Turk J Elec Eng & Comp Sci
() : –

doi:10.3906/elk-

Clustered Mobile Data Collection in WSNs: An Energy-Delay Trade-off

Izzet Fatih SENTURK*
Department of Computer Engineering, Bursa Technical University, Bursa, Turkey,
ORCID iD: https://orcid.org/0000-0002-1550-563X

Received: .201 • **Accepted/Published Online:** .201 • **Final Version:** ..201

Abstract: Wireless sensor networks enable monitoring remote areas with limited human intervention. However, the network connectivity between sensor nodes and the base station (*BS*) may not be always possible due to the limited transmission range of the nodes. In such a case, one or more mobile data collectors (MDCs) can be employed to visit nodes for data collection. If multiple MDCs are available, it is desirable to minimize the energy cost of mobility while distributing the cost among the MDCs in a fair manner. Despite availability of various clustering algorithms, there is no single fits all clustering solution when different requirements and performance metrics are considered. Depending on the available wireless communication technology, the MDCs may or may not be required to visit the *BS* to forward the collected data. This paper considers both cases and suggests clustering algorithms for various performance metrics including the energy consumption and the maximum travel time.

Key words: Wireless Sensor Networks, Mobile Data Collection, Traveling Salesman Problem, Clustering, Energy Consumption

1. Introduction

The technological advancements of the past decade in the field of microelectromechanical systems have enabled the prevalence of Wireless Sensor Networks (WSNs). WSNs are composed of low-cost nodes with sensing, processing, and wireless communication capabilities. The nodes are typically powered using batteries with a limited energy capacity. Usually, it is infeasible to recharge batteries considering the tendency of deploying WSNs in hostile terrains and the sheer number of nodes in the network. Therefore, WSNs are expected to be functional over extended periods in an autonomous manner without external intervention. To extend the network lifetime, energy consumption must be reduced or the network should generate energy from the surrounding environment. While energy harvesting mechanisms are available, due to the variability of energy sources, harvested energy is subject to intermittent outages and current solutions are still immature [1].

WSNs are often complemented with a powerful base station (*BS*), also known as sink, to provide connectivity between the network and the remote user. The limited transmission range of nodes imposes inter-node collaboration to access the *BS*. In a multi-hop wireless communication scheme, the generated data is relayed through the nodes and collected at the *BS*. Permanent loss of connectivity with the *BS* has adverse effects on network operations and even leads to network partitioning if the failed nodes serve exclusively on the routing paths of other nodes [2]. Such node(s) are classified as cut-vertices. Due to environmental exposure, the nodes are prone to physical damage and hardware malfunction. Battery depletion and the initial node

*Correspondence: izzet.senturk@btu.edu.tr

deployment strategy are other common factors leading to connectivity problems in WSNs.

In recent years, mobility has been used extensively to improve certain network performance metrics including lifetime, connectivity, coverage, and fault-tolerance in WSNs. Mobility can be enabled in the network by using various mobile entities such as mobile robots, vehicles, moving objects and humans, unmanned aerial vehicles (UAVs), etc [3–5]. Mobility can be provided to existing network elements (e.g. *BS*, nodes) or can be introduced to the network as part of a new element. According to the control level of mobility, existing mobility solutions can be broadly classified into three groups as random, predictable, and controlled. In this study, we assume the *BS* and the nodes are stationary and employ controlled mobility through multiple MDCs to provide intermittent connectivity between network partitions. We assume a mobile robot such as NetCar [6] to act as an MDC. Consequently, the MDCs are assumed to be equipped with limited onboard batteries. Considering the excessive energy cost of mobility, the primary goal of this study is minimizing the energy consumption of the MDCs while distributing the mobility load between the MDCs in a fair manner.

The MDCs traverse between partitions and the *BS* to collect and forward data. If there are multiple MDCs, the set of nodes are divided into groups and each subset is visited by one MDC. It is desirable to minimize the distance traveled by MDCs to reduce the energy cost of mobility. Therefore, the MDCs follow a closed tour (i.e. cycle) by visiting each node in the respective subset once and returning to the starting node. Considering the availability of wireless communication, the MDCs do not have to visit the exact location of a node to collect its data. Furthermore, depending on the network layout and the employed transmission range, it can be possible to collect data from multiple partitions upon visiting carefully selected data collection points (DCPs) in the network. This study employs *SZP* [7] to determine the set of DCPs. The main goal of *SZP* is ensuring coverage to all partitions while minimizing the number of DCPs. Next, the DCPs are clustered into groups as many as the number of MDCs and one MDC is assigned to each group.

This study considers two different use cases for the MDC-*BS* connectivity depending on the wireless communication technology available for the MDCs. In the first use case, the MDCs are equipped with Low Power Wide Area (LPWA) technology to communicate with the *BS* over longer ranges. Therefore, the MDCs are assumed to be able to communicate with the *BS* independent from their location. In the second use case, the MDCs employ the same short-range wireless technology with the nodes for data collection. Due to the constrained transmission range, the MDCs are required to visit the *BS* in the second use case. This additional requirement is not addressed by existing clustering schemes. Therefore, this study presents a novel clustering algorithm considering the mentioned requirement. Obtained results indicate that the proposed solution reduces the energy consumption compared to existing clustering strategies at the expense of increased maximum travel time.

For both use cases, the MDCs follow a cycle in the respective cluster periodically until their battery is depleted. The order of the DCPs to be visited in each cycle is determined by modeling the problem as the Traveling Salesman Problem (TSP). In the TSP model, each DCP denotes a city and the MDC acts as a salesman. The TSP considers the DCPs, that are defined as discrete points in the two-dimensional plane, while designating the cycle. On the other hand, *SZP* provides a continuous Steiner zone associated with each DCP. A Steiner zone indicates an area where communication is possible with the corresponding nodes in one or more partitions. After obtaining the TSP solution, Steiner zones are used instead of the DCPs to identify the exact locations to be visited in the cycle. Considering a continuous area rather than a discrete point provides flexibility in designating the cycle. As a result, the TSP solution is improved by reducing the cycle length.

2. Problem statement

$n-1$ sensors and one BS are deployed randomly in an application area of $W \times H$ meters. Sensor nodes and the BS have a common transmission range of R meters to communicate with each other. The BS is also equipped with an LPWA technology (e.g. LTE-M) to connect remote users to the network. Due to the random deployment, the network connectivity is not assured. The network is assumed to be divided into multiple disjoint partitions. Each partition consists of one or more nodes. Intra-partition connectivity is possible, however, partitions cannot establish connectivity with the rest of the network. To provide intermittent connectivity between partitions and the BS, in a delay tolerant network, m MDCs are employed. The MDCs use the same short-range wireless technology to communicate with nodes and share the same transmission range of R. In certain cases, the MDCs can also be equipped with an LPWA technology so that they do not have to approach the BS to forward the collected data. We consider both use cases as two different application scenarios. Availability of the wireless communication technology enables remote data collection from nodes. Based on R and the network layout, it is possible to collect data from one or more nodes upon stopping at a DCP. Considering the availability of m MDCs, the DCPs are clustered into m groups and each cluster is associated with an MDC. The MDCs visit the DCPs in the respective cluster in a certain order as a cycle periodically. Considering the excessive energy cost of mobility, the goal of this study is minimizing the energy consumption of the MDCs occurred due to mobility.

3. Background

3.1. Connectivity Maintenance in WSNs

Environmental conditions, hardware malfunction, battery life, and deployment characteristics often lead to WSNs partitioned into multiple disjoint segments. A significant amount of work has been devoted to fault-tolerance in WSNs and readers are referred to [2] for a detailed discussion. We will briefly touch upon three common classes of connectivity restoration schemes for WSNs. The first class of solutions considers the possibility of relocating nodes in the network. The main idea of these solutions is to restructure the network topology so that the partitions are reconnected. Some typical goals are minimizing the number of relocated nodes and minimizing the movement distance. The main drawback of this type of solutions is the additional mobility cost for each node. Since the nodes to be relocated cannot be known in advance, each node should be mounted on a mobile platform during the network setup.

An alternative approach is the deployment of additional nodes to the network. The new nodes introduced to the network is often referred to as relay nodes. Relay nodes can be less resource restricted to minimize their number or the same type of nodes can be used as well. The most common performance metrics are the number of relay nodes deployed to the network and their movement distance after deployment until reaching their final position. This type of solutions avoids the hardware cost of mobility for each node in the network. However, the main limitation of these solutions is the requirement of intervening the network or the application area. Unless an intervention is possible, such solutions will be infeasible. Another major challenge is the difficulty of computing the exact number of relay nodes before deployment. Using too many relay nodes will increase the deployment cost. On the other hand, the solution will be invalid unless sufficient number of relay nodes are deployed.

The third class of solutions to address connectivity issues in WSNs employ mobile entities. In the literature such mobiles are referred to as MDCs as in this study or data mules, mobile robots, etc. More recent studies employ UAVs as well. The MDCs are similar to relay nodes connecting partitions in the second class of solutions. However, unlike relay nodes which remain stationary in their final locations (possibly after relocating

from their initial deployment locations), the MDCs maintain active motion until their battery is depleted. The most common pattern is following a pre-determined static tour repeatedly and stopping at certain locations for a short time to collect data from nearby nodes. Some common performance metrics are tour length and travel time. When the locations to be visited are known, determining the optimal tour can be formulated as the TSP. On the other hand, nodes are equipped with transceivers and the MDCs do not have to visit the exact node locations to collect data from them. Remote data collection provides an opportunity to improve the tour. On the other hand, it complicates an already NP-complete problem [8].

3.2. Mobile Data Collection in WSNs

[9] presents a network connectivity solution for WSNs by introducing relay nodes and an MDC to the network. Relay nodes act as a cluster head where each sensor node sends data directly to the corresponding relay node. The MDC visits relay nodes to collect data. Two main goals are minimizing the number of relay nodes and the tour lengths of the MDCs. Considering the availability of transceivers, [9] formulates the problem as a special case of the traveling salesman problem with neighborhoods (TSPN). In this case, the MDC is able to collect data from relay nodes if its tour passes within a certain distance. Unlike [9], we assume multiple MDCs. To restore connectivity in a spatially separated WSN, a mobile data gathering scheme is presented by [10]. The proposed tour planning algorithm assumes a single MDC but an extension for multiple MDCs is also provided through employing k-means clustering on the subnetworks. The objective is minimizing the tour lengths of the MDCs while ensuring a certain threshold of energy consumption is not exceeded. Despite similarities with our problem such as using multiple MDCs and employing a clustering method, this solution does not consider remote data collection from nodes and requires the MDCs to visit the exact node location to collect data.

[11] defines sojourn points to be visited by the MDCs to collect data from nodes. The main goal of this study is minimizing the number of sojourn points while assuring connectivity coverage for all nodes. MDC tours are determined considering the buffer size of the nodes. Sojourn points are determined by dividing the network into triangles and defining a circumcenter for each triangle. The tour of the MDC is designated in a greedy fashion. [11] leads to increased number of DCPs (sojourn points) compared to SZP as indicated in [7]. A multi-objective optimization genetic algorithm is presented by [12] using multiple MDCs to collect data from partitioned segments. In this study, the MDCs visit the BS to forward the collected data from partitions. Considering the MDCs as salesmen, the problem is formulated as multiple traveling salesmen problem. Initially, Fuzzy c-means clustering algorithm is employed to determine the segments. It is assumed that the nodes in the same cluster can communicate with each other. Unlike [12], we identify the segments according to node connectivity. In other words, each segment is a connected component in our work.

Besides mobile robots, UAVs are also used to collect data from WSNs. [13] exploits multiple base stations mounted on UAVs and presents a joint optimization scheme for UAV tour planning and transmit power control. The problem is formulated as a mixed integer nonconvex optimization problem. Despite similarities, our problem is different in the sense that one of the use cases defined in this paper requires the MDCs to visit the BS. Another UAV-based data collection solution is provided by [14]. The objective is designating the tour of a UAV while maximizing the minimum data collection rate among all nodes. Unlike our study, one MDC is employed.

3.3. Low-rate Wireless Technologies

IEEE 802.15.4/Zigbee protocol stack enables low-cost low-power WSNs. The standard specifies three unlicensed frequency bands for WSNs to operate. 2.4 GHz band provides the highest data rate. On the other hand, WSNs

operating at 2.4 GHz band often experience interference from various devices including microwave ovens and Bluetooth devices. Other emerging low-power wireless technologies that can be used with WSNs involve LTE-M, LoRa, and SigFox which provide connectivity over longer ranges [15–17].

The low-rate wireless technology enabling wide area networks is classified as LPWA technology and typically used to interconnect battery-operated IoT devices over long ranges. Keeping data rates low is an essential part of these technologies to reduce power consumption and extend the battery life. LTE-M exploits existing cellular wireless infrastructure. Therefore, it has a better network coverage compared to SigFox and LoRa. Having operated on a well-known licensed spectrum, LTE is considerably robust. On the other hand, LTE-M requires equipping end devices with a SIM card to access the cellular network. This requirement increases hardware and maintenance costs.

SigFox and LoRa use unlicensed sub-GHz bands and do not require SIM cards to operate. Both of these technologies provide connectivity around 10 km in urban areas. SigFox requires end devices to send their data to SigFox servers through SigFox base stations for processing. The cost of using SigFox depends on the number of messages sent per day. LoRa requires an annual license fee. Both SigFox and LoRa impose strict requirements on employed data rates.

3.4. Clustering

Clustering is a well-studied technique in WSNs applied to organize sensor nodes into a set of groups (i.e. clusters) in order to improve efficiency of the network operations. Some predefined criteria (e.g. distance, cluster size, network load, etc.) are identified to designate clusters. Clustering can be regarded as a topology management mechanism [18] to improve network lifetime [19], load balancing [20], throughput, fault tolerance, mobility management [21], etc. In this study, we employ clustering to group nodes which can span over multiple partitions. Subsequently, one MDC is assigned to each group (i.e. cluster) for data collection. In this study, the main performance goal of clustering is to minimize energy consumption of mobility.

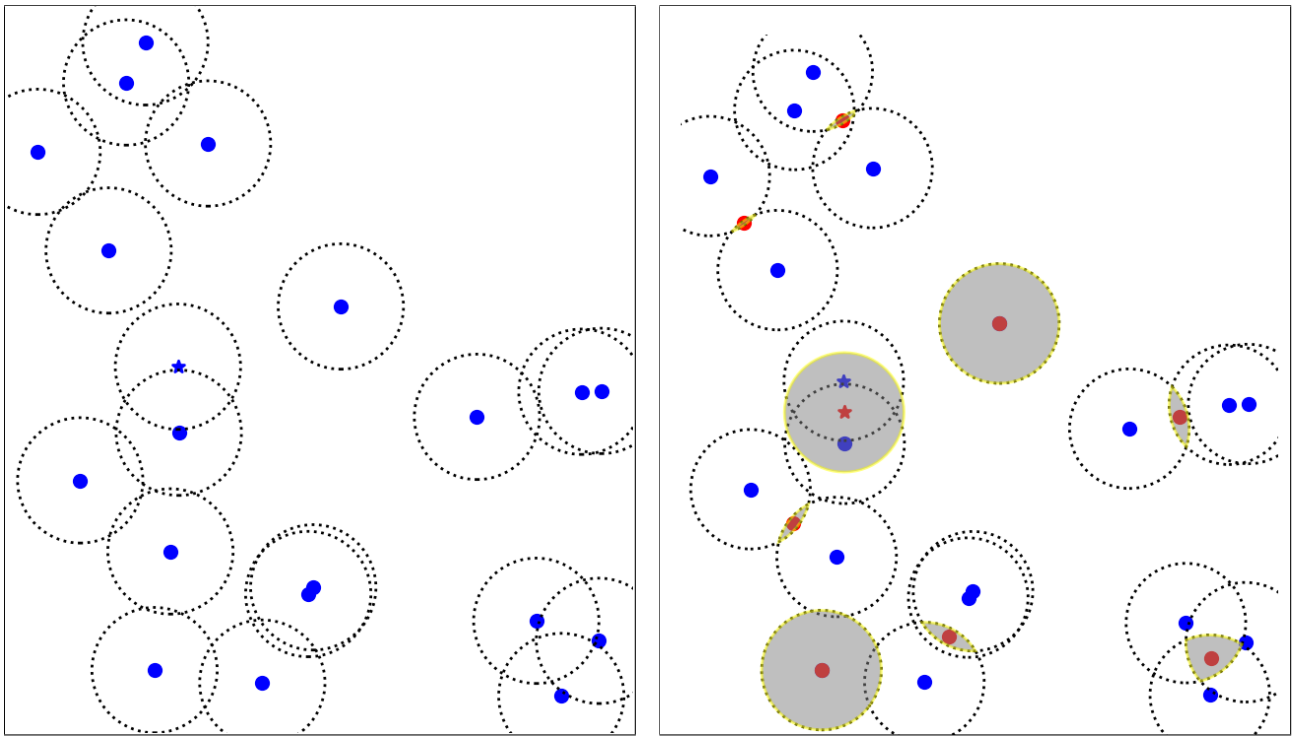
4. The Proposed Solution

First, we detail how the DCPs are identified. Then we introduce the proposed clustering solution, *The Closest Centers (CC)*, and discuss how the DCPs are assigned to the MDCs. Finally, we elaborate how the MDC tours are designated and then improved using a novel scheme.

4.1. Identifying the DCPs

Availability of the wireless communication deem it unnecessary to visit the exact node locations for data collection. Assuming a unit disk radio model [22], we define a circular coverage disk to represent the area where data collection is possible from a node. The disk is centered at the corresponding node location and the radius of the disk is equal to R. We assume that an MDC can collect data from a node while residing in the respective coverage disk. The MDCs are assumed to stop moving during data collection. Such stopping locations are referred to as DCPs.

Depending on R and the network layout, it is possible to have overlapping coverage disks. The degree of the overlap denotes the number of nodes that can be communicated with. It is desirable to determine the DCPs within the disk overlaps with a high overlapping degree in order to minimize the number of DCPs. Because, typically (but not necessarily), the number of DCPs is inversely related with the length of the tour visiting



(a) A sample network topology. (b) The DCPs identified by *SZP*.

Figure 1: The blue points represent node locations. The circles denote the coverage disks where communication is possible with the respective nodes. Steiner zones are signified with a gray background color. Each Steiner zone comprises a single DCP represented with a red point. While the blue star denotes the BS, the red star is the respective DCP for the BS. $R = 75$ meters.

these DCPs. Various solutions exist in the literature to identify data collection points [11, 23, 24]. This paper employs *SZP* [7] to identify the DCPs.

SZP models the mobile data collection problem as the Close-enough Traveling Salesman Problem (CETSP) and considers the degrees of disk overlaps (i.e. Steiner zones) similar to [23]. The number of coverage disks overlapping in a Steiner zone denotes the degree of the respective Steiner zone. Following an iterative approach, the Steiner zone with the highest degree is selected at each step. A representative point is determined in the selected Steiner zone to be used as a DCP. Covered nodes and their coverage disks are removed from the network and *SZP* proceeds with the next step by selecting a new Steiner zone until all nodes are covered with identified DCPs. [23] collects data from individual nodes and does not consider connected components (i.e. partitions) in the network. A connected component is a partitioned set of nodes that can communicate with each other but separated from the rest of the network. *SZP* assumes availability of multi-hop routing and addresses the problem of data collection from partitions comprising multiple nodes as well. For more information, the reader is referred to [7]. A sample network topology along with the DCPs determined using *SZP* can be found in Figure 1.

Considering the fact that the coverage disks have a common radius of R, the likelihood of disk intersections increases when the transmission range is extended. To demonstrate the relationship between the transmission range and the number of DCPs identified by *SZP*, we vary R between 25 and 100 and report

the average number of DCPs in Table 1. We have employed 30 different topologies as discussed in Section 5.1. The readers are referred to Section 5.1 for a detailed discussion on the topologies used in the experiments. According to Table 1, the average number of DCPs declines when R is extended. Considering the network size, the maximum number of DCPs cannot be more than 20. When R is increased to 100 meters, the number of DCPs declines as low as 6.

R (meters)	Minimum num. DCPs	Maximum num. DCPs	Average num. DCPs
25	16	20	18.03
50	12	16	14.03
75	8	14	10.73
100	6	11	8.03

Table 1: The minimum, maximum, and average number of DCPs identified by *SZP* with respect to R. The network size is 20. Please see Section 5.1 for a detailed discussion on the topologies.

4.2. DCP-MDC Assignment

The set of DCPs obtained in the previous step provides intermittent network coverage to the whole network. *SZP* assures that all of the nodes can communicate with one of the MDCs through the DCPs either directly or through multi-hop routing in the respective partition. If there is only one MDC in the network, obtained DCPs can be provided to the next step to designate a tour so that the DCPs are visited in a particular order. We assume the network is delay-tolerant. However, if multiple MDCs are available, we can use this opportunity to reduce the energy consumption and latency. Given the availability of m MDCs and the set of DCPs to be visited by exactly one MDC, the problem is assigning each DCP to an MDC so that the energy consumption and latency can be minimized.

To assign DCPs to m different groups, one of the clustering strategies can be employed. Clustering is regarded as the process of grouping objects into classes of similar objects [25]. Thus, objects within the same cluster will be more similar than objects in other clusters. As discussed in Section 3.4, several clustering methods exist in the literature. In this study, we employ *k-means*, *p-center*, and some of the hierarchical clustering algorithms as baselines. In the experiments, we consider two different use case scenarios. In the first scenario, the MDCs are assumed to be equipped with LPWA technology and therefore form clusters independent from the BS. However, in the second scenario, we assume availability of a short-range wireless technology for the MDCs. Consequently, the MDCs include the respective DCP to communicate with the BS as well. Experiments reveal that the performances of the well-known clustering methods decline for the second scenario.

In order to address the specific requirement of the second use case scenario, we present a novel clustering algorithm so that the objects (i.e. DCPs) are clustered while the specific ones (e.g. BS) are included by multiple clusters. *CC* obtains the set of DCPs (V) from *SZP*. One of the elements of V is the corresponding point to collect data from the BS. This DCP is regarded as V_{BS} and can be found based on its distance to the BS. *CC* forms m clusters (C) for m MDCs. Since V_{BS} will be included in all clusters, it is excluded from the set of DCPs ($V' = V - V_{BS}$). Next, the DCPs in V' are sorted according to their distance to V_{BS} in an increasing order. Considering the order of elements in V', one DCP is assigned to each MDC along with the V_{BS} in the first round. In the second round, the rest of the elements in V' are considered one by one and each of them is assigned to the closest cluster (C_j). The formal algorithm of *CC* can be found in Algorithm 1.

Algorithm 1 CC(V, m, V_{BS})

$V' = V - V_{BS}$ ▷ Set of DCPs but the BS
$C = \emptyset$ ▷ Set of clusters
Sort $V_i \in V', i = \{1, 2, \ldots, |V'|\}$ according to $D(V_i, V_{BS})$ in an increasing order. D is the Euclidean distance function
for $i = \{1, 2, \ldots, m\}$ **do**
    $C_i.append(V_i)$ ▷ $C_i \in C$
    $C_i.append(V_{BS})$
end for
for $i = \{1, 2, \ldots, |V'| - m\}$ **do**
    Find $C_j \in C$, the closest cluster to V'_i
    $C_j.append(V'_i)$
end for

4.3. Designating MDC Tours

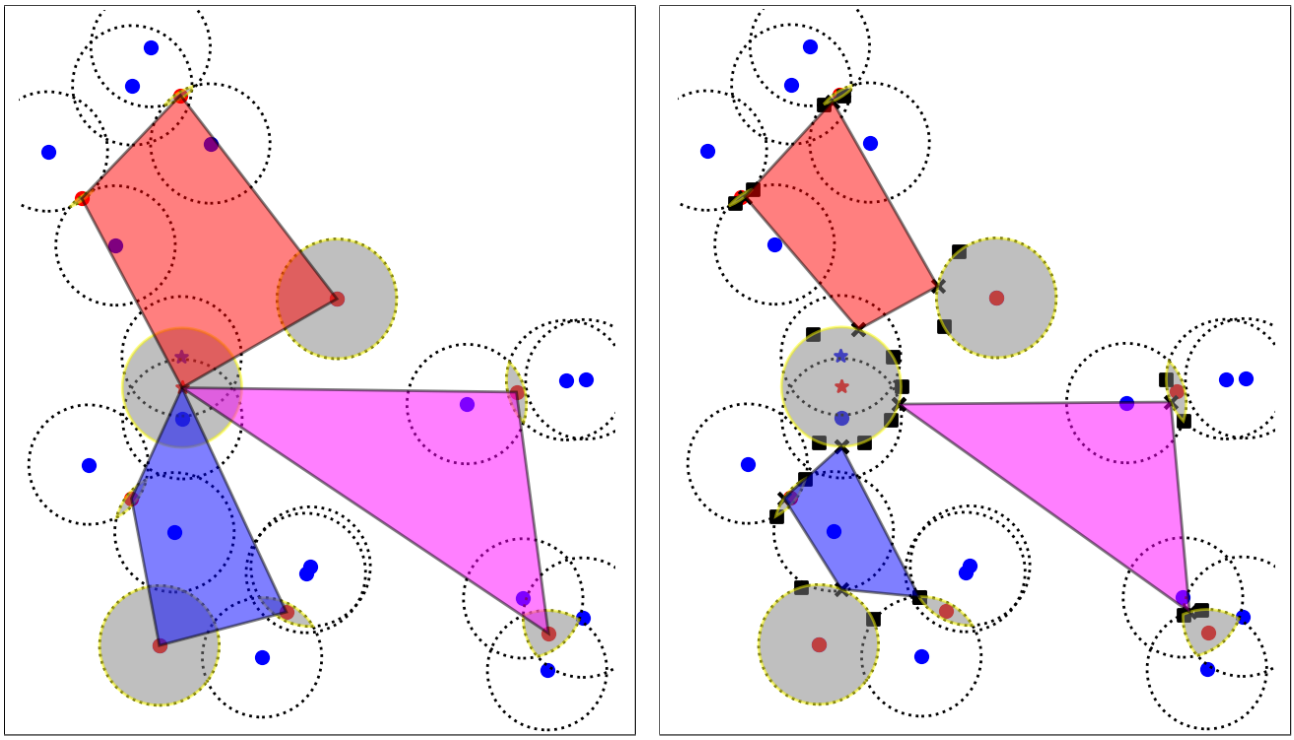
The applied clustering method provides m different groups of DCPs. Each DCP refers to a coordinate in the two-dimensional application area. The final step is to determine tours to be followed by each MDC. Given a set of coordinates, the problem of designating the optimal path visiting each coordinate once and returning to the original location is known as the TSP. Considering the similarities, the problem of determining the best tour for each MDC can be modeled as the TSP. The mobile robot represents the salesman in our case. Instead of cities, DCPs are visited. The optimal path refers to the closed tour (i.e. cycle) with the least mobility cost. Various cost metrics can be defined such as energy consumption, delay, etc. This study emphasizes minimizing the energy consumption given the latency tolerance of the delay tolerant networks. Unlike TSP, the MDCs follow the same tour recurrently until the battery is depleted. In this study, we report the cost of one tour.

Several variations of the TSP exist in the literature. This study regards the Euclidean distance between the DCPs and therefore employs Euclidean TSP. The TSP is solved using Google OR-Tools [26]. The distance matrix is calculated first using the Euclidean distance between the DCPs. The number of salesmen is one because one MDC is assigned to each cluster. In Scenario I, the MDCs are able to communicate with the BS without approaching it. In this case, the depot is the first DCP in each cluster and there are m depots in total. In scenario II, the MDCs are required to visit the V_{BS} which is the respective DCP for the BS. Therefore, V_{BS} is the single and common depot for all MDCs in the second scenario.

4.4. Tour Improvement

The Euclidean TSP requires discrete points to represent cities to be visited. Therefore, as discussed earlier, each Steiner zone is represented with a discrete DCP selected within the respective Steiner zone. After obtaining the TSP solution, we are no longer limited with the representative DCPs to collect data. Instead, we can consider the continuous Steiner zones for the corresponding DCPs to improve the MDC tours further. For the sample topology given in Figure 1a, three MDCs are employed. The DCPs are clustered into three groups and the TSP solution is obtained for each MDC. Computed TSP tours are illustrated in Figure 2a. To improve the obtained TSP tours further, the respective Steiner zones are considered in each TSP tour. First, the closest points are determined between consecutive Steiner zones in the tour. Then, median points between the identified closest points within the Steiner zones are determined and used as the actual DCPs in the improved TSP tours. The improved TSP tour can be found in Figure 2b.

In order to assess the performance of the tour improvement, we consider 30 topologies discussed Section 5.1



(a) Initial TSP tours.

(b) Improved TSP tours.

Figure 2: Three MDCs are employed in the topology given in Figure 1. Each polygon represents a separate MDC tour. The blue points represent node locations. The circles denote the coverage disks where communication is possible with the respective nodes. Steiner zones are given with a gray background color. Each Steiner zone comprises a single DCP represented with a red point. While the blue star denotes the BS, the red star is the respective DCP for the BS. In Figure 2b, the black squares indicate the closest points in consecutive Steiner zones within a TSP tour. The black cross signs represent the actual DCPs in the improved TSP tours. $R = 75$ meters.

and employ a single MDC (i.e. no clustering) for data collection. We report the average tour lengths before and after improvement with respect to R in Table 2. Table 2 shows that the average tour lengths decline for both cases when R is extended. This can be attributed to the increased coverage disk overlaps (i.e. Steiner degrees). Consequently, the number of DCPs decline as can be seen in Table 1. Table 2 also indicates that the average tour length decreases after the tour improvement. The cost decline slightly increases when R is extended and reaches up to 19% when $R = 100$ meters.

5. Performance

5.1. Experimental Setup

We have implemented a simulator using Python programming language. To simulate data collection from sensors, we have deployed n sensor nodes randomly in an application area of $W \times H$ meters. For statistical significance, 30 different topologies are used in the experiments. Generated topologies are publicly available [27] to ensure repeatability of the experiments. We assume a flat two-dimensional plane without obstacles. To enable mobile data collection, we have employed m MDCs. We assume availability of a mobile robot platform such as NetCar [6] to be used as an MDC. Experiments simulate the energy model (E) of NetCar with a constant

R (meters)	The average tour length (meters)	
	Before improvement	After improvement (change %)
25	2984,74	2642,84 (-11%)
50	2832,63	2364,43 (-17%)
75	2613,10	2138,55 (-18%)
100	2353,50	1909,53 (-19%)

Table 2: The average tour lengths before and after improvement with respect to R. The network size is 20. Please see Section 5.1 for a detailed discussion on the topologies.

velocity (V).

Two different use cases are considered in the experiments based on the wireless technologies available for the MDCs. In the first scenario, the MDCs use a short-range wireless technology (e.g. IEEE 802.15.4, 6LoWPAN, etc.) to collect data from nodes and an LPWA technology (e.g. LTE-M, SigFox, LoRa, etc.) to deliver the collected data to the BS. In this case, the MDCs do not have to approach the BS for data delivery. Thus, the MDCs only visit the DCPs in their respective cluster.

In the second scenario, LPWA is not available for the MDCs. Therefore, the MDCs use a short-range wireless technology to communicate with the sensor nodes and the BS. In this case, the BS is included in all clusters to impose the MDCs to visit the BS. In the experiments, a transmission range of R is used for the short-range wireless connectivity.

Table 3 summarizes the default parameters used in the simulations. To evaluate the impact of R and m on the performance metrics individually, we have varied m between 2 and 5 while $R = 75$ meters and then varied R between 25 meters and 100 meters when m is fixed to 3. Other parameters (e.g. n, W, H, E) are held constant in the experiments.

n	20
$W = H$	800 meters
m	2-5
R	25-100 meters
V	0.627 m/s
E	6,590.1 mJoule/m

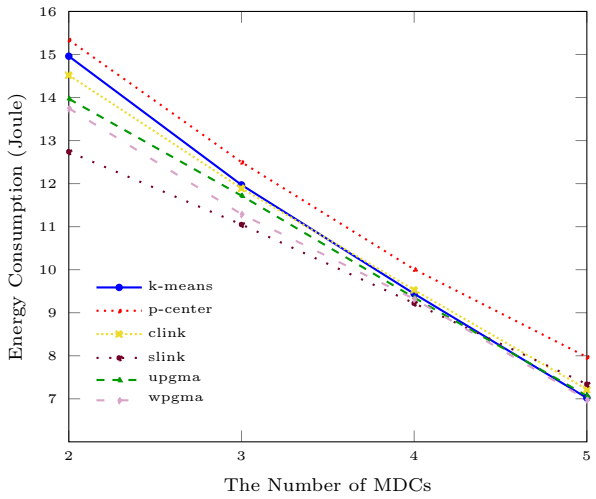
Table 3: Default parameters used in the experiments.

5.2. Performance Metrics

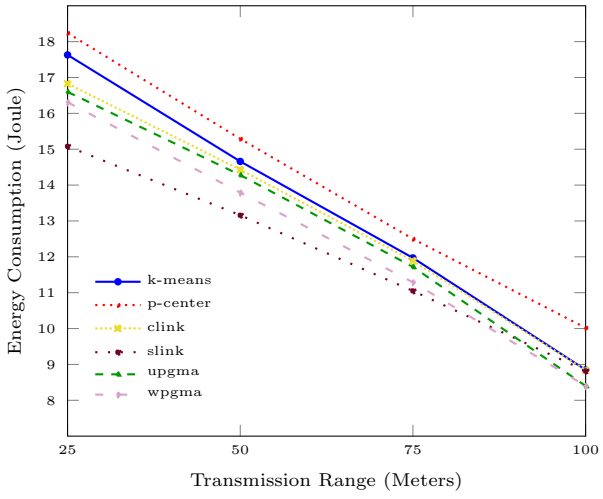
We have employed 2 different cost indicators to assess the performance of the proposed approach.

- *Energy Consumption* reports the total energy consumed due to the mobility of the MDCs. Energy consumption is given in Joules and calculated as $\sum_{i=1}^{m} TTL_i * E$ where TTL_i is the total tour length of each MDC in meters. E is the energy model when a fixed velocity (V) is applied. The energy model is obtained from [6]. The value of E can be found in Table 3.

- *Maximum Travel Time* indicates the longest time interval to complete a tour among the MDCs. Considering a fixed velocity (V) applied to the MDCs, the maximum travel time can be calculated as



(a) Energy consumption with respect to m. $R = 75$.



(b) Energy consumption with respect to R. $m = 3$.

Figure 3: The energy cost of mobility with varying number of MDCs and the transmission range.

$\max(\{\frac{TTL_i}{V}\} : i = 1, 2, \ldots, m)$. V is determined according to the energy model provided in [6]. The value of V can be found in Table 3.

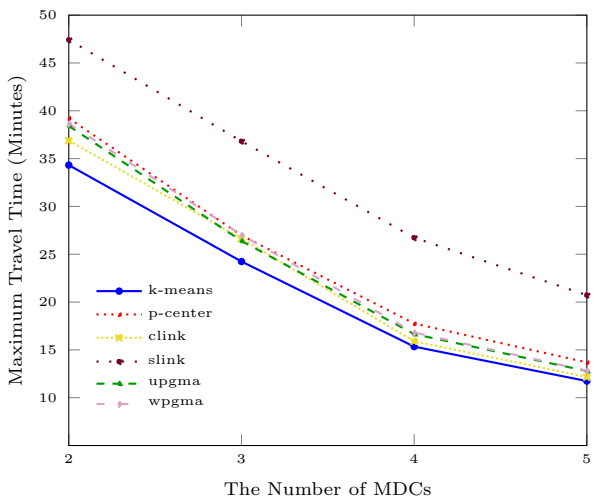
5.3. Results for Scenario I

Figure 3a depicts the change in the energy consumption when the number of MDCs employed in the network varies. Figure 3a indicates that the energy cost of mobility declines for all approaches when the number of MDCs is increased.

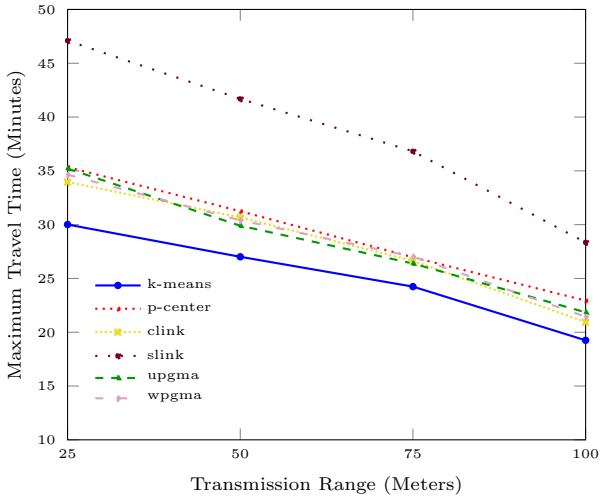
When the number of MDCs is increased, the average number of DCPs assigned to an MDC is expected to decrease. Also, in this scenario, the MDCs do not have to approach the BS to forward their data. Consequently, the MDCs designate tours including fewer DCPs. Tours with fewer DCPs are likely to result shorter tours incurring reduced cost of energy. For *slink*, the best performing approach in this experiment, energy consumption declines 42% by increasing the MDC count from 2 to 5. For *p-center*, the worst performing approach, the decline is 48%.

k-means is one of the most popular clustering methods in the literature to distribute the mobility load among mobile entities in a fair manner [28, 29]. On the other hand, *k-means* is not among the top performers in this experiment. *slink* provides the best results in most of the cases (m = 2,3, and 4) while *wpgma* outperforms the rest when $m = 5$. *p-center* incurs the highest cost among all. *wpgma* performs slightly better than *upgma*. *upgma* outperforms *k-means* when m is less than 5. *k-means* and *clink* perform similar but *clink* provides better results when m is less than 4. Compared to *p-center*, *slink* reduces energy consumption up to 17%.

Figure 3b shows the energy consumption results when R is changed. It can be observed from Figure 3b that the energy consumption declines for all approaches when R is extended. Since R is used as the radius of the coverage disks, when R is increased, the area of the coverage disks expands as well. Expanded coverage disks leads to higher degrees of disk overlaps (i.e. Steiner zones). Consequently, the number of DCPs required to cover the network declines as indicated by Table 1. The number of DCPs represents the cities that must be visited in a TSP tour. Thus, when the number of DCPs declines, it is likely to designate a tour incurring



(a) Maximum travel time with respect to m. $R = 75$.



(b) Maximum travel time with respect to R. $m = 3$.

Figure 4: Maximum travel time with varying number of MDCs and the transmission range.

reduced energy cost. For *slink*, the energy consumption reduces 41% when R is increased from 25 to 100. For other approaches, the decline is between 41% and 50% when R is increased from 25 to 100.

slink provides the best results when R is lower than 100 meters. When $R = 100$, *wpgma* performs slightly better than *slink*. *p-center* results the highest cost in all cases. *wpgma* and *upgma* perform similar but *wpgma* is slightly better. *k-means* performs worse than all approaches excluding *p-center*. Compared to *p-center*, *slink* reduces the energy cost up to 17%.

Figure 4a portrays how the maximum travel time changes depending on the available MDC count. Figure 4a suggests that the increased availability of the MDCs can alleviate the maximum travel time. As discussed earlier, the average number of DCPs assigned to a single MDC declines when there are more MDCs in the network. Fewer DCPs typically (but not necessarily) lead to shorter MDC tours. Consequently, the maximum travel time drops on average when the number of MDCs is increased. In this experiment, *k-means* provides the best results while *slink* leads to the highest maximum travel time. When the number of MDCs is increased from 2 to 5, the maximum travel time declines 66% and 56% for *k-means* and *slink* respectively.

According to Figure 4a, *k-means* outperforms other approaches for this cost metric. *clink* provides the best results after *k-means*. Compared to *clink*, *k-means* reduces the maximum travel time 7% when two MDCs are employed. However, the gap declines to 3% when the number of MDCs is increased to five. *clink*, *upgma*, *wpgma*, and *p-center* have similar performances. However, when the number of MDCs is four and more, *clink* performs slightly better and *p-center* performs slightly worse than others. Unlike energy consumption results, *slink* provides the worst performance for this metric. Compared to *k-means*, *slink* leads to higher travel times up to 43%.

Considering both results depicted in Figures 3a and 4a, it can be concluded that *slink* tends to form longer tours for some of the MDCs while *k-means* provides a better cost distribution among the MDCs. On the other hand, sharing mobility cost in a fair manner does not necessarily leads to improved energy consumption. On the contrary, energy consumption is found to be higher for *k-means* compared to *slink* as given in Figure 3a.

Figure 4b depicts how the maximum travel time changes with a varying R. It can be observed from

Figure 4b that the maximum travel time declines for all approaches when R is increased. This is expected for two reasons. First, the number of segments (i.e. partitions) is expected to decline when R is higher. Decreased partition count implies reduced demand for data collection. Second, extended R leads to larger communication disks and fewer DCPs. A reduction in the number of DCPs does not always results shorter tours. However, obtained results suggest that, on average, shorter tours can be formed with fewer points to be visited especially in Scenario I where clusters are formed without considering the BS.

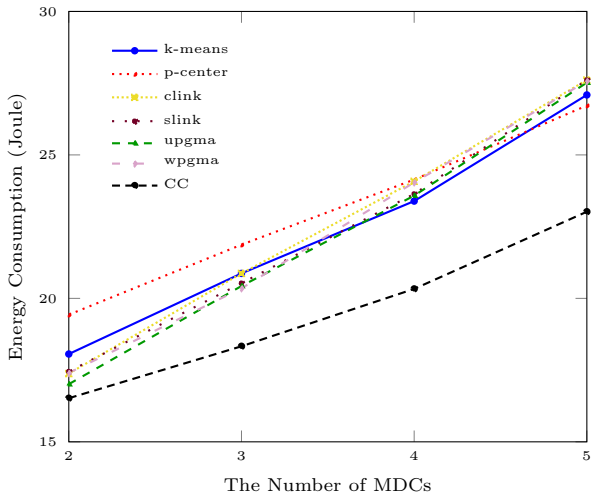
For *k-means*, the best performing approach, and *slink*, the worst performing approach, the maximum travel times decline 36% and 40% respectively when R is increased from 25 to 100. *clink*, *upgma*, *wpgma*, and *p-center* have similar performances. However, *clink* performs slightly better and *p-center* performs slightly worse than others when R is 100. Compared to *slink*, *k-means* reduces the maximum travel time up to 36%.

5.4. Results for Scenario II

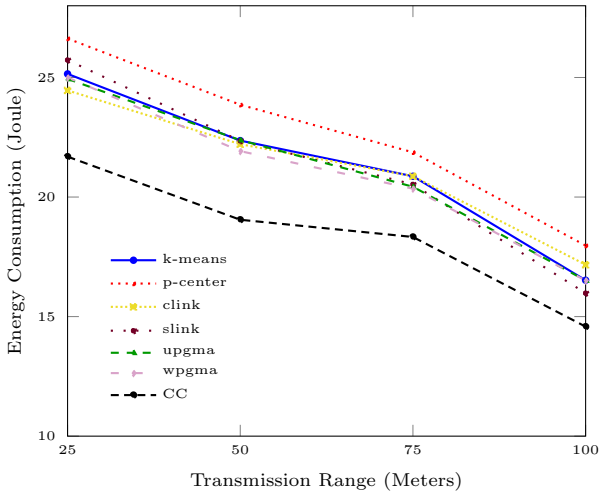
In this scenario, the MDCs are not equipped with an LPWA technology. Therefore, the MDCs employ the same transmission range, R, similar to sensor nodes to communicate with the BS. Such a constraint requires the MDCs to visit the V_{BS} to connect the BS and forward the collected data. In other words, in this scenario, each cluster is required to include the V_{BS}. This additional requirement makes it necessary to modify the existing clustering schemes. One possible solution is to apply clustering without considering V_{BS} and then append it to all clusters. We follow this approach to adapt the existing clustering methods that we employ in this scenario. Also, we employ a novel clustering algorithm, *CC*, particularly designed to group DCPs while considering the special case of the BS in WSNs.

Figure 5a presents how the energy cost changes for *CC* and other clustering solutions when the number of MDCs is varied. Figure 5a suggests that the energy consumption increases when there are more MDCs in the network. Such a relationship between the number of MDCs and the energy consumption can be counterintuitive considering the results obtained for Scenario I. Recall that the increased MDC count alleviates the energy cost for Scenario I as given in Figure 3a. However, a key difference between these two scenarios changes the cost pattern completely. To understand the change in the cost pattern, both scenarios should be studied with an edge case. Let us assume that there are $p+1$ DCPs in the network and $m = p$. In Scenario I, none of the MDCs visits V_{BS}. The remaining p DCPs are assigned to MDCs exclusively. Since MDCs can communicate with the BS using LPWA in Scenario I, the MDCs do not move after they are positioned to their respective DCPs. Therefore, the total energy cost will be zero in this case. It can be concluded that the energy consumption increases when the number of MDCs is decreased in Scenario I. On the other hand, in Scenario II, a separate tour will be formed between each DCP and V_{BS}. Requiring each MDC to visit V_{BS} leads to redundant movement when the network comprises more MDCs. A demonstrative example is a line topology of DCPs where V_{BS} is at the edge of the topology. When $m = 1$, one MDC follows the line and collect data. On the other hand, for $m = p$, each MDC follows their tour separately and the redundant mobility increases the energy consumption. In conclusion, the requirement of visiting V_{BS} is a constraint which reduces the flexibility of designating optimal tours. According to obtained results, the energy consumption increases 38% for *p-center* when the number of MDCs is increased from two to five. For the same case, the cost increase reaches 50% for *k-means* and 58% for *clink*.

It can be observed from Figure 5a that *CC* performs the best and provides the lowest energy consumption. *CC* reduces the energy consumption up to 14% and 15% compared to *p-center* and *k-means* respectively. While the performance gap is more steady between *CC* and *p-center*, *CC* increases the performance gap with *k-means* when the number of MDCs is higher. *p-center* results the highest energy consumption initially, but outperforms



(a) Total energy consumption with respect to m.



(b) Total energy consumption with respect to R.

Figure 5: The energy cost of mobility with varying number of MDCs and the transmission range.

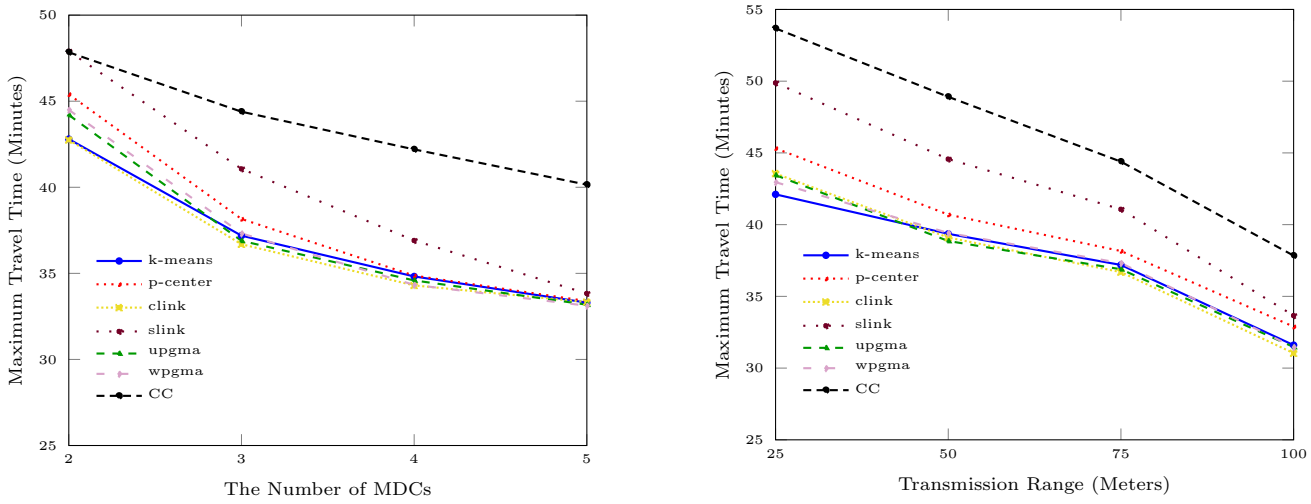
k-means, *clink*, *slink*, *upgma*, and *wpgma* when five MDCs are used. *k-means* follows a similar pattern with *p-center*. Despite its initial poor performance, *p-center* performs relatively better when the MDC count is four or more. *clink*, *slink*, *upgma*, and *wpgma* perform similar. For five MDCs, *clink* and *slink* provide the worst results.

Figure 5b presents the energy consumption results for Scenario II when R is varied. It can be noticed from Figure 5b that *CC* provides the best results while *p-center* performs the worst among the considered clustering schemes. For all solutions, the energy consumption declines when R is increased. The inverse relationship between R and the energy consumption can be attributed to the increased coverage disk size. Since the radius of the coverage disk is equal to R, the degrees of disk overlaps are likely to increase when R is extended. For *CC*, the energy consumption declines 21% when R is increased from two to five. For *k-means* and *p-center* the decline reaches 24% and 26% respectively.

Compared to *p-center*, *CC* reduces the energy consumption up to 20%. *k-means* performs better than *slink* when R is 25. However, *slink* incurs less energy than *k-means* when R is more than 25. *clink*, *slink*, *upgma*, and *wpgma* perform similar. On the other hand, *slink* outperforms the other three approaches when R is 100. Contrary, despite its initial performance, *clink* performs worse than these three solutions when R is 100.

Figure 6a reveals how the maximum travel time changes for Scenario II with varying number of MDCs. Similar to Scenario I, the maximum travel time declines when there are more MDCs in the network. To understand the cost decline, let us consider the edge case where there are $p+1$ DCPs and $m = p$. In this case, each DCP, but V_{BS}, is assigned to an exclusive MDC. The designated MDC tours include the respective DCP and V_{BS}. If m is fewer than p and reduced even further, the resulting tours cannot be shorter than the ones in the edge case. Therefore, the decline in the maximum travel time is expected when m is increased in Scenario II as well. For *k-means*, the maximum travel time declines 23% when the number of MDCs is increased from two to five. For *slink*, which performs initially worse, the cost decline reaches 29% for the same case.

clink and *k-means* perform similar but *clink* provides decreased travel times when m is less than five. Despite its superior performance in energy consumption, *CC* leads to higher travel times. Compared to *clink*, *CC* increases the maximum travel time up to 20%. *upgma* and *wpgma* perform almost similar and provide



(a) Maximum travel time with respect to m.

(b) Maximum travel time with respect to R.

Figure 6: Maximum travel time with varying number of MDCs and the transmission range.

better results than *slink*.

Compared to the maximum travel time results in Scenario I (Figure 4a), one of the notable outcomes of applying Scenario II is the common elevation in maximum travel times for all approaches. For *k-means*, the maximum travel time increases from 34.3 minutes to 42.8 minutes (25% increase) for two MDCs. When m is five, the cost increase reaches 184% (from 11.7 minutes to 33.3 minutes). The higher maximum travel times in Scenario II, can be attributed to the extended tours after including V_{BS} in clusters.

Figure 6b presents the maximum travel times for Scenario II when R is varied. Figure 6b demonstrates that the maximum travel time declines for all approaches when R is extended. This is expected since R also denotes the radius of the coverage disks and extended R indicates larger coverage disks with a higher chance of disk overlaps. Consequently, the average number of DCPs assigned to clusters is expected to decline. Since m is constant, fewer DCPs results shorter travel times.

For *k-means*, the maximum travel time declines 24% when m is increased from two to five. For *CC*, the decline reaches 30% for the same case. *k-means* performs the best when m is two. However, *clink* and *upgma* outperform *k-means* when m is more than two. *p-center* provides better results than *slink* and *CC*. *CC* results higher travel times compared to other approaches.

6. Conclusion

WSNs are expected to operate unattended in a self-configuring manner over extended periods. However, the network can be subject to partitioning due to initial topology or random node failures. One of the possible solutions that can be applied to restore connectivity is employing MDCs. The main task of the MDCs is to traverse partitions to collect their data and then deliver to the *BS*. Considering the limited onboard batteries of the MDCs and the excessive energy cost of mobility, it is desired to minimize the energy consumption of mobility. Given the availability of multiple MDCs, a certain level of load balancing should be provided as well. This paper follows a three-step approach to designate MDC tours. Initially, the DCPs are identified according to Steiner zones defined by the overlapping coverage disks. Subsequently, identified data collection points are grouped as many as the number of MDCs using various clustering methods. In this step, one MDC

is assigned to each cluster and MDC tours are designated using the TSP. The final step improves the obtained MDC tours by converting discrete DCPs into corresponding continuous Steiner zones. This paper considers two different scenarios for the BS-MDC connectivity depending on the wireless technology available for the MDCs. One of the scenarios assumes an LPWA network between the MDCs and the BS. In this scenario, the MDCs are assumed to be able to deliver their data to the BS anywhere in the network. The other scenario dictates the MDCs to visit the BS for communication. It is shown that existing clustering methods are not effective in terms of energy consumption for the second scenario and a novel clustering algorithm, *The Closest Centers (CC)*, is suggested to minimize the energy cost of mobility. The performance of the proposed solution is validated through simulations. Obtained results indicate that CC reduces the energy consumption at the expense of increased maximum travel time.

Acknowledgment

This work was supported by the Scientific and Technical Research Council of Turkey (TUBITAK) under Grant No. EEEAG-117E050.

References

[1] Shaikh FK, Zeadally S. Energy harvesting in wireless sensor networks: A comprehensive review. Renewable and Sustainable Energy Reviews 2016; 55: 1041-1054.

[2] Younis M, Senturk IF, Akkaya K, Lee S, Senel, F. Topology management techniques for tolerating node failures in wireless sensor networks: A survey. Computer Networks 2014, 58: 254-283.

[3] Ren Y, Wang T, Zhang S, Zhang J. An intelligent big data collection technology based on micro mobile data centers for crowdsensing vehicular sensor network. Personal and Ubiquitous Computing 2020: 1-17.

[4] Ghosh N, Banerjee I, Sherratt RS. On-demand fuzzy clustering and ant-colony optimisation based mobile data collection in wireless sensor network. Wireless Networks 2019, 25(4): 1829-1845.

[5] Zhan C, Zeng Y, Zhang R. Energy-efficient data collection in UAV enabled wireless sensor network. IEEE Wireless Communications Letters 2017, 7(3): 328-331.

[6] Korkmaz RB, Kizilirmak K, Senturk IF. NetCar: A testbed for mobile sensor networks. In: 2nd International Conference on Advanced Technologies, Computer Engineering and Science; Alanya, Turkey; 2019. pp. 334–338.

[7] Senturk IF. A steiner zone approach for mobile data collection in partitioned wireless sensor networks. International Journal of Informatics Technologies 2020, 13(3): 217–224.

[8] Papadimitriou CH. The Euclidean travelling salesman problem is NP-complete. Theoretical Computer Science 1977, 4(3): 237-244.

[9] Bari A, Chen Y, Roy D, Jaekel A, Bandyopadhyay S. Designing hierarchical sensor networks with mobile data collectors. Pervasive and Mobile Computing 2011, 7(1): 128-139.

[10] Wu FJ, Tseng YC. Energy-conserving data gathering by mobile mules in a spatially separated wireless sensor network. Wireless Communications and Mobile Computing 2013, 13(15): 1369-1385.

[11] Ghosh N, Banerjee I. An energy-efficient path determination strategy for mobile data collectors in wireless sensor network. Computers & Electrical Engineering 2015, 48: 417-435.

[12] Kang Z, Zeng H, Hu H, Xiong Q, Xu G. Multi-objective optimized connectivity restoring of disjoint segments using mobile data collectors in wireless sensor network. EURASIP Journal on Wireless Communications and Networking 2017, 2017(1): 1-12.

[13] Wu Q, Zeng Y, Zhang R. Joint trajectory and communication design for multi-UAV enabled wireless networks. IEEE Transactions on Wireless Communications 2018, 17(3): 2109-2121.

[14] Luo C, Wu L, Chen W, Wang Y, Li D, Wu W. Trajectory optimization of UAV for efficient data collection from wireless sensor networks. In: International Conference on Algorithmic Applications in Management; 2019. pp. 223–235.

[15] Roque G, Padilla VS. LPWAN based IoT surveillance system for outdoor fire detection. IEEE Access 2020, 8: 114900-114909.

[16] Mekki K, Bajic E, Chaxel F, Meyer F. A comparative study of LPWAN technologies for large-scale IoT deployment. ICT express 2019, 5(1): 1-7.

[17] Wixted AJ, Kinnaird P, Larijani H, Tait A, Ahmadinia A, et al. Evaluation of LoRa and LoRaWAN for wireless sensor networks. In: 2016 IEEE SENSORS; 2016. pp. 1-3.

[18] Shahraki A, Taherkordi A, Haugen Ø, Eliassen F. Clustering objectives in wireless sensor networks: A survey and research direction analysis. Computer Networks 2020, 180: 107376.

[19] Heinzelman WR, Chandrakasan A, Balakrishnan H. Energy-efficient communication protocol for wireless microsensor networks. In: 33rd annual Hawaii international conference on system sciences; 2000. pp. 10.

[20] Younis O, Fahmy S. Distributed clustering in ad-hoc sensor networks: A hybrid, energy-efficient approach. In: IEEE INFOCOM 2004; 2004.

[21] Banerjee S, Khuller S. A clustering scheme for hierarchical control in multi-hop wireless networks. In: Twentieth Annual Joint Conference of the IEEE Computer and Communications Society; 2001. pp. 1028-1037.

[22] Clark BN, Charles JC, Johnson DS. Unit disk graphs. Discrete Mathematics 1990, 86(1): 165–177.

[23] Gulczynski DJ, Heath JW, Price CC. The close enough traveling salesman problem: A discussion of several heuristics. In Perspectives in Operations Research: 271-283. Boston, MA, USA: Springer, 2006.

[24] Sharma U, Krishna CR, Sharma TP. An efficient mobile data collector based data aggregation scheme for wireless sensor networks. In: 2015 IEEE International Conference on Computational Intelligence & Communication Technology; 2015. pp. 292-298.

[25] Han J, Lee JG, Kamber M. An overview of clustering methods in geographic data analysis. Geographic Data Mining and Knowledge Discovery 2009, 2: 149-170.

[26] Google. Traveling Salesman Problem, Example: Solving a TSP with OR-Tools, https://developers.google.com/optimization/routing/tsp, Accessed: 06/07/2021.

[27] Robot and Sensor Networks Lab. Network topologies, SZP-20, http://ceng.btu.edu.tr/rosenet/software.html, Accessed: 06/07/2021.

[28] Kumar AK, Sivalingam KM, Kumar A. On reducing delay in mobile data collection based wireless sensor networks. Wireless networks 2013, 19(3): 285-299.

[29] Allam AH, Taha M, Zayed HH. Enhanced Zone-Based Energy Aware Data Collection Protocol for WSNs (E-ZEAL). Journal of King Saud University-Computer and Information Sciences 2019.