Comparative Study of K-Power Means, Ant Colony Optimization, Kernel Power Density-based Estimation, and Gaussian Mixture Model for Wireless Propagation Multipath Clustering

Antipas Teologo Jr.[1,2], Lawrence Materum[1,4], Jojo Blanza[1,3], Takuichi Hirano[4]
[1]Department of Electronics and Communications Engineering, De La Salle University, 2401 Taft Ave., Malate, Manila, 0922, Philippines, antipas_jr_teologo@dlsu.edu.ph
[2]Department of Electrical and Electronics Engineering, FEU Institute of Technology, Paredes St., Sampaloc, Manila, 1008, Philippines
[3]Electronics Engineering Department, University of Santo Tomas, Espana Blvd., Sampaloc, Manila, 1008, Philippines
[4]Tokyo City University, 1-28-1 Tamazutsumi, Setagaya, Tokyo, 158-8557, Japan

ABSTRACT

In radio communications, channel modeling has a very significant impact. Most of the system's performance relies on the behavior and characteristics of a radio channel. Due to its mobility features, radio channels are time-variant that change over time, making channel characterizations to be dynamic and very challenging. To ensure maximum data rate and reliable communication, accurate channel models are necessary, which requires a correct grouping of wireless multipaths. Accurate clustering of wireless multipath components (MPCs) is essential for cluster-based channel modeling resulting in a more reliable wireless channel system. Currently, determining the best clustering technique is still a challenge as there is no standard way of evaluating and comparing the performance of the various clustering algorithms. This work presents the comparative study on the accuracy performance of the four clustering algorithms namely K-Power Means (KPM), Ant Colony Optimization (ACO), Kernel Power Density-based Estimation (KPD), and Gaussian Mixture Model (GMM) in grouping the wireless MPCs using datasets generated from COST 2100 channel model (C2CM) which represent different Indoor and Semi-urban channel scenarios. Using the Jaccard index, an evaluation metric that compares the calculated data with the reference data, the accuracy of each algorithm is determined as well as their corresponding computational duration. A comparison of their performance is presented, and results show that KPM outperforms other clustering techniques in all channel scenarios of Indoor and Semi-urban environments, making it a right candidate for further improvements to develop a more accurate and computationally efficient clustering technique for wireless propagation multipaths.

Keywords: k-power means, ant colony optimization, gaussian mixture model, kernel power density estimation, clustering techniques, channel modeling

1. INTRODUCTION

In a wireless communication system, an accurate channel model plays a significant role in ensuring a reliable system design and good performance evaluation. An accurate channel model can be achieved if the various wireless multipath components (MPCs) can be grouped correctly according to their similarities as the channel modeling's primary goal is to characterize the MPCs in various environments. An MPC is characterized in delay and angular domains by its delay (τ), angle of departure (Azimuth of Departure (AoD), Elevation of Departure (EoD)), and angle of arrival (Azimuth of Arrival (AoA), Elevation of Arrival (EoA)). Grouping of these MPCs into different multipath clusters is based on the similarity of their angles and delay.

Over the past two decades, clustering of MPCs [1]–[2] gained much attention in the research community and is now the basis of various channel models nowadays such as the European Cooperation in Science and Technology (COST) 259, COST 2100, 3GPP Spatial Channel Model, and the European Wireless World Initiative New Radio (WINNER) due to the correspondence of the majority of the signals apart from the line-of-sight to form as significant clusters of signal energy. Various parameters of MPCs such as the angular spreads, delay, number, and position are used in the clustered structure modeling to group MPCs into clusters. Wider bandwidth requirement, which is increasing in demand, in fifth-generation (5G) systems can be achieved by utilizing the cluster-based channel model.

Figure 1 presents the clustering scenario of channel multipaths between a base station (BS) and a mobile station (MS). Each circle that contains various dots represents a cluster. A cluster is a group of propagation multipaths having the same properties. Several multipath clustering algorithms have already been proposed, which are either manual (visual

inspection of data) or automatic, and some are the combination of both. Visual inspection has been widely used in the past [3]–[4], but because of its limitations most especially in clustering high-dimensional data, there was the development of the different automatic clustering algorithms for better channel modeling. However, finding an efficient and accurate clustering algorithm is still a challenge.

In the field of machine learning, clustering analysis is very well discussed [5], but on how to use these findings in the clustering of MPCs in wireless channels is still a challenge. MPC contains a lot of attributes such as delay, power, and angle, which are independent of each other. Incorporating the impacts of these attributes is the primary concern in MPC clustering. Some algorithms consider only the power and delay attributes, while others consider all (power, delay, angles). There are several developments already in automatic clustering but at the expense of the system's accuracy and complexity. Most of these existing algorithms suffer their accuracy in clustering those MPCs because of the failure in considering the natural time-variant characteristics of MPC.

Attributes of real-world MPCs should be incorporated in the clustering algorithm to make sure that it is indeed considering the actual set up in a real-world environment. Prior information such as the number of clusters is also required for these clustering techniques, and this is as well a big challenge for automatic clustering. Another limitation, many user-specified parameters are still needed for most clustering algorithms [6]. These methodologies have their respective advantages and disadvantages, depending on the intrinsic and site-specific characteristics of the data to be analyzed, which makes choosing among these methods or techniques a difficult one. Knowing the best of these clustering techniques is very helpful in channel modeling. If the identification of multipath clusters is incorrect, this leads to a cluster-based wireless channel description that is inaccurate. The resulting unrealistic performance evaluation would mean to the user that his data rate is never assured. However, selecting the best performing clustering technique is still a big question due to the unavailability of a standardized evaluation of the clustering performance. Various ground-truth data and validation indices are used in numerous studies, making it difficult to compare the performance of these clustering techniques.

The remaining of this paper is organized in the following way. Section 2 talks about the four clustering techniques. Section 3 gives information about the COST 2100 datasets, while Section 4 discusses the validation index used. The different results generated will be presented in Section 5, and Section 6 concludes the work.

2. THE CLUSTERING TECHNIQUES

The algorithms considered in this study are recently used in the field of wireless multipath clustering. These four algorithms are the ones discussed in this study as they represent different types of clustering algorithms about the
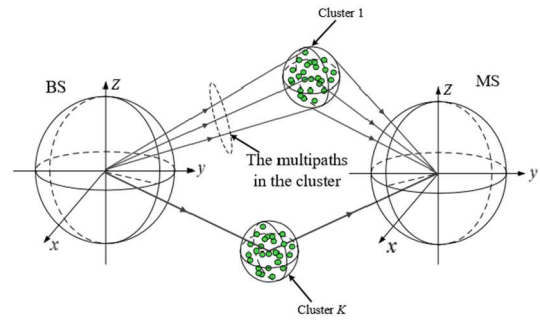

Figure 1: Image of multipath clustering between the BS and MS [7]

clustering taxonomies presented by [5]. Each algorithm belongs to a unique category, and there is no duplication in the clustering type. These are the recent algorithms in their respective clustering category that has improved performance in terms of clustering wireless multipaths. These four techniques also utilize multiple parameters in their input data, which are essential if the MIMO channel environment will be considered as COST 2100 channel model is based on MIMO channel models.

2.1 K-Power Means Framework

KPM algorithm is based on the K-Means algorithm with MPCs power being incorporated. K-Means is a popular method but must be initialized with the number of clusters present in the radio channel. This information is a priori unknown, and improvement to solve this problem gave rise to the introduction of K-Power Means in the different studies [7]–[8]. Being one of the unsupervised learning algorithms, KPM clustering requires that the clusters' initialization has to be known a priori. Below are the main steps [6]:

1. Initialize randomly K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K$, i.e., the K centroid positions are chosen to be as independent events of equal probability (without replacement) from the data set Φ.
2. Assign each MPC sample x to the reasonable cluster centroid μ_j: for each set x, set

$$c^{(k)} := \underset{j}{\operatorname{argmin}}\left\{\alpha_x \cdot d_{\mathrm{MPC}}(x, \mu_j^{(k)})\right\} \qquad (1)$$

where superscript (k) indicates the k-th iteration and c serves as the store indices in the k-th iteration of MPC clustering.

3. Update the cluster centroids: for each j, set

$$\mu_j^{(k+1)} := \frac{\sum_{x\in\Phi} 1\{c^{(k)}=j\}\alpha_x \cdot x}{\sum_{x\in\Phi} 1\{c^{(k)}=j\}\alpha_x} \qquad (2)$$

4. Repeat steps 2 and 3 until convergence.

The addition of the power weighting to determine the MPC distance, d_{MPC}, is the one that sets KPM to be different from the standard K-Means.

2.2 Ant Colony Optimization

Ant colony optimization algorithm is a kind of swarm-intelligent method that is inspired by the behavior of the ant colonies. This relatively new ant-based clustering algorithm came from the observation on how ants perform their clustering of corpses and sorting of their larval. The basic principle of the algorithm is very straightforward. Ants are modeled as the agents that randomly move in their given environment. They can pick up, transport, and drop objects scattered in the environment. The sorting and clustering of the elements are done in the way that ants tend to pick up objects that are either isolated or different from its neighbors and drop these objects in the vicinity wherein they have the similar ones [9].

A more efficient ant-based data clustering is given in the following formulation [10]. First, every MPC is modeled as *virtual* ant-agent $\mathbf{O}_i$ and being plotted onto a 2-D *amplitude-time* plane (x, y), which is considered to be a virtual workspace. Likewise, $O_i^{(k)} = [x_i^{(k)}, y_i^{(k)}]$ gives the position of the i-th agent during the k-th iteration. An environment similarity, $\Lambda_r(O_i^{(k)}, O_j^{(k)})$, corresponds to the local-region fitness (or suitability), which will serve as a guide in the agents' movement.

The virtual local environment of the ith agent $\mathbf{O}_i^{(k)}$, denoted by $L(O_i^{(k)}, r)$, is specified by its sensing range r and the neighbor agents $\mathbf{O}_j^{(k)}$ satisfying $|\mathbf{O}_i^{(k)} - \mathbf{O}_j^{(k)}| \le r$. Hence, r specifies the largest radius distance between the current agent $\mathbf{O}_i^{(k)}$ and its neighbors of interest. In every iteration, the local environment similarity $\Lambda_r(\mathbf{O}_i^{(k)}, \mathbf{O}_j^{(k)})$ of the ith agent $\mathbf{O}_i^{(k)}$ will be evaluated first, which will then decide on what region it should move into.

Specifically, if the local similarity is very weak, then the agent $\mathbf{O}_i^{(k)}$ tends to look for other feasible regions. On the other hand, if there is a significant value for the similarity, this simply indicates the possibility that the current surrounding agents are sharing the same unique large scale effects. In this case, the agent will choose to move closer to this region and needs to update its position $\mathbf{O}_i^{(k+1)} = f(\mathbf{O}_i^{(k)})$, where $f(O_i^{(k)})$ gives the position update function from the kth iteration to the $(k + 1)$th iteration.

2.3 Kernel Power Density-Based Estimation

KPD is also a density-based clustering algorithm, just like the density-based spatial clustering for applications with noise (DBSCAN) but with some notable differences. KPD uses the kernel density instead of samples. It also uses the relative density, and a threshold for the determination of the connectivity of two clusters and power is incorporated in the clustering as well.

Considered as a novel clustering framework, KPD takes into account the kernel density and MPCs' power with only the K nearest MPCs being considered. Some studies utilized KPD, such as in [6], where a KPD-based algorithm achieved the best performance compared with KPM and DBSCAN.In this study, KPD's performance was compared with KPM by using a varying number of clusters.

The following are the steps for KPD-based clustering [6].

1. *Calculating Density*: For every sample of MPC, x, compute the density ρ using the K nearest MPCs as follows:

$$\begin{aligned}
\rho_x &= \sum_{y \in K_x} \exp(\alpha_y) \times \exp\left(-\frac{|\tau_x - \tau_y|^2}{(\sigma_\tau)^2}\right) \\
&\times \exp\left(-\frac{|\Omega_{T,x} - \Omega_{T,y}|}{\sigma_{\Omega_T}}\right) \times \exp\left(-\frac{|\Omega_{R,x} - \Omega_{R,y}|}{\sigma_{\Omega_R}}\right) \\
&\times \exp\left(-\frac{|\Theta_{T,x} - \Theta_{T,y}|}{\sigma_{\Theta_T}}\right) \times \exp\left(-\frac{|\Theta_{R,x} - \Theta_{R,y}|}{\sigma_{\Theta_R}}\right)
\end{aligned} \tag{3}$$

where y is an arbitrary MPC $y \ne x$. K_x is the set of the K nearest MPCs for the MPC x. $\sigma_{(\cdot)}$ is the standard deviation of the MPCs in the domain of $(\cdot)$.

2. *Calculating Relative Density*: For each MPC sample, calculate the relative density ρ^* using the K nearest MPCs' density, as follows:

$$\rho_x^* = \frac{\rho_x}{\max\limits_{y \in K_x \cup \{x\}} \rho_y} \tag{4}$$

3. *Searching Key MPCs*: For each MPC x, if $\rho_x^* = 1$, label it as the key MPC $\hat{x}$. Obtain the set of key MPCs as follows:

$$\widehat{\Phi} := \{x | x \in \Phi, \rho_x^* = 1\} \tag{5}$$

The key MPCs can be considered as the initial cluster centroids.

4. *Clustering*: For each non-key MPC x, define its high-density-neighboring MPC $\tilde{x}$ as:

$$\tilde{x} := \underset{y \in \Phi, \rho_y > \rho_x}{\arg\min} d(x, y) \tag{6}$$

where d represents the Euclidean distance. The high-density neighboring MPC allows constructing a directed graph $\zeta_1 = (\Phi, E_1)$, where each vertex is an MPC, and an arc exists from a non-key MPC x to its high-density neighboring $\tilde{x}$. Namely,

$$E_1 = \{(x, \tilde{x}) | x \in \Phi \backslash \widehat{\Phi}\} \tag{7}$$

where $(x, \tilde{x})$ denotes an arc (ordered pair) from x to $\tilde{x}$, and $\Phi \backslash \widehat{\Phi}$ is the relative complement of set $\widehat{\Phi}$ in set Φ. In ζ_1, starting from any non-key MPC and following the arcs, eventually, a key MPC will be reached. Those MPCs which reach the same key MPC are grouped as one cluster.

5. *Cluster Merging*: For each MPC, add edges between it

and its K nearest MPCs. Thus, acquiring a graph of $\zeta_2 = (\Phi, E_2)$, where

$$E_2 = \{\{x, y\} | x \in \Phi, y \in K_x\} \tag{8}$$

and the unordered pair $\{x, y\}$ is an undirected edge between x and y. If there exists a path where any MPC has $\rho^* > \chi$ between two key MPCs in ζ_2, where χ is a density threshold, merge the two key MPCs' clusters as one new cluster.

2.4. Variational Gaussian Mixture Model

GMM is a generative model that assumes that all data consist of various Gaussian distributions in different proportions. With high accuracy, GMM is capable of approximating any given dataset. In the perspective of probability, each multipath can be assigned to several groups or clusters. This is what also can be found in GMM for it to be considered as a soft clustering community. In GMM, channel multipaths'statistical characteristics are the basis for clustering. Composed of different Gaussian distributions, every distribution contains its own mean and covariance, and the K basic Gaussian distributions are linearly combined. By adjusting the means, covariances, and coefficients, and by utilizing an adequate number of Gaussian distributions, GMM can be able to give a close approximation of any continuous density function.

Shown in Figure 2 is a graphical model where nodes in the form of circles represent the random variables, while those in squares pertain to the model's parameters. Nodes in double circles are representing the observed random variables, while those random variables which are hidden are in the form of single-circled nodes. Variable Z represents the component that has been selected to produce an observed sample x. The detailed expression of the Gaussian probability density function is as follows:

$$N(x) = \left(\exp\left((\tfrac{-1}{2})(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right) \right) / \left(\sqrt{(2\pi)^D \det \Sigma_k} \right) \tag{9}$$

Then the joint probability density function (pdf) of X and Z is

$$p(X, Z) = p(X|Z)p(Z) \tag{10}$$

Then the marginal probability function is

$$p(X = x) = \sum_{k=1}^{K} p(X = x|Z = k)p(Z = k) = \sum_{k=1}^{K} \pi_k p_k(x) \tag{11}$$

where the density of the k-th component is $p_k(x) = N(x; \mu_k, \Sigma_k)$, in which the π_k are the weights (mixing coefficients), $k=1,2,...,K$. Then,

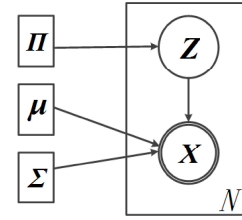$$p(x) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k) \tag{12}$$

Figure 2: Graphical model for GMM [7]

With the above distributions, the posterior probability can be computed using the Bayes theorem as follows:

$$p(k|x) = \frac{p(x|Z=k)p(Z=k)}{p(x)} = \frac{\pi_k N(x;\mu_k,\Sigma_k)}{\sum_{\ell=1}^{K} \pi_\ell N(x;\mu_\ell,\Sigma_\ell)} \tag{13}$$

By assigning each multipath x to the component of the maximum posterior, dataset X can be clustered into K components.

3. COST 2100 DATASETS

The datasets utilized in this study were generated using the COST 2100 channel model (C2CM). This channel model is capable of reproducing MIMO channels' stochastic properties over time, space, and frequency. Originally, C2CM was proposed for simulating the radio channel between a static multiple-antenna BS and a multiple-antenna MS. Its implementation in MATLAB supports indoor (285 MHz) and semi-urban (5.3 GHz) channel scenarios of MIMO channel access both for single-link and multiple-links and a more detailed Matlab implementation with massive MIMO extensions can be obtained from [11]. Meanwhile, an overview of the C2CM can be found in [12], whereas the channel model's detailed description is presented in [13]. In [14], a complete detail on C2CM parametrization in indoor scenarios is given, while for the semi-urban scenarios, it is being presented in [15]. MATLAB implementation of C2CM can be found in [16], while the semi-urban channel scenario is implemented in [15].

Implemented in MATLAB, the eight datasets generated for multipath clustering and used in this study are the following:

1. Indoor, B1, line-of-sight, single link(channel scenario 1)
2. Indoor, B2, line-of-sight, single link (channel scenario 2)
3. Semi-urban, B1, line-of-sight, multiple links (channel scenario 3)
4. Semi-urban, B1, line-of-sight, single link (channel scenario 4)
5. Semi-urban, B1, non-line-of-sight, single link (channel scenario 5)
6. Semi-urban, B2, line-of-sight. multiple links (channel scenario 6)
7. Semi-urban, B2, line-of-sight, single link (channel scenario 7)
8. Semi-urban, B2, non-line-of-sight, single link (channel scenario 8)

These datasets were pre-processed first before being utilized

by the different clustering techniques. These pre-process include directional cosine transform (DCT), clusterability, and whitening transform. A detailed description of each pre-process can be found in [17]. Each dataset consists of 30 sets of data since it is generated with 30 trials and is presented in an Excel file format. The datasets can be found and downloaded from IEEE DataPort.

4. CLUSTER VALIDATION

The Jaccard index η_{jac} is used to evaluate the accuracy of the four clustering techniques. Its numeric values can range from 0 to 1, with one being the highest, which means a perfect match. This Jaccard index or score represents the 'intersection over union' between the considered algorithm and the 'real' clustering [18]. The use of the Jaccard index is an objective way of showing the match and mismatches between the simulated classification vectors and the experimental one. Jaccard index is just one of the many external comparison indices. Being designed to measure the similarity between two partitions, external indices only consider the distribution of points in the various clusters. Metrics used are given as follows:

$$\eta_{jac} = \frac{n_{11}}{n_{11}+n_{10}+n_{01}} \qquad (14)$$

where

n_{11} is the number of pairs that are classified together correctly (Case 1).

n_{01} is the number of pairs that are not classified together correctly by the algorithm (Case 2).

n_{10} is the number of pairs that are incorrectly classified together when they are not supposed to (Case 3).

5. RESULTS AND DISCUSSION

This section will present the different results obtained and gathered during the study. Using the eight datasets generated from C2CM and the Jaccard index for the evaluation of the accuracy, the four clustering techniques were implemented in Matlab, and their performances were being analyzed and compared.

5.1 The Distribution of the Jaccard Accuracy Scores

Table 1 shows the summary of the generated mean Jaccard scores of the four clustering techniques in every channel scenario. The Jaccard scores are varying for each clustering technique in all channel scenarios, and it can be observed that Indoor channels obtain a wider range of Jaccard values compared with the Semi-urban environments. For channel scenario 1 (Indoor, B1, LOS, Single Link), KPM obtains the highest performance, followed by KPD. GMM comes third while ACO is the last, both obtaining low scores with almost close Jaccard mean values. KPM still dominates in the channel scenario 2, followed again by KPD. However, ACO and GMM switch positions, but still with Jaccard mean scores close with each other. For channel scenario 3, it is very evident that there is a significant drop in the values of Jaccard

Table 1: Summary of the mean Jaccard index values

Channel Scenario	KPM	ACO	KPD	GMM
Indoor, B1, LOS, Single Link	0.8915	0.2776	0.7459	0.2981
Indoor, B2, LOS, Single Link	0.8446	0.2906	0.6389	0.2530
Semi-urban, B1, LOS, Multiple Links	0.1206	0.0296	0.0547	0.0830
Semi-urban, B1, LOS, Single Link	0.1190	0.0387	0.0882	0.0926
Semi-urban, B1,NLOS, Single Link	0.1170	0.0176	0.0619	0.0932
Semi-urban, B2, LOS, Multiple Links	0.1206	0.0290	0.0517	0.0856
Semi-urban, B2, LOS, Single Link	0.1168	0.0373	0.0861	0.0834
Semi-urban, B2,NLOS, Single Link	0.1162	0.0394	0.0598	0.0997

indices. However, KPM maintains the lead, but this time followed by GMM. KPD drops to third while ACO still obtains the least score. Moving on, KPM still holds the highest score in channel scenario 4. GMM still in the second spot followed closely by KPD. ACO still obtains the lowest Jaccard index far from the other three clustering techniques.

In channel scenarios 5 and 6, almost the same trend can be seen in channel scenario 4. In channel scenario 7, KPM still performs the best, but this time KPD comes next with only a small difference from the GMM, which is in third place. ACO continues to exhibit the least performance. KPM continues to outperform other clustering techniques in the last channel scenario with GMM comes next, followed by KPD and ACO, with still the lowest performance. Comparing the mean Jaccard scores, it shows that KPM outperforms all other clustering candidates in all types of channel scenarios, both in Indoor and in Semi-urban. It is also worth noting that Jaccard indices drop drastically from an Indoor environment to a Semi-urban one.

5.2 Relative Frequency Distribution of the Jaccard Accuracy Index

Figure 3 to Figure 10 present the generated histogram plots of the various clustering techniques. The relative frequency is given for each corresponding Jaccard score. In channel scenario 1 shown in Figure 3, it can be observed that the majority of the trials (19 out of 30) or 63.33% in KPM obtain a perfect clustering with only very few Jaccard scores (6.67%) lying on low values. For ACO, most of the scores (93.33%) fall on the lower portion, ranging from 0.14 to 0.44. Only 6.67% of the total trial obtains a value greater than 0.5, indicating poor performance. KPD, on the other hand, offers an average performance with the majority of the Jaccard scores (93.33%) converge in the range of 0.56 to 0.98. For GMM, scores are scattered below 0.5, with only one trial (3.33%) above 0.5. In Figure 4 is another Indoor environment (Indoor, B2, LOS, Single Link), 60% of the KPM trials obtain a perfect score. This is contrary to the ACO performance, in
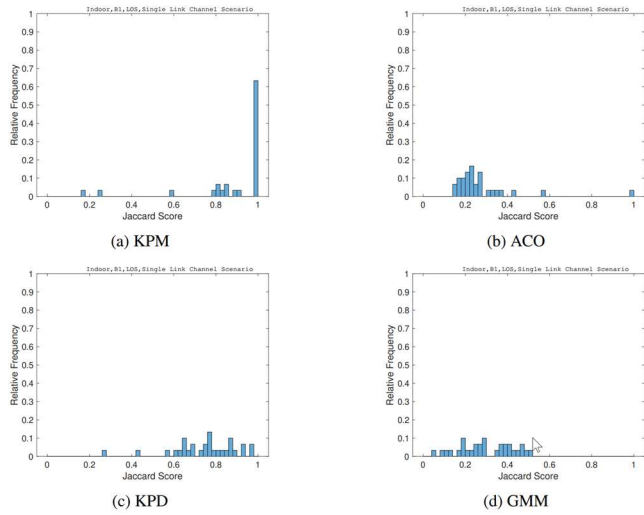
Figure 3: The histogram plots for channel scenario 1 of (a) KPM (b) ACO (c) KPD and (d) GMM
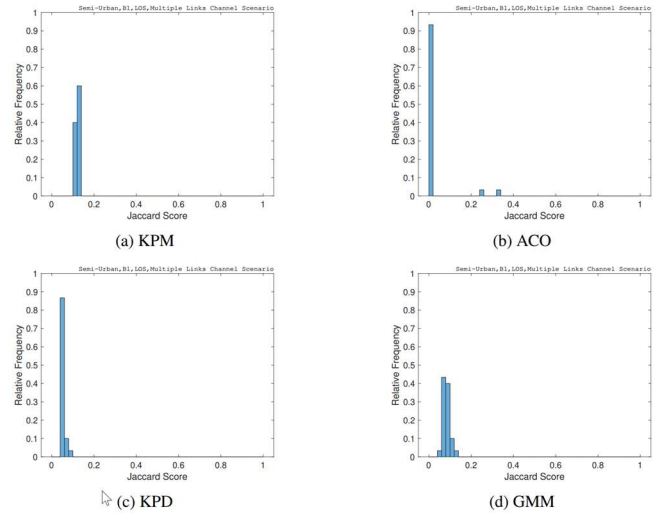

Figure 4: The histogram plots for channel scenario 2 of (a) KPM (b) ACO (c) KPD and (d) GMM

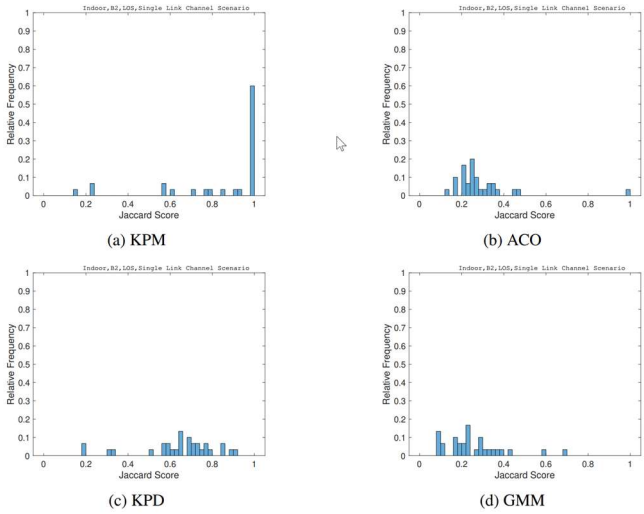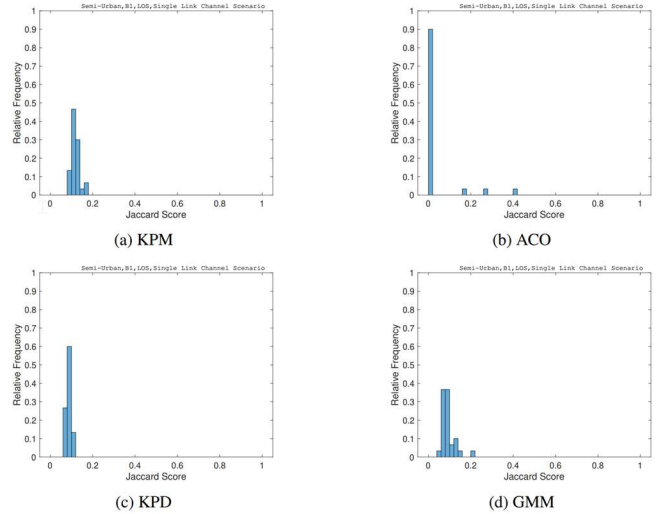**Figure 5:** The histogram plots for channel scenario 3 of (a) KPM (b) ACO (c) KPD and (d) GMM

Figure 6: The histogram plots for channel scenario 4 of (a) KPM (b) ACO (c) KPD and (d) GMM

which the majority of the scores (90%) fall below the 0.4 marks. Meanwhile, 86.67% of the KPD Jaccard scores are above 0.5, making it as well as the right choice for channel scenario 2. GMM, on the other hand, shows poor clustering performance as 93.33% of its scores are below 0.5.

Histogram plots of channel scenario 3 are presented in Figure 5. There is a significant drop in Jaccard scores of the four clustering techniques. KPM Jaccard scores converge in the range of 0.10 to 0.14 while in ACO, 93.33% of the scores fall below 0.02. There is also a converging of values in KPD as 86.67% of its trials obtain scores in the range of 0.04 to 0.06 while in GMM, most of the scores (83.33%) can be found in 0.06 to 0.1 range. Jaccard scores continue to be low in Semi-urban, B1, LOS, Single link channel scenario, as shown in Figure 6. Most of the Jaccard scores (76.67%) are situated from 0.10 to 0.14 in KPM, while 90% of ACO scores are below 0.02. For KPD, 60% of scores are in the range of 0.08 to 0.1 while in GMM, 73.34% of scores can be found from 0.06 to 0.1. Almost the same trend of values can be observed

in channel scenario 5 as shown in Figure 7. KPM still generates low Jaccard scores, with 76.67% of it in the range 0.10 to 0.14 same in channel scenario 4. ACO, meanwhile, continues to exhibit low results as 96.67% of its scores are below 0.02. On the other hand, converging of values can be observed in KPD as 93.34% of its trials are in the 0.04 to 0.08 range, unlike in the GMM where there is some spreading of scores with 46.67% of it landing in 0.08 to 0.10.

KPM scores continue to showcase consistency in the 0.10 to 0.14 range as 93.33% of its trials are now in this range as being depicted in Figure 8 for Semi-urban, B2, LOS, Multiple links channel scenario. ACO in this channel scenario also maintains an almost consistent performance, as 93.33% of its scores are still below 0.02. However, converging of scores can be observed in KPD in this channel scenario, with 90% of its values are in the 0.04 to 0.06 range while GMM exhibits spreading of values with 50% of its scores are in the range of 0.06 to 0.08. In channel scenario 7, given in Figure 9, clustering techniques continue to produce low results. More
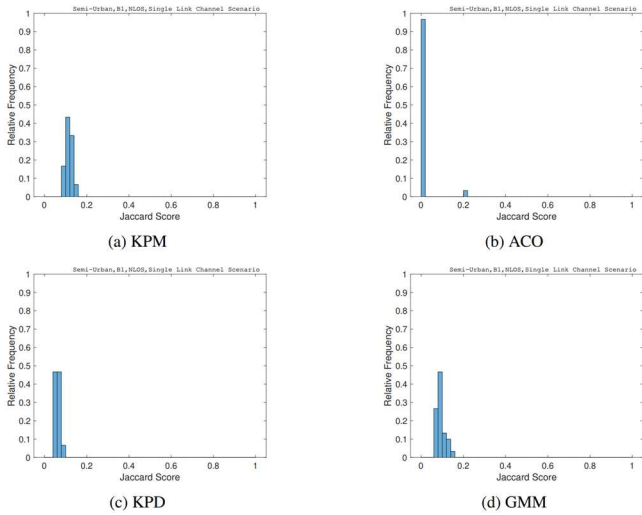
Figure 7: The histogram plots for channel scenario 5 of (a) KPM (b) ACO (c) KPD and (d) GMM
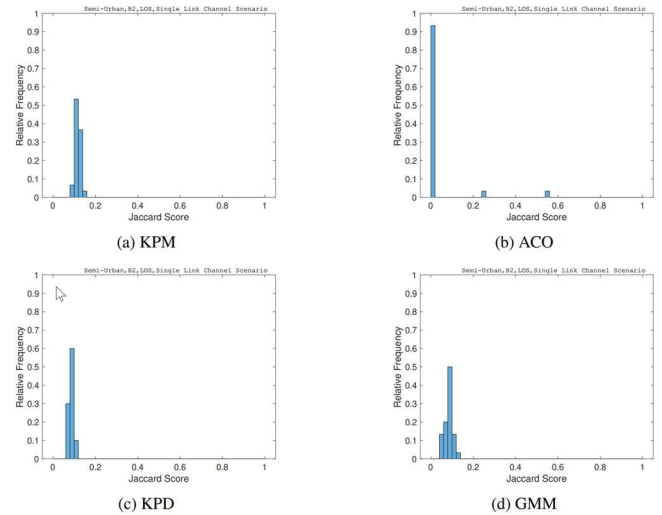

Figure 9: The histogram plots for channel scenario 7 of (a) KPM (b) ACO (c) KPD and (d) GMM

Figure 8: The histogram plots for channel scenario 6 of (a) KPM (b) ACO (c) KPD and (d) GMM
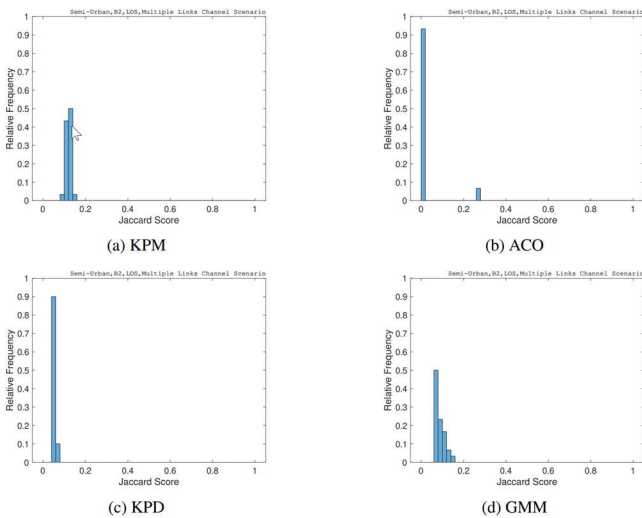

Figure 10: The histogram plots for channel scenario 8 of (a) KPM (b) ACO (c) KPD and (d) GMM

than half of the trials (53.33%) in KPM are in the 0.10 to 0.12 range, while the majority (93.33%) of scores in ACO are still under 0.02. For KPD, most scores (60%) are in the 0.08 to 0.10, while in GMM, half of its scores (50%) are in the same range of values. Almost the same trends in data can be observed for the last channel scenario of Semi-urban, B2, NLOS, Singe link, as shown in Figure 10. KPM shows a modest increase in the number of scores in 0.10 to 0.12 range as it goes up to 56.56% from 53.33% in channel scenario 7. Same as well in channel scenario 7, 93.33% of ACO scores are still below 0.02. For KPD, it shows a degrading performance as 60% of its scores is now in the 0.04 to 0.06 range from the 0.08 to 0.10 range in channel scenario 7. GMM scores, meanwhile, still mostly (56.67%) fall in the 0.08 to 0.10 range.

In summary, it proves that KPM is superior to the other three clustering techniques in all channel scenarios. KPM performs very well in the Indoor channel scenarios as compared to the other clustering candidates. For semi-urban channel
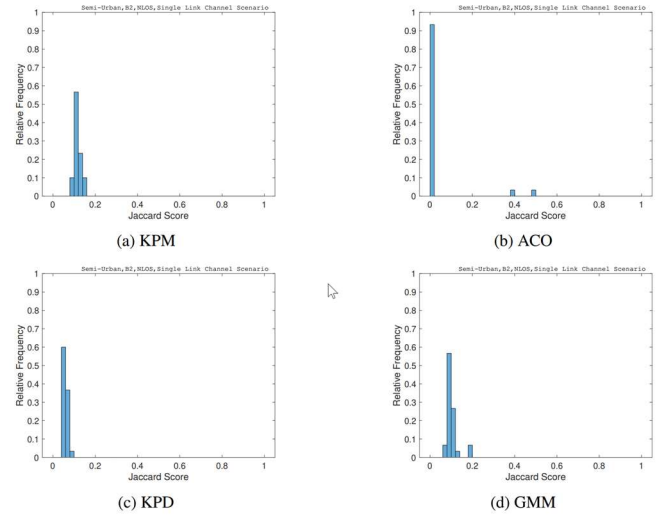
environments, KPM is still dominating, but it can be observed that other clustering techniques also generate Jaccard scores close to the KPM values.

5.3 Computational Duration

Presented in Table 2 is the mean computational duration of the four clustering techniques for each channel scenario. This gives an idea of the complexity of each clustering algorithm. Run in Matlab 2019a with 64-bit Windows 10 HP Envy, Intel Core i7-8550U CPU at 1.8 GHz processor, and 8.00 GB memory, and the duration was obtained by using the tic-toc function in Matlab. It can be observed that in all the four clustering techniques, indoor scenarios have a shorter duration as compared to the semi-urban environments. This just simply indicates that the computational complexity of a particular clustering technique greatly depends on the number of multipaths. As the number of multipaths increases, the duration of clustering also increases. It can also be observed that in the case of semi-urban channel scenarios, multiple

Table 2: Mean computational duration (in seconds) in each channel scenario

Channel Scenario	KPM	ACO	KPD	GMM
Indoor, B1, LOS, Single Link	3.22	380.80	1.81	0.38
Indoor, B2, LOS, Single Link	3.14	381.82	2.16	0.44
Semi-urban, B1, LOS, Multiple Links	104.13	385.15	705.43	1.77
Semi-urban, B1, LOS, Single Link	18.23	392.59	126.24	0.90
Semi-urban, B1, NLOS, Single Link	37.85	358.84	519.95	1.82
Semi-urban, B2, LOS, Multiple Links	105.63	387.03	742.00	2.00
Semi-urban, B2, LOS, Single Link	19.89	393.05	150.76	1.26
Semi-urban, B2, NLOS, Single Link	81.46	368.63	604.08	1.23

links exhibit higher duration as compared to the single link as the former contains more multipaths than the latter.

Table 3 shows the summary of the mean Jaccard accuracy scores and the mean computational duration of the four clustering algorithms. Based on the presented results, it can be said that GMM offers the best computational duration among the four techniques but at the expense of its accuracy performance. GMM seems to be not the best option if accuracy performance is concerned. Meanwhile, KPM shows to be the next option if the computational duration is to be considered noting as well that it performs best among the four techniques in terms of clustering multipaths correctly. On the other hand, KPD exhibits a slightly lower duration in its indoor channel scenarios compared with KPM but shows a higher duration in the semi-urban environments as compared with KPM. In the case of ACO, it can be observed that it exhibits high duration both in indoor and semi-urban scenarios taking note as well its poor accuracy performance. There is somewhat a trade between the accuracy performance and computational complexity of a clustering algorithm. Higher accuracy means higher computational duration, and with this, it can be said that a more accurate yet computationally efficient multipath clustering method needs to be developed.

6. CONCLUSIONS AND RECOMMENDATIONS

This work presents a comparative study on the accuracy performance of the four clustering algorithms, namely K-Power Means (KPM), Ant Colony Optimization (ACO), Kernel Power Density-based Estimation (KPD), and Gaussian Mixture Model (GMM). Using the datasets generated from C2CM as the common ground-truth data representing various Indoor and Semi-urban channel scenarios and the Jaccard index for the validation, the accuracy performance of the four clustering techniques was determined successfully, and their corresponding computational duration was determined.

Table 3: Summary of the mean Jaccard accuracy and mean computational duration of the four clustering techniques

Clustering Technique	Mean Jaccard Accuracy (%)		Mean Computational Duration (s)	
	Indoor	Semi-urban	Indoor	Semi-urban
KPM	89.15	12.06	3.22	105.63
ACO	29.06	3.94	381.82	393.05
KPD	74.59	8.82	2.16	742.00
GMM	29.81	9.97	0.44	2.00

For an Indoor environment (channel scenarios 1 and 2), the highest Jaccard score that KPM achieves is 0.8915 in channel scenario 1, which is equivalent to 89.15% accuracy in clustering wireless multipaths. This is followed by KPD with a 0.7459 Jaccard score or 74.59% also in channel scenario 1. GMM's highest Indoor channel scenario accuracy is 0.2981 or 29.81% in channel scenario 1, while ACO achieves only 0.2906 or 29.06% in channel scenario 2. Meanwhile, considering the six-channel scenarios (channel scenarios 3 to 8) for the Semi-urban environment, KPM's highest accuracy is only 0.1206 or 12.06% both from channel scenarios 3 and 6. GMM comes next with 0.0997 or 9.97% in channel scenario 8. KPD places third with 0.0882 or 8.82% from channel scenario 4, while ACO obtains the least accuracy of only 0.0394 or 3.94% coming from channel scenario 8. On the other hand, considering the computational duration, GMM offers the least computational complexity among the four clustering techniques followed by KPM. KPD is in the third rank, while ACO shows to be the most complex one.

Based on the results, it shows that KPM is the best option for both Indoor and Semi-urban channel environments if accuracy performance is concerned at the little expense of its computational duration. Although KPM obtains an acceptable accuracy performance in the Indoor scenario, however, it should be noted that its performance drastically drops when it comes to the Semi-urban setup. It can also be observed that other clustering techniques have a significant drop in their accuracy when the Semi-urban environment is concerned. This is due to the reason that there are a lot of scatterers in the Semi-urban environment compared in Indoor, which can significantly affect the characteristics of propagation multipaths, making it difficult to group or cluster. Taking note of all of these results, it can be said that an improvement in grouping MPCs is necessary, most notably in a Semi-urban scenario. Findings in this study could be a useful reference in developing a new clustering algorithm for multipath propagations that could offer an improved accuracy while considering as well its computational complexity. A more accurate yet computationally efficient multipath clustering method is necessary and needs to be developed.

REFERENCES

1. J. Blanza and L. Materum. **Wireless propagation multipath clustering: On simultaneously solving the membership and the number of clusters**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 5, pp. 1914-1919, 2019.
https://doi.org/10.30534/ijatcse/2019/15852019

2. M.A.Roque and L. Materum. **Interface for the factor-inclusion weighting approach in determining the number of multipath propagation clusters**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 4, pp. 1768-1776, 2019.
https://doi.org/10.30534/ijatcse/2019/107842019

3. L. Vuokko, P. Vainikainen, and J. Takada. **Clusters extracted from measured propagation channels in macrocellular environments**, *IEEE Trans. Antennas Propag.*, vol. 53, no. 12, pp. 4089-4098, 2005.
https://doi.org/10.1109/TAP.2005.859763

4. C. Oestges and B. Clerckx. **Modeling outdoor macrocellular clusters based on 1.9-GHz experimental data**, *IEEE Trans. Veh. Technol.*, vol. 56, no. 5, pp.2821-2830, 2007.
https://doi.org/10.1109/TVT.2007.900391

5. R. Xu and D. Wunsch. **Survey of clustering algorithms**, *IEEE Trans. Neural Netw.*, vol.16, no.3, pp. 645-678, 2005.
https://doi.org/10.1109/TNN.2005.845141

6. R. He, Q. Li, B. Ai, Y. Geng, A. Molisch, V. Kristem, Z. Zhong, and J. Yu. **A kernel-power-density-based algorithm for channel multipath components clustering**, *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7138-7151, 2017.
https://doi.org/10.1109/TWC.2017.2740206

7. Y. Li, J. Zhang, Z. Ma, and Y. Zhang. **Clustering analysis in the wireless propagation channel with a variational Gaussian mixture model**, *IEEE Trans. Big Data*, vol. 6, no. 2, pp. 223-232, 2020.
https://doi.org/10.1109/TBDATA.2018.2840696

8. P. Hanpinitsak, K. Saito, J. Takada, M. Kim, and L. Materum. **Multipath clustering and cluster tracking for geometry-based stochastic channel modeling**, *IEEE Trans. Antennas Propag.*, vol. 65, no. 11, pp. 6015-6028, 2017.
https://doi.org/10.1109/TAP.2017.2754417

9. W.Gao. **Improved ant colony clustering algorithm and its performance study**, *Comput. Intel. Neurosc.*, 2016.
https://doi.org/10.1155/2016/4835932

10. B. Li, C. Zhao, H. Zhang, Z. Zhou, and A. Nallanathan. **Efficient and robust cluster identification for ultra-wideband propagations inspired by biological ant colony clustering**, *IEEE Trans. Commun.*, vol. 63, no. 1, pp. 286-300, 2015.
https://doi.org/10.1109/TCOMM.2014.2377120

11. COST 2100 channel model, 2018. [Online]. Available: https://github.com/cost2100/cost2100/tree/master/matlab

12. L. Liu, C.Oestges, J.Poutanen, K. Haneda, P. Vainikainen, F.Quitin, F.Tufvesson,and P. De Doncker.**The COST 2100 MIMO channel model**, *IEEE Wireless Commun.*, vol. 19, no. 6, pp.92-99, 2012.
https://doi.org/10.1109/MWC.2012.6393523

13. R. Verdone and A. Zanella.**Pervasive Mobile and Ambient Wireless Communications: COST Action 2100**, Springer Science & Business Media, 2012, ch. 3.

14. J. Poutanen, K. Haneda, L. Liu, C. Oestges, F. Tufvesson, and P. Vainikainen. **Parameterization of the COST 2100 MIMO channel model in indoor scenarios**, in *Proc. 5th Eur. Conf. Antennas Propag. (EuCAP)*, 2011, pp. 3606-3610.

15. M. Zhu, G. Eriksson, and F. Tufvesson. **The COST 2100 channel model: Parameterization and validation based on outdoor MIMO measurements at 300 MHz**, *IEEE Trans. Wireless Commun.*, vol. 12, no. 2, pp.888-897, 2013.
https://doi.org/10.1109/TWC.2013.010413.120620

16. L. Liu, N. Czink, and C. Oestges. **Implementing the COST 273 MIMO channel model**, in *Proc. NEWCOM-ACoRN Joint Workshop*, 2009.

17. J. Blanza, A. Teologo, and L. Materum. **Datasets for multipath clustering at 285 MHz and 5.3 GHz bands based on COST 2100 MIMO channel model**, in *9th Int. Symp. Multimedia Commun. (ISMAC 2019)*, Aug. 2019, pp. 1-5.
https://doi.org/10.1109/ISMAC.2019.8836143

18. R. Varshavsky, M.Linial, and D. Horn. **COMPACT: A comparative package for clustering assessment**, in Proc. Parallel Distrib. Process. Appl., 2005, pp. 159-167.
https://doi.org/10.1007/11576259_18