Article

Enhancement of Advanced Metering Infrastructure Performance Using Unsupervised K-Means Clustering Algorithm

Daisy Nkele Molokomme ^{1,*} , Chabalala S. Chabalala ² and Pitshou N. Bokoro ¹

¹ Department of Electrical and Electronic Engineering Technology, University of Johannesburg, Johannesburg 2028, South Africa; pitshoub@uj.ac.za

² School of Electrical and Information Engineering, University of the Witwatersrand, Johannesburg 2050, South Africa; chabalala.chabalala@wits.ac.za

* Correspondence: 201132400@student.uj.ac.za

Abstract: Data aggregation may be considered as the technique through which streams of data gathered from Smart Meters (SMs) can be processed and transmitted to a Utility Control Center (UCC) in a reliable and cost-efficient manner without compromising the Quality of Service (QoS) requirements. In a typical Smart Grid (SG) paradigm, the UCC is usually located far away from the consumers (SMs), which has led to a degradation in network performance. Although the data aggregation technique has been recognized as a favorable solution to optimize the network performance of the SG, the underlying issue to date is to determine the optimal locations for the Data Aggregation Points (DAPs), where network coverage and full connectivity for all SMs deployed within the network are achieved. In addition, the main concern of the aggregation technique is to minimize transmission and computational costs. In this sense, the number of DAPs deployed should be as minimal as possible while satisfying the QoS requirements of the SG. This paper presents a Neighborhood Area Network (NAN) placement scheme based on the unsupervised K-means clustering algorithm with silhouette index method to determine the efficient number of DAPs required under different SM densities and find the best locations for the deployment of DAPs. Poisson Point Process (PPP) has been deployed to model the locations of the SMs. The simulation results presented in this paper indicate that the NAN placement scheme based on the ageless unsupervised K-means clustering algorithm not only improves the accuracy in determining the number of DAPs required and their locations but may also improve the network performance significantly in terms of network coverage and full connectivity.

Citation: Molokomme, D.N.; Chabalala, C.S.; Bokoro, P.N. Enhancement of Advanced Metering Infrastructure Performance Using Unsupervised K-Means Clustering Algorithm. *Energies* **2021**, *14*, 2732. <https://doi.org/10.3390/en14092732>

Academic Editor: André Madureira

Received: 9 March 2021

Accepted: 2 April 2021

Published: 10 May 2021

Keywords: advanced metering infrastructure (AMI); data aggregation points (DAPs); neighborhood area network (NAN); smart meters (SMs); poisson point process (PPP); unsupervised K-means clustering

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The evolution of advanced Information and Communication Technology (ICT) has been a great motivation behind the improvement of operational infrastructure in a wide variety of business domains. These include education, healthcare, transportation, and energy domains. As a result, the concept of “smart cities” has gained significant attention as an interesting field of study in academia and research industry [1]. Generally, the concept of smart cities has been used extensively as an umbrella term for the incorporation of advanced ICT in these business domains with the aim of enhancing economy and Quality of Experience (QoE) for all the involved stakeholders (e.g., utility providers and end-users). The main objective of this paper is particularly on the changes that ICT has brought into the energy sector, especially the electricity grid, in terms of reliability and efficiency from a communication perspective. With the integration of ICT into power grid, the traditional power grid that has revolutionized daily lives of consumers since its inception has been

transitioned into a “Smart Grid (SG)” paradigm. This paradigm entails the integration of advanced ICT, computer intelligence, and control optimization methods throughout the network domains from power generation to transportation up to end-users [2]. Compared to the traditional power grid, the SG paradigm will have intelligent and improved communication networks. For instance, SG promises to provide frequent and timely two-way interaction between the end-users and Utility Control Center (UCC), as well as remote monitoring to a vast majority of power grid [3,4]. The terms end-users and consumers will be used interchangeably in this paper.

As ICT gains momentum in the area of power grids, the existing obsolete communication infrastructure is prone to major challenges such as efficiency and security, among others [5]. Complicating the issues further is the exponential increase in the demand for wireless applications and services driven by Internet of Things (IoT) smart devices which yields huge data volume. Since the existing communication infrastructure was not designed for 21st Century challenges, it is confronted with the major issues of managing, processing, and storing such amounts of generated data. Fundamentally, this has changed the data traffic of wireless networks [6]. Considering the interest in the innovation of SG in both the developing and developed countries, recent works in the literature have devoted their attention to designing and planning the network architecture that supports the level of Quality of Service (QoS) requirements emerging with diverse SG applications [4]. Despite the SG’s promising benefits, power grids have always been known by their complex design nature [7]. However, the path towards the realization of SG as an effective solution to these issues seems promising since it is still an ongoing innovation. Contributing to the benefits emerging with SG deployment is the Advanced Metering Infrastructure (AMI), which is perceived as an initial step towards SG innovation. An approach to exploit the limited data from AMI is presented in [8]. The proposed approach is concerned with mitigating the underlying limitations such as communication, storage, and computing power.

Therefore, this paper aims to evaluate the importance of AMI in SG deployment by proposing a placement Neighborhood Area Network (NAN) scheme based on the combination of unsupervised K-means clustering algorithm and Poisson Point Process (PPP). A typical AMI, as depicted in Figure 1, consists of large-scale deployment of Smart Meters (SMs), sensors, smart IoT devices, and two-way communication, among others.

Figure 1. Network Topology of Advanced Metering Infrastructure (AMI).

In this configuration, the SMs are used as an essential component to collect the power consumption data from all smart electrical appliances within the premises and allow frequent monitoring of the power grid status in real-time using sensors. In this manner, UCC may be able to control and optimize the load control commands, which can lead to effective decision-making. The Data Aggregation Points (DAPs) distributed within NANs are employed to serve as intermediary nodes between the SMs and UCC as well as enabling effective and reliable communication. For this reason, the efficient number of DAPs employed as well as their optimal locations have a great impact on the given QoS communication requirements for each SM within the network and the overall

network performance of NAN. This may be in terms of coverage probability, latency, packet error probability, and outage probability, among others [4,5]. Considering the limited communication capacity of each DAP within the network, NAN may be confronted with huge data traffic, which may result in packet losses and latency if not managed in a timely manner. The data traffic in SG may be classified into fixed scheduling and event-driven data traffic depending on the type of data transmitted. For ease of our explanation, each DAP within NAN has a limited maximum number of SMs that it can be associated with for communication purpose, which needs to be taken into consideration during network planning in SGs. Furthermore, the limited bandwidth within NAN may contribute to design issues as it is not guaranteed to support the level of QoS requirements from SMs [5].

Several methods have been proposed in the literature to mitigate the limited resource issues in large-scale networks, such as the allocation model and scheduling algorithms, among others [9]. Therefore, this paper aims to compliment the existing works in the literature by proposing a hybrid NAN scheme without compromising the given QoS requirements of SMs in terms of coverage probability and connectivity. Given the large-scale random distribution of SMs within NAN, our proposed scheme will consider an effective heuristic model with the capability of grouping SMs into sets of clusters to mitigate the limited capacity of DAPs and minimize the transmission costs within NAN. In addition, the scheme uses the silhouette index method to measure how similar the SMs assigned to one DAP are to one another in terms of closeness and also how dissimilar they are to the SMs assigned to the neighboring DAPs (compactness). This may result in minimized data traffic and energy consumption within the network. In a nutshell, the problem in this paper has been formulated as the optimization problem, where the efficient number of DAPs required for a given SM's density is determined and also their best locations in which the coverage probability and connectivity for SMs is satisfied [9].

Unlike the existing works in the literature, a unique approach to solve the constraints of power independence, power robustness, and communication robustness was presented in [10]. The optimization problem that has resulted as an NP-hard problem has been formulated with the aim of minimizing the installation costs of DAPs with QoS requirements taken into consideration. In [11], a similar approach has been taken with the optimization problem formulated as an integer problem. In addition, the K-means clustering algorithm has been adopted to increase the effectiveness and minimize the complexity of mitigating the DAPs placement problem. In [12], to minimize computational complexities in solving DAPs placement problem such as the Set Covering Problem (SCV), the proposed model divides the major problem into set of subproblems that are optimally solved. The proposed model is called Memory Oriented Split using K-means with Post Optimization Unification (MOSKOU) [12]. The optimization of DAPs placement problem has been extensively studied in the literature [13,14]. In [13], authors approached the DAPs placement problem from the angle of mitigating the latency and overhead issues in AMI. Both the wireless and wired communication technologies have been incorporated into the infrastructure to enhance the information exchange between consumers and microgrid. Therefore, following the approach taken in the aforementioned works, we also propose to solve the DAPs placement problem as a clustering problem. This paper considers, an effective model based on the K-means clustering algorithm to subdivide SMs into groups where the minimum distance is taken as a similarity metric. As such, we propose to simulate the effectiveness of selected DAP locations using the coverage probability as the performance metric. Compared to the prior works in the literature, the scheme proposed in this paper takes the randomness of SM locations into consideration using PPP. In prior works, K-means clustering algorithm has indicated satisfactory and reliable results in solving the DAPs placement problem. The clustering, also referred to as cluster analysis, has been applied in a wide variety of application domains such as business intelligence, psychology and social science, information retrieval, pattern classification, bioinformatics, localization, and service segmentation, among others [15–17].

The main aim of clustering in its application is to identify groups of similar objects referred to as clusters and, in return, help to discover the distribution of patterns and interesting correlations in large data sets [18]. As such, clustering may be classified into different types of algorithms, namely, prototype-based, density-based, graph-based, hybrid, and algorithm-independent methods [19]. Usually, these clustering methods are distinguished by their unique structural designs. A clear distinction among these methods has been extensively studied in [15]. Despite their unique benefits in their applications, prototype-based clustering algorithms are the most adopted technique, particularly K-means clustering [20,21]. In recent works, the K-means clustering algorithm has been utilized as a benchmark method in a wide variety of both unsupervised and supervised learning algorithms to improve the robustness and effectiveness of the model [19]. Despite the variety of works based on K-means clustering algorithm in the literature, the work by Sinaga and Yang [20] caught our attention as an appropriate model to address our DAPs placement problem formulated as clustering problem [20]. The proposed algorithm in this paper considers the Expectation–Maximization (EM) algorithm to address the initialization problem in the traditional K-means clustering algorithm, since it is very sensitive to such a problem. Unlike the works in the literature that rely on knowing the number of clusters beforehand, the proposed algorithm automatically finds the number of clusters by incorporating the entropy method into their proposed model. A detailed description of the entropy method has been highlighted in [15]. The precise locations of SMs in high-dimensional environments such as SG NANs are usually unknown. Therefore, the attention in modeling the locations of SMs prior to clustering has been diverted to stochastic geometry tools such as PPP [22]. In addition, this has been driven by the increase in the randomness and irregularity of SM locations in NANs. The key contributions of this paper are outlined as follows:

- A self-sustainable placement NAN scheme based on unsupervised K-means clustering is proposed to determine the efficient number of DAPs for a given SM density [23,24].
- To address the randomness of SM locations, the SMs are assumed to be deployed according to PPP.
- The silhouette index method is deployed to measure the accuracy of the proposed clustered scheme. This method measures how accurate the membership of each SM with respect to their associated DAP is compared to that of the neighboring DAPs.
- The network coverage and connectivity are used to evaluate the performance of each SM within the network under the corresponding DAP.

The rest of this paper is organized as follows: Section 2 presents the review on the existing and recent works in relation to mitigating the DAPs placement problem. In Section 3, the system model proposed in this work is extensively discussed. The problem formulation based on unsupervised K-means clustering algorithm along with the summary of the algorithm is presented in Section 4. In Section 5, the simulation of the system model and performance evaluation is presented. The results achieved in Section 5 are further discussed in Section 6. Finally, the paper is concluded in Section 7.

2. Related Works

There are several approaches proposed in the literature in the effort to solve the DAPs placement problem in wireless networks [5,10–14]. That is, modeling the optimal locations of the DAPs within the network, with the aim of ensuring that all the QoS requirements of the SMs are satisfied. The location modeling for DAPs is of key importance for the improvement of NAN performance in terms of coverage optimization and minimized QoS requirements such as communication delay, packet losses, and many more [25]. In the literature, [5] developed an analytical model to quantify the QoS metrics of SMs distributed within the network with the minimum number of concentrators deployed. In addition, the authors proposed mitigating the limited bandwidth issues by developing a channel allocation between the groups of SMs headed by the serving concentrators. The terms DAPs, transmitters, and concentrators are used interchangeably in the literature.

In communication networks with many channels such as wireless networks (e.g., cellular, fiber optics), the stochastic geometry approach has been suggested in prior works as an effective solution to address the randomness among distributed nodes [26]. The most significant challenge in NAN is to ensure an effective interoperability among the smart IoT devices without excessive costs for the communication infrastructure [27]. In this manner, the high cost of communication with respect to computation cost in wireless networks necessitates the importance of aggregation [28]. A cost-minimization DAP placement (CMDP) as a constrained optimization problem has been formulated to minimize the cost of DAPs placement in NAN [10], while an extensive investigation based on the DAPs placement problem in AMI was conducted in [11]. The simulation results achieved in this work have highlighted the impact of selected DAP locations on SM density at a given communication range.

Among the issues arising in cellular network analysis is modeling the maximum number of users to be associated with a relevant base station (BS) [29]. Unlike the works in the literature, the maximum capacity of SMs each DAP may be able to accommodate has been considered in [10,11]. Due to the evolution of machine learning in various business domains, clustering has been realized as an effective tool to address the localization problem in wireless networks [17]. In [14], an investigation on the application of partition-based clustering algorithms—namely, K-means, Fuzzy C-means (FCM), and Self-Organizing Map (SOM)—has been conducted. In addition, multihop shortest path distance (MSPD) has been used as a similarity metric to evaluate the performance of these algorithms. Considering the large-scale deployment of smart IoT devices in SG, the sensor placement problem has been mitigated by FCM with the effect of extracting a high-dimension dataset from the network with minimum communication delay [30]. As such, the fundamental concept of FCM is similar to that of traditional K-means clustering, which is concerned with clustering or grouping datasets with huge similarity but also huge dissimilarity with datasets belonging in other clusters.

Despite the benefits emerging with the K-means clustering algorithm, initialization of numbers beforehand is a challenging issue, especially in high-dimensional datasets [31]. An alternative approach to mitigating the initialization of traditional K-means clustering has been addressed in [20]. A novel unsupervised K-means clustering algorithm has been developed to automatically find the cluster number based on the insight discovered within the datasets. The EM algorithm, which is very sensitive to initialization, has been incorporated in parallel with the well-known and oldest K-means clustering algorithm. One distinguishing factor of this proposed novel scheme by Sinaga and Yang [20] from the existing works in the literature is its ability to automatically find the cluster number. The works such as [16] have adopted the optimized hybrid approach of K-means to cluster the learners with the aim of improving the education system.

3. System Model

The communication model of AMI is considered in this paper as depicted in Figure 1, where each customer's premises consist of vast amount of smart electrical appliances embedded within them. Each premises is assumed to have an SM installed that is used to collect the data from smart electrical appliances periodically or timely depending on the latency sensitivity of the data collected. For example, power consumption data may be transmitted periodically due to its latency tolerance, whereas monitoring and controlling data may be transmitted in a timely manner. Since the UCC is usually located far away from the customer's premises, the communication system may not be able to satisfy the QoS requirements from SMs, which may yield degradation in network performance. In this sense, the DAPs are deployed within NAN to serve as the communication gateways between the premises and remote control center to allow reliable data transmission.

In this paper, we consider the two-dimensional geographical area A , which consists of a random distribution of SMs and DAPs. Due to the large-scale environment of a typical AMI, the precise locations of the SMs are unknown beforehand. In this manner, the set

of $\mathcal{D}$ DAPs and $\mathcal{S}$ SMs are considered to be modeled according to PPP with intensities of β_d and β_s , respectively. The typical distribution of SMs according to PPP is depicted in Figure 2. Each SM within the network is assumed to be associated with the DAP whose signal strength is the highest to achieve the desired network coverage. In this sense, we consider the location of DAP $d \in \mathcal{D}$ to be denoted by a_k .

Figure 2. An example of the random distribution of Smart Meters (SMs) according to the Poisson Point Process (PPP).

For simplicity, the DAPs deployed within this network are assumed to be communicating at the same time with a transmitting power of P . For each $d \in \Phi_d$, we consider an SM located at $x_i \in \Phi_s$ that forms a typical communication pair. In this sense, the received signal-to-interference-plus-ratio (SINR) by the SM, located at distance r from the corresponding DAP, is given by

$$\text{SINR}_x = \frac{P\sigma_a\ell(r)^{-\alpha}}{1 + \sum_{a \in \Phi \setminus x} P\sigma_a\ell(r)^{-\alpha}}. \quad (1)$$

The distance r may be computed as the Euclidean distance denoted by $r = \|x_i - y_i\|$, where σ_a and $\ell(\cdot)$ denotes the channel gain and the distance-dependent path loss function with the path loss exponent given by $\alpha > 2$, respectively. The channel gain experienced by the SM located at y from its corresponding DAP at x is assumed to follow Rayleigh fading with zero mean and unit additive white Gaussian noise (AGWN) variance [24]. Since the DAPs within the network are assumed to be communicating at the same time, we also consider the interfering signals that may be experienced by the DAP of interest, as mathematically expressed by $\sum_{x \in \Phi_m y} P\sigma_x\ell(r)^{-\alpha}$. The typical link of each communication pair in the network is used to measure the coverage probability of each SM within the network as defined below:

Definition 1. *Coverage probability: The probability that the SM located at y_i has received a signal from the associated DAP located at x_i that is above the given threshold value τ . In other words, the SM installed within the premises may be considered to be covered by the associated DAP, if and only if the received SINR from the DAP is greater than the given signal threshold value ($\text{SINR} > \tau$).*

In this sense, the coverage probability of an SM located at x_i may be modeled according to Definition 1.

4. Problem Formulation

The main objective of this paper is to determine the best locations for DAPs deployment within the given $\mathbb{R}^2$ and concurrently determine the minimum number of DAPs to be deployed for the given SM's density β_m . Given the locations of the SMs represented by $x = \{x_i, \dots, x_n\}$ in Section 3, we take x as an input parameter of our proposed UK-means clustering algorithm borrowed from [20]. The summary of our proposed model is presented in Algorithm 1 below. Unlike the traditional K-means clustering algorithm, which requires the number of clusters beforehand, this algorithm automatically finds the optimal number of clusters required based on the dataset given. Due to the randomness of DAPs within the dimensional space, we consider $a = \{a_k, \dots, a_c\}$ to represent the random locations of the DAPs within the space, whose density is represented by β_d , as highlighted in Section 3. Let the association problem for the i -th SM and its corresponding k -th DAP be formulated as the membership function $z = [z_{ik}]_{n \times c}$, where $z_{ik} \in [0, 1]$ is used as an indicator function and $n \times c$ denotes the membership matrix. In this case, the membership matrix is made up of the total number of SMs distributed on the $\mathbb{R}^2$, denoted by n and their corresponding cluster labels c . Since we have knowledge regarding SM location within the $\mathbb{R}^2$, with a_k denoting the reference location of the serving DAP, we can further express the association problem as follows:

$$z_{ik} = \begin{cases} 1, & \text{if } \|x_i - a_k\|^2 = \min_{1 \leq k \leq c} \|x_i - a_k\|^2 \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where $\|x_i - a_k\|$ denotes the Euclidean distance from the i -th SM to the associated k -th DAP. The constraint of this model is that the SM within the geographical area can only be considered to be associated with the relevant DAP if and only if its signal strength is the highest compared to other DAPs within the same geographical area. Thus, the SM is associated to the nearest DAP. To avoid the allocation of SMs to DAPs more than once, we use the indicator function $-\ln a_k$ to represent the probability of having an SM already assigned to the relevant DAP—thus, ensuring that the SM is assigned to the DAP with the highest signal strength over the neighboring DAPs on network. We further adopt the QoS requirements in terms of successful probability for information transmission with respect to network coverage to analyze the performance of the associated DAP, as defined in Section 3.

Due to the large-scale environment of NAN, the most efficient number of DAPs is unknown; we consider any node within the dimensional space to represent the random location of the DAP a_k to serve as the reference point. In this context, we can consider a_k as the initial location of the DAP modeled as the Poisson point [24]. This can also be derived using the d-variate model $\sum_{k=1}^c a_k$ as the average SM location information from the geographical area. This is termed as the entropy according to [32]. Therefore, the objective function of this algorithm, which is concerned with obtaining a minimum distance between the SMs and serving DAP can be formulated as

$$J(z, a) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^c a_k \ln a_k, \quad (3)$$

where $\beta > 0$ is the constant and the membership for each SM within the network, which is based on the minimum distance as the objective function, is denoted by z_{ik} , as defined in Equation (3). Additionally, the total number of SMs per km^2 and number of clusters is denoted by n and c , respectively. To avoid the initialization of cluster number, we consolidate the EM algorithm into the traditional K-means algorithm, which allows the

computation of the groups of SMs headed by serving DAP belonging within the k -th class. Therefore, the proportions of a_k can be given by

$$a_k = \frac{\sum_{i=1}^n z_{ik} x_i}{\sum_{k=1}^c z_{ik}}. \quad (4)$$

Algorithm 1 Unsupervised K-Means Clustering Algorithm.

- 1: **Input:** Initial locations for SMs x_i and their associating DAPs a_k , given the initial values: $t = 0, \epsilon > 0$.
 - 2: **Output:** Membership for each SM on the network z_{ik}^* and the best locations for DAPs deployment a_k^* .
 - 3: Randomly select K to determine the structure of the given x SMs, which serve as the initial cluster centers (DAPs).
 - 4: Compute the Euclidean distance between the i -th SM and the associated k -th DAP within $\mathbb{R}^2$, $\|x_i - a_k\|$.
 - 5: Compare the distance values obtained in Step (4).
 - 6: Compute the membership (z_{ik}) of all SMs using Steps (4) and (5) until all SMs are assigned to relevant clusters.
 - 7: Divide the SMs into sets of clusters based on the highest membership.
 - 8: Using z_{ik} , compute the initial locations of centroids $a_k^{(t)}$ and the cluster numbers $c^{(t)}$.
 - 9: Update the membership for all SMs ($z_{ik}^{(t+1)}$) after each t iteration until the termination condition of iteration is satisfied.
 - 10: Compute the arithmetic mean for all the SMs belonging within the same cluster.
 - 11: Update the locations for all the centroids ($a_k^{(t+1)}$) given by the results in Step (9).
 - 12: Using Steps (9) and (11), update the cluster number from $c^{(t)}$ to $c^{(t+1)}$, discarding the neighboring clusters whose distance is very minimum.
 - 13: Finally, determine a_k^* and z_{ik}^* by comparing the difference between the initial number of clusters in Step (8) and the final results obtained in Step (11).
 - 14: **if then** $\max_{2 \leq k \leq c^{(t)}} \|a_k^{(t+1)} - a_k^{(t)}\| < \epsilon$, STOP
 - 15: **end if**
 - 16: Increment the iteration until convergence, $t = t + 1$.
-

The introduced entropy is concerned with the maximization likelihood of having an SM assigned to the relevant DAPs based on the minimum distance. As such, the maximization of $-\sum_{k=1}^c a_k \ln a_k$ can be equivalent to the minimization of $\sum_{k=1}^c a_k \ln a_k$, thus, we adopt $\sum_{k=1}^c a_k \ln a_k$ as our next entropy term. This yields to the objective function of the UK-means clustering expressed as

$$J(z, a) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^c a_k \ln a_k - \gamma \sum_{i=1}^n \sum_{k=1}^c z_{ik} \ln a_k, \quad (5)$$

where β and γ can be used as the learning parameters of the proposed algorithm. Based on the objective of our proposed scheme, which is concerned with determining the candidate locations of the DAPs at the minimum QoS from SMs, we can use these two parameters to learn the data traffic within the network, which is proportional to the QoS requirements. By reducing the packet delay, which can be achieved at minimum distance, we can use β in this instance, whereas γ can be used to ensure the successful transmission probability of the data packet sent within the obtained minimum distance at minimal to zero packet error probability. Therefore, the Lagrangian of (4) with constraint of $\sum_{k=1}^c a_k = 1$ can be formulated as

$$\tilde{J}(z, a, \lambda) = \sum_{i=1}^n \sum_{k=1}^c z_{ik} \|x_i - a_k\|^2 - \beta n \sum_{k=1}^c a_k \ln a_k - \gamma \sum_{i=1}^n \sum_{k=1}^c z_{ik} \ln a_k - \lambda \left(\sum_{k=1}^c a_k - 1 \right). \quad (6)$$

With all the SMs within the dimensional space assigned to their relevant clusters headed by centroid, we can now calculate the arithmetic mean of all the SMs clustered around each centroid. Therefore, the candidate location of the DAP along with the new membership of the SM after each iteration can be formulated by taking the partial derivative of Equation (6) with respect to λ , z_{ik} , and a_k . Firstly, we let the derivative to be equated to zero:

$$\begin{aligned} \frac{\partial \tilde{f}}{\partial a_k} &= -\beta n (\ln a_k + 1) - \gamma \sum_{i=1}^n \frac{z_{ik}}{a_k} - \lambda = 0 \\ \text{and } -\beta n a_k (\ln a_k + 1) - \gamma \sum_{i=1}^n z_{ik} - \lambda a_k &= 0 \end{aligned} \quad (7)$$

This yields

$$-\sum_{k=1}^c \beta a_k \ln a_k - \sum_{k=1}^c n \beta a_k - \gamma \sum_{k=1}^c \sum_{i=1}^n z_{ik} - \sum_{k=1}^c \lambda a_k = 0, \quad (8)$$

where the parameter λ can be given as

$$\lambda = -\beta n \sum_{k=1}^c a_k \ln a_k - \beta n - n\gamma. \quad (9)$$

Substituting Equation (9) into Equation (7), the left-hand side of Equation (7) results in

$$-\beta n a_k (\ln a_k + 1) - \gamma \sum_{i=1}^n z_{ik} - \left(\beta n \sum_{k=1}^c a_k \ln a_k - \beta n - n\gamma \right) a_k = 0. \quad (10)$$

With the candidate locations of the DAPs given by the a_k , we can now update its best location by computing the arithmetic mean of all the SMs surrounding the DAP denoted by z_{ik} . Therefore, the updated location of the k -th DAP may be expressed as

$$a_k^{(t+1)} = \sum_{i=1}^n z_{ik} \ln a_k + (\beta \setminus \gamma) a_k^{(t)} \left(\ln a_k^t - \sum_{s=1}^c a_s^{(t)} \ln a_s^{(t)} \right). \quad (11)$$

To avoid assigning the SMs to the DAP with the lowest signal strength, which is located far away, the SMs are further subdivided into clusters to ensure that each cluster consists of the SMs with the highest membership. This implies that the SMs belonging within each cluster should be as close as possible. Therefore, the updated membership of each SM after each iteration can be given as

$$z_{ik} = \begin{cases} 1, & \text{if } \|x_i - a_k\|^2 = \min_{1 \leq k \leq c} \|x_i - a_k\|^2 - \lambda \ln a_k \\ 0, & \text{otherwise} \end{cases}. \quad (12)$$

Based on the updated equations formulated in Equations (11) and (12), representing the membership of all SMs within the given area and the candidate locations of the DAPs, we can now estimate the desired number of clusters for the given SM density β_m as follows:

$$c^{(t+1)} = c^{(t)} - |\{a_k^{(t+1)} | a_k^{(t+1)} < 1/n, k = 1, \dots, c^{(t)}\}|, \quad (13)$$

where t denotes the iteration number and $|\{\cdot\}|$ represents the maximum capacity of SMs that the DAP has been assigned after each iteration. When the assignment of SMs to the relevant DAPs along with their candidate locations have converged, we can now denote the best location of the DAP, where the network coverage probability, as highlighted in Definition 1, is satisfied.

5. Simulation and Performance Evaluation

The network coverage is a reliable performance metric that has been used to analyze the performance of wireless networks. Using the network coverage, in this paper, we consider a 2D dataset that consists of two features (latitude and longitude) to represent the locations of the SMs on a given geographical area. In addition, an area of (3.0×3.0) km² is designed as the testing bed for finding the best locations for DAPs deployment, where all the SMs are covered within their transmission range. For the purpose of analysis, the different SM densities per km² have been considered, $n = 500$ and $n = 800$, as depicted in Figure 3.

Figure 3. Random distribution of SMs on a given geographical area: (a) Random distribution of SMs with a given $n = 500$; (b) Random distribution of SMs with a given $n = 800$.

The main objective of this paper is to determine the best locations for DAPs deployment, where the efficient number of DAPs may be required to cover all the SMs within their communication distance. Firstly, an initial value of K as highlighted in Step (3) of the summarized framework has been given $2 \leq K \leq 9$ to determine the initial structure of the given SM's density. In this sense, the initial number of clusters is determined from the initial structure, which also gives the initial locations of the centroids denoted by a_k . The method of the Euclidean distance is employed to compute the distance between the i -th SMs and the corresponding k -th DAP. Figure 4 shows the analytical results when $n = 500$ SMs. From the distance values obtained, all the SMs within the network denoted by the random points are assigned to the nearest cluster. For each cluster, the arithmetic mean is computed to determine the candidate locations for centroids (DAPs), $a_k^{(t)}$. The maximum iteration number of $t = 10$, to iteratively search for the best locations for DAPs deployment by updating the membership of each SM through the membership function $z_{ik}^{(t+1)}$, and also the efficient number of clusters suitable for the given β_m . Figure 4 illustrates the analytical results of $K = 4$ and $K = 9$ for $\beta_m = 500$ SMs/area. The memberships for each SM are distinguished by different colors with the DAP locations and their corresponding labels denoted by the numbers. From the simulation results, it can be seen that when $K = 4$, the SM capacities assigned to each cluster are more or less the same. This implies that the number of SMs assigned to each other are balanced. Whereas, for $K = 9$, the probability of having an SM assigned to the wrong cluster is high as it can be seen on the cluster labels (0, 4, and 5) that there is overlapping of clusters.

Figure 4. SM's density ($n = 500$) vs. efficient number of Data Aggregation Points (DAPs): (a) The determined structure for $n = 500$, when $K = 4$; (b) The determined structure for $n = 500$, when $K = 9$.

However, the results of this will be validated through the silhouette analysis method. Figure 5 illustrates the structure for a given SM density of $n = 800$ when $K = 2$ and $K = 4$, respectively. From the simulation results, both results are satisfactory as there is no overlapping of the clusters and the SM capacities for all the clusters are balanced.

Figure 5. SM's density ($n = 800$) vs. efficient number of DAPs: (a) The determined structure for $n = 800$, when $K = 2$; (b) The determined structure for $n = 800$, when $K = 4$.

To analyze the effectiveness and satisfaction for each SM's density, as illustrated in Figures 4 and 5, in terms of the number of clusters in Equation (13) and the optimal locations for each centroids as defined in Equation (11), the UK-means clustering algorithm with the additional silhouette index is employed. The analytical results, as depicted in Figure 6a,b, show the efficient results when $K = 4$ compared to when $K = 9$. The silhouette index method is concerned with validating the accuracy of the clustering results, i.e., by analyzing the membership of each SM within the assigned cluster with respect to the neighboring clusters. In this sense, the silhouette coefficient is computed to determine how close the i -th SM is to the SMs belonging within the same cluster and how far it is from the SMs belonging within the neighboring clusters. The silhouette coefficient is computed to measure the goodness in the membership of each SM within the network. From the analytical results depicted in Figure 6 of $K = 9$ with the cluster label of 4, there are outliers that show there are SMs that have been wrongly assigned to cluster number 4. In addition, this may degrade the performance of the network, as those SMs may not be covered by their corresponding DAP. As a result, the communication infrastructure leads to the communication delay. In summary, for a 500 SM density, 4 DAPs may be efficient to meet the coverage probability requirement for each SMs within the dimensional space with full connectivity guaranteed.

Figure 6. Silhouette Plot for $n = 500$ SMs: analyze the membership for each SM assigned to its corresponding cluster: (a) Cluster label vs. silhouette index, for $K = 4$; (b) Cluster label vs. silhouette index, for $K = 9$.

In Figure 7, the analytical results for $n = 800$ is illustrated. From the silhouette plot, for $n = 800$, both $K = 2$ and $K = 4$ indicate efficient results. In this sense, we may conclude that for 800 SMs, the number of DAPs required to meet the network coverage probability for each SM may range between $2 \leq K \leq 4$.

Figure 7. Silhouette Plot for $n = 800$ SMs: analyze the membership for each SM assigned to its corresponding cluster: (a) Cluster label vs. silhouette index, for $K = 2$; (b) Cluster label vs. silhouette index, for $K = 4$.

6. Results and Discussion

In this section, the results presented in Section 5 above are discussed. As mentioned in previous chapters, the DAPs placement scheme based on unsupervised K-means clustering algorithm [20] is proposed in this paper. In addition, the criterion to determine the optimal locations for DAPs deployment, where the efficient number of DAPs is required to achieve full connectivity for all the SMs within the network, has been developed. The Poisson process has been employed to model the randomness and irregularity of the SM locations in the given geographical area. To validate the clustering results, the silhouette analysis method is adopted. For each number of SMs distributed on the network, the highlighted steps are followed to determine the best locations for DAPs, where the efficient number of DAPs is required to guarantee network coverage and connectivity. From the visualization point of view, it can be concluded for the 500 distributed SMs, 4 DAPs are required to ensure that all the SMs are covered. Under the same scenario with 800 randomly distributed on the dimensional space, a minimum of 4 is also required. In addition, the K-means clustering indicated reliable results that have been validated through silhouette analysis for the DAPs placement scheme.

7. Conclusions

In this paper, the scheme based on the model based on the “ageless” K-means clustering algorithm has been presented to determine the optimal number of DAPs for different SM densities and on a given geographical area. Consequently, the silhouette index method has been used to evaluate the performance of the developed scheme, thus using the silhouette index score. The silhouette coefficient is computed by measuring how close the SMs within the same cluster are as compared to the SMs belonging within the neighboring clusters. This method is known as one of the most popular and uses the external indices technique, which is concerned with how different the node (SMs) belonging within the cluster of interest is to that of the neighboring clusters. Since the central issue of this paper revolves around the communication effectiveness between the SMs and the DAPs, the network coverage and the information efficiency have also been utilized as the performance metrics of this scheme. To extend the work presented in this paper in future, a simulation model of the packet transmitted from the DAP to its associated SMs can be developed. Further, this model can be used as a cross validation for the clustering results presented in this paper. The successful probability of transmitting the data can be an additional performance metric of the model.

Author Contributions: Conceptualization, D.N.M.; methodology, D.N.M., C.S.C. and P.N.B.; validation, D.N.M., C.S.C., and P.N.B.; formal analysis, D.N.M., C.S.C., and P.N.B.; writing—original draft preparation, D.N.M.; writing—review and editing, C.S.C., and P.N.B.; visualization, D.N.M., C.S.C., and P.N.B.; supervision, C.S.C., and P.N.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: not applicable.

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

References

1. He, Y.; Yu, F. R.; Zhao, N.; Leung, V. C.; Yin, H. Software-defined networks with mobile edge computing and caching for smart cities: A big data deep reinforcement learning approach. *IEEE Commun. Mag.* **2017**, *12*, 31–37. [\[CrossRef\]](#)
2. Molokomme, D.N.; Chabalala, C.S.; Bokoro, P.N. A Review of Cognitive Radio Smart Grid Communication Infrastructure Systems. *Energies* **2020**, *12*, 3245. [\[CrossRef\]](#)
3. Ye, F.; Qian, Y.; Hu, R.Q. *Smart Grid Communication Infrastructures: Big Data, Cloud Computing, and Security*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
4. Molokomme, D.N.; Chabalala, C.S.; Bokoro, P. A survey on information and communications technology infrastructure for smart grids. In Proceedings of the 2019 IEEE 2nd Wireless Africa Conference (WAC), Pretoria, South Africa, 1–6 August 2019.
5. Kong, P.Y. Wireless neighborhood area networks with QoS support for demand response in smart grid. *IEEE Trans. Smart Grid* **2015**, *4*, 1913–1923. [\[CrossRef\]](#)
6. Ali, A.; Yaqoob, I.; Ahmed, E.; Imran, M.; Kwak, K.S.; Ahmad, A.; Hussain, S.A.; Ali, Z. Channel clustering and QoS level identification scheme for multi-channel cognitive radio networks. *IEEE Commun. Mag.* **2018**, *4*, 164–171. [\[CrossRef\]](#)
7. Gellings, C.W. *The Smart Grid: Enabling Energy Efficiency and Demand Response*; CRC Press: Boca Raton, FL, USA, 2020.
8. Della Giustina, D.; Rinaldi, S.; Robustelli, S.; Angioni, A. Massive Generation of Customer Load Profiles for Large Scale State Estimation Deployment: An Approach to Exploiting AMI Limited Data. *Energies* **2021**, *14*, 1277. [\[CrossRef\]](#)
9. Chabalala, C.S.; Takawira, F. Hybrid channel assembling and power allocation for multichannel spectrum sharing wireless networks. In Proceedings of the 2017 IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, USA, 19–22 March 2017.
10. Kong, P.Y. Cost efficient data aggregation point placement with interdependent communication and power networks in smart grid. *IEEE Trans. Smart Grid* **2017**, *1*, 74–83. [\[CrossRef\]](#)
11. Aalamifar, F.; Shirazi, G.N.; Noori, M.; Lampe, L. Cost-efficient data aggregation point placement for advanced metering infrastructure. In Proceedings of the 2014 IEEE International conference on smart grid communications (SmartGridComm), Venice, Italy, 3–6 November 2014; pp. 344–349.
12. Rolim, G.; Passos, D.; Albuquerque, C.; Moraes, I. Moskou: A heuristic for data aggregator positioning in smart grids. *IEEE Trans. Smart Grid* **2017**, *6*, 6206–6213. [\[CrossRef\]](#)

13. Tavasoli, M.; Yaghmaee, M.H.; Mohajerzadeh, A.H. Optimal placement of data aggregators in smart grid on hybrid wireless and wired communication. In Proceedings of the 2016 IEEE Smart Energy Grid Engineering (SEGE), Oshawa, ON, Canada, 21–24 August 2016; pp. 332–336.
14. Hassan, A.; Zhao, Y.; Pu, L.; Wang, G.; Sun, H.; Winter, R.M. Evaluation of clustering algorithms for DAP placement in wireless smart meter network. In Proceedings of the 2017 9th International Conference on Modelling, Identification and Control (ICMIC), Guiyang, China, 10–12 July 2017; pp. 1085–1090.
15. Wu, J. *Advances in K-Means Clustering: A Data Mining Thinking*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
16. Montazer, G.A.; Rezaei, M.S. A new approach in e-learners grouping using hybrid clustering method. In Proceedings of the International Conference on Education and e-Learning Innovations, Sousse, Tunisia, 1–3 July 2012; pp. 1–5.
17. Yang, H.; Xie, X.; Kadoch, M. Machine learning techniques and a case study for intelligent wireless networks. *IEEE Netw.* **2020**, *3*, 208–215. [\[CrossRef\]](#)
18. Halkidi, M.; Batistakis, Y.; Vazirgiannis, M. On clustering validation techniques. *J. Intell. Inf. Syst.* **2001**, *2*, 107–145. [\[CrossRef\]](#)
19. Nasraoui, O.; N’Cir, C.E.B. Clustering Methods for Big Data Analytics. In *Techniques, Toolboxes and Applications*; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; p. 192.
20. Sinaga, K.P.; Yang, M.S. Unsupervised K-means clustering algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [\[CrossRef\]](#)
21. Lee, S.G.; Lee, C. Developing an Improved Fingerprint Positioning Radio Map using the K-Means Clustering Algorithm. In Proceedings of the 2020 International Conference on Information Networking (ICOIN), Barcelona, Spain, 7–10 January 2020; pp. 761–765.
22. Haenggi, M. *Stochastic Geometry for Wireless Networks*; Cambridge University Press: Cambridge, UK, 2012.
23. Błaszczyszyn, B.; Karray, M.K. Spatial distribution of the SINR in Poisson cellular networks with sector antennas. *IEEE Trans. Wirel. Commun.* **2015**, *1*, 581–593. [\[CrossRef\]](#)
24. Wang, Y.; Zhu, Q. Modeling and analysis of small cells based on clustered stochastic geometry. *IEEE Commun. Lett.* **2016**, *3*, 576–579. [\[CrossRef\]](#)
25. Samarasinghe, T.; Inaltekin, H.; Evans, J.S. Optimal SINR-based coverage in poisson cellular networks with power density constraints. In Proceedings of the 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), Las Vegas, NV, USA, 2–5 September 2013; pp. 1–5.
26. Azimi-Abarghouyi, S.M.; Makki, B.; Haenggi, M.; Nasiri-Kenari, M.; Svensson, T. Stochastic geometry modeling and analysis of single-and multi-cluster wireless networks. *IEEE Trans. Commun.* **2018**, *10*, 4981–4996. [\[CrossRef\]](#)
27. Rolim, G.; Passos, D.; Moraes, I.; Albuquerque, C. Modelling the data aggregator positioning problem in smart grids. In Proceedings of the 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool, UK, 26–28 October 2015; pp. 632–639.
28. Kumar, M.; Verma, S.; Singh, P.P. Clustering approach to data aggregation in wireless sensor networks. In Proceedings of the 2008 16th IEEE International Conference on Networks, New Delhi, India, 12–14 December 2008; pp. 1–6.
29. George, G.; Lozano, A.; Haenggi, M. Distribution of the number of users per base station in cellular networks. *IEEE Wirel. Commun. Lett.* **2018**, *2*, 520–523. [\[CrossRef\]](#)
30. Yin, H.; Zhang, Y.; Peng, Z. Optimal sensor placement based on Fuzzy C-means clustering algorithm. In Proceedings of the 2018 International Conference on Sensor Networks and Signal Processing (SNSP), Xi’an, China, 28–31 October 2018; pp. 92–98.
31. Hämmäläinen, J.; Kärkkäinen, T.; Rossi, T. Improving Scalable K-Means++. *Algorithms* **2021**, *14*, 6. [\[CrossRef\]](#)
32. Yang, M.S.; Lai, C.Y.; Lin, C.Y. A robust EM clustering algorithm for Gaussian mixture models. *Pattern Recognit.* **2012**, *11*, 3950–3961. [\[CrossRef\]](#)