IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 21, 2022, accepted April 3, 2022, date of publication April 8, 2022, date of current version April 18, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3165799

Unsupervised Clustering for 5G Network Planning Assisted by Real Data

M. UMAR KHAN[1], MOSTAFA AZIZI[2], A. GARCÍA-ARMADA[3], (Senior Member, IEEE), AND J. J. ESCUDERO-GARZÁS[3,4]

[1]Center for Advanced Studies in Telecommunication (CAST), COMSATS University Islamabad, Islamabad 45550, Pakistan
[2]École Supérieure de Technologie (ESTO), Université Mohammed Premier Oujda (UMP), Oujda 60000, Morocco
[3]Department of Signal Theory and Communications, Universidad Carlos III de Madrid (UC3M), 28911 Leganés, Spain
[4]Galician Research and Development Center in Advanced Telecommunications (GRADIANT), 25971 Vigo, Spain

Corresponding author: M. Umar Khan (umar_khan@comsats.edu.pk)

This work was supported by the Spanish National Project IRENE-EARTH (PID2020-115323RB-C33/AEI/10.13039/501100011033).

ABSTRACT The fifth-generation (5G) of networks is being deployed to provide a wide range of new services and to manage the accelerated traffic load of the existing networks. In the present-day networks, data has become more noteworthy than ever to infer about the traffic load and existing network infrastructure to minimize the cost of new 5G deployments. Identifying the region of highest traffic density in megabyte (MB) per km^2 has an important implication in minimizing the cost per bit for the mobile network operators (MNOs). In this study, we propose a base station (BS) clustering framework based on unsupervised learning to identify the target area known as the highest traffic cluster (HTC) for 5G deployments. We propose a novel approach assisted by real data to determine the appropriate number of clusters k and to identify the HTC. The algorithm, named as *NetClustering*, determines the HTC and appropriate value of k by fulfilling MNO's requirements on the highest traffic density MB/km^2 and the target deployment area in km^2. To compare the appropriate value of k and other performance parameters, we use the Elbow heuristic as a benchmark. The simulation results show that the proposed algorithm fulfills the MNO's requirements on the target deployment area in km^2 and highest traffic density MB/km^2 with significant cost savings and achieves higher network utilization compared to the Elbow heuristic. In brief, the proposed algorithm provides a more meaningful interpretation of the underlying data in the context of clustering performed for network planning.

INDEX TERMS 5G, network planning, machine learning, network clustering, network data acquisition, cluster analysis, elbow method.

I. INTRODUCTION

The 5G and beyond is ideated for the provisioning of use cases defined by 3GPP, from ultra-reliable low latency communication (URLLC) services to enhanced mobile broadband (eMBB) and massive machine type communications (mMTC) [1]. These use cases are offered as services that should be capable to sustain the tight requirements needed by applications like virtual reality, vehicle-to-all (V2X) and mission critical communications. Inevitably, network infrastructures supporting 5G services are being planned. In this respect, network planning assisted by real network data is of utmost importance as the process of identifying the highest traffic region (per km^2) at cluster level, determining the appropriate number of clusters and to provide lower cost per bit. Network planning not only provides the intended coverage and capacity to the subscribers [2], [3] but it is also an effective way to reduce capital and operational expenditure (CAPEX and OPEX, respectively) for the MNOs [4], [5].

A substantial attention has been devoted to network planning for 3G/4G networks [6]–[10], though it is confronted with several challenges for new 5G deployments [11], including the cost-efficiency [12], the service and electromagnetic fields (EMFs) constraints [13], the adoption of numerology and bandwidth part (BWP) using new radio (NR) [14]–[16], identification of highest traffic region in ultra-dense networks (UDNs) [17], [18], machine learning (ML) and data-driven decision making capabilities of 5G networks (see for instance [19]–[22]).

The associate editor coordinating the review of this manuscript and approving it for publication was Lorenzo Mucchi.

Traditionally, network planning utilizes forecast data or estimated traffic demand, to provide extended coverage or capacity enhancements in the target area [23]–[25]. However, the requirements of new 5G use-cases demand data-driven network planning to deliver ultra-low latency, higher data rates, ultra-reliable network deployments [26] with lower cost per bit. The network data acquisition for traffic and infrastructure information of an MNO endows network planning to deploy network with lower cost per bit [27] by identifying the densest traffic region. Therefore, the proposals for 5G network planning should not only explore how to deal with data acquisition but also include data-driven decision making by adopting appropriate ML techniques.

Different studies investigate ML and big data aspects for next generation networks (NGNs) [28]–[30]. The NGNs are expected to be highly complex systems due to heterogeneity in devices, networks, services and application requirements. ML techniques have versatile accomplishments in adapting big data analytics, data-driven decision making, correct parameter estimation and multi-objective optimization problems [31], [32]. Both ML and big data approaches can be applied to NGN scenarios and techniques including massive multiple-input multiple-output (mMIMO), machine to machine (M2M), vehicular ad-hoc networks (VANETs) or internet of things (IoT) [33]. At the same time, ML techniques can play a very important role in bringing new frameworks of data analytics for efficient control, operation, optimization and network planning of 5G and beyond [34], [35]. In particular, clustering techniques have received much interest in the academic community to handle network problems in a variety of settings, like CR [36], mobile ad-hoc networks [37], VANETs [38], [39], wireless sensor networks [40], [41], IoT [42], [43], fog computing alongside with small cells [44], [45] and 5G [46].

In this work, we propose a network planning algorithm to perform network clustering and provide highest traffic cluster (HTC) as a deployment area for new 5G services fulfilling the MNO's requirements of the traffic density and lower cost per bit. We utilize available real mobile data [47] with an unsupervised clustering technique to identify the HTC of minimum cost per bit. This study contains substantial contributions that distinguish it from the existing work on clustering-based network planning. In the first place, real open data are used with the k-means clustering technique to clusterize the MNO's area into k clusters. Second, an algorithm is proposed to identify the HTC and appropriate value of k, based on MNO's requirements which are not previously investigated in the existing literature.

The manuscript is organized as follows. In Section II, we provide the state of the art for clustering techniques adopted in several network problems. Consequently, we contextualize clustering for network planning in Section III. In Section IV, we develop our approach of proposed framework for clustering and analysis to determine appropriate value of k and to identify the HTC. Section V contains a detailed explanation of the developed algorithm while the traditional Elbow method is discussed in Section VI. Finally, we present our results and conclusions in Sections VII and VIII, respectively.

II. RELATED WORK AND CONTRIBUTION

In mobile communication networks, clustering has been mainly applied from two viewpoints, namely to associate users and BSs according to a defined criterium to cluster them. These criteria may range from interference minimization to throughput maximization and spectral efficiency improvement. In this section we compile the most relevant areas for network-related clustering.

From a user-centric perspective, a number of works have addressed throughput maximization [48]–[50]. The authors of [48] study resource allocation to maximize user throughput and the clusters are formed by taking into account the physical distance and social ties between the users, also ensuring fairness among the clusters. In [49], a joint clustering plus scheduling algorithm is proposed to maximize throughput by limiting the cluster size. The cluster size is associated with the increase in the number of users in terms of fairness and throughput degradation. To balance the trade-off between throughput and fairness a dynamic power optimization and user allocation problem is investigated in [50] with a limit of two users per cluster to enhance throughput, resulting in fixed members and a large number of clusters.

Users clustering has also been investigated to improve spectral and energy network efficiency [51]–[53]. In [51], a two-step clustering scheme first divides small cells into disjoint cell clusters according to the neighboring relationship and then the UEs in each cell cluster are further grouped into UE groups with the target of minimizing intra-cluster interference. A final two-step power allocation scheme maximizes the network energy efficiency. A statistical framework has been proposed in [52] to improve both spectral and energy efficiency, being the cluster size sensitive to changes in user and BS densities. In [53], the authors propose users clustering to enhance energy efficiency by lowering signaling overhead. In this work, the user with the best channel quality will communicate with the cellular BS on behalf of the whole cluster to reduce overhead and minimize energy consumption. However, as a large cluster size means high intra-cluster signaling overhead, the cluster size is bounded by the overhead generated inside the cluster. Besides, VANETs offer a suitable scenario to study vehicles clustering to face some challenges that characterize vehicular networks. The clustering of vehicles is explored in [54] based on their moving speeds, the number of hops and road conditions where provisioning of desired data rates is achieved by serving users with different clusters. In [55], the authors propose vehicle clustering up to three hops to get stable clusters in terms of low latency and high packet delivery. By limiting the number of clusters, they overcome the handover problem and achieve better connectivity. In [56], an SDN-based scenario is investigated where the clustering technique is adapted to cluster vehicles based on data acquisition of real-time road conditions. The

results show that the packet delivery significantly improves by forming clusters based on real-time road conditions.

Interference management has long been a popular issue investigated in the context of BS clustering. The authors of [57] propose a clustering strategy to mitigate inter and intra-cluster interference. It is investigated that the higher threshold interference formulates smaller-size clusters and also minimizes the overhead and network latency. In [58], a greedy approach is used to form clusters by randomly selecting an initial BS as a cluster head and adding nearby BSs until the cluster size threshold is reached. The analysis shows lower interference when the cluster sizes are small and a higher cluster size can be seen as a compromise with the inter-cell interference. The optimization technique of BS clustering proposed in [59] makes use of antenna down-tilt to limit the interference to a small number of cells where the BSs that are jointly serving the users are clustered together. The authors conclude to recommend smaller tilt of the antennas residing in the same cluster and large tilt for antennas of the other clusters.

Transversal to clustering areas is how to determine the ideal cluster size or number of clusters, a critical factor for solving a particular network problem. For instance, the problem of congestion due to large scale data communication is addressed in [60]. The appropriate number of clusters is determined by simulation and analyzing the variations in packet delay, packet loss and the number of vehicles in corresponding clusters. Increasing the number of clusters reduces the number of collisions, lowers the packet loss ratio and prevents congestion. The authors of [61] propose a novel scheme to adjust the size and the number of clusters in a large-scale cloud radio access network (C-RAN) where the processing and computational complexity of large scale channel metrics is enhanced by clustering. In [62], a clustering technique is proposed to regulate cluster size such that users need to take permission from BS before joining the cluster. When the cluster size is small, there is insufficient multichannel diversity, which reduces the transmitter's gain. When the cluster size increases, the multichannel diversity improves, and thus enhances the retransmission efficiency.

Foregone in review of the above literature, our question remains unanswered, that is what should be the appropriate size or number of clusters from a network planning perspective. Though some works have addressed clustering-based network planning in the past [63]–[66], they cannot be straightforwardly applied to 5G networks due to the very different requirements. Some previous studies have incorporated clustering to network planning [67]–[70]. The authors of [67] used clustering approach for cell planning such that the signaling overhead is minimized. In [68], a clustering approach is used to automate the 5G network planning by identifying the appropriate geographical locations for BS placements based on service quality targets and regulatory constraints. Another work in [69], proposes a network planning tool based on a clustering algorithm to optimize service quality in order to meet the desired targets. The work of [70] proposes a framework to process input data from several sources and perform learning-based clustering to enhance self-planning, self-healing and self-optimization capabilities of 5G networks. However, traffic density, network utilization, MNO's requirements and cost per bit discussions in the context of BS clustering are not jointly addressed in the present-day literature. In the next subsections, the takeaways from the literature review are provided followed by authors' contributions.

A. TAKEAWAYS

Based on the above discussion, we find that clustering for users and BSs is being conducted on many performance measures like throughput, spectral and energy efficiency, interference, latency and high packet delivery concerning particular use cases. The same performance indicators regulate the cluster size or number of clusters. Traditionally, the clustering techniques have been examined in diverse network problems, though, disclosing insights from real network data such as traffic load and infrastructure information of the BSs is not addressed. To overcome the gaps we consider real network data and new performance metrics to cater the cost minimization and regulate the number of clusters. The new technological knowledge sought in this study is the correlation of clustering, network planning and performance metrics like target area (km^2), traffic density (MB/km^2), cost per MB ($) and network utilization (%). These performance metrics are considered in the context of MNO's financial and technical requirements in the proposed clustering framework.

B. CONTRIBUTIONS

This study proposes a clustering methodology in the context of network planning. In brief, this paper brings the following contributions:

1) We propose the utilization of an unsupervised clustering technique assisted by real network data to determine the 5G deployment area.
2) In contrast to the conventional method, a new methodology is proposed that incorporates the MNO's criteria in computing the appropriate number of clusters denoted by k.
3) We develop and propose a learning based network clustering algorithm to identify the highest traffic density cluster (HTC) which serves as 5G deployment area to offer new services with minimum cost per MB.

III. CLUSTERING FOR NETWORK PLANNING

The advent of 5G in recent years suggests that the new network deployments will be carried out by identifying the densest traffic area per km^2 in order to minimize cost per bit by achieving higher network utilization. We consider the cost per bit metric that is being utilized in several disciplines e.g., electronics, information theory, satellite systems, optical and communication networks to model and analyze the cost associated with the delivery or transfer of data (see [71]–[75]

for details) in order to decide cost-effective solutions. It is an efficient metric to model different parameters and conduct comparisons across MNO's network, technologies and spectrum. The cost per bit values are usually considered as absolute e.g., $15 per MB or GB depending on the considered model and scenario. In this study, we consider delivery of data between user and BS whose associated cost per MB is considered corresponding to the geographical area of cluster (km^2) based on an appropriate value of k, traffic density MB/km^2 and the network utilization (%). The network utilization is based on the MNO's traffic density which reveals different traffic loads with respect to geography. The MNO's traffic and infrastructure information can be acquired through the network data with the aim to determine the highest traffic region at cluster level. The utilization of the legacy network data in this regard can be beneficial to reveal traffic and infrastructure correlations. As an alternative to conventional data sources, crowdsourced real network data of the MNO can be utilized for planning of 5G and beyond to infer about network traffic and infrastructure.

When the data is available, the following step is to choose which learning technique is best suited to the data's labeled or unlabeled nature. Supervised learning is defined by its use of labeled datasets for classification or regression problems [76]. Given the unlabeled nature of the crowdsourced data, we believe that unsupervised clustering assisted by real network data can reveal insights into the highest traffic density area. In unsupervised learning, clustering refers to revealing unseen patterns from the unlabeled data in the form of k clusters [77]. It consists of the organization of data in a way that there is high similarity in intra-cluster compared to lower inter-cluster similarity [78]. For clustering problems, the k-means clustering technique is widely adopted for traffic analysis, network analytics, data mining and pattern recognition [79]. When compared to the algorithms of the same class, k-means ensures convergence, reconfigurability, automation, scalability, high computing efficiency and low computational complexity with large datasets [80]. Researchers show that the k–means clustering provides exceptional results in network traffic classification with a precision of up to 90% [81]. However, choosing the right value for k i.e., the total number of required clusters plays a significant role in revealing the accurate insights from the considered data.

The widely used legacy method to determine the value of k is Elbow method [82], [83] that does not interpret the data insights while suggesting the k value. However, we believe that determining the value of k is subjective in the context of what one is trying to achieve with the given targets and constraints. Therefore, considering the crowdsourced network data, we scrutinize the value of k determined by Elbow method in the context of achieving the MNO's requirements. The appropriate value of k has two fold importance for the MNO. First, the value of k corresponds to the coverage area of the clusters that compromises the deployment cost related to the adequate number of radio sites. Second, the number of clusters k correlates with the traffic density MB/km^2 which decides the potential network utilization of the newly deployed gNBs (Next Generation Node-B).

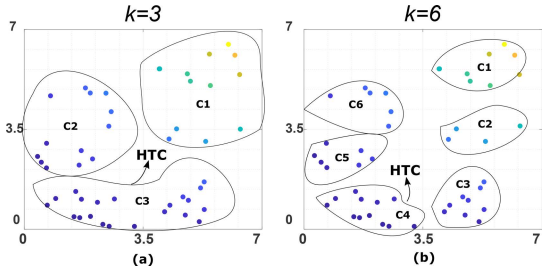

FIGURE 1. Illustration of clustering for network planning with smaller (a) and larger (b) value of k clusters in 7 × 7 km^2 boundary. The dots in the figure represent the legacy BSs where solid boundaries represent clusters. To simplify the HTC identification, we consider that a unit traffic sample is captured per BS and each BS provides coverage in a unit area.

Consider the illustration of small and large values of k as provided in Fig.1(a & b), respectively. In the case of smaller value for $k = 3$, the corresponding clusters (C1 to C3) are bigger in size as compared to the clusters (C1 to C6) for $k = 6$. For $k = 3$, the accumulative traffic samples of BSs in C3 is 21 which identifies it as an HTC compared to C2 and C3. On the other hand, for $k = 6$ the total number of traffic samples for HTC C4 is 12. The HTC of $k = 3$, namely C3, has a larger coverage area compared to HTC of $k = 6$, namely C4, as shown in Fig.1(a & b), respectively. Therefore, a large number of gNBs will be required for C3 to deploy new 5G services which will result in a higher deployment cost and the network may be over-budgeted if the financial limit of the MNO is exceeded. Besides, traffic density MB/km^2 will be lowered which may yield to the under-utilization of the network. As a result, the under-utilization of the network will eventually raise the cost per MB for the MNO. In contrast, C4 of Fig.1(b) has a smaller coverage area which means that MNO will be offering its new 5G services to a very limited area with a smaller number of deployed gNBs compared to C3 of Fig.1(a). However, the traffic density MB/km^2 of C4 will be higher, which means traffic demand is higher and may result in an over-utilized network. In this case, the deployment cost may be under-budgeted and the cost per MB will be decreased but at the same time, MNO will not be able to reach a larger number of subscribers due to the smaller coverage area. Therefore, clustering for network planning should be incorporated in terms of cost-effectiveness and improved network utilization. An adequate strategy is required to handle not only the MNO's budgetary limits, but also network over/under-utilization. The proposed clustering framework is developed on real mobile data to ensure that the network utilization is improved and the cost per MB is minimized.

IV. NETWORK CLUSTERING FRAMEWORK

In this section, we introduce the vision of network clustering which will be fully developed in Sections IV-A to IV-D following the scheme of Fig.2. In this study, we use network data from OpenCellID [47] database to reveal insights of the

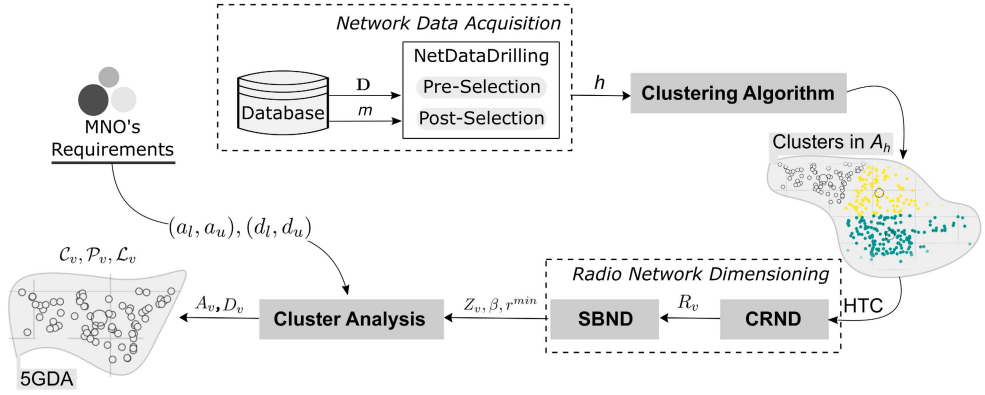

FIGURE 2. Network clustering framework with Network Data Acquisition entity along with Clustering and Radio Network Dimensioning algorithms to provide corresponding parameters for Cluster Analysis to provide 5G deployment area (5GDA).

MNO's network related to traffic density and infrastructure. In our previous study [27], we address the acquisition, selection, cleaning, features, size and format of the database. The database matrix $\mathbf{D}$ for Spain contains the raw data, being E the number of samples in millions, and each row $\mathbf{d}_e$ of $\mathbf{D}$ represents a data point $\mathbf{d}_e = (r_e, m_e, t_e, n_e, l_e, s_e)$. Hence, $\mathbf{D} := \{\mathbf{d_e}\}_{e=1}^{E} \; \forall \; \mathbf{d_e} \in \mathbb{R}^v$, where $v = 6$ represents the number of data variables:

$$\mathbf{D} = \begin{pmatrix} r_1 & m_1 & t_1 & n_1 & l_1 & s_1 \\ r_2 & m_2 & t_2 & n_2 & l_2 & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_e & m_e & t_e & n_e & l_e & s_e \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ r_E & m_E & t_E & n_E & l_E & s_E \end{pmatrix},$$

- r_e: radio access technology (RAT) of row e.
- m_e: mobile network code (MNC) of row e.
- t_e: tracking area code (TAC) of row e.
- n_e: cell ID (CID) of row e.
- l_e: bi-dimensional location of the cell with latitude and longitude $l_e = (l_e^1, l_e^2)$.
- s_e: number of samples of row e.

The network data acquisition entity of Fig.2 processes the database $\mathbf{D}$ through two phases, i.e., Pre-Selection and Post-Selection. The Pre-Selection phase acquires the MNO's data while the Post-Selection phase ensures reliability of the acquired data. Consisting of these two phases, the network data $\mathbf{D}$ of the MNO m is processed by the *NetDataDrilling* Algorithm (see [27] for details) to find h (highest traffic TAC ID), as shown in Fig.2. The corresponding geographical area of the TAC ID h is represented by A_h. The network data corresponding to h is then fed to the clustering algorithm which is used to divide the area A_h into k clusters. For clustering we used the k-means learning algorithm, to be discussed in Section IV-A. The clusters obtained from the k-means technique are evaluated based on the traffic samples to determine the HTC (labeled as v), described in Section IV-B. Next, we perform radio network dimensioning to determine relevant parameters of the HTC (see Fig.2). The conventional radio network dimensioning (CRND) is performed for HTC to determine the site range R_v and the area A_v covered by the HTC, to be discussed in Section IV-C. At the same time, network dimensioning for new 5G services is performed for the area under HTC with the objective to determine the offered capacity β_v, minimum offered data rate r^{min} and the required radio sites Z_v. Finally, cluster analysis is performed that is the nucleus of our study to determine the appropriate k fulfilling the MNO's requirements of coverage area A_v and the highest traffic density D_v MB/km^2, to be discussed in Section IV-D. The MNO's requirements are provided as bounds on coverage area (a_l, a_u) and the traffic density (d_l, d_u). In cluster analysis, we formulate the problem of estimating the potential network utilization $\mathcal{P}_v$ and the cost per MB $\mathcal{C}_v$ of the HTC given that the MNO's requirements are fulfilled. The result of the cluster analysis is the 5G deployment area A_v (km^2) and the traffic density MB/km^2 of the HTC (see Fig.2) with lowest cost per MB $\mathcal{C}_v$ and highest network utilization $\mathcal{P}_v$.

The different entities of this framework are developed in the subsequent sections.

A. CLUSTERING ALGORITHM

The k-means clustering is an efficient and unsupervised ML algorithm widely used to clusterize data into k clusters [84]. The k-means method consists of a twofold mechanism; first, it selects k data points known as centroids and other data points are assigned to the corresponding closest centroids based on Euclidean distance. Second, once the clusters are formed, re-computations are performed for the centroids of each cluster. This mechanism iterates until the cluster formation converges. The k-means algorithm tends to minimize the following objective function:-

$$\sum_{j=1}^{k} \sum_{i=1}^{E} ||x_i^{(j)} - c_j||^2, \tag{1}$$

where $||x_i^{(j)} - c_j||^2$ represents the distance between data point $x_i^{(j)}$ and the centroid c_j of the jth cluster. In our study, there are E data points that represent the geographical coordinates of the cell towers inside the area A_h. In the database $\mathbf{D}$, $x_i^{(j)}$ represents the latitudes and longitudes (l_e^1, l_e^2) of the cell tower i inside the jth cluster. The latitudes and longitudes (l_e^1, l_e^2) of the cell towers are fed to the k-means to clusterize the area A_h with respect to the Euclidean distance.

B. HIGHEST TRAFFIC CLUSTER

The HTC refers to the cluster with the highest number of traffic samples corresponding to the cell towers residing inside that cluster. The objective to find the HTC is to determine the geographical area of the subscribers with the highest traffic demands. Thus, when the clusters are obtained from the clustering algorithm we aim to find the HTC among the k clusters. Compared to other clusters, HTC carries highest traffic density MB/km^2 which maximizes the network utilization and decreases the cost per MB for the MNO. The computation of the HTC is based on the aggregated number of samples of all the existing cell towers as:-

$$\begin{aligned} v &= \arg_j \max V^j, \\ V^j &= \sum_{i=1}^{E} s^{j,i}, \quad \forall j = 1, \ldots, k, \end{aligned} \tag{2}$$

where v is the cluster label of the HTC which contains a unique set of cell towers $\mathcal{N} = \{n_1, \ldots, n_e, \ldots, n_E\}$, where n_e represents the CID and $N = |\mathcal{N}|$ is the total number of towers in HTC. The term $s^{j,i}$ refers to the traffic samples of the cell tower i inside the cluster j where as the variable V^j represents the total number of traffic samples of the jth cluster.

C. RADIO NETWORK DIMENSIONING

Radio network dimensioning (RND) is an essential phase of network planning process to determine the required number of radio sites and cell range such that the coverage and capacity requirements are fulfilled depending on path loss, transmit power, data rates and frequencies [27]. For 4G BSs the CRND approach is used where the corresponding 4G frequency bandwidth and related data rate are considered to obtain coverage area of legacy sites. We determine 4G coverage area based on CRND as the current users are being provided coverage by 4G. In this way, determination of required radio sites is achieved in order to provide coverage by new gNBs.

In this study, RND is performed for the HTC in order to calculate its area based on the coverage provided by the 4G cells. The coverage area of the HTC enables network planners to decide the radio and capacity requirements of the new network deployments. Since, the network data belongs to the legacy network we use the CRND approach of LTE. We perform CRND to determine the site range R_v of the cell tower inside the HTC. We compute the site range R_v in order to estimate the coverage area A_v of the HTC based on LTE service model [6], [8], [27] as:-

$$A_v = W_{aux} \cdot A \cdot R_v^2, \tag{3}$$

where W_{aux} is the number of cell towers within the HTC and $A = 1.95$ is the area coefficient.

On the other hand, we perform service-based network dimensioning (SBND) with *NetDimensioning* algorithm [27] for the HTC area A_v based on frequency and data rates of 5G. In contrast to CRND, 5G SBND is performed with the NR parameters to obtain the required gNB sites to provide coverage in A_v and to determine the required capacity in the area. The SBND algorithm [27] provides the capacity β of a gNB, minimum data rate r^{min} and the number of radio sites Z_v required in HTC area. Thus, the cost Y_v associated with the deployment of gNB sites under the HTC area is determined as:-

$$Y_v = Z_v \cdot \vartheta, \tag{4}$$

where ϑ is the cost of deployment per gNB in dollars [85]. The aggregated network capacity of the HTC β_v in MB can be determined in the same manner as:-

$$\beta_v = r^{min} \cdot \beta \cdot Z_v, \tag{5}$$

where r^{min} is the minimum data rate offered by the designed capacity β when all the users per gNB are active in the HTC. The designed capacity β is obtained by probabilistically characterizing the 5G radio resource control (RRC) states such that the data rates are guaranteed for MNO's defined percentage of time (see the capacity model of [27]).

D. CLUSTER ANALYSIS

The cluster analysis is the core of our framework to determine the network parameters and adapt the appropriate value of k such that the MNO's requirements of the traffic density MB/km^2 and deployment area km^2 are fulfilled. The MNO's requirements are not only considered to cater for the financial constraint but also to improve the efficiency in the context of network utilization. The financial constraint controls the deployment cost of the target area where new 5G services will be deployed. At the same time, these services likely to be offered within the area of highest traffic density in MB/km^2, thus, achieving the higher network utilization with lowest cost per MB for the MNO.

We define traffic density of the HTC as D_v in MB/km^2. The traffic density D_v depends on the number of samples $s^{j,i}$ of the cell tower i of cluster j and the coverage area A_v of the cluster. Thus traffic density D_v is computed as:-

$$D_v = \frac{\sum_{i=1}^{E} s^{j,i}}{A_v}, \quad \forall j = 1, \ldots, k. \tag{6}$$

Next, we compute the network utilization which is a very important parameter in determining the cost per MB of the HTC. Supported by SBND, each user in the HTC will experience data rate equal to r^{min}. Hence, we consider that the minimum volume of traffic transferred between user and the

BS is represented by the conversion of traffic samples ($s^{j,i}$) into current network utilization T_v in MB as:-

$$T_v = \alpha \cdot r^{min} \cdot \sum_{i=1}^{E} s^{j,i}, \quad \forall j = 1, \ldots, k, \tag{7}$$

where there are $\alpha = 0.125$ bytes in a bit which is used to convert bits into bytes and the term $\left(\sum_{i=1}^{E} s^{j,i}\right)$ represents the total number of samples of cluster j. Thus, the cost C_v per MB of the HTC is a function of the total deployment cost Y_v of the gNBs and the current network utilization T_v as:-

$$C_v = Y_v / T_v. \tag{8}$$

To determine the potential network utilization P_v percentage of the HTC, we can use the aggregated network capacity β_v in (5) as:-

$$P_v = \left(\frac{T_v}{\beta_v}\right) \cdot 100 \tag{9}$$

where T_v is the current network utilization in (7).

The two relevant metrics in the proposed framework are the coverage area A_v and the traffic density D_v MB/km^2 of the HTC. The constraints to determine the appropriate value of k are translated as MNO's requirements for new 5G deployments which are controlled by bounds on the deployment area ($a_l \leqslant A_v \leqslant a_u$) km^2 and on the traffic density ($d_l \leqslant D_v \leqslant d_u$) MB/km^2 of the HTC, respectively. These upper and lower bounds are imposed qualitatively but not quantitatively, though, the translation may not be obvious. The coverage area A_v in km^2 is constrained by the MNO to control the financial aspect of the deployment as larger clusters imply higher deployment cost. Therefore, over or under-budgeting for the new 5G deployments is handled with the introduction of A_v. On the other hand, traffic density D_v in MB/km^2 is the second requirement of the MNO to achieve adequate level of network utilization to handle cost per MB. As the higher D_v, the higher the traffic and the lower the cost per MB, and vice versa. We need to ensure that the network is not under or over-utilized, thus bounds on D_v are imposed accordingly. To introduce these MNO's requirements within the proposed framework a network clustering algorithm is introduced that is to be discussed in the next section.

V. NETWORK CLUSTERING ALGORITHM

This section presents the network clustering algorithm which is the implementation of the proposed framework presented in Fig.2 and whose pseudocode is given in Algorithm 1. We recall that, to the best of our knowledge, this is the first network clustering proposal taking into account radio dimensioning, MNO's requirements of coverage area A_v and traffic density D_v MB/km^2 assisted by real network data. In this section, we develop the proposed algorithm, named as *NetClustering*, to clusterize A_h and to determine the appropriate value of k according to the MNO's requirements. First, we use k-means to clusterize the geographical area A_h in to k clusters and identify the HTC. We perform the CRND technique to determine LTE site range R_v and compute the area A_v of the HTC. We also perform SBND technique by *NetDimensioning* algorithm to get the required number of radio sites Z_v, gNB site capacity β and the minimum data rate r^{min} for the HTC. We develop the mechanism to analyze the clusters and corresponding network traffic from real mobile data following the MNO's requirement to determine the appropriate value of k.

Algorithm 1: Network Clustering

```
procedure NetClustering(D, α, ϑ, r_c, a_l, a_u, b_l, b_u)
V_aux ← 0
W_aux ← 0
result ← 0
[h] ← NetDataDrilling(D)
D(t_e) = h
for all k ← K_min to K_max do
    [L] ← k-means(k, D(te, l_e))
    for all n_e ∈ L
    p ← q : n_e = L(q) do
        if n_e = D(n_e) then
            V_aux(p) ← V_aux(p) + s_e
            W_aux(p) ← W_aux(p) + n_e
        end
    end
    Find index v : max V_aux
    L_v ← L(v)
    [R_v] ← CRND(r_c)
    A_v ← W_aux(v) · 1.95 · R_v^2
    [Z_v, β, r^min] ← NetDimensioning(A_v)
    Y_v ← Z_v · ϑ
    T_v ← α · r_x^min · V_aux(v)
    D_v ← V_aux(v)/A_v
    β_v ← r^min · β · Z_v
    C_v ← Y_v/T_v
    P_v ← (T_v/β_v) · 100
    if (a_l ⩽ A_v ⩽ a_u) ∧ (d_l ⩽ D_v ⩽ d_u) then
        update result;
    end
end
result ← [C_v, P_v, L_v]
```

The proposed *NetClustering* (Algorithm 1) is developed to work out two problems. First, it performs the data acquisition by the *NetDaraDrilling* procedure and determines the area A_h. Second, based on MNO's requirements it acquires the appropriate value of k and determines the HTC for new 5G deployments. It requires the following inputs:

- The network database **D**
- The conversion constant α
- The deployment cost ϑ of a gNB
- The peak datarate r_c for LTE
- The lower and upper bounds on A_v as (a_l, a_l)
- The lower and upper bounds on D_v as (d_l, d_l)

The proposed algorithm is designed to be evaluated on a range of $k = [K_{min}, K_{max}]$. We introduce some auxiliary variables to store intermediate results. V_{aux} represents the summation of the samples of those cells within HTC for each k, then having a dimension equal to k. W_{aux} represents the summation of cell towers inside the HTC (with dimension equal to the number of cells forming the HTC).

First, we call the *NetDataDrilling* procedure [27] to obtain the TAC ID h of the highest traffic TAC area A_h (step 5). The data points corresponding to h are represented in $\mathbf{D}(t_e)$ in step 6. Next, network clustering is performed for each k (*for* loop in step 7). In step 8, clustering is performed by calling the algorithm for the corresponding value of k. We pass the coordinates of the cell towers ($l_e = (l_e^1, l_e^2)$) of the area A_h as $\mathbf{D}(t_e, l_e)$. The output of the k-means algorithm is given by the set of cluster's label in $\mathcal{L}$ having dimension equal to k. To sum the traffic samples per cluster, we begin the loop) in step 9, keeping in p the index of the n_e CIDs (step 10) belonging to the k clusters. In step 11, if CID n_e belongs to $\mathbf{D}(n_e)$ we save the aggregated traffic samples per cluster in V_{aux} (step 12) and the number of cell towers per clusters are provided in W_{aux} (step 13). In step 16, we find the index v of the HTC and the corresponding label is represented by $\mathcal{L}_v$ (step 17).

Next, CRND procedure is called (step 18) to obtain the cell tower range R_v in the HTC area. Based on R_v, coverage area A_v of the HTC is computed in step 19. In step 20, the HTC area A_v is provided to *NetDimensioning* algorithm [27] to get the required number of radio sites Z_v, the gNB capacity β and r^{min} for HTC. In step 21, we compute the total cost of deployment Y_v for Z_v radio sites to be deployed in the HTC area A_v. Next, in step 22 we compute the current traffic T_v of the HTC. To obtain the traffic density D_v of the HTC we utilize (6) in step 23. Next, we compute network capacity β_v (step 24) of the HTC based on the required number of radio sites Z_v and the data rate r^{min}. Based on the previous computations, we then determine C_v (the estimated cost per MB of the HTC) in step 25 along with the potential network utilization percentage P_v in step 26. Finally, the MNO's criterion is introduced (in step 27) along with bounds $[(a_l, a_u), (d_l, d_u)]$ on A_v and D_v, respectively. Given that the MNO's criterion is fulfilled, the results are updated in step 28. Finally, the results of the HTCs for the range of $k = [K_{min}, K_{max}]$ are compiled in step 30.

The complexity of Algorithm 1 is mainly based on three aspects. First, it depends on the size of the database $\mathbf{D}$ represented by the number of samples E in millions. The *NetDataDrilling* procedure in step 5 process $\mathbf{D}$ with a finite number of iterations (see [27] for details). Thus the complexity term for *NetDataDrilling* can be represented as $\mathcal{O}(E) + \mathcal{O}(E \times |t_e|) + \mathcal{O}(|\mathbf{D}(t_e)|)$, where $|\mathbf{D}(t_e)|$ represents the number of TACs in $\mathbf{D}$. Second aspect of complexity reclines on the *k-means* algorithm (step 8) to cluster the area A_h in k clusters with a finite number of iterations given by the *for* loop in step 7. Thus the complexity term is represented as $\mathcal{O}(k \times |\mathbf{D}(t_e)| + |\Delta K|)$, where ΔK represents the granularity to increment k for the next iteration, respectively. The third aspect is subject to the complexity of the RND algorithm given by $\mathcal{O}(B \times |\Delta P| + |\Delta Q|)$. The term B represent the bandwidth where cellular services are configured and evaluated for the transmit power ΔP and cell load ΔQ granularities for next iteration [27]. Note that the RND algorithms are independently executed for CRND (step 18) and NetDimensioning (step 20), respectively.

VI. ELBOW METHOD

The Elbow method is primarily based on k-means learning technique that computes the sum of squared distances (distortions) from each point to its assigned centroid as a function of k [86]. The appropriate value of k is selected by running the k-means algorithm across a range of k. The method plots the distortion as a function of k and choses the k at the point where distortion drops drastically forming the smallest angle. The distortions can be calculated using (1) as explained in the Section IV-A. The pseudocode of the Elbow heuristic is given in Algorithm 2. We start by initializing the value of $k = 2$ in step 1 and clustering is performed for a range of $k = [K_{min}, K_{max}]$ (*for* loop in step 3). We measure the distortions by using (1) and values are stored in $\mathcal{V}_k$ having dimension equal to k (step 4). In step 5, all the distortion values are updated to the *result* for each k and finally *result* is returned in step 7.

Algorithm 2: Elbow Heuristic

```
1 k ← 2
2 result ← 0
3 for all k ← K_min to K_max do
4 |   [V_k] ← Calculate distortions with (1)
5 |   update result;
6 end
7 result ← [V_k]
```

The complexity of Algorithm 2 is similar to the k-means as the distortion values are computed independently by finite number of iterations of the *for* loop in step 3. Thus, the complexity of Elbow heuristic is given as $\mathcal{O}(k \times |\mathbf{D}(t_e)| + |\Delta K|)$, where ΔK represents the granularity of k.

VII. RESULTS AND ANALYSIS

This section presents the results and the corresponding analysis of the considered parameters of our study, i.e., cost per MB C_v, potential network utilization P_v (%), traffic density D_v (MB/km^2) and the 5G deployment area A_v (km^2). The simulation results of the proposed *NetClustering* algorithm are presented and compared with the Elbow heuristic.

The simulation includes the area A_h shown in Fig.3 obtained from the *NetDataDrilling* procedure [27], where LTE cell towers are located across the area. The simulation curves presented in this section have been obtained by averaging the results from 1,000 executions, each corresponding to one independent, random and uniform users distribution. In this study, we investigate two scenarios with different

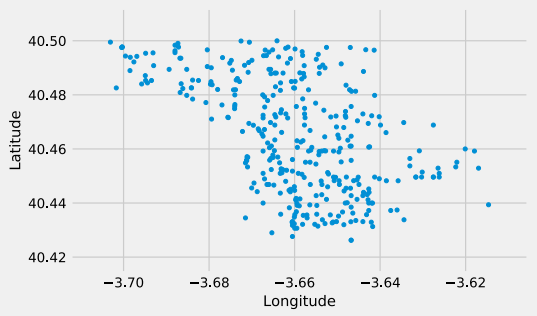

FIGURE 3. Highest traffic TAC area A_h obtained from *NetDataDrilling* procedure [27] where the blue circles represent cell towers corresponding to their longitudes and latitudes, respectively.

TABLE 1. Simulations parameters.

Symbol	Description	Value	Unit
k	Range of clusters K_{min} to K_{max}	$[2, 100]$	–
B	5G New Radio Bandwidths	$[30, 50]$	MHz
r^{min}	Minimum data rate for 5G [27]	$[37, 42.8]$	Mbps
r_c	Data rate offered by LTE @ 10 MHz	1	Mbps
(a_l, a_u)	Lower and Upper bounds on A_v	(20,25)	km^2
(d_l, d_u)	Lower and Upper bounds on D_v	(50,100)	MB/km^2
σ	Shadow fading	$\ln(\sigma) \sim \mathcal{N}(0, 7)$	dB
ϑ	Deployment cost per gNB [85]	40500	USD

bandwidth $B = \{30, 50\}$ MHz, while other simulation parameters are provided in Table 1.

Reference to the Algorithm 1, the *NetDataDrilling* procedure [27] provides the TAC ID h of the area A_h. The database corresponding to the ID h is fed to the Elbow heuristic, the Elbow heuristic suggests the number of clusters $k = 3$, as shown in the elbow curve of Fig.4. These three clusters are obtained from the clustering algorithm and are shown in Fig.5, where HTC is shown with the towers presented with white circles. On the other hand, the same database corresponding to the ID h feeds to the proposed *NetClustering* algorithm for a range of clusters K_{min} to K_{max}. The curves of cost per MB (C_v) and potential network utilization (P_v) are presented in Fig.6. The network utilization curve has

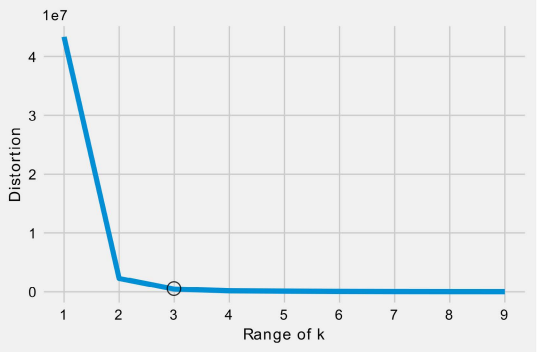

FIGURE 4. Sum of squared distances or distortion curve obtained from Elbow heuristic where marked point forming the smallest angle shows the recommended value of $k = 3$.

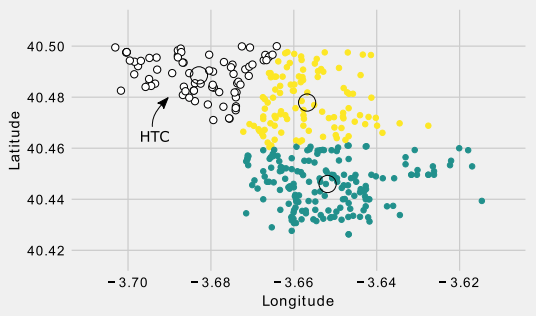

FIGURE 5. Clustering performed in the area A_h for recommended value of $k = 3$ by Elbow heuristic, where white circles shows the HTC and big circles are the centroids.

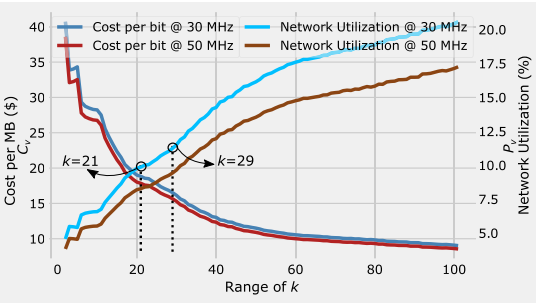

FIGURE 6. Cost per MB (C_v) and potential network utilization (P_v) as a function of k for two bandwidth scenarios of $B = [30, 50]$, respectively. The appropriate range of k (21 to 29) is shown by the dotted lines.

a non-decreasing behavior with the increase in number of clusters from 2 to 100. As k increases, the area A_h tends to divide into smaller coverage areas per cluster thus increases both the traffic density D_v MB/km^2 and network utilization P_v, as the services are offered in a limited area within a smaller cluster size. However, the overall achievable value of P_v is not more than 20% and 17.4% for $B = \{30, 50\}$ MHz, respectively.

The reason for a lower percentage value of P_v is due to the fact that the designed bandwidth provides more capacity than the current requirement of the subscribers per cluster. On the other hand, cost per MB C_v curves tend to decrease with smaller size clusters as the coverage is provided into a more concentrated area of subscribers with higher traffic demands. When the k value is large, it means that the area is divided into multiple smaller regions, therefore, it becomes convenient for the MNOs to deploy an adequate number of radio sites fulfilling the current requirement of the subscriber's traffic. The smaller size cluster means that the MNO has to deploy new radio sites in a smaller area, thus decreases the deployment cost Y_v. Network planning driven on the cluster level minimizes C_v for the MNOs, however, the opportunity cost is paid in offering new services within the limited geographical area. Besides, if the traffic density D_v and P_v are not considered while deciding the appropriate k, the deployed network may become over-utilized over time for larger values of k as more number of subscribers will be acquiring new services

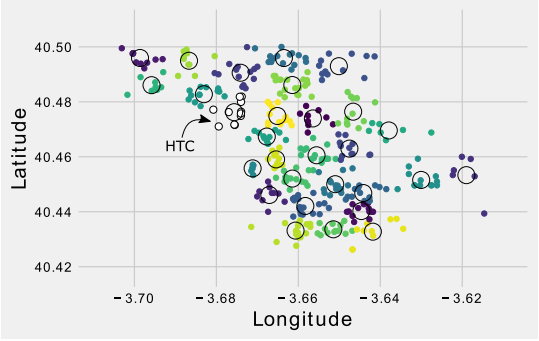

FIGURE 7. Clusters of the area A_h based on the value of $k = 29$ provided by the *NetClustering* algorithm where towers in HTC are shown with white circles.

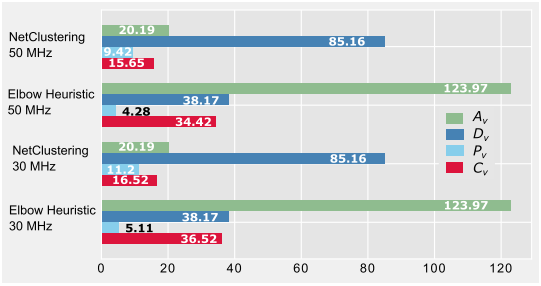

FIGURE 8. Comparison of different parameter's values of the HTCs obtained from Elbow heuristic ($k = 3$) and *NetClustering* algorithm ($k = 29$) for $B = [30, 50]$ MHz cases.

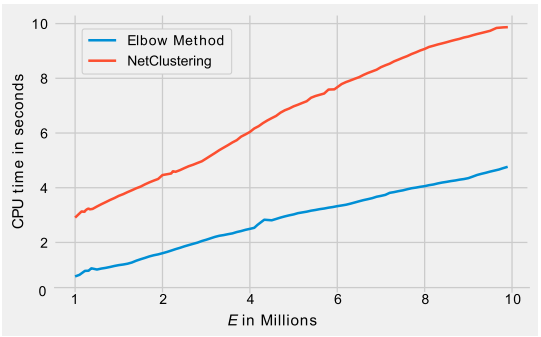

FIGURE 9. Elbow method vs. proposed *NetClustering* computational complexity in terms of CPU time (seconds).

within a smaller concentrated area. In contrast, the deployed network may become under-utilized with the smaller values of k, as can be seen in the trend followed by the P_v curves for both bandwidth scenarios in Fig.6.

The accuracy of *NetClustering* is evaluated for $(20 \leq A_v \leq 25)$ and $(50 \leq D_v \leq 100)$. Under these bounds the algorithm reveals the appropriate value of k under the range of 21 to 29, as shown with dotted lines in Fig.6. The algorithm fulfills the MNO's requirements of A_v and D_v under this range according to the values provided in Table 1. Within this range, the highest potential network utilization and the lowest cost per MB cluster is provided by $k = 29$ (see dotted lines in Fig.6) for both bandwidth cases. The clustered area A_h for $k = 29$ is presented in Fig.7, with cell towers plotted with white circles. It is clear that larger value of $k = 29$ results in smaller size clusters (Fig.7) compared to larger size clusters for $k = 3$ (Fig.5). However, the geographical region of the highest traffic remains the same under the HTC obtained from both Elbow heuristic and *NetClustering* algorithm, respectively.

The comparison between appropriate values of k suggested by Elbow method and *NetClustering* algorithm is based on cost per MB C_v, potential network utilization P_v, traffic density D_v MB/km^2 and the deployment area A_v (km^2) is provided in Fig.8 for the given bandwidths. The deployment area A_v of HTC is independent of the bandwidth, thus reveals $A_v = 123.97$ km^2 for both bandwidth cases as shown with the green bars in Fig.8. In case of *NetClustering*, the deployment area is $A_v = 20.19$ km^2 and is under the bounds provided by the MNO. The results (for $k = 3$, Elbow heuristic) show under-utilization of the network because coverage is provided in a larger area while the potential network utilization is lower with $P_v = 5.11\%$ and $P_v = 4.28\%$ for $B = [30, 50]$ MHz cases, respectively. On the other hand, HTC (for k=29, *NetClustering*) has better network utilization of $P_v = 11.2\%$ and $P_v = 9.42\%$.

The traffic density values for the Elbow heuristic based HTC are very low i.e., $D_v = 38.17$ MB/km^2 for both bandwidth cases, respectively. The lower D_v value uncovers the fact that the formation of the large size clusters (for $k = 3$) are not suitable in this region as the traffic load is not much. On the other hand, the HTC obtained by *NetClustering* has better traffic density with $D_v = 85.16$ MB/km^2 (for $k = 29$) as shown in Fig.8 by sky blue bars, for both bandwidth cases. Moreover, the cost per MB values $C_v = [16.52, 15.65]$ from *NetClustering* algorithm are lower compared to $C_v = [36.52, 34.52]$ values obtained from the Elbow heuristic, as shown in Fig.8. The lower cost per MB is due to the fact that the appropriate value of $k = 29$ results in smaller size clusters which saves deployment cost of the MNO and provides cheaper clusters in terms of cost per MB. Deploying the new 5G gNBs within the smaller size clusters for $k = 29$ seems to be a better choice for the MNOs. It saves the deployment cost Y_v and at the same time a lower cost C_v per MB is achieved with better network utilization P_v and higher traffic density D_v MB/km^2. However, the price paid is the computational complexity in terms of CPU time, as shown in Fig.9. The complexity curves are presented by increasing the number of samples E or data points $\mathbf{d}_e$ by 1 up to 10 million. The *NetClustering* algorithm consumes more CPU time, but it is still linear with E. In this case, the computational complexity is significantly less essential than the cost, as clustering can be performed off-line on non-real-time basis.

VIII. CONCLUSION

Revealing insights about the current traffic loads from the existing network infrastructure assist the network planning to reduce the cost for the MNOs. In this paper, we show

the network planning problem of identifying the HTC area A_v of the highest traffic density D_v MB/km^2 by employing clustering based on ML. The appropriate value of k and corresponding HTC is identified by the *NetClustering* algorithm fulfilling the MNO's requirements on A_v and D_v. We show that the proposed algorithm determines the value of k such that the potential network utilization P_v is higher while the cost per MB C_v is minimized. The performance comparison is evaluated based on cost per MB C_v, traffic density D_v, deployment area A_v and the network utilization P_v of the HTC. We compare these parameters and observe that the *NetClustering* algorithm not only attains up to 45% cost savings per MB but achieves higher network utilization P_v compared with the Elbow heuristic. We have evaluated our proposed algorithm on the two bandwidth scenarios of 30 and 50 MHz and our algorithm shows consistent performance.

As future research lines, we are committed to exploring the ML applications in the context of network data combined with radio dimensioning of mmWave for 5G and beyond. In sixth-generation (6G), one of the primary use cases is ultra-massive machine type communication (umMTC) with a density of 10^7 devices per km^2. The proposed clustering framework can be evolved to optimize spectrum and energy efficiency in large-scale IoT scenarios for newly defined MNO's requirements.

REFERENCES

[1] *IMT Vision-Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document ITU-R M.2083, 2015.

[2] D. Cao, S. Zhou, and Z. Niu, "Optimal base station density for energy-efficient heterogeneous cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2012, pp. 4379–4383.

[3] Z. Niu, S. Zhou, Y. Hua, Q. Zhang, and D. Cao, "Energy-aware network planning for wireless cellular system with inter-cell cooperation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1412–1423, Apr. 2012.

[4] M. Zheng, P. Pawelczak, S. Stanczak, and H. Yu, "Planning of cellular networks enhanced by energy harvesting," *IEEE Commun. Lett.*, vol. 17, no. 6, pp. 1092–1095, Jun. 2013.

[5] S. Wang and C. Ran, "Rethinking cellular network planning and optimization," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 118–125, Apr. 2016.

[6] M. Jaber, Z. Dawy, N. Akl, and E. Yaacoub, "Tutorial on LTE/LTE-A cellular network dimensioning using iterative statistical analysis," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1355–1383, 2nd Quart., 2016.

[7] A. Taufique, M. Jaber, A. Imran, Z. Dawy, and E. Yacoub, "Planning wireless cellular networks of future: Outlook, challenges and opportunities," *IEEE Access*, vol. 5, pp. 4821–4845, 2017.

[8] A. Elnashar, *Coverage and Capacity Planning of 4G Networks*. Hoboken, NJ, USA: Wiley, 2014, ch. 6, pp. 349–444. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118703434

[9] A. Elnashar and M. A. El-Saidny, "Looking at LTE in practice: A performance analysis of the LTE system based on field test results," *IEEE Veh. Technol. Mag.*, vol. 8, no. 3, pp. 81–92, Sep. 2013.

[10] M. U. Khan, M. Azizi, A. G. Armada, and J. J. E. Garzás, "Heuristic for network planning based on 5G services," in *Advances in Smart Technologies Applications and Case Studies* (Lecture Notes in Electrical Engineering), vol. 684. Cham, Switzerland: Springer, 2020.

[11] F. J. Martin-Vega, J. C. Ruiz-Sicilia, M. C. Aguayo, and G. Gomez, "Emerging tools for link adaptation on 5G NR and beyond: Challenges and opportunities," *IEEE Access*, vol. 9, pp. 126976–126987, 2021.

[12] E. J. Oughton and A. Jha, "Supportive 5G infrastructure policies are essential for universal 6G: Assessment using an open-source techno-economic simulation model utilizing remote sensing," *IEEE Access*, vol. 9, pp. 101924–101945, 2021.

[13] L. Chiaraviglio, C. Di Paolo, and N. B. Melazzi, "5G network planning under service and EMF constraints: Formulation and solutions," *IEEE Trans. Mobile Comput.*, early access, Jan. 26, 2021, doi: 10.1109/TMC.2021.3054482.

[14] A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, and I. Da Silva, "Waveform and numerology to support 5G services and requirements," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 90–98, Nov. 2016.

[15] *Technical Specification: 5G NR-User Equipment (UE) Radio Transmission and Reception*, document TS 38.101-1 Version 15.2.0 Release 15, 3GPP, 2018.

[16] V. N. Ha, T. T. Nguyen, L. B. Le, and J.-F. Frigon, "Admission control and network slicing for multi-numerology 5G wireless networks," *IEEE Netw. Lett.*, vol. 2, no. 1, pp. 5–9, Mar. 2020.

[17] A. L. Rezaabad, H. Beyranvand, and M. Maier, "Ultra-dense 5G small cell deployment for fiber and wireless backhaul-aware infrastructures," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12231–12243, Dec. 2018.

[18] F.-H. Tseng, L.-D. Chou, H.-C. Chao, and J. Wang, "Ultra-dense small cell planning using cognitive radio network toward 5G," *IEEE Wireless Commun.*, vol. 22, no. 6, pp. 76–83, Dec. 2015.

[19] M. Gheisarnejad, A. Mohammadzadeh, H. Farsizadeh, and M.-H. Khooban, "Stabilization of 5G telecom converter-based deep type-3 fuzzy machine learning control for telecom applications," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 69, no. 2, pp. 544–548, Feb. 2022.

[20] F. B. Mismar and J. Hoydis, "Unsupervised learning in next-generation networks: Real-time performance self-diagnosis," *IEEE Commun. Lett.*, vol. 25, no. 10, pp. 3330–3334, Oct. 2021.

[21] S. Bakri, P. A. Frangoudis, A. Ksentini, and M. Bouaziz, "Data-driven RAN slicing mechanisms for 5G and beyond," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 4, pp. 4654–4668, Dec. 2021.

[22] B. M. ElHalawany, S. Hashima, K. Hatano, K. Wu, and E. M. Mohamed, "Leveraging machine learning for millimeter wave beamforming in beyond 5G networks," *IEEE Syst. J.*, early access, Jul. 2, 2022, doi: 10.1109/JSYST.2021.3089536.

[23] C. Y. Lee and H. G. Kang, "Cell planning with capacity expansion in mobile communications: A Tabu search approach," *IEEE Trans. Veh. Technol.*, vol. 49, no. 5, pp. 1678–1691, Sep. 2000.

[24] T. Bauschert, C. Büsing, F. D'Andreagiovanni, A. M. C. A. Koster, M. Kutschka, and U. Steglich, "Network planning under demand uncertainty with robust optimization," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 178–185, Feb. 2014.

[25] H. Ghazzai, E. Yaacoub, M.-S. Alouini, Z. Dawy, and A. Abu-Dayya, "Optimized LTE cell planning with varying spatial and temporal user densities," *IEEE Trans. Veh. Technol.*, vol. 65, no. 3, pp. 1575–1589, Mar. 2016.

[26] *Technical Specification. NR and NG-RAN Overall Description; Stage 2 (Release 16)*, document TS 38.300 V16.0.0, 3GPP, 2019.

[27] M. U. Khan, A. Garcia-Armada, and J. J. Escudero-Garzás, "Service-based network dimensioning for 5G networks assisted by real data," *IEEE Access*, vol. 8, pp. 129193–129212, 2020.

[28] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[29] Y. Fu, H. H. Yang, K. N. Doan, C. Liu, X. Wang, and T. Q. S. Quek, "Effective cache-enabled wireless networks: An artificial intelligence- and recommendation-oriented framework," *IEEE Veh. Technol. Mag.*, vol. 16, no. 1, pp. 20–28, Mar. 2021.

[30] H. B. Yilmaz, C.-B. Chae, Y. Deng, T. O'Shea, L. Dai, N. Lee, and J. Hoydis, "Special issue on advances and applications of artificial intelligence and machine learning for wireless communications," *J. Commun. Netw.*, vol. 22, no. 3, pp. 173–176, Jun. 2020.

[31] M. S. Hadi, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, "Patient-centric HetNets powered by machine learning and big data analytics for 6G networks," *IEEE Access*, vol. 8, pp. 85639–85655, 2020.

[32] R. Alkurd, I. Y. Abualhaol, and H. Yanikomeroglu, "Personalized resource allocation in wireless networks: An AI-enabled and big data-driven multi-objective optimization," *IEEE Access*, vol. 8, pp. 144592–144609, 2020.

[33] J. Wang, C. Jiang, H. Zhang, Y. Ren, K.-C. Chen, and L. Hanzo, "Thirty years of machine learning: The road to Pareto-optimal wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1472–1514, 3rd Quart., 2020.

[34] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE Access*, vol. 6, pp. 32328–32338, 2018.
[35] M. Polese, R. Jana, V. Kounev, K. Zhang, S. Deb, and M. Zorzi, "Machine learning at the edge: A data-driven architecture with applications to 5G cellular networks," *IEEE Trans. Mobile Comput.*, vol. 20, no. 12, pp. 3367–3382, Dec. 2021.
[36] M. M. A. Osman, S. K. S. Yusof, N. N. N. A. Malik, and S. Zubair, "A survey of clustering algorithms for cognitive radio ad hoc networks," *Wireless Netw.*, vol. 24, pp. 1451–1475, Jul. 2018.
[37] J. Y. Yu and P. H. J. Chong, "A survey of clustering schemes for mobile ad hoc networks," *IEEE Commun. Surveys Tuts.*, vol. 7, no. 1, pp. 32–48, 1st Quart., 2005.
[38] X. Liu, "A survey on clustering routing protocols in wireless sensor networks," *Sensors*, vol. 12, no. 8, pp. 11113–11153, 2012.
[39] S. Vodopivec, J. Bester, and A. Kos, "A survey on clustering algorithms for vehicular ad-hoc networks," in *Proc. 35th Int. Conf. Telecommun. Signal Process. (TSP)*, Jul. 2012, pp. 52–56.
[40] M. Gheisari, A. A. Abbasi, Z. Sayari, Q. Rizvi, A. Asheralieva, S. Banu, F. M. Awaysheh, S. B. H. Shah, and K. A. Raza, "A survey on clustering algorithms in wireless sensor networks: Challenges, research, and trends," in *Proc. Int. Comput. Symp. (ICS)*, Dec. 2020, pp. 294–299.
[41] O. Boyinbode, H. Le, A. Mbogho, M. Takizawa, and R. Poliah, "A survey on clustering algorithms for wireless sensor networks," in *Proc. 13th Int. Conf. Netw.-Based Inf. Syst.*, 2010, pp. 358–364.
[42] A. M. Ortiz, D. Hussein, S. Park, S. N. Han, and N. Crespi, "The cluster between Internet of Things and social networks: Review and research challenges," *IEEE Internet Things J.*, vol. 1, no. 5, pp. 206–215, May 2014.
[43] L. Xu, R. Collier, and G. M. P. O'Hare, "A survey of clustering techniques in WSNs and consideration of the challenges of applying such to 5G IoT scenarios," *IEEE Internet Things J.*, vol. 4, no. 5, pp. 1229–1249, Oct. 2017.
[44] S. Wang, X. Zhang, Y. Zhang, L. Wang, J. Yang, and W. Wang, "A survey on mobile edge networks: Convergence of computing, caching and communications," *IEEE Access*, vol. 5, pp. 6757–6779, 2017.
[45] E. Balevi and R. D. Gitlin, "A clustering algorithm that maximizes throughput in 5G heterogeneous F-RAN networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
[46] M. F. Khan, K. L. A. Yau, R. M. Noor, and M. A. Imran, "Survey and taxonomy of clustering algorithms in 5G," *J. Netw. Comput. Appl.*, vol. 154, Mar. 2020, Art. no. 102539. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804520300138
[47] (2019). *Exisiting 4G Infrastructure From OpenCelliD by Uniwired Labs*. [Online]. Available: https://opencellid.org/
[48] X. Wang, Z. Qian, S. Zhai, and X. Wang, "Social-aware resource allocation for multicast device-to-device communications underlying UAV-assisted networks," *Comput. Commun.*, vol. 153, pp. 367–374, Mar. 2020.
[49] L. Liu, Y. Zhou, V. Garcia, L. Tian, and J. L. Shi, "Load aware joint CoMP clustering and inter-cell resource scheduling in heterogeneous ultra dense cellular networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2741–2755, Mar. 2018.
[50] M. Song, H. Shan, H. H. Yang, and T. Q. S. Quek, "Joint optimization of fractional frequency reuse and cell clustering for dynamic TDD small cell networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 398–412, Jan. 2022.
[51] Y. Sun, G. Guo, S. Zhang, S. Xu, T. Wang, and Y. Wu, "A cluster-based energy-efficient resource management scheme with QoS requirement for ultra-dense networks," *IEEE Access*, vol. 8, pp. 182412–182421, 2020.
[52] U. S. Hashmi, S. A. R. Zaidi, and A. Imran, "User-centric cloud RAN: An analytical framework for optimizing area spectral and energy efficiency," *IEEE Access*, vol. 6, pp. 19859–19875, 2018.
[53] N. Merabtine, D. Djenouri, and D.-E. Zegour, "Towards energy efficient clustering in wireless sensor networks: A comprehensive review," *IEEE Access*, vol. 9, pp. 92688–92705, 2021.
[54] J. Zhu, M. Zhao, and S. Zhou, "An optimization design of ultra dense networks balancing mobility and densification," *IEEE Access*, vol. 6, pp. 32339–32348, 2018.
[55] Z. Khan and P. Fan, "A multi-hop moving zone (MMZ) clustering scheme based on cellular-V2X," *China Commun.*, vol. 15, no. 7, pp. 55–66, Jul. 2018.
[56] W. Qi, Q. Song, X. Wang, L. Guo, and Z. Ning, "SDN-enabled social-aware clustering in 5G-VANET systems," *IEEE Access*, vol. 6, pp. 28213–28224, 2018.
[57] W. Hao, O. Muta, H. Gacanin, and H. Furukawa, "Dynamic small cell clustering and non-cooperative game-based precoding design for two-tier heterogeneous networks with massive MIMO," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 675–687, Feb. 2018.
[58] M. Nikooroo and Z. Becvar, "Optimal positioning of flying base stations and transmission power allocation in NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1319–1334, Feb. 2022.
[59] Y. Dai, J. Liu, M. Sheng, N. Cheng, and X. Shen, "Joint optimization of BS clustering and power control for NOMA-enabled CoMP transmission in dense cellular networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 2, pp. 1924–1937, Feb. 2021.
[60] N. Taherkhani and S. Pierre, "Centralized and localized data congestion control strategy for vehicular ad hoc networks using a machine learning clustering algorithm," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 11, pp. 3275–3285, Nov. 2016.
[61] C. Fan, Y. J. Zhang, and X. Yuan, "Dynamic nested clustering for parallel PHY-layer processing in cloud-RANs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1881–1894, Mar. 2016.
[62] B. Zhou, H. Hu, S. Q. Huang, and H. H. Chen, "Intracluster device-to-device relay algorithm with optimal resource utilization," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 2315–2326, Jun. 2013.
[63] J. L. Gallardo, M. A. Ahmed, and N. Jara, "Clustering algorithm-based network planning for advanced metering infrastructure in smart grid," *IEEE Access*, vol. 9, pp. 48992–49006, 2021.
[64] G. Yang, A. Esmailpour, N. Nasser, G. Chen, Q. Liu, and P. Bai, "A hierarchical clustering algorithm for interference management in ultra-dense small cell networks," *IEEE Access*, vol. 8, pp. 78726–78736, 2020.
[65] X. Jia, P. Ji, and Y. Chen, "Modeling and analysis of multi-tier clustered millimeter-wave cellular networks with user classification for large-scale hotspot area," *IEEE Access*, vol. 7, pp. 140278–140299, 2019.
[66] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, and J. Henry, "UAV trajectory planning in wireless sensor networks for energy consumption minimization by deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9540–9554, Sep. 2021.
[67] J. Moysen and M. García-Lozano, "Learning-based tracking area list management in 4G and 5G networks," *IEEE Trans. Mobile Comput.*, vol. 19, no. 8, pp. 1862–1878, Aug. 2020.
[68] C. Bektas, S. Böcker, B. Sliwa, and C. Wietfeld, "Rapid network planning of temporary private 5G networks with unsupervised machine learning," in *Proc. IEEE 94th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2021, pp. 1–6.
[69] J. Moysen, L. Giupponi, and J. Mangues-Bafalluy, "A machine learning enabled network planning tool," in *Proc. IEEE 27th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–7.
[70] J. Pérez-Romero, O. Sallent, R. Ferrús, and R. Agustí, "Knowledge-based 5G radio access network planning and optimization," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2016, pp. 359–365.
[71] D. Mennie, "Computer report V: Power supplies: How to trim the budget: Orphaned in the world of decreasing cost-per-bit, power conditioning equipment cries for cost reduction," *IEEE Spectr.*, vol. 11, no. 2, pp. 57–61, Feb. 1974.
[72] V. Chandar, A. Tchamkerten, and D. Tse, "Asynchronous capacity per unit cost," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1213–1226, Mar. 2013.
[73] Y. Guan, F. Geng, and J. H. Saleh, "Review of high throughput satellites: Market disruptions, affordability-throughput map, and the cost per bit/second decision tree," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 34, no. 5, pp. 64–80, May 2019.
[74] H. Dong and J. D. Downie, "Applications for spatial divison multiplexing fiber and associated cost per bit," in *Proc. IEEE Photon. Soc. Summer Topicals Meeting Ser. (SUM)*, Jul. 2020, pp. 1–2.
[75] L. Blair and S. Thiagarajan, "Impact of moving to 100 Gbps on the cost per bit of a USA national network," in *Proc. Nat. Fiber Opt. Eng. Conf.*, Mar. 2010, pp. 1–3.
[76] F. Musumeci, L. Magni, O. Ayoub, R. Rubino, M. Capacchione, G. Rigamonti, M. Milano, C. Passera, and M. Tornatore, "Supervised and semi-supervised learning for failure identification in microwave networks," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 2, pp. 1934–1945, Jun. 2021.
[77] R. Xu and D. C. Wunsch, "Survey of clustering algorithms," *IEEE Trans. Neural Netw.*, vol. 16, no. 3, pp. 645–678, Jun. 2005.
[78] P. Berkhin, *A Survey of Clustering Data Mining Techniques*. Berlin, Germany: Springer, 2006, pp. 25–71, doi: 10.1007/3-540-28349-8_2.
[79] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network anomaly detection: Methods, systems and tools," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 303–336, 1st Quart., 2014.

[80] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Ann. Data Sci.*, vol. 2, pp. 165–193, Jun. 2015.
[81] Y. Liu, W. Li, and Y. Li, "Network traffic classification using K-means clustering," in *Proc. 2nd Int. Multi-Symp. Comput. Comput. Sci. (IMSCCS)*, Aug. 2007, pp. 360–365.
[82] F. Liu and Y. Deng, "Determine the number of unknown targets in open world based on elbow method," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 5, pp. 986–995, May 2021.
[83] X. Song, W. Li, D. Ma, Y. Wu, and D. Ji, "An enhanced clustering-based method for determining time-of-day breakpoints through process optimization," *IEEE Access*, vol. 6, pp. 29241–29253, 2018.
[84] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002.
[85] E. J. Oughton, K. Konstantinos, E. Fariborz, K. Dritan, and C. Jon, "An open-source techno-economic assessment framework for 5G deployment," *IEEE Access*, vol. 7, pp. 155930–155940, 2019.
[86] J. Zeng, J. Wang, L. Guo, G. Fan, K. Zhang, and G. Gui, "Cell scene division and visualization based on autoencoder and K-means algorithm," *IEEE Access*, vol. 7, pp. 165217–165225, 2019.

M. UMAR KHAN received the B.S. degree in telecommunication engineering from COMSATS University Islamabad, Pakistan, in 2012, the master's degree in telematics engineering from the Universidad Carlos III de Madrid (UC3M), in 2015, and the Ph.D. degree from the Department of Signal Theory and Communications, UC3M, in 2021. He was a Lecturer with the Center for Advanced Studies in Telecommunications (CAST), COMSATS University Islamabad, from 2015 to 2016, where he is currently a Lecturer. His research interests include planning, optimization, and dimensioning of next generation networks (NGNs) assisted by data based decision-making.

MOSTAFA AZIZI received the Diploma of State Engineering degree in electrical and computer engineering from the Mohammadia School of Engineers, Morocco, in 1993, and the Ph.D. degree in computer science from the University of Montreal (DIRO-FAS), Montreal, Canada, in 2001. He is currently a Professor with Université Mohammed Premier Oujda, Oujda, Morocco. He teaches several courses in the domain of computer science such as OOP, IA, RT-systems, distributed systems, TCP/IP, WEB, computers security, data structures, and algorithmics. He also supervises a number of master's/Ph.D. students. His research interests include verification/coverification of real-time and embedded systems; data communication and security; and computer-aided management of industrial processes, artificial intelligence, and smart things.

A. GARCÍA-ARMADA (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the Polytechnic University of Madrid, in February 1998. She is currently a Professor with the University Carlos III of Madrid, Spain, where she is leading the Communications Research Group. She has been a Visiting Scholar with Stanford University, Bell Labs, and the University of Southampton. She has participated (and coordinated most of them) in more than 30 national and ten international research projects as well as 20 contracts with the industry, all of them related to wireless communications. She has published around 150 papers in international journals and conference proceedings and holds four patents. She has contributed to international standards organizations, such as ITU and ETSI. She is a member of the Expert Group of the European 5G PPP and a member of the Advisory Committee 5JAC of the ESA as an expert appointed by Spain on 5G. She was awarded the third place by Bell Labs Prize 2014 for shaping the future of information and communications technology. She has served on the editorial board of several journals, such as IEEE Communications Letters and IEEE Transactions on Communications. She is the Secretary of the IEEE ComSoc SPCC Committee. She was the Secretary and the Chair of the IEEE ComSoc Women in Communications Engineering Standing Committee, from 2016 to 2017 and from 2018 to 2019, respectively.

J. J. ESCUDERO-GARZÁS received the Ph.D. degree in electrical engineering from the Universidad Carlos III de Madrid (UC3M). From 1997 to 2002, he was a Provisioning Engineer with Spanish telcos and the Head of the Communications Network Maintenance Department. He has been a Faculty Member with the Department of Signal Theory and Communications, UC3M. He was a Postdoctoral Fellow with the Universitat Autónoma de Barcelona, from 2010 to 2012. He was also a Researcher with the University of Virginia, in 2012; the University of Florida, in 2017; and Texas A&M University, from 2018 to 2019. He is currently a Senior Researcher with the Galician Research and Development Center in Advanced Telecommunications (GRADIANT). His research interests include wireless communication systems and resource management for 5G and beyond 5G networks.