A Constrained k-means and Nearest Neighbor Approach for Route Optimization: With an application to the Bale Collection Problem

David S. Zamar [*] **Bhushan Gopaluni** [*] **Shahab Sokhansanj** [*,**]

[*] *Department of Chemical and Biological Engineering, University of British Columbia, Vancouver, BC, V6T 1Z3, Canada (e-mail: zamar.david@gmail.com, bhushan.gopaluni@ubc.ca, shahabs@chbe.ubc.ca)*

[**] *Resource and Engineering Systems Group, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, 37831-6422*

Abstract: The bale collection problem (BCP) appears after harvest operations of agricultural crops. Its solution defines the sequence of collecting bales which lie scattered over the field. Current technology on navigation systems in autonomous agricultural vehicles and machines are able to provide accurate data for reliable bale collection planning. This paper presents a constrained k-means algorithm and nearest neighbor approach to the BCP, which minimizes travel time and hence fuel consumption. The constraints imposed on the k-means procedure is not the usual condition that certain groups of objects must be clustered together, but rather that the cluster centers must lie on valid locations, which may be specified as functions or sets of points. The algorithmic route generation provides the basis for a navigation tool dedicated to loaders and bale wagons. The approach is experimentally tested on a simulated study area similar to those found in real situations.

Keywords: constraint satisfaction problems, optimization, logistics planning, autonomy, agriculture

1. INTRODUCTION

The agricultural industry is now capable of collecting comprehensive real-time data regarding their field operations. Proper use of this data compels the formulation of novel methods to help improve the management of tasks involving the coordination of agricultural machines and vehicles. These technologies can provide accurate information for precision agriculture (PA) decision support systems in farm management.

PA is conceptualized by a system approach to reorganize farm management systems towards a low-input, high-efficiency, sustainable practice. PA benefits from a suite of technologies, such as global positioning system (GPS), geographic information system, automatic control, remote sensing, miniaturized computer components, mobile computing, advanced information processing, and telecommunications (Gracia et al., 2013).

Major field operations are performed throughout the planned coordination of different farm equipment. The bale collection problem (BCP) appears after harvest and baling operations of a crop and consists of defining the sequence in which bales spread over a field are collected. Once the harvesters have operated throughout the field, the mowed crop is left behind in windrows to be compressed and compacted by balers into bales that are convenient to handle and transport. Bales remain scattered on the surface of the field awaiting their collection by loaders and transported to the roadside by wagons (either self-propelled or pull-type).

The BCP is concerned with the collaborative operation of several machines and vehicles. Therefore, planned management becomes necessary to coordinate the various tasks efficiently. Usually the collection is decided by the operator based on his skills and experience. The inconsistent and subjective nature of decisions based on operator judgment tend to produce suboptimal solutions (Milkman et al., 2009). On the other hand, an accurate bale collection plan and its proper execution are achievable. Balers, loaders and bale wagons can be provided with positioning-system based devices enabling geo-referenced information (Amiama et al., 2008) which makes it possible to know the exact location of bales and track vehicles on predetermined paths. The BCP can be modeled and solved efficiently by applying optimization techniques, and thereafter be integrated as a part of a farm management decision support system (Gracia et al., 2013).

Solving the BCP involves determining the optimal roadside storage sites and the bale collection routes while taking into account the maximum capacity of the wagons as well as the distance traveled in transporting the bales to the roadside. The general procedure is that bales are transported from the field to the roadside, where they will be collected and taken to more permanent storage

Preprint submitted to 20th IFAC World Congress. Received November 11, 2016.

locations (i.e., silos, storage bunkers or barns). Hence, it becomes necessary to describe such operations by mathematical models that can be used for optimal allocation, route planning and timing.

The BCP belongs to a class of operational research problems known as the vehicle routing problem (VRP), which has been widely studied. The vehicle routing problem (VRP) is a combinatorial optimization and integer programming problem and is a generalization of the traveling salesman problem (TSP). The goal of the VRP is to find the optimal set of routes for a fleet of vehicles delivering goods or services to a set of geographically dispersed locations or customers. Eksioglu et al. (2009) have developed a taxonomic review of the abundant literature published on VRP related research. Despite the fact that field tasks involve the collaborative use of vehicles, these concepts have only recently been transferred to the agricultural environment (Bochtis et al., 2013; Gracia et al., 2013). According to the theory of computational complexity, this class of decision problems is nondeterministic polynomial time complete (NP-complete).

Commonly used techniques for solving VRPs focus on the use of algorithmic methods based on the application of heuristics or meta-heuristics, because no exact algorithm can be guaranteed to find the optimal route in reasonable computing time when the number of customers is large. Heuristic methods perform a relatively limited exploration of the search space and typically produce good quality solutions within modest computing times. Examples of heuristic procedures applied to VRP include particle swarm optimization (Lei et al., 2014), artificial optimized performance of bees (Szeto et al., 2011), ant colony optimization (Yu and Li, 2012), constraint programming algorithms (Rego, 2006) and genetic algorithms (Gracia et al., 2013). This paper presents a heuristic algorithm to efficiently solve the BCP appearing after mowing and harvesting operations. The proposed algorithm aims to increase the overall field efficiency of collection operations by providing the basis for a navigation tool dedicated to loaders and bale wagons. The low computational requirements of our method makes it feasible for integration in large scale operations. The proposed method has two main parts. The first part involves identifying the optimal roadside storage sites where bales are to be transported. This translates to a constrained cluster analysis optimization problem as the storage sites must lie on the side of the road. The second part involves identifying efficient collection routes for transporting the bales from the field to their respective roadside storage site. This is a capacitated VRP (CVRP) as the number of bales that a wagon can pick up on any given route is limited by the wagons capacity.

2. METHODS

2.1 Part I: Constrained Cluster Analysis of Bales

Problem Description The identification of roadside storage sites where bales are to be transported can be expressed as a cluster analysis problem, where the aim is to partition the bales into k clusters in which each bale belongs to the cluster with the nearest mean, resulting in a partition of the bales on the field. If the location of the cluster centers were not constrained to lie on the roadside, then the k-means algorithm (Hartigan, 1975), originally used for signal processing, may be directly applied. Even so, this standard cluster analysis problem is known to be computationally difficultly (NP-hard). The fact that the storage sites are constrained to lie on the roadside makes this part of the BCP an even greater challenge to solve. We believe that this is the first work that shows how to optimize this class of constrained cluster analysis problem.

Given a set of points $\mathcal{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$, where each point is a d-dimensional real vector, k-means clustering aims to partition these n points into $k \leq n$ sets $S = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster sum of squares. In the BCP we also have m compact sets in $\mathbb{R}^d$ denoted $\{T_1, \dots, T_m\}$, such that the valid positions for the cluster centers are in $T = T_1 \cup \cdots \cup T_m$. We introduce the auxiliary functions:

$$g_j(\boldsymbol{x}) = \min_{\boldsymbol{t} \in T_j} ||\boldsymbol{x} - \boldsymbol{t}||^2, \quad j = 1, \dots, m$$

to help formulate our constrained cluster analysis (CCA) problem shown in Equations (1a) to (1d). The function $g_j(\boldsymbol{x})$ calculates the minimum Euclidean distance between a given point, $\boldsymbol{x}$, and any point in the set T_j.

$$\underset{(\boldsymbol{\mu_1}, S_1), \dots, (\boldsymbol{\mu_k}, S_k)}{\text{minimize}} \quad \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in S_i} ||\boldsymbol{x} - \boldsymbol{u_i}||^2 \tag{1a}$$

$$\text{subject to} \quad \prod_{j=1}^{m} g_j(\boldsymbol{u_i}) = 0 \quad \text{for} \quad i = 1, \dots, k \tag{1b}$$

$$S_1 \cup S_2 \cup \cdots \cup S_k = S \tag{1c}$$

$$S_i \cap S_j = \emptyset \quad \forall i \neq j. \tag{1d}$$

The constraint given in Equation (1b) ensures that each cluster center is located in a valid position. Equations (1c) and (1d) enforce strict partitioning clustering and a many to one mapping of bales to clusters, respectively. The CCA problem represented in Equations (1a) to (1d) is a quadratically constrained quadratic program that is NP-hard.

Step 1 Choose k random points from S to be the initial cluster centers, $\boldsymbol{u} = (\boldsymbol{u_1}, \boldsymbol{u_2}, \dots, \boldsymbol{u_k})$. The relaxed solution of the k-means algorithm is a good initial starting point.

Step 2 Assign points to clusters based on their Euclidean distance to the cluster centers:

$$S_i = \left\{ \boldsymbol{x} \in \mathcal{X} : ||\boldsymbol{x} - \boldsymbol{u_i}||^2 < ||\boldsymbol{x} - \boldsymbol{u_j}||^2, \;\; j \neq i \right\}.$$

Step 3 For $i = 1, \dots, k$, update the cluster centers by solving the following minimization problem:

$$\boldsymbol{u_i} = \arg\min_{\boldsymbol{u}} \left\{ f_i(\boldsymbol{u}) = \sum_{\boldsymbol{x} \in S_i} ||\boldsymbol{x} - \boldsymbol{u} + \gamma \cdot g_i(\boldsymbol{u})||^2 \right\}$$

Step 4 Repeat Steps 2 and 3 until there is no significant change in the clustering criteria, $\sum_{i=1}^{k} f_i(\boldsymbol{u_i})$.

Fig. 1. The CCA Algorithm

Solution Approach The back-fitting algorithm, described in Figure 1, may be used to solve the CCA problem represented in Equations (1a) to (1d). A sufficiently large constant weight parameter γ must be given to ensure the solution satisfies the CCA constraint of Equation (1b) as accurately as necessary.

2.2 Part II: Within Cluster Route Optimization

As in the VRP, this part of the BCP may be formulated as a graph theory problem. Let $G = (N, A)$ be an undirected graph, where N is the set of nodes and A is the set of edges. In our case, $N = \{0, 1, \ldots, n\}$ is an index set for the n bales and the roadside storage node, denoted as 0. $A = \{(i, j) \mid i, j \in N;\ i < j\}$ represents the set of $(n+1)(n+2)/2$ existing edges connecting the n bales and the storage site.

A weight, q_i, is assigned to each bale i, $1 \leq i \leq n$ $(q_0 = 0)$. Each edge has an associated cost, $c_{ij} > 0$, of sending a vehicle from node i to node j. The c_{ij} are assumed to be symmetric and proportional to the Euclidean distance, d_{ij}, between any two nodes, thus $c_{ij} = c_{ji} \propto d_{ij}$, $i, j \in N$. The collection process is to be carried out by a fleet of v vehicles, $v \geq 1$, with equal capacity, $\kappa \geq \max\{q_i \mid 1 \leq i \leq n\}$. Notice that each vehicle may represent an autoloader trailer or a loader accompanied by a transport wagon.

The problem is to determine the set of routes that minimize the total travel cost within each cluster identified by the CCA algorithm in Section 2.1. As time is not being considered, the number of vehicles does not effect the problem solution as the routes are independent of one another and can be completed either in series or in parallel. The following are some additional constraints associated with the problem:

i) The roadside storage node can only be visited at the start and at the end of each route.
ii) All routes begin and end at the roadside storage node.
iii) No two routes visit the same bale.
iv) All bales are visited exactly once.
v) No vehicle can be loaded exceeding its maximum capacity.

The decision vector is $\boldsymbol{x} = (x_{ijr})$, where $i, j \in N$, $r \in R = \{1, 2, \ldots, \tau\}$ and $\tau = \lceil n/\kappa \rceil$ is the number of routes needed in order to pick up all of the n bales:

$$x_{ijr} = \begin{cases} 1 & \text{if route } r \text{ contains edge (i,j)} \\ 0 & \text{otherwise,} \end{cases}$$

Here we assume that, with exception to the last route, vehicles are loaded to their maximum capacity. The mathematical formulation of the route optimization part of the BCP is as shown in Equations (2a) to (2e). Equation (2b) is to make sure that each bale is assigned to exactly one route. Equation (2c) states capacity constraints, so that the sum of all bales collected in a route is less than or equal to the loading capacity of the vehicle. Finally flow constraints are shown in Equations (2d) and (2e) to ensure that each route begins and ends at the roadside storage site and that the inflow and outflow of edges must be equal for all the nodes.

$$\underset{\boldsymbol{u}}{\text{minimize}} \quad \sum_{r \in R}^{t} \sum_{(i,j) \in A} c_{ij} x_{ijr} \tag{2a}$$

$$\text{subject to} \quad \sum_{r \in R} \sum_{j \in N} x_{ijr} = 1 \qquad \forall i \in N \tag{2b}$$

$$\sum_{i \in N} \sum_{j \in N} x_{ijr} \times q_j \leq \kappa \qquad \forall r \in R \tag{2c}$$

$$\sum_{j \in N} x_{0jr} = 1 \qquad \forall r \in R \tag{2d}$$

$$\sum_{i \in N} x_{ijr} = \sum_{i \in N} x_{jir} \qquad \forall j \in N,\ r \in R \tag{2e}$$

Solution Approach The VRP is an NP-hard problem, as is the route optimization part of the BCP, which explains why most research efforts have focused on heuristic approaches. Various approaches to solve the classical VRP have been investigated over the past decades. These range from the use of pure optimization methods for solving small size problems to the use of heuristics and meta-heuristics that provide near-optimal solutions for medium and large-size problems with complex constraints (Gracia et al., 2013). We provide a simple, yet efficient heuristic, summarized in Figure 2, for identifying near-optimal solutions for the second part (i.e., within cluster route optimization) of the BCP. Our heuristic works by identifying the most isolated bale, b^*, based on the number of times it is selected as a nearest neighbor, in Euclidean distance, by its peers. For each bale that selects b^* as a nearest neighbor, we solve a corresponding TSP that visits all of that bales $\kappa - 1$ nearest neighbors (which must include b^* by definition) as well as the storage node. Recall that κ represents the capacity of the wagon. The

Step 1 Set $k = min(\kappa, |N| - 1)$, where κ is the capacity of the wagon . Compute the set of $k - 1$ nearest neighbors for each bale $b \in N$. Denote the set of $k - 1$ nearest neighbors of bale b as Q_b.

Step 2 Define the function

$$M(x, S) = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise} \end{cases}$$

For each bale, $b \in N$, compute

$$m_b = \sum_{i \in N} M(b, Q_i)$$

Let $b^* = \arg\min_b \{m_b : b \in N\}$.

Step 3 Among the sets

$$\mathcal{Q} = \{Q_b \mid b^* \in Q_b, b \in N\},$$

which contain b^* select the one that has the shortest cycle, including the roadside storage node, and denote this set as Q^*.

Step 4 Update the set N by:

$$N = N \setminus Q^*.$$

If $N \neq \emptyset$, then go to Step 1.

Fig. 2. Minimax Route Optimization Algorithm

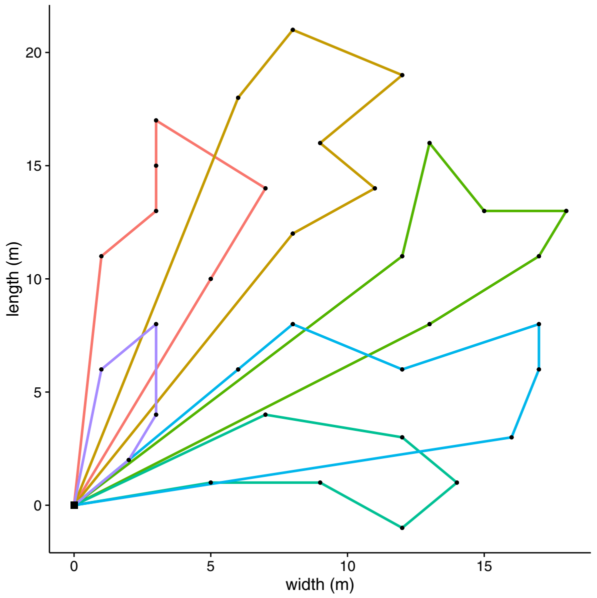


Fig. 3. Obtained tours by Minmax-ROA for a problem previously proposed by Grisso et al. (2007).

TSP with the shortest path is chosen as a route and the corresponding bales are removed from the set of nodes, N. The process is repeated until all bales have been picked up. The proposed heuristic is called the minimax route optimization algorithm (Minmax-ROA) because at each iteration it identifies the most isolated bale and minimizes the length of the route that picks up this bale as described in Step 3 in Figure 2. The Minmax-ROA algorithm reduces a very large CVRP into several much smaller TSP problems. This is justified as the number of bales to be collected are typically several orders of magnitude larger than the capacity of the vehicles. In such case, the proposed heuristic will dramatically reduce the complexity of the problem and the computing time necessary to solve it. There are almost no previous references to solving this part of the BCP in the literature. Grisso et al. (2007) raised a simple instance in which 34 bales scattered over a field should be collected with a vehicle capacity of 6 bales. This problem was subsequently solved by Gracia et al. (2013) with a hybrid genetic algorithm yielding a 6% reduction in the total travel distance. The solution found by our route optimization algorithm also achieves a 6% improvement in the travel distance. Figure 3 shows the solution obtained by Minmax-ROA. Each tour is depicted with a different color and the storage site is located at the origin.

3. APPLICATION

In order to test the proposed algorithm in a realistic situation, it is required to develop a problem generator able to produce problem instances from a certain set of parameters. A problem will be defined by the capacity constraint of the vehicle, κ, the location of roads and the n exact locations of the bales in the field. As such, the problem generator will need to simulate the number and location of bales on the field.

If we consider a uniform yield (kg ha^{-1}) throughout the field, the distribution of bales follows a constant distance pattern easily obtained using a Poisson process with rate (or intensity):

$$\lambda = \frac{\mu \times 10000}{\gamma \times \omega}, \tag{3}$$

where λ is the average travel distance in meters (m) by a baler until it packs a bale, μ is the mass of one bale in kilograms (kg), γ is the production level of wheat (kg ha^{-1}) and ω is the working width of balers in meters. We focus on farmlands in Western Canada, where the majority of farmlands have been divided into quarter (square) sections of approximately 64 hectares. The hectare is the area of $10,000$ m^2.

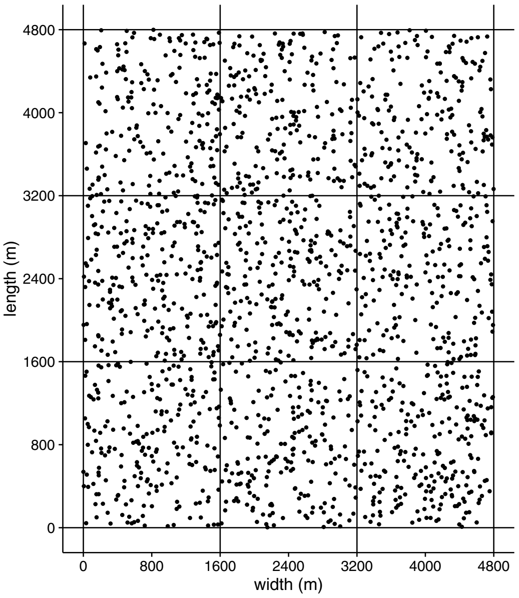


Fig. 4. Distribution of wheat bales in a study area composed of 9 sections. The sections are separated by access roads represented as lines.

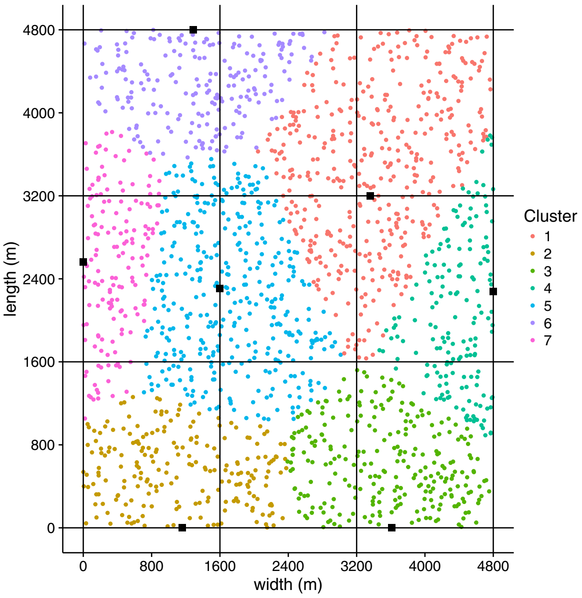
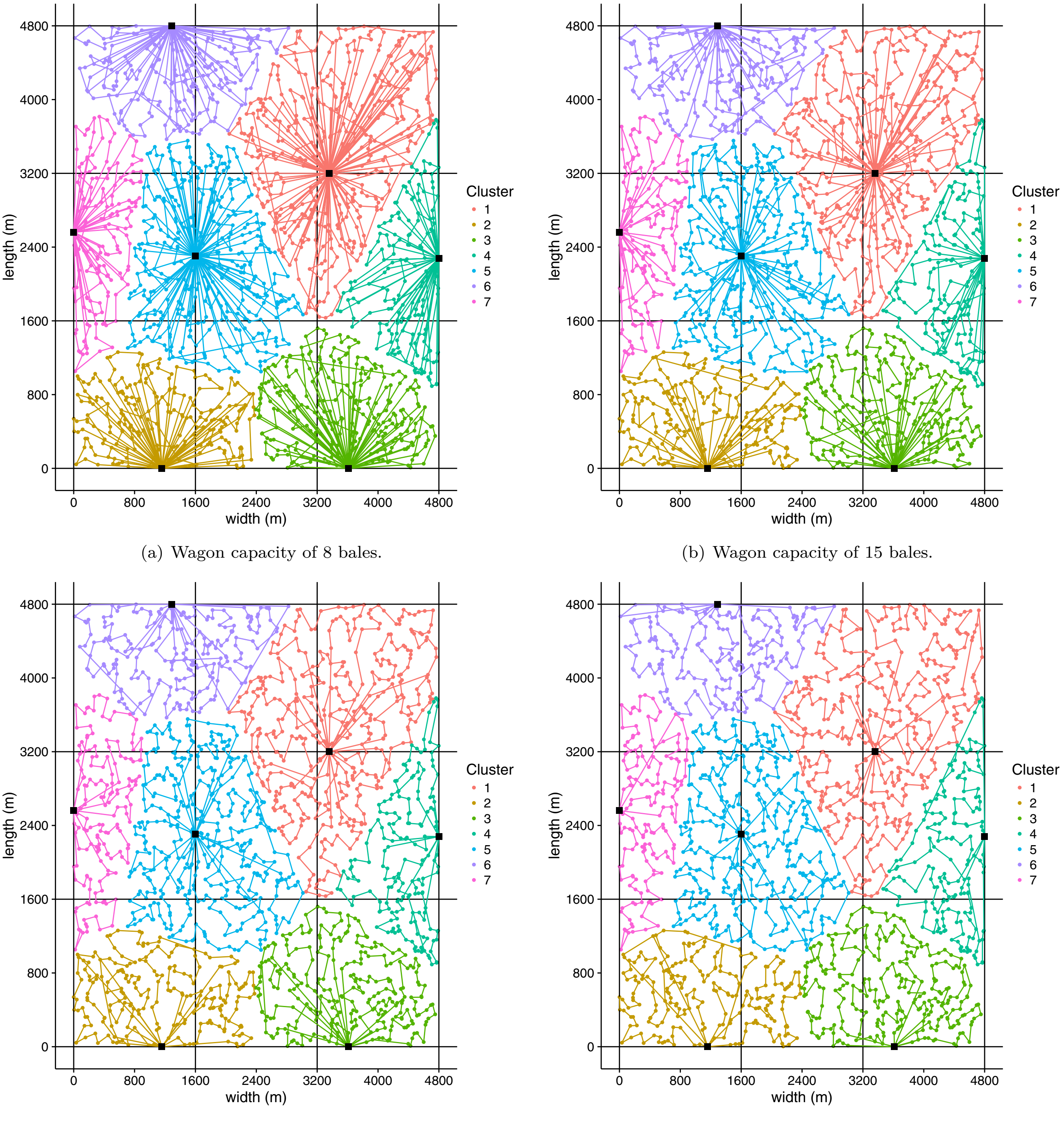


Fig. 5. Bale clusters and roadside storage sites identified by the CCA method.



(a) Wagon capacity of 8 bales.

(b) Wagon capacity of 15 bales.

(c) Wagon capacity of 35 bales.

(d) Wagon capacity of 70 bales.

Fig. 6. The Minmax-ROA solutions for wagon capacities of 8, 15, 35 and 40 bales. Larger wagon capacities provided more optimal solutions as the total distance traveled was found to decrease logarithmically with an increase in the wagon capacity.

Table 1. Model Parameter Values

Parameter	Value	Unit
Average yield	1500	kg ha^{-1}
Working width of baler	6	m
Average mass of bales	450	kg
Wagon capacity	8, 15, 35, 70	bales

We consider a study area with a dimension of 4800 m × 4800 m that is composed of 9 sections of $64 \times 4 = 192$ ha each. Figure 4 shows the distribution of bales in the study area.

The input parameters for the first and second parts of the BCP are comprised in Table 1, which includes information on the average yield, bale, and machinery characteristics. Different capacity constraints appear depending on the wagon used. There is a wide range of either self-propelled or pull-type bale wagons with different loading capacities depending on the dimension of the bales. Four different wagon capacities are considered in Table 1.

Regarding the first part of the BCP, we chose to divide the bales into seven clusters. As such, we apply the CCA algorithm with $k = 7$ in order to identify the bale clusters and their corresponding roadside storage sites, which are constrained to lie on the grid of roads shown in Figure 4. The seven bale clusters and corresponding constrained centers identified by the CCA method are shown in Figure 5. Other choices for the number of clusters could be considered during the optimization process, but this is beyond the scope of this paper.

The optimal routes identified by the Minmax-ROA for the four wagon capacities that were considered (8,15,35, and 70) are shown in Figure 6. A summary of the solution obtained for each wagon capacity is shown in Table 2. We implement our algorithms using the R programming language (R Core Team, 2016). The optimal path for each route is calculated using the TSP package in R (Hahsler and Hornik, 2007). As expected, the number of routes required decreases with an increase in the wagon capacity. The travel distance also decreases as the wagon capacity increases, but at a logarithmic rate. For this application, the recommended wagon capacity is 35 bales as the rapid decline in the total distance traveled by loaders and bale wagons reaches a plateau at this value. We can also see this lower plateau in the total travel distance by observing the similarity between the route densities in Figure 6 (c) and (d).

Table 2. Summary of Minmax-ROA Solutions

Wagon Capacity (bales)	Number of Routes	Distance (m)
8	229	525,855
15	124	341,712
35	55	221,361
70	30	184,562

4. CONCLUSION

We present a two-part optimization approach for solving the bale collection problem (BCP). The first part of our approach is to identify the optimal locations for the roadside storage sites where bales are to be temporarily piled for future transport to their final destination, such as a silo, silage bunker or barn. We assume that the number of roadside storage sites is given, but their locations are to be optimized. This results in a constrained cluster analysis problem as the storage sites must be located on the roadside. This first part of the BCP is solved using a new constrained k-means cluster analysis algorithm (CCA). This is not to be confused with previous work on constrained k-means procedures where the constraints consist of groups of points that must be clustered together. On the other hand, our constraints pertain to the locations of the cluster centers, which must be situated on the roadside. The optimization of the number of roadside storage sites should be determined to minimize the total travel distance in the second part of the BCP problem. This will be considered in future research.

The second part of our approach to solving the BCP is to identify the optimal collection routes for bringing the bales within each cluster to their corresponding roadside storage site, which have already been determined by the CCA algorithm. We developed an algorithm, called the Minmax-ROA, which approximately solves this NP-complete problem by sequentially identifying the most isolated bales and optimizing their collection routes. The Minmax-ROA heuristic was shown to give comparable results in a test case proposed by Grisso et al. (2007), which was subsequently solved by Gracia et al. (2013) with a hybrid genetic algorithm yielding a 6% reduction in the total travel distance. The solution identified by the Minmax-ROA algorithm achieves the same improvement.

The potential benefits of our approach to solving the BCP is its scalability and ease of implementation, which allow it to tackle much larger problems. The Minmax-ROA algorithm implements a divide and conquer strategy that breaks down a complex CVRP into several smaller TSPs that can be approximately solved using well-known and efficient heuristic procedures.

ACKNOWLEDGEMENTS

We would like to acknowledge the financial support from MITACS, NSERC, and BioFuelNet NCE (BFN).

REFERENCES

Amiama, C., Bueno, J., Álvarez, C.J., and Pereira, J.M. (2008). Design and field test of an automatic data acquisition system in a self-propelled forage harvester. *Computers and Electronics in Agriculture*, 61(2), 192–200.

Bochtis, D.D., Dogoulis, P., Busato, P., Sørensen, C.G., Berruto, R., and Gemtos, T. (2013). A flow-shop problem formulation of biomass handling operations scheduling. *Computers and Electronics in Agriculture*, 91, 49–56.

Eksioglu, B., Vural, A.V., and Reisman, A. (2009). The vehicle routing problem: A taxonomic review. *Computers & Industrial Engineering*, 57(4), 1472–1483.

Gracia, C., Diezma-Iglesias, B., and Barreiro, P. (2013). A hybrid genetic algorithm for route optimization in the bale collecting problem. *Spanish Journal of Agricultural Research*, 11(3), 603–614.

Grisso, R.D., Cundiff, J.S., and Vaughan, D.H. (2007). Investigating Machinery Management Parameters with Computer Tools. *ASABE Conf, Paper 071030.*

Hahsler, M. and Hornik, K. (2007). Tsp – Infrastructure for the traveling salesperson problem. *Journal of Statistical Software*, 23(2), 1–21. URL `http://www.jstatsoft.org/v23/i02/`.

Hartigan, J.A. (1975). *Clustering algorithms.* John Wiley & Sons, New York.

Lei, K., Zhu, X., Hou, J., and Huang, W. (2014). Decision of multimodal transportation scheme based on swarm intelligence. *Mathematical Problems in Engineering*, 2014, 1–10.

Milkman, K.L., Chugh, D., and Bazerman, M.H. (2009). How Can Decision Making Be Improved? *Perspectives on Psychological Science*, 4(4), 379–383. doi: 10.1111/j.1745-6924.2009.01142.x.

R Core Team (2016). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rego, C. (2006). *Operations Research/Computer Science Interfaces Series: Metaheuristic Optimization via Memory and Evolution : Tabu Search and Scatter Search.* Springer US.

Szeto, W.Y., Wu, Y., and Ho, S.C. (2011). An artificial bee colony algorithm for the capacitated vehicle routing problem. *European Journal of Operational Research*, 215(1), 126–135.

Yu, S.P. and Li, Y.P. (2012). An improved ant colony optimization for vrp with time windows. *Applied Mechanics and Materials*, 263-266, 1609.