \(\hat{W}\) to denote the final set of waypoints, as marked by Eq. (17), as the union of the waypoints issued in both Phase I and Phase II. The assignment plan \(\Delta\) can be constructed as a matrix of indicator variables \(1_{\Delta}(w,d)\), indicating whether waypoint \(w\) was assigned to depot \(d\), as illustrated in Eq. (14). As evident from Eq. (15) and (16), we specify the constraints of the way points as the assignment plan \(\Delta\) and the transpose of the assignment plan \(\Delta^T\). Each row of the assignment plan \(\Delta\) refers to a specific waypoint, \(w\), and each column, a depot \(d\). As a small non-convention, but for the sake of clear simplicity we define the row-sum operator \([\Delta]^+\) a new vector containing the matrix row sums of the matrix \(\Delta\). Allowing us to clearly specify the constraints.

\[
\Delta = \begin{bmatrix}
1_{\Delta}(w,d) & \ldots & 1_{\Delta}(w,d) \\
\vdots & \ddots & \vdots \\
1_{\Delta}(w,d) & \ldots & 1_{\Delta}(w,d)
\end{bmatrix} \quad (14)
\]

\[
[\Delta]^+ = \begin{bmatrix}
1_{\Delta}(w,d) + \ldots + 1_{\Delta}(w,d) \\
\vdots & \ddots & \vdots \\
1_{\Delta}(w,d) + \ldots + 1_{\Delta}(w,d)
\end{bmatrix} = \begin{bmatrix} 1 \end{bmatrix} \quad (15)
\]

\[
[\Delta^T]^+ = \begin{bmatrix} n_k & \ldots & n_k \end{bmatrix}^T \quad (16)
\]

Several industry standard solutions to the IP problem exist for the NP-hard Hitchcock Problem. We present some preliminary background on such methods, however, we will not go into heavy theoretical details, as it is not the focus of our thesis. The solutions to IP problems concern the Cutting Planes method (Gomory, 1958), which includes a solution to the separation problem, that is, finding an inequality that separates the optimal value from the convex hull, known as a cut. Many cuts are found until the non-integer optimal solution is no longer feasible. The cutting plane method, was in its introduced in the 1950’s had impractical applications due to numerical instabilities. Another solution to the IP problem is known as Branch-and-Bound (Land and Doig, 1960). In principle, the Branch-and-Bound algorithm involves iterative and selective computations to produce subspaces of the feasible solution after elimination and eliminating such subspaces by remembering the bounds of the solution space pertaining so such a subspace, and eliminating them accordingly. This
algorithm expedites the computation of the iterative search involved in solving for the global optimal.

We option to apply the Branch-and-Cut (Balas and Matthew J. Saltzman, 1991) method as the solution to our IP, on the grounds that it combines the effectiveness of both the Cutting Planes and Branch-and-Bound methods, improving the numerical instability of the Cutting planes method, while also improving the subspace elimination capabilities of the original Branch-and-Bound. This serves currently, as a state-of-the-art solution for most Mixed Integer Programs today. Therefore, provided the objective functions and constraints illustrated in Eq. (11), (15), and (15), we proceed to solve the system of equations using the Branch-and-Cut method (Elf et al., 2001). We use the software implementation from the open source package for Branch-and-Cut developed by COIN-OR, Computational Infrastructure for Operations Research (Forest et al., 2020).

3.3 Two-Phase Algorithm

The full algorithm proposed in this work begins with a set of initial waypoints, \( W \), and then we are subsequently introduced new waypoints, \( \tilde{W} \). As outlined, we define two phases of optimization, the balanced clustering, and Assignment Phases. In both phases, the depot locations, \( D \) remains constant, but are positioned in Phase I. In reality, \( D \) can represent the required placement of distribution centers that must be placed to serve such locations. The initial set of waypoints, \( W \), can represent, for example, store locations or service centers, that are initially planned by a company or organization. The second phase of waypoints, \( \tilde{W} \), can represent a subsequent addition of store locations planned by a company. Typically, the planning of these waypoints are perhaps strategic, however, we assume the locations of these stores to be unknown in Phase I. In our expriment, we simulate this strategic placement of new stores, by randomly sampling from a set of \( \hat{W} \) all store locations, randomly allocating \( W \) to the balanced clustering phase, and \( \tilde{W} \) to the Assignment phase, with the percentage of allocation respecting \( \gamma = 0.05 \), as described in Section 1.1.

\[
\hat{W} = W \cup \tilde{W}
\]

The balanced clustering phase utilizes the iterative algorithm outlined in Algorithm (1) to generate a prescribed set of depot locations. In this phase, the depot locations fall anywhere on an \( \mathbb{R}^2 \) plane. Inevitably, the assignment, \( \Delta \), is simply the closest depot, \( d \), to each \( w \in W \). In the second phase, we perform the assignment algorithm illustrated in Section (3.2). Where, as described, we assign new waypoints, \( \hat{W} \) to \( \tilde{D} \). The second phase can represent new stores or service locations that are planned for the future. Since the depots have been already constructed, we are only allowed to solve the assignment problem with \( \psi(w,d) \) serving as the loss measure, of which we minimize. Algorithm 2 summarizes this Two-Phase strategy. We also provide source code for implementing this algorithm on (Liu, 2020).
Algorithm 2 Two-Phase Algorithm

1: \( \hat{W} \) is given.
2: Compute \( \Delta^* \) using Algorithm (1)
3: \( \hat{W} \) is given.
4: Compute \( \Delta^* \) using Integer Programming solution outlined in Section (3.2).

---

4. Simulation

In order to measure the performance of the Two-Phase algorithm, we use the Mean Squared Error (MSE) of distance function \( \psi \) of each \( w \in \hat{W} \) to its assigned depot, as designated as \( \Delta^*[w] \).

\[
MSE = \frac{1}{|\hat{W}|} \sum_{w \in \hat{W}} \psi(w, \Delta^*[w])^2
\]  

Eq. (18) produces a measure of the average of cost from the each waypoint, \( w \) to its depot mapping \( \Delta^*[w] \). This is a common metric for measuring clustering performance, as we wish to reduce the Euclidean distance from \( w \) to its optimally mapped depot, \( d = \Delta^*[w] \).

As evidence from the performance from Table (2), we see that the percent deviation from Phase I to Phase II in terms of MSE is under on average of 3%, and no greater than 10 % for up \( K = 10 \) starting from \( K = 3 \). This indicates that, even if mobility of depots is not allowed in Phase II, the Assignment Algorithm using Integer Programming still produces a mapping \( \Delta \) that does not differ far from a supposedly more flexible, and thereby more optimal Phase I algorithm. In general, this strategy is simple to understand and implement, and provides strong empirical results.
5. Conclusion

In our research we present a basic logistics problem, a two-phase facility planning problem, with a different set of constraints in each phase. In each of the two phases, we present an application of pre-existing methods to solve it. Notably, the same size K-means algorithm (Schubert et al., 2015) for the balanced clustering phase, and an Integer Programming solution to the Hitchcock Transportation problem utilizing the Branch-and-Cut algorithm. Furthermore, we present a unique mathematical formulation that ties the two phases together in a unified notation. Subsequently, we prove the capability of our proposed two phase algorithm in a simulation framework on real world data. We acknowledge that further research can be done with relation to some of the hyperparameters of the project. Hyperparameters include the amount of waypoints given at each phase, as determined by $\gamma$, and also for example, the maximum number of iterations of the same size K-means algorithm. We can also potentially study the efficacy of this two-phase algorithm beyond two dimensions. Or also, increasing the number of depots, $K$, and of course, attempting to run such an algorithm on other datasets. Nevertheless, this paper serves as a proposal for a possibility of applying such a two-phase algorithm any multi-phase planning of facilities as illustrated in our work.

Acknowledgements

This research was conducted without external funding, resulting in no conflicts of interests. The sole author would like to thank the additional efforts of Dr. Seyed Ali Hesammohseni from the University of Waterloo for providing extensive comments regarding the scientific direction of the work, as well as the Kaggle organization for providing the open source data set.

Appendix A.

We illustrate a problem statement for alternate solution to the Hitchcock Transportation problem, using the Hungarian Method (Kuhn, 1955). This method was purposed to find a minimum cost solution for assignment of workers to tasks, given the capacity of the workers, and the requirements for the each of the tasks. Notably, the original Hungarian Algorithm
requires that the number of workers equals the number of tasks, where as in our scenario the number of waypoints greatly exceed the number of depots. It applies a series of algebraic manipulations to compute an optimal solution for the Assignment Problem. This involves only a one worker to one task solution based on a matrix that contains the cost per worker per task. In order to address this mismatch, we can duplicate the number of depots by \( n_k \) times, increasing the number of columns repetitively, to create a \( N \) by \( K \cdot n_k \) square matrix. Let \( M \) represent the distance matrix from each \( w \in W \) to \( d \in D \). Therefore we have an \( N \times K \) distance matrix, which we denote as \( \tilde{M} \).

\[
M = \begin{bmatrix}
\psi(w,d) & \ldots & \psi(w,d) \\
\vdots & \ddots & \vdots \\
\psi(w,d) & \ldots & \psi(w,d)
\end{bmatrix}
\]  

(19)

In order to build a matrix where the Hungarian can be applied we must construct \( \tilde{M} \), we simply replicate \( M \) column-wise a total of \( n_k \) times. In this respect, we can obtain an \( N \times K \) matrix.

\[
\tilde{M} = [M \ldots M]
\]

(20)

Because of the repetition created in Eq. (20), we opt to not apply the Hungarian as a viable solution for our experiment when \( \Delta^* \) in the Assignment Phase.

References

Karen Aardal. Capacitated facility location: Separation algorithms and computational experience. *Math. Program.*, 81:149–175, 1998. doi: 10.1007/BF01581103. URL https://doi.org/10.1007/BF01581103.

Egon Balas and Grard Cornujols Matthew J. Saltzman. An algorithm for the three-index assignment problem. *Oper. Res.*, 39(1):150–161, 1991. doi: 10.1287/opre.39.1.150. URL https://doi.org/10.1287/opre.39.1.150.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977. URL http://web.mit.edu/6.435/www/Dempster77.pdf.

Matthias Elf, Carsten Gutwenger, Michael Jünger, and Giovanni Rinaldi. *Branch-and-Cut Algorithms for Combinatorial Optimization and Their Implementation in ABA-CUS*, pages 157–222. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-540-45586-8. doi: 10.1007/3-540-45586-8_5. URL https://doi.org/10.1007/3-540-45586-8_5.

Sándor P. Fekete, Joseph S. B. Mitchell, and Karin Beurer. On the continuous fermat-weber problem. *CoRR*, cs.CG/0310027, 2003. URL http://arxiv.org/abs/cs/0310027.

L. R. Ford and D. R. Fulkerson. Solving the transportation problem. *Management Science*, 3(1):24–32, 1956. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2627172.
Algorithm for Two-Phase Facility Planning

Jon Forest, Stefan Vigerske, Ted Ralphs, Lou Hafer, jpfasano, Haroldo Gambini Santos, Matthew Saltzman, h-i gassmann, Bjarni Kristjansson, and Alan King. coin-or/clp: Version 1.17.6, April 2020. URL https://doi.org/10.5281/zenodo.3748677.

Ralph E. Gomory. Outline of an algorithm for integer solutions to linear program. Bulletin of the American Mathematical Society, 64(5):275–278, September 1958.

S. L. Hakimi. Optimum locations of switching centers and the absolute centers and medians of a graph. Operations Research, 12(3):450–459, 1964. URL https://EconPapers.repec.org/RePEc:inm:oropre:v:12:y:1964:i:3:p:450-459.

Frank L. Hitchcock. The distribution of a product from several sources to numerous localities. Journal of Mathematics and Physics, 20(1-4):224–230, 1941. doi: 10.1002/sapm1941201224. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sapm1941201224.

Harold W. Kuhn. The hungarian method for the assignment problem. Naval Research Logistics Quarterly, 2:83–97, 1955.

A. H. Land and A. G. Doig. An automatic method of solving discrete programming problems. Econometrica, 28(3):pp. 497–520, 1960. ISSN 00129682.

Larkin Liu. Two-phase algorithm for facility planning. https://github.com/larkz/equal_clustering, 2020.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281–297. University of California Press, 1967.

Mikko I. Malinen and Pasi Fränti. Balanced k-means for clustering. In Pasi Fränti, Gavin Brown, Marco Loog, Francisco Escolano, and Marcello Pelillo, editors, Structural, Syntactic, and Statistical Pattern Recognition, pages 32–41, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg. ISBN 978-3-662-44415-3.

Erich Schubert and Arthur Zimek. ELKI: A large open-source library for data analysis - ELKI release 0.7.5 "heidelberg”. CoRR, abs/1902.03616, 2019. URL http://arxiv.org/abs/1902.03616

Erich Schubert, Alexander Koos, Tobias Emrich, Andreas Züflle, Klaus Arthur Schmid, and Arthur Zimek. A framework for clustering uncertain data. Proc. VLDB Endow., 8(12):1976–1979, 2015. doi: 10.14778/2824032.2824115. URL http://www.vldb.org/pvldb/vol8/p1976-schubert.pdf.

Grigory Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier membre. sur quelques propriétés des formes quadratiques positives parfaites. Journal für die reine und angewandte Mathematik, 1908. URL https://www.deutsche-digitale-bibliothek.de/item/KDI2RGX0022HGVX6KZCPNRI522TZCHVG
A. Weber. *Ueber den Standort der Industrien: Reine Theorie des Standorts, mit einem mathematischen Anhang, von Georg Pick.* Ueber den Standort der Industrien. J.C.B. Mohr (Paul Siebeck), 1909. URL https://books.google.de/books?id=FSrZAAAAMAAJ