Hybrid Computational Approach applied to the Positioning of Small Cells in 5G Networks considering Real Data

Flávio Henry Ferreira Corresp., 1 , **Fabrício Barros** 1 , **Miércio Alcantara Neto** 1 , **Evelin Cardoso** 2 , **Carlos Renato Francês** 1 , **Jasmine Araújo** Corresp. 1

1 PostGraduate Program in Electrical Engineering, Institute of Technology of Federal University of Pará, Federal University of Pará, BELEM, PA, Brasil

2 Computer Systems Department, Federal Rural University of the Amazon, Capitão Poço, Pará, Brasil

Corresponding Authors: Flávio Henry Ferreira, Jasmine Araújo
Email address: henryferreira014@gmail.com, jasmine@ufpa.br

One of the key technologies in smart cities is the use of next generation networks such as 5G Networks. Mainly because this new mobile technology offers massive connections in dense populated areas in smart cities, thus playing a crucial role for numerous subscribers anytime and anywhere. Indeed, all the most important infrastructure to promote a connected world is being related to next generation networks. Specifically, the Small Cells transmitters is one of the 5G tecnologies more relevant to provide more connections and to attend the high demand in smart cities. . In this paper, a smart small cell positioning is proposed in the context of a smart city. The work proposal aims to do this through the development of a hybrid clustering algorithm with meta-heuristic optimizations to serve users, with real data, of a region satisfying coverage criteria. Furthermore, the problem to be solved will be the best location of the small cells, with the minimization of attenuation between the Base Stations and its users. The possibilities of using multi-objective optimization algorithms based on bioinspired computing, such as Flower Polination and Cuckoo Search, will be verified. It will also be analyzed by simulation which power values would allow the continuity of the service with emphasis on three 5G spectrums used around the world: 700 MHz, 2.3 GHz and 3.5 GHz.

Hybrid Computational Approach Applied to the Positioning of Small Cells in 5G Networks Considering Real Data

Flávio Henry Ferreira[1], Fabrício Barros[2], Miércio Alcântara Neto[3], Evelin Cardoso[4], Renato Francês[5], and Jasmine Araújo[6]

[1]**henryferreira014@gmail.com**
[2]**fjbbrito@gmail.com**
[3]**miercio@ufpa.br**
[4]**evelin.cardoso@ufra.edu.br**
[5]**renato.frances2010@gmail.com**

Corresponding author:
Jasmine Araújo[6]

Email address: jasmine@ufpa.br

ABSTRACT

One of the key technologies in smart cities is the use of next generation networks such as 5G Networks . Mainly because this new mobile technology offers massive connections in dense populated areas in smart cities, thus playing a crucial role for numerous subscribers anytime and anywhere. Indeed, all the most important infrastructure to promote a connected world is being related to next generation networks. Specifically, the Small Cells transmitters is one of the 5G tecnologies more relevant to provide more connections and to attend the high demand in smart cities. . In this paper, a smart small cell positioning is proposed in the context of a smart city. The work proposal aims to do this through the development of a hybrid clustering algorithm with meta-heuristic optimizations to serve users, with real data, of a region satisfying coverage criteria. Furthermore, the problem to be solved will be the best location of the small cells, with the minimization of attenuation between the Base Stations and its users. The possibilities of using multi-objective optimization algorithms based on bioinspired computing, such as Flower Polination and Cuckoo Search, will be verified. It will also be analyzed by simulation which power values would allow the continuity of the service with emphasis on three 5G spectrums used around the world: 700 MHz, 2.3 GHz and 3.5 GHz.

INTRODUCTION

The emerging concept of Smart Cities is being promoted with the deployment of 5G technologies. Considering the increasing amount of industrialization with urbanization, the huge demand for resources and their ubiquitous use are catalyzing the emergence deployment of smart city technologies and applications. All urbanization enablers like transportation and mobility, health care, natural resources, electricity and energy, homes and buildings, commerce and retail, society and workplace, industry, agriculture and the environment, are hugely dependant of a suitable communication and safe capabilities to support these smart cities application domains.

Meanwhile, the fifth generation of wireless communications, the 5G system, is currently being integrated, with an extensive range of applications and frequency channels for its operation. According to Dahlman et al. (2014), 5G operation aims at a 1,000 times greater traffic capacity and a pulled bandwidth capacity capable of working with a latency response of 1 ms with data rates in the order of 1 Gbps to 10 Gbps. The development of 5G systems is divided into indoor and outdoor spreads. Generally, the sub-6 GHz band is applied outdoors because it is easier to transmit and propagate, and communication companies are already testing and applying 5G systems in this band for commercial purposes, in different parts of the world, as reports from Viavi (2022) have shown. In Brazil, it is no different - all major operators already have 5G systems in operation for several capitals in the country. AgênciaBrasil (2021) and Reuters (2021) inform that the main frequencies to operate 5G in the country, according to government body Anatel, are 700 MHz, 2.5 GHz and 3.5 GHz for outdoor systems, and 26 GHz for indoor deployments.

Along the difficulties tied to the deployment of 5G heterogeneous networks (5G HetNet), are the challenges to optimize their user coverage and user capacity, allowing for an increased number of services provided whilst keeping network costs low. And, as Rappaport et al. (2013) has stated, 5G coverage should be available everywhere, to anyone. That is, user coverage needs to be as closer to a hundred percent as ever.

And so, a number of studies and surveys have dealt with the necessity of coverage optimization. Agiwal et al. (2021), for instance, dedicates a whole survey on the applications of 4G-5G inter-operations, and how those can be better achieved. This is because implementation of 5G networks till this day are much costing, and changes cannot be applied overnight. Meanwhile, 4G-LTE cells can provide service coverage while 5G is still expanding. Another survey that is worth noticing, written by Shayea et al. (2020), focuses on user mobility management and how user equipments (UEs) are prone to disconnect if there are no satisfactory solutions to coverage, capacity and handoff problems and challenges.

The study herein proposed aims to provide a solution to the coverage problem for future 5G network applications, focusing especially in the range of Small Cells. The positioning of Small Cells is a key concept of densification offering a potential solution for the ultra-dense traffic in Smart Cities. Otherwise, to add to the traditional cell planning in this work two types of computational intelligence techniques will be tested, i.e., metaheuristic optimization through the utilization of a bioinspired computing algorithm (BIC) and a clustering technique. It is possible to group a set of users into an intelligent network coverage system, that aims to not only optimize the number of Small Cells but also deal with energy efficiency measures (such as controlling the transmitted power used in the cells).

Bioinspired computational methods are mainly based on natural selection. They are set to mimic the natural behavior of nature, in which the best and most surviving individuals prevail. With that in mind, these bioinspired algorithms serve as good optimization methods for mathematical and engineering problems, especially those where metaheuristic techniques (trial and error) can be applied to achieve one or more concrete goals. They have been applied to a multitude of areas where non-linear, multimodal optimization is required. E.g. Li et al. (2022) cite some areas of robotics where they might be useful, Nguyen et al. (2020) exposes some challenges in smart energy management that can be overcome with bioinspired solutions, and Gill and Buyya (2019) shows that some of them are even used on big data analysis and as aid to digitization of important documents into digital libraries.

The clustering method chosen for this application is K-Means, which is extensively utilized in the literature for its simplicity and efficiency Ahmed et al. (2020). As for the bioinspired methods, two are to be tested in conjunction with K-Means: the Cuckoo Search (CS) and the Flower Pollination Algorithm (FPA). As the aim of the work is to both maximize user coverage and minimize transmitted power, it is needed to use their multi-objective counterparts (that is, MOCS and MOFPA).

By using real data from open and free database OpenCellid, see Khan et al. (2020), it is possible to pre-select cells with the greatest amount of user traffic in 4G-LTE in order to plan out how future 5G Small Cells shall behave in order to provide good coverage of service to users. More details about this are to be explained in the Methodology section.

The proposed hybridization of clustering and bioinspired algorithms is to be tested in simulations to determine an optimal user coverage for the three aforementioned frequencies that have been auctioned to operate in Brazil: 700 MHz, 2.3 GHz and 3.5 GHz. In total, two hybrid algorithms have been produced: MOCS + K-Means (MOCS-KM) and MOFPA + K-Means (MOFPA-KM).

The main contributions of this study are the provision of a fairly simple metaheuristic method to achieve optimal network coverage with low-power Small Cells, and to provide data that can be adapted to an densely urban but with a rainforest climate such as the city of Belém, Brazil - which is where our OpenCellid data is from. A considerable area of the city has been selected to test the intelligent UE clustering simulations for 5G Small Cell implementation.

The paper is organized into the following sections: Related Works discourses about some of the state-of-the-art solutions for coverage and capacity optimization for 5G as well as bioinspired/clustering algorithm hybrids; Methodology explains how the study was conducted, the theory behind the algorithms utilized and gives information on the propagation model chosen for path loss and user coverage modeling; Results demonstrate the simulation of the algorithms for the different frequency ranges; and Conclusions expose our final considerations of the study.

RELATED WORKS

This section is to be divided into two topics: Coverage and Capacity Optimizations for applications in 5G systems, and Bioinspired Algorithm Hybridizations focusing on clustering of data.

Coverage and Capacity Optimization of Future 5G Networks

An analytical study on the Coverage, Handoff and Cost optimization for Heterogeneous 5G Networks (5G-HetNet) has been written by Ouamri et al. (2020). A path loss model is suggested for both Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) situations, with different path loss exponents, and the handling of coverage and handoff probability is done by Stochastic Geometry with values in SINR (Signal to Interference-Plus-Noise Ratio). By the method of Cellular Network Planning (CNP), network investment cost issues are optimized.

Khan et al. (2020) talks about a heuristic method to predict the employment of 5G base-stations in Spain by obtaining user and traffic data from LTE networks, using the free database of OpenCellid. This dimensioning of a 5G network takes into consideration the various aspects of this technology: heterogeneous architecture, the necessity for a high Quality of Experience (QoE) and smart resource allocation. Their objectives are achieved by separating the highest traffic areas, and then deciding where 5G cells are to be deployed.

Khan et al. (2022), however, improves on their model of 5G planning by assigning a clustering algorithm to the task of deploying 5G base-stations into high data traffic areas. A K-Means algorithm has been utilized, using the Elbow heuristic as a benchmark. Not only does it deal with the coverage characteristics of the network, but also demonstrates an extensive study on its capacity dimensioning. The goal of the study is to decrease the network cost, as well as provide a more robust way to interpret the data for network planning decision.

Another paper that talks about coverage area optimization for 5G is found in Wang et al. (2020). It is mathematically complex, as the goal is to provide numerical solutions for the employment of Macro and Small Cells in 5G. Therefore, different clustering approaches are suggested by the authors, as well as models to prevent noise and interference. Numerical case examples are given for different scenarios and, as it deals with heterogeneous networks, the coverage simulations can be seen with generally one or two Macro Cells and several Small Cells around them.

Bioinspired Algorithm Hybridizations for Data Clustering

Jensi and Jiji (2015) have accomplished to generate a FPA hybrid with K-Means, a similar method to the study presented herein. In the paper, they present this technique to test and clusterize eight different datasets, with promising results. It is attested that the hybridization has given better average fitness results for all tests, thus proving it to be more efficient.

However, the hybridization process shown in the paper takes the current best solution in the FPA and executes a local search around it using the K-Means. This is different from the implementation we have provided, to be further explained in the Methodology section.

Hatamlou (2017) is another paper that deals with nature-inspired algorithms, with their proposed technique of utilizing Particle Swarm Optimization (PSO) with the Big Bang-Big Crunch (BB-BC). The objective, like in Jensi and Jiji (2015), is to provide an efficient method of data clustering, and both of them use five equal datasets (Iris, Wine, Glass, CMC, Cancer) in order to test their techniques. However, for all the same datasets, the PSO-BB-BC hybrid has shown better average fitness than its FPA-KM counterpart.

Another robust work in the subject is found in Logesh et al. (2020). Based on TripAdvisor travel recommendations data, the authors propose several hybrid solutions. These are based on swarm intelligence algorithms, such as an improved PSO model, Brainstorm Optimization (BSO), Quantum-Behaved Brainstorm Optimization (QBSO) and the Immune Genetic (IG) technique. In most cases, the best optimizer is the novel QBSO-IG hybrid that they have proposed themselves. This study also provides many details on recommendation systems and draws a comparison on many other papers that have dealt with its data clustering optimizations.

Khan et al. (2019) have proposed a clustering technique for Flying Ad-Hoc Networks (FANETs), named Bio-Inspired Clustering Scheme for FANETs (BICSF), that consist of organizing drones in the air based on their positioning and energy management. The goal is to reduce energy consumption, elevate the air time of the devices and minimize the time that the drones take to organize themselves into clusters.

The hybrid is between the glowworm swarm optimization (GSO) and krill herd (KH) algorithms, and it is tested against more commonly found bioinspired ones - Grey Wolf Optimization (GWO) and Ant Colony Optimization (ACO), to specify.

Pitchaimanickam and Murugaboopathi (2020) also proposes a hybrid approach for network clustering, however, with an emphasis on ameliorating battery life and information management in Wireless Sensor Networks. The creation of clusters itself is done by a LEACH-C algorithm, which is then optimized with a conjuction of a PSO and the Firefly Algorithm (FA). HFAPSO, as the hybridization is abbreviated, proves to be considerably more efficient than the separated techniques in a simulation of 100 sensors and an area of (100x100) m^2. Results have shown that not only does it better coordinates the clusters, but it sustains the battery life of sensor nodes for longer.

Cao et al. (2021) demonstrates a heterogeneous Wireless Sensor Network (WSN) coverage area optimization, in which a Chaos-Improved Social Spider Optimization (CSSO) is used to accomplish the task. The solution aims to deploy the sensors by saving as much energy consumption and network cost as possible, as well as reducing coverage redundancy and trying to cover blind spots as best as possible.

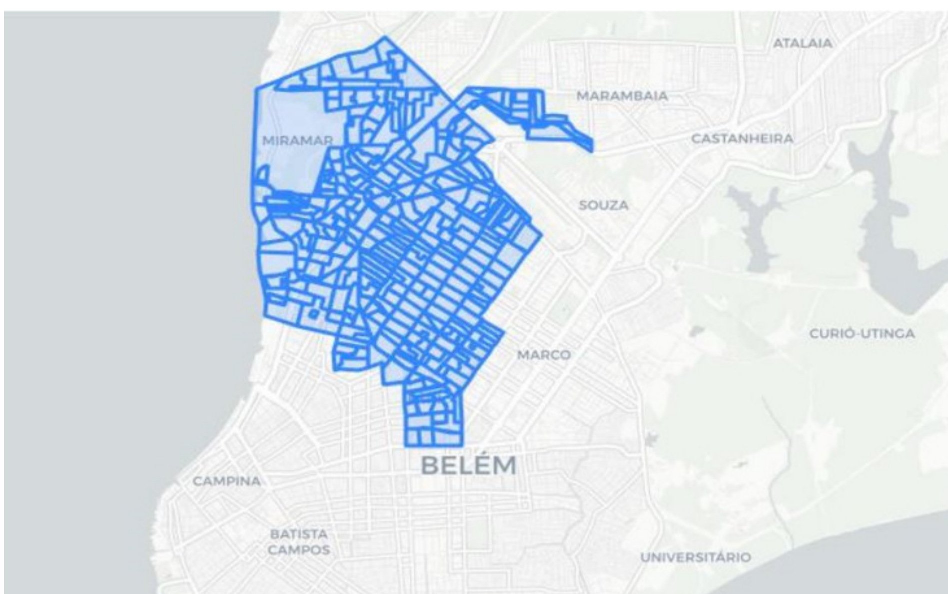
Lastly, a study that also uses K-Means in a hybrid bioinspired computing setting is conducted by Aswani et al. (2018). It is combined with a Firefly Algorithm, but also with a modified, Lévy flight and chaos-induced version of it (LFA-chaos). This clustering technique is then tested against a Fuzzy C-Means (FCM) algorithm, and its application is to find spam accounts on social media website Twitter. It uses many significant behavioral factors of the user as input to solve this problem, such as its hashtag frequency, number of mentions, tweet count, follower count and many more, up to a total of 13. The Lévy flight applied to this FA is similar to the one used in the FPA exposed in the Methodology section.

METHODOLOGY

In this section, details about the objectives, the methods and optimization techniques used in this study are elucidated.

In general terms, the aim of the study presented herein is to acquire real-life LTE user data from OpenCellid to aid in the deployment of 5G cells in a four-neighborhood area from the city of Belém, Brazil. The involved neighborhood names are Sacramenta, Barreiro, Miramar and Pedreira, and the area is represented in Figure 1.

Figure 1. Area of interest inside Belém, Brazil.



For that matter, by selecting the LTE cells with the greater traffic according to OpenCellid data, and inspired by the data mining process of Khan et al. (2022), a total of 19 LTE cells were selected to provide the highest data traffic areas and aid our deployment proposal. Therefore, by taking learning data from these cells, we have made a hybrid clustering/bioinspired technique that aims to solve the coverage problem in 5G frequency bands of 700 MHz, 2.3 GHz and 3.5 GHz.

Two problems to be optimized show up when thinking of providing coverage for a medium-to-high density urban area so large such as this: the number of k cells (or clusters of users) that is required to achieve a good user coverage percentage, and also, for matters of energy efficiency, how much power P_t should be applied to these cells in order to reach the best coverage by spending less energy. Given that our approach aims to minimize both these variables, network costs should also be minimized throughout the process. So, a simulation with an N number of users normally distributed in the aforementioned area is drawn to test how much coverage the proposed metaheuristic setup will achieve.

The Methodology from here is separated in various subsections. Firstly, the bioinspired algorithms shall be revised, then an explanation on the K-Means clustering algorithm is given. Lastly, the hybridization process is explained, as well as the path loss propagation model for producing the objective functions.

In order to the utilize multi-objective functions, the simplest way is to involve all objectives into a single mathematical sentence, as shown in (1):

$$f = \sum_{i=1}^{n} w_i f_i, \sum_{i=1}^{n} w_i = 1, \tag{1}$$

in which w_i are the weights given to each objective, f_i are the single objectives and the sum of all weights must be equal to 1.

Multi-Objective Cuckoo Search (MOCS)

The Cuckoo Search algorithm (CS) has been coined by Yang and Deb (2009), proving itself to be a very effective metaheuristic algorithm for all sorts of applications in mathematics, industry and engineering, according to Shehab et al. (2017). The main bioinspired idea behind this method is the computational modeling of the parasitic behavior of cuckoo-type birds, that often lay their eggs inside the nests of other types of birds. This is a natural occurrence in nature, as the host species often does not perceive the cuckoo's "alien" egg inside the nest, or either choose to ignore it completely or abandon the nest altogether, choosing another place to lay its eggs on. However, if the egg is ignored and left to grow inside the host bird's nest, the cuckoo hatchling is born, and reaches maturity much faster than the other eggs, pushing the host bird's eggs outward. Thus, the cuckoo baby bird expels the other eggs from the nest, resulting in a higher food share for it, and becoming well-fed.

Thus, the whole process is based upon three major rules:

1. Each cuckoo lays one egg at a time, and deposits it in a random nest. Each egg is considered a potential solution – metaheuristic

2. The best nests carry the best eggs (solutions), and these will survive the next generations due the parasitic nature of the cuckoo hatchlings – elitism.

3. The number of available nests is constant, and defined by the code developer. The probability of a cuckoo egg being discovered by the host bird is defined as $Pa \in [0,1)$. After this, the bird may choose to discard this egg or abandon its nest – discarding the worst solutions.

The latter rule can also be described by indicating that a probability fraction Pa from the various n nests of host birds are replaced by new nests, presenting randomized solutions.

The cuckoo birds in this method move according to the so-called Lévy Flights. This device provides a random flight path that which each cuckoo (in a number of i cuckoo birds) will trek to find nests. Equations (2) and (3) are a mathematical representation of the Lévy Flight and Lévy Distribution, respectively, to be implemented in code form:

$$X_i^{t+1} = X_i^t + \alpha \oplus Levy(\beta) \tag{2}$$

$$Levy(\beta) \cong u = t^{-(\beta+1)}; (1 < \beta \leq 3) \tag{3}$$

In which i is the maximum number of cuckoo birds in the current generation, and t is the current code iteration. Constant α is the step size to be utilized in the code, and is adaptable to the developer's need – it must always be greater than zero, and in this study the value is set to $\alpha = 1$.

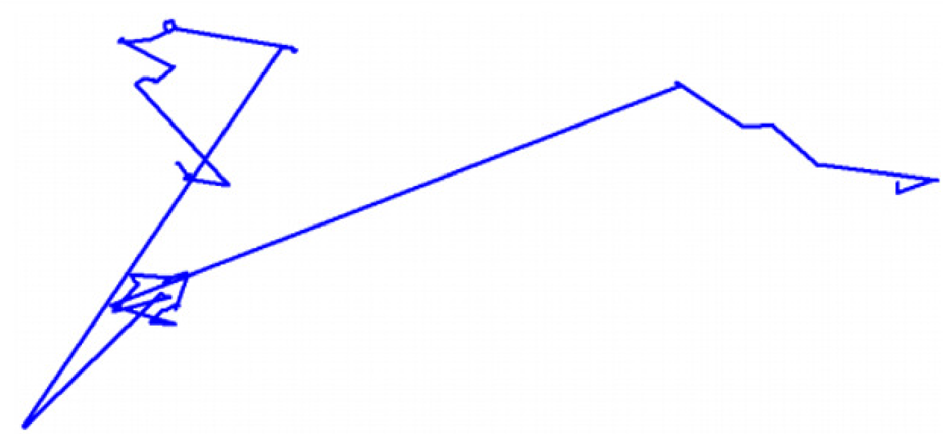
In equation (2), the Lévy Distribution is associated to the Lévy Flight with the product $\oplus$, which means "entrywise multiplication" – a kind of product between two matrices of the same size. In the PSO algorithm, a similar product can be found, however, for the Lévy Flight method, the search space can be much better harnessed.

As for equation (3), this is about the Lévy Distribution. It possesses infinite variance and average values. The variable β is the random step length, needed for providing a variable magnitude to the random walk performed by the Lévy Flight.

Given that the Lévy distribution presents infinite variance, the search space is virtually limitless, meaning that the length of the flight taken could be very short or incredibly long. However, generally, the new solutions are generated through the Lévy method around the best obtained solution on a given instant, accelerating the process and concentrating the computational effort in a part of the search space. Oftentimes, however, solutions are generated randomly across the space – this is good to prevent the algorithm from getting stuck in local optima, which are not the globally best possible solution.

An illustration showing a Lévy flight with around a hundred steps can be found in 2.

Figure 2. Representation of a 100-step Lévy flight in its search space



A pseudocode of the CS algorithm, as well as its multi-objective counterpart, can be found in the article in which it was proposed by Yang and Deb (2009). Below, in algorithm 1, a transcription of this code is presented.

Algorithm 1 Multi-Objective Cuckoo Search Algorithm

Define the objective functions as $f_n(x), x = (x_1, ..., X_d)^t$
Generate the initial population of n host nests $x_i (i = 1, 2, ..., n)$

while ($t <$ number of iterations) or (stop criterion) **do**
  Select a cuckoo randomly via Lévy Flight (2)
  Evaluate the cuckoo's fitness (represented by F)
  Draft a random nest (say, j) out of the n available

  if $F_i < F_j$ **then**
    Replace randomly drafted j by the new solution x_i
  end if

  Discard a fraction P_a of worse nests and build new ones
  Keep the best / better quality solutions
  Rank the best nest set and find the current best

end while

Post-process and visualize results

Multi-Objective Flower Pollination (MOFPA)

A solution based on the behavior of natural flower pollinators has been proposed by Yang (2012), the same author of the cuckoo search, which also utilizes the Lévy flight method of optimal space search. Its efficiency, in many single and multi-objective applications is proven to be greater than Particle Swarm and Genetic optimization algorithms. Its multi-objective counterpart has been proposed in 2013, also by Yang et al. (2013).

As for the stages of the algorithm, four rules are defined thusly:

1. Biotic and cross-pollination is considered as global pollination process with pollen-carrying pollinators performing Lévy flights.
2. Abiotic and self-pollination are considered as local pollination.
3. Flower constancy can be considered as the reproduction probability and is proportional to the similarity of two flowers involved.
4. Local pollination and global pollination is controlled by a switch probability p $\in$ [0, 1]. Due to the physical proximity and other factors such as wind, local pollination can have a significant fraction p in the overall pollination activities.

Two kinds of pollination are considered and simulated: global and local pollination. This assures that the code does not only fall for local solutions, healthily seeking to encounter a global solution to the objectives. For simplicity manners, the algorithm is based on the idea that every plant possesses only one flower and can pollinate also just one other flower at a time, when in true biological terms they can hold a few flowers and millions of pollinating gametes. This is so that a plant / flower / pollinating gamete are all considered to be part of one solution altogether.

Hence, the first rule (global pollination) and third rule (flower constancy) of FPA are mathematically represented as shown in (4):

$$X_i^{t+1} = X_i^t + L(X_i^t - g*) \tag{4}$$

Where X_i^t is the pollen i or solution vector X_i at iteration t, and g* is the current best solution among all solutions at the current iteration.

Pollination strength L is dealt via Lévy flight, in which it is a measure of each flight's step size, as denoted in (5). Here, the flights symbolize the path of insects and pollinator animals in a given area -

in the algorithm, this area is the optimization's global search space. However, the equation used in this algorithm differs from the one found in cuckoo flights, as it is based on producing Lévy flights via the Mantegna algorithm. This is basically a technique to generate pseudo-random step sizes via normal distributions in order to provide an optimal performance whilst still maintaining the demands of the Lévy distribution.

$$L \approx \frac{\lambda \Gamma(\lambda) \sin(\pi\lambda/2)}{\pi s^{1+\lambda}}, \tag{5}$$

in which $\Gamma(\lambda)$ is the standard, classic gamma function found in Lévy flights, and other probabilistic and complex number applications.

The Mantegna step size algorithm can be explained as the equation (6).

$$s = \frac{U}{|V|^{\frac{1}{\lambda}}}, \tag{6}$$

with s being the step size, U being drawn from a Gaussian distribution of variance σ^2 and V also being drawn from a Gaussian distribution but with unitary variance, as can be verified in Yang et al. (2014). Generally, the lambda is treated as a parametric value and it is safe to assume that it is a constant with a possible value of around $\lambda \in [0.5, 1.5]$. When $\lambda = 1$, the variance also equals 1, and results are in such case easier to predict.

For the purpose of Rule 2 (local pollination), the flower constancy is mimicked for a limited neighborhood near to the reproductive flower's position. It is represented as

$$X_i^{t+1} = X_i^t + \varepsilon(X_j^t - X_k^t), \tag{7}$$

where X_j^t and X_k^t are pollens from different flowers of the same plant species.

The fourth rule is a probabilistic switch between global and local pollination, and the probability p can be parametrically and singularly adjusted to improve optimization performance, depending on the need of the objective function.

All stages of the algorithm are represented in pseudocode form by the recommendations in Yang (2012), which are transcribed in algorithm 2. Some details previously discussed can be noticed, such as an if/else switch for global and local pollination, which are done by Lévy flights and random selection, respectively.

Algorithm 2 Multi-Objective Flower Pollination Algorithm

```
Define the objective functions as f_n(x), x = (x_1, ..., X_d)^t
Generate an initial population of flowers/pollens as random solutions
Find, within this population, the best solution g_*
Define the switch probability p ∈ [0,1]

while (t < number of iterations) or (stop criterion) do
    for all n flowers in the population do
        if rand < p then
            Generate a step vector L which obeys a Lévy distribution
            Execute global pollination as in (4)
        else
            Pick a uniformly distributed number from ε = [0,1]
            Randomly choose individuals j and k from all solutions
            Do local pollination according to (7)
        end if

        Evaluate new solutions
        Update solutions that are better into the population

    end for

    Find the best current solution, represented by g_*

end while
Post-process and visualize results
```

The K-Means Algorithm

An unsupervised learning method of computational intelligence, the K-Means clustering algorithm has the goal of grouping similar data points. Each of those k amount of clusters, then, possesses a centroid, which is the mean value M of all positions of the data points. Due to being fast and easy to reproduce, it is one of the most popular clustering algorithms. Sinaga and Yang (2020) defines the objective function of K-Means according to the following equation:

$$J(A,Z) = \sum_{i=1}^{n} \sum_{k=1}^{c} z_{ik} ||x_i - a_k||^2 \tag{8}$$

in which x_i is a data point i belonging to a dataset $X = \{x_1, x_2, ..., x_n\}$ spread over an Euclidean space $\mathbb{R}^d$ of d-dimensions, $a_k \in A = \{a_1, a_2, ..., a_n\}$ is the centroid of the k-th cluster and z_{ik} is a binary variable (either 0 or 1) that signals if the user i belongs to cluster k. Given that the objective function needs to be minimized, the algorithm then proceeds to alter the centroids by also minimizing the Euclidean distance between them and the data points that belong to them, according to (9):

$$a_k = \frac{\sum_{i=1}^{n} z_{ik} x_i}{\sum_{i=1}^{n} z_{ik}}, \quad z_{ik} = \begin{cases} 1, & \text{if } ||x_i - a_k||^2 = \min_{(1 \le k \le c)} ||x_i - a_k||^2. \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where $||x_i - a_k||^2$ is the Euclidean distance.

The ECC-33 Propagation Model

A channel path loss model is required to evaluate which users are connected to the network. This propagation model should both belong to a frequency range where the 5G-NR channels are going to be propagated, as well as adapt to the environment in which it shall be applied. For those reasons, an ECC-33 model was chosen, which is adapted to the urban but forested environments of the city of Belém, as proposed by de Carvalho et al. (2021).

ECC-33 is derived from the famous Okumura-Hata model, but taking into consideration the behavior of higher frequencies, thus extending the frequency range. In Mollel and Michael (2014), experimental results have been drawn by comparing different path loss models conducted in Dar es Salaam, Tanzania. The COST-231 and ECC-33 models had the lesser Root Mean Square Error (RMSE), meaning that - for urban, suburban and rural settings - they had the most approximate results to measured signal data.

Therefore, the model used in this work is defined thusly:

$$PL_{ECC} = A_{fs} + A_{bm} - GF_t - GF_r \tag{10}$$

in which A_{fs} is the free-space attenuation, A_{bm} is the base median path loss and GF_t and GF_r are the transmitting antenna gain factor and received antenna gain factor, respectively. The values used for the model correspond to the large city scenario, as all acquired data correspond to an urban region of medium-to-high density. The following equations (11), (12), (13) and (14) further detail the variables of the model:

$$A_{fs} = 92.4 + 20log(d) + 20log(f_c) \tag{11}$$

$$A_{bm} = 20.41 + 9.83log(d) + 7.89log_(f_c) + 9.56(log(f_c))^2 \tag{12}$$

$$GF_t = log(\frac{h_{BS}}{200})(13.958 + 5.8log(d))^2 \tag{13}$$

$$GF_r = 0.759(h_{UE}) - 1.862 \tag{14}$$

in which f_c is the central frequency, h_{BS} is the base-station antenna height, h_{UE} is the user device's antenna height. While d is the distance between a user and a base-station, that in this study is calculated by the haversine distance formula, as it appears on Gade (2010) and Chopde and Nichat (2013):

$$d = 2R \arcsin \sqrt{sin^2(\frac{\mu_2 - \mu_1}{2}) + cos(\mu_1)cos(\mu_2)sin^2(\frac{\lambda_2 - \lambda_1}{2})} \tag{15}$$

where λ are the latitudinal positions, μ are the longitudinal positions, and R is the approximate radius of the Earth ($6.371.10^6$ meters). Furthermore, the received power for an user is represented in (16):

$$P_r = P_t - PL_{ECC} + G_t + G_r, \tag{16}$$

in which G_t, G_r are the transmitting (Tx) antenna gain and receiving (Rx) antenna gain.

The Hybridization Process

In real-life scenarios, not often is the number of clusters to achieve optimal results is known beforehand. For that, there are some heuristic methods that can be utilized to get approximations of how many clusters one might need to satisfy an application. For instance, Khan et al. (2022) have chosen the Elbow heuristic. However, given that one might need a more complex and sophisticated way of deciding how to cluster, and how much to cluster, like in a scenario of network dimensioning and coverage area optimization, these heuristics may not provide optimal solutions. This is why a metaheuristic method, like a bioinspired algorithm, is more suited for this task. The hybridization of a clustering algorithm like K-Means and a metaheurisitc BIC means that both of those methods benefit from each other. The K-Means provides the base-station coordinates on a given area, and the BIC provides the values of the ideal number of clusters, taking into consideration multiple inputs that will decide how much clustering there must be. Figure 3 demonstrates the whole process of the study. Please notice how the K-Means processes a geographical

Algorithm 3 The Hybrid K-Means + BIC Algorithm

Define the constant values of h_{BS}, h_{UE}, f_c and antenna gains
Initialize one of the bioinspired optimization techniques (MOCS or MOFPA)
Generate an initial population with random solutions for k clusters and P_{BSt} transmitted power

while (Iterations $\leq$ Max Iterations) **do**

  Initialize the K-Means for each k given by the BIC

  for (all the clusters k in the K-Means) **do**
    for (all the users associated to cluster k) **do**
      Calculate the distance between the user and its clusters' centroid via (15)
      Calculate the ECC path loss, present in (10)
      Calculate the received power P_r as in (16), with P_{BSt} given by the BIC

      if ($P_r \geq -90$dBm) **then**
        Count the user as within the coverage area
      end if

    end for
  end for

  Calculate the percentage of connected users in relation to the total number of users

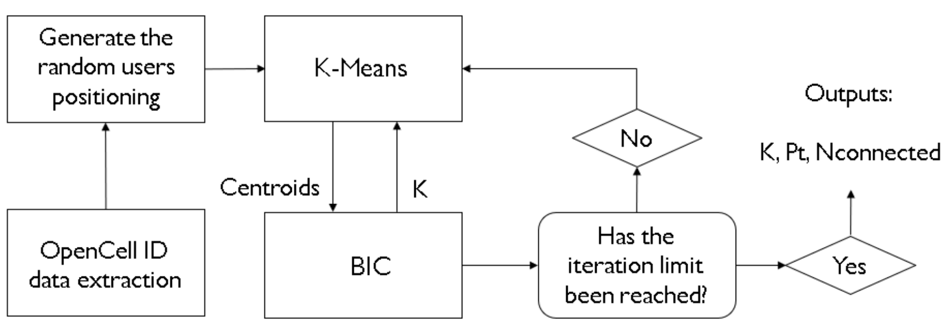
  Evaluate, with the BIC, the outputs in a single objective function
  Update the best solutions into the population

end while

Post-process and visualize results

position for cluster centers and the BIC returns it a number of clusters to attempt and see if an optimal result is reached. This process has the goal, in our study, to provide an optimal value of k clusters, as well as a satisfactory transmitted power value.

Figure 3. Flowchart of the processes involved in this study



A pseudocode displaying the structure of the hybridization can be found in 3. For each iteration of the BIC, the K-Means will run alongside it, providing cluster center (or centroid) values that the BIC shall take and calculate the received power for each and every user, one cluster at a time. If the user has a received power of -90 dBm or greater, it will be considered as connected to the network and, thus, covered.

RESULTS

A total of 2000 users have been randomly generated within the simulated area where the Small Cells shall actuate. This is to ensure that there are users in every corner of the area, and to assure that the clusters can manage to provide coverage for them.

It is worth to denote that the received power threshold to consider a user as connected to its respective cluster is $P_r \geq -90$dBm. This value has been chosen as a reasonable target for 5G mobile, heterogeneous networks to main good signal strength and capacity. Furthermore, as per stated in Ayad et al. (2022), values of between $-100 \leq P_r \leq -80$ dBm are acceptable for this kind of application.

The objective function for the simulations can be defined as:

$$f = 0.7N_{outage} + 0.3P_{diff}, \quad N_{outage} = \frac{N_{users} - N_{connected}}{N_{users}}, \quad P_{diff} = P_t - 30, \tag{17}$$

in which P_{diff} is the difference of transmitted power to the ideal minimum value of 30 dBm, N_{users} is the total number of users to be attended by the network, $N_{connected}$ are the number of users connected to the Small Cells and N_{outage} is the percentage of users which are experiencing outage (i.e. disconnected from any cell). Therefore, the main objectives are to minimize the number of users that are excluded from the network whist maintaining the lowest transmitted power possible, that is, $N_{outage} = 0$ and $P_{diff} = 0$.

The search space bounds on the number of clusters and transmitting power are: $L_b = (1$ cluster, 30 dBm) and $U_b = (100$ clusters, 40 dBm). These power values for P_t are the limits generally found in Small Cells, as stated by Khan et al. (2020). So, the objective for P_t is to approach its minimum value, that is, $P_t = 30dBm$. It also can be noticed that, since the main goal of the study is to achieve better coverage for users in 5G networks, there is a 0.7-0.3 weight towards coverage over transmitted power optimization in the objective function (17).

Table 1 refers to the constant and variable values that have been coded into the algorithms and the propagation model:

Parameters	Values
Tx antenna height (h_{BS})	40 m
Rx antenna height (h_{UAV})	1.6 m
Tx antenna gain	3 dBi
Rx antenna gain	0 dBi
Central frequencies (fc)	[700 MHz, 2.3 GHz, 3.5 GHz]
Number of clusters (k) bounds	1 to 100 clusters
Transmitted power (P_t) bounds	30 to 40 dBm
Nest discard probability (for MOCS)	0.25
Switch probability (for MOFPA)	0.5

Table 1. Values of interest for the simulations

The simulation and its algorithms have been coded in MATLAB, and are run on a personal computer with these characteristics: AMD Ryzen 3 CPU with 3.7 GHz clock and 16GB DDR4 RAM.

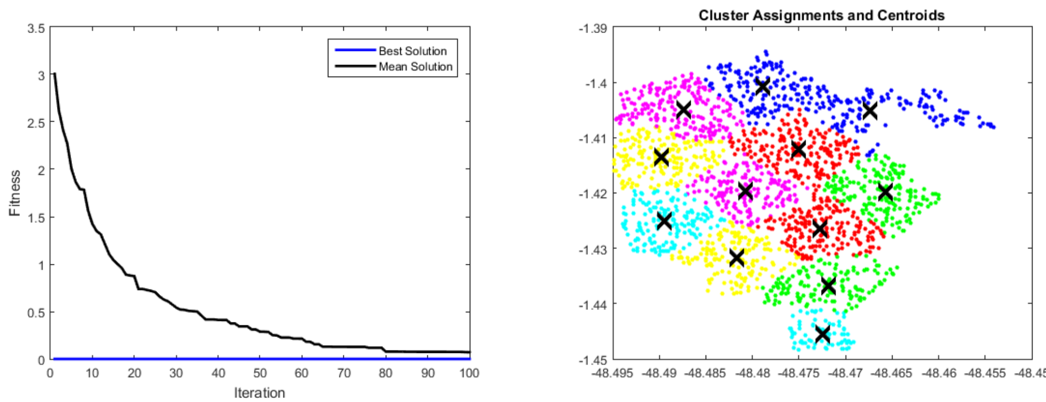
Results for 700 MHz

Firstly, Figures 4 and 5 display the results achieved for the 700 MHz frequency band for, respectively, the MOCS-KM and MOFPA-KM. This is a band that possesses a fairly low data rate and bandwidth, but since it is one of the auctioned bands for 5G operation in Brazil, it is included here as means of comparison and to test the BIC + K-Means hybrids.

The X symbols in the figures displaying the clusters are referent to the position of centroids, and the colored dots correspond to the users.

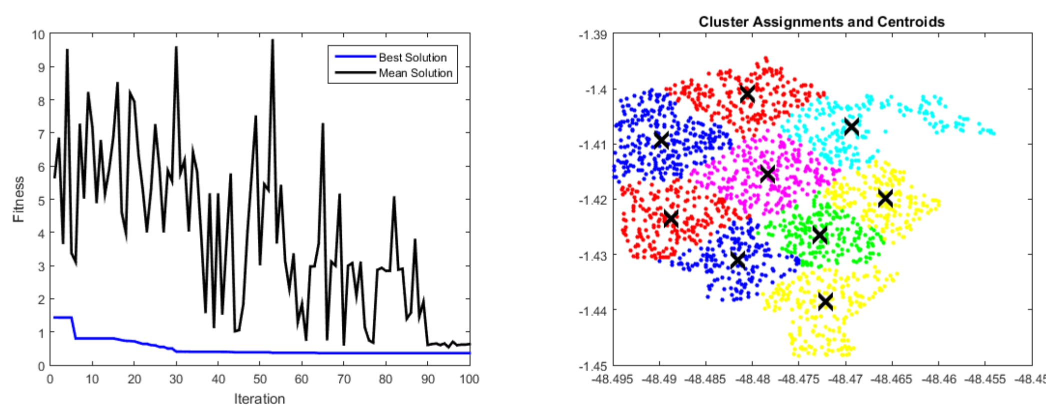
Here, the algorithm had no problems whatsoever in optimizing for zero fitness, needing only a few iterations to converge. This is because path loss is less intense for lower frequencies, and so a small group of small cells can already provide perfect coverage.

The best solver in this case is MOCS-KM, with a solution of 12 clusters and power of 30dBm. However, MOFPA-KM comes very close to a zero fitness value as well, with an inferior number of clusters of only 9 and 30 dBm power.



(a) **(b)**

Figure 4. Results for 700 MHz: (a) Fitness plot of MOCS-KM and (b) Cluster formation given by the technique

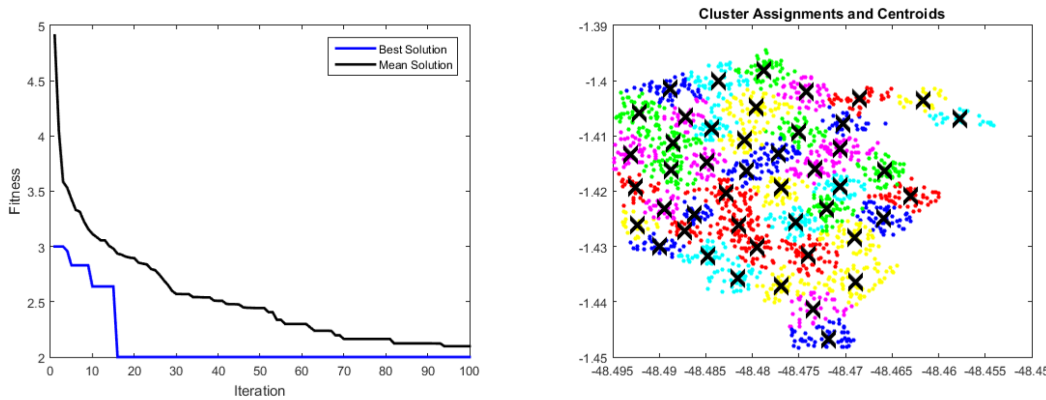


(a) **(b)**

Figure 5. Results for 700 MHz: (a) Fitness plot of MOFPA-KM and (b) Cluster formation given by the technique

Results for 2.3 GHz

The band that promises to function both in LTE and in 5G is shown next. Figures 6 and 7 are the results for the 2.3 GHz frequency band. It is expected to operate in smaller cities as a means of digital inclusion, providing more bandwidth to 4G-LTE demands, or in urban centers as a complementary traffic alternative to the main band of 5G, which is 3.5 GHz.



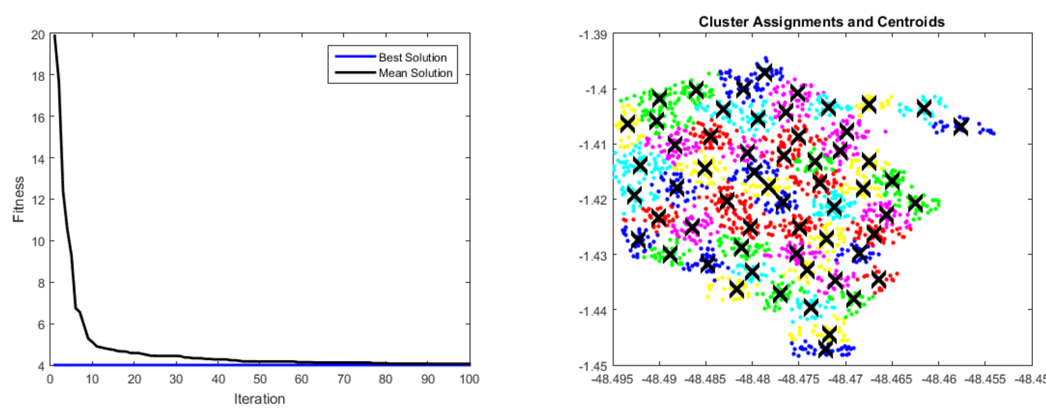
(a) **(b)**

Figure 6. Results for 2.3 GHz: (a) Fitness plot of MOCS-KM and (b) Cluster formation given by the technique

Results for this band are already harder to solve than 700 MHz, with the best fitness results coming from MOCS-KM once again. It has achieved $f_{best} = 2$, with 46 clusters and a satisfactory transmitted power of 31 dBm. MOFPA-KM has proposed a higher P_t value with less clusters, resulting in $f_{best} = 2.4267$, $k = 34$, and $P_t = 33.3352$ dBm.

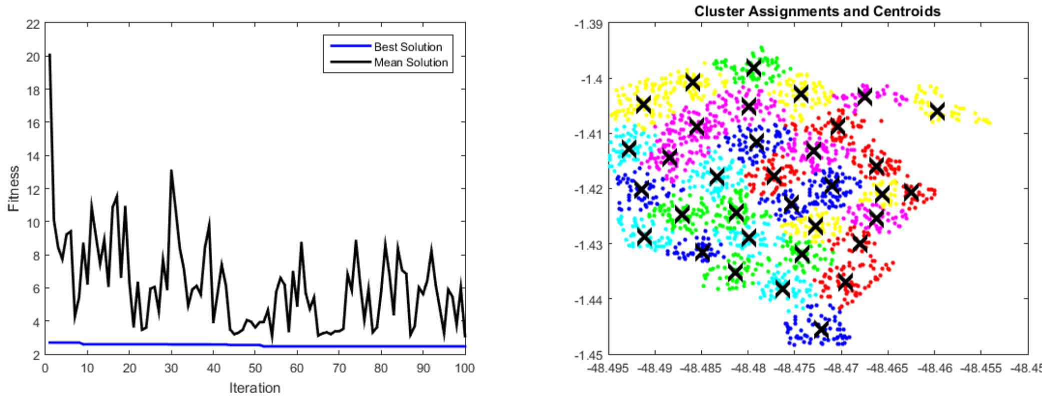
Results for 3.5 GHz

Finally, the main and most used frequency band of 5G. The 3.5 GHz spectrum possesses up to 80 MHz of band per operator in Brazil, and is set to expand in highly-urbanized areas, as it can handle greater traffic and provide faster data rates than the other two. Figures 8 and 9 demonstrate the results of the simulation for this frequency.



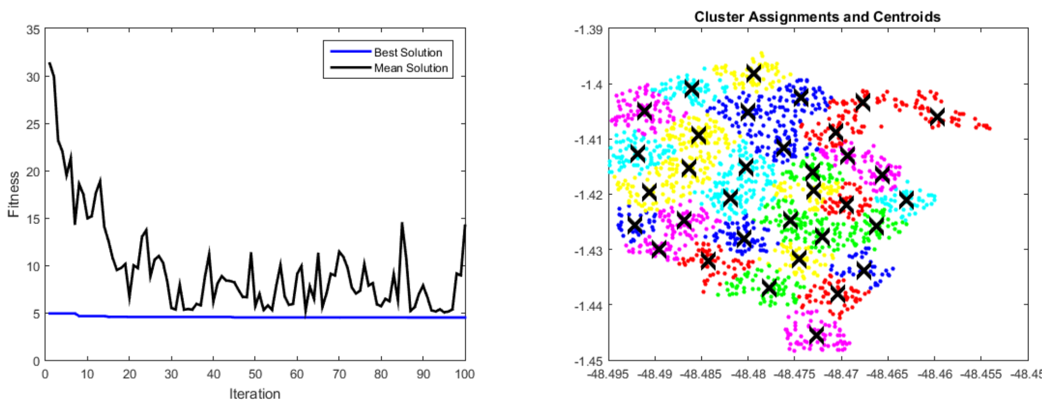
(a) **(b)**

Figure 8. Results for 3.5 GHz: (a) Fitness plot of MOCS-KM and (b) Cluster formation given by the technique



(a) (b)

Figure 7. Results for 2.3 GHz: (a) Fitness plot of MOFPA-KM and (b) Cluster formation given by the technique



(a) (b)

Figure 9. Results for 3.5 GHz: (a) Fitness plot of MOFPA-KM and (b) Cluster formation given by the technique

For this case, the results are basically the same in terms of coverage, but fitness values turn greater because of the choice in both algorithms to utilize more transmitted power instead of increasing the number of clusters too much. In terms of fitness, MOCS-KM marginally wins again, with $f_{best} = 4.006$. That being said, MOFPA-KM has converged to a solution with 34 clusters again - the same number from the 2.3 GHz simulation. It has, however, increased the transmitted power to its maximum value of 40 dBm in order to achieve that.

Table of Results

In Table 2, results for the optimization processes of MOCS-KM and MOFPA-KM hybrids are shown. For the best generated population (the one which produced f_{best}), it records the maximum, medium and best fitness values, as well as the outputs and the running time of the respective simulations.

CONCLUSIONS

Smart cities will bring forth many future challenges, and this work proposed a novel Small Cell positioning system using a hybrid approach to provide better user coverage and to save more energy. The presented scheme deals with a method of clustering and optimizing for the implementation of 5G Small Cells according to user traffic, using the city of Belém, Brazil as a simulational example. Optimization was provided by the two hybrid methods, MOCS-KM and MOFPA-KM.

Simulation	k	P_t	N_{outage}	P_{diff}	f_{max}	f_{avg}	f_{best}	Time (s)
700 MHz (MOCS-KM)	12	30	0	0	0.7892	0.1327	0	886
700 MHz (MOFPA-KM)	9	30	0.5133	0	1.9397	0.6075	0.3593	992
2.3 GHz (MOCS-KM)	46	31	2.4285	1	10.9882	2.1649	2	1542
2.3 GHz (MOFPA-KM)	34	33.3352	2.0373	3.3352	4.7515	2.7564	2.4267	1558
3.5 GHz (MOCS-KM)	59	38	2.2857	8	5.5686	4.1163	4.006	1637
3.5 GHz (MOFPA-KM)	34	40	2.0421	10	6.3954	4.7312	4.4295	2804

Table 2. Results for each best iteration of the hybrid techniques.

In general the simulations were satisfactory, as user outage is never greater than 2.5% for all cases, making the amount of connected users over 97%. Transmitted power levels have been kept within Small Cell ranges, providing good implementation opportunities. That being said, the techniques have a sort of preference that differentiate themselves. In Table 2, it can be denoted that MOCS-KM prefers to increase the amount of clusters to reduce the transmitted power, whilst MOFPA-KM is the opposite.

Therefore, the usage of said hybridizations should be a matter of preference for the network planner professional that intends to use them. For coverage optimization both present similar and adequate results, but MOFPA-KM keeps cluster numbers to a bare minimum and optimizes for greater power usage; and MOCS-KM is a bit more precise and uses less power per cluster but produces a higher amount of clustering. So, these variables are bound to limitations and cost of implementation issues that are up to which of those are less costly to produce.

Further challenges to enrich the study would be to implement more 5G capacity variables into the simulation. The focus of this paper has been mostly on coverage problems, but there is much to add in concern to capacity dimensioning according to the problems found in many papers present throughout our study, mainly in Shayea et al. (2020) and Ouamri et al. (2020). Further opportunities for research involve further minimization and measurement of network cost and maximizing throughput of data per area, as per the work of Khan et al. (2022) manages to accomplish.

ACKNOWLEDGMENTS

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel—CAPES, the National Council for Scientific and Technological Development—CNPq, and the Support Program for Qualified Production—PROPESP/UFPA (PAPQ), for supporting the research carried out by the authors of this paper.

REFERENCES

Agiwal, M., Kwon, H., Park, S., and Jin, H. (2021). A survey on 4g-5g dual connectivity: road to 5g implementation. *IEEE Access*, 9:16193–16210.

AgênciaBrasil (2021). Brazil's first 5G auction brings in $8.4 bi.

Ahmed, M., Seraj, R., and Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, 9(8):1295.

Aswani, R., Kar, A. K., and Vigneswara Ilavarasan, P. (2018). Detection of spammers in twitter marketing: a hybrid approach using social media analytics and bio inspired computing. *Information Systems Frontiers*, 20(3):515–530.

Ayad, M., Alkanhel, R., Saoudi, K., Benziane, M., Medjedoub, S., and Ghoneim, S. S. (2022). Evaluation of radio communication links of 4g systems. *Sensors*, 22(10):3923.

Cao, L., Yue, Y., Cai, Y., and Zhang, Y. (2021). A novel coverage optimization strategy for heterogeneous wireless sensor networks based on connectivity and reliability. *IEEE Access*, 9:18424–18442.

Chopde, N. R. and Nichat, M. (2013). Landmark based shortest path detection by using a* and haversine formula. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2):298–302.

Dahlman, E., Mildh, G., Parkvall, S., Peisa, J., Sachs, J., and Selén, Y. (2014). 5g radio access. *Ericsson review*, 6(1).

de Carvalho, A. A., Batalha, I., Neto, M., Castro, B., Barros, F., Araujo, J., and Cavalcante, G. (2021). Adjusting large-scale propagation models for the amazon region using bioinspired algorithms at 1.8

and 2.6 ghz frequencies. *Journal of Microwaves, Optoelectronics and Electromagnetic Applications*, 20:445–463.

Gade, K. (2010). A non-singular horizontal position representation. *The journal of navigation*, 63(3):395–417.

Gill, S. S. and Buyya, R. (2019). Bio-inspired algorithms for big data analytics: a survey, taxonomy, and open challenges. In *Big data analytics for intelligent healthcare management*, pages 1–17. Elsevier.

Hatamlou, A. (2017). A hybrid bio-inspired algorithm and its application. *Applied Intelligence*, 47(4):1059–1067.

Jensi, R. and Jiji, G. W. (2015). Hybrid data clustering approach using k-means and flower pollination algorithm. *arXiv preprint arXiv:1505.03236*.

Khan, A., Aftab, F., and Zhang, Z. (2019). Bicsf: Bio-inspired clustering scheme for fanets. *IEEE Access*, 7:31446–31456.

Khan, M. U., Azizi, M., García-Armada, A., and Escudero-Garzás, J. (2022). Unsupervised clustering for 5g network planning assisted by real data. *IEEE Access*, 10:39269–39281.

Khan, M. U., García-Armada, A., and Escudero-Garzás, J. (2020). Service-based network dimensioning for 5g networks assisted by real data. *IEEE Access*, 8:129193–129212.

Li, J., Xu, Z., Zhu, D., Dong, K., Yan, T., Zeng, Z., and Yang, S. X. (2022). Bio-inspired intelligence with applications to robotics: a survey. *arXiv preprint arXiv:2206.08544*.

Logesh, R., Subramaniyaswamy, V., Vijayakumar, V., Gao, X.-Z., and Wang, G.-G. (2020). Hybrid bio-inspired user clustering for the generation of diversified recommendations. *Neural Computing and Applications*, 32(7):2487–2506.

Mollel, M. and Michael, K. (2014). Comparison of empirical propagation path loss models for mobile communication.

Nguyen, T.-H., Nguyen, L. V., Jung, J. J., Agbehadji, I. E., Frimpong, S. O., and Millham, R. C. (2020). Bio-inspired approaches for smart energy management: state of the art and challenges. *Sustainability*, 12(20):8495.

Ouamri, M. A., Oteşteanu, M.-E., Isar, A., and Azni, M. (2020). Coverage, handoff and cost optimization for 5g heterogeneous network. *Physical communication*, 39:101037.

Pitchaimanickam, B. and Murugaboopathi, G. (2020). A hybrid firefly algorithm with particle swarm optimization for energy efficient optimal cluster head selection in wireless sensor networks. *Neural Computing and Applications*, 32(12):7709–7723.

Rappaport, T. S., Sun, S., Mayzus, R., Zhao, H., Azar, Y., Wang, K., Wong, G. N., Schulz, J. K., Samimi, M., and Gutierrez, F. (2013). Millimeter wave mobile communications for 5g cellular: It will work! *IEEE access*, 1:335–349.

Reuters (2021). Telecom giants shell out $1.27 bln in first day of Brazil's 5G auction.

Shayea, I., Ergen, M., Azmi, M. H., Çolak, S. A., Nordin, R., and Daradkeh, Y. I. (2020). Key challenges, drivers and solutions for mobility management in 5g networks: A survey. *IEEE Access*, 8:172534–172552.

Shehab, M., Khader, A. T., and Al-Betar, M. A. (2017). A survey on applications and variants of the cuckoo search algorithm. *Applied Soft Computing*, 61:1041–1059.

Sinaga, K. P. and Yang, M.-S. (2020). Unsupervised k-means clustering algorithm. *IEEE access*, 8:80716–80727.

Viavi (2022). 635 New 5G Cities in 2021; 1,947 5G Cities Globally, According to VIAVI.

Wang, C.-H., Lee, C.-J., and Wu, X. (2020). A coverage-based location approach and performance evaluation for the deployment of 5g base stations. *IEEE Access*, 8:123320–123333.

Yang, X.-S. (2012). Flower pollination algorithm for global optimization. In *International conference on unconventional computing and natural computation*, pages 240–249. Springer.

Yang, X.-S. and Deb, S. (2009). Cuckoo search via lévy flights. In *2009 World congress on nature & biologically inspired computing (NaBIC)*, pages 210–214. IEEE.

Yang, X.-S., Karamanoglu, M., and He, X. (2013). Multi-objective flower algorithm for optimization. *Procedia Computer Science*, 18:861–868.

Yang, X.-S., Karamanoglu, M., and He, X. (2014). Flower pollination algorithm: a novel approach for multiobjective optimization. *Engineering optimization*, 46(9):1222–1237.