Concurrent and Spectral Clustering of Wireless Waves

Jojo Blanza[1,*] and Lawrence Materum[2,3]

[1]Electronics Engineering Department, University of Santo Tomas, Philippines
[2]Department Electronics and Computer Engineering, De La Salle University, Philippines
[3]International Centre, Tokyo City University, Tokyo, Japan
Email: jfblanza@ust.edu.ph (J.B.); materuml@dlsu.edu.ph (L.M.)
*Corresponding author
Manuscript received September 4, 2023; revised October 26, 2023; accepted December 6, 2023; published January 30, 2024

***Abstract*—Unsupervised clustering is generally used to identify groupings of wireless waves from estimated multipath parameters in order to optimize the cluster count and membership. The estimated parameters exclude cluster labels. For a better comparison of clustering approaches, cluster-based wireless channel models provide the labels. This work proposes that using the Three-Constraint Affinity Matrix (3CAM) in formulating the affinity or similarity matrix improves the clustering accuracy. Datasets generated from the European Cooperation in Science and Technology (COST) 2100 Channel Model (C2CM) were used and subjected to directional cosine and whitening transforms. Simultaneous Clustering and Model Selection Matrix Affinity (SCAMSMA), 3CAM-SCAMSMA, Spectral Clustering (SC), and 3CAM-SC were used to concurrently determine cluster count and membership. Various studies on multipath clustering give only the number of clusters. Others would state only the validity index of the membership of clusters. The problem with such an approach is that the correctness of the number of clusters is not an assurance that the membership of the clusters is accurate. The four clustering approaches solve this problem by determining the number of clusters and their membership. Thus, knowing each technique's performance is essential. In the algorithms of all the clustering approaches, cluster count aims to ensure that the target cluster count is within the vicinity of the reference clusters. The cluster count and membership accuracy are computed through the cluster-wise Jaccard index of the multipath membership to their clusters. The performance of the clustering approaches was validated using the Jaccard index by comparing the calculated data with the reference data. The results show that 3CAM-SCAMSMA improved the clustering accuracy of SCAMSMA by an average of 5.218% in semi-urban scenarios. At the same time, 3CAM-SC increased the performance of SC in indoor scenarios by an average of 44%. 3CAM-SC is the most robust clustering approach, registering the highest accuracy and slightest variation.**

***Keywords*—5G, channel model, clustering algorithms, data handling, Multiple-Input Multiple-Output (MIMO)**

I. Introduction

Clustering is a process that analyses data by classifying groups with similar structures. Clustering aims to categorize the data into several clusters such that points in the same group are similar while that of the other groups are dissimilar. Datasets for wine [1], data mining [2, 3], breast cancer [4], and metagenomic sequences [5] have been clustered over the years. The clustering of wireless propagation multipaths gained interest due to the widespread application of Multiple-Input Multiple-Output (MIMO) antennas in wireless communications systems. MIMO systems are developed to increase data rates and ensure wireless transmission reliability, and cluster-based channel models have been used extensively to describe the MIMO propagation channel.

The Channel Impulse Response (CIR) is one of the common ways of characterizing the most crucial portion of communications systems design. Among the popular channel models are 3rd Generation Partnership Project (3GPP) [6], International Mobile Telecommunications-2020 (IMT-2020) [7], QUAsi Deterministic RadIo channel GenerAtor (QuaDRiGa) [8], and Cooperation in Science and Technology (COST) 2100 [9]. The communications signals propagate in multiple directions as they move from the transmitter to the receiver. The Multipath Components (MPCs) are grouped in clusters, as shown in Fig. 1. The MPCs in different environments must be characterized to determine the accuracy of the channel model. Multipath clusters with similar parameters of the MPCs, such as delay, azimuth, and elevation of arrival and departure, are considered to describe the propagation channel using a clustering technique accurately. Traditionally, clusters are identified through human visual inspection [10]. It works well when the dataset is small. However, this approach is subjective and tedious for large datasets [11]. Such reasoning prompts the use of automatic clustering approaches, which have become popular to remove the bias of visually clustering the multipaths and eliminate the problem of accurately clustering large datasets.

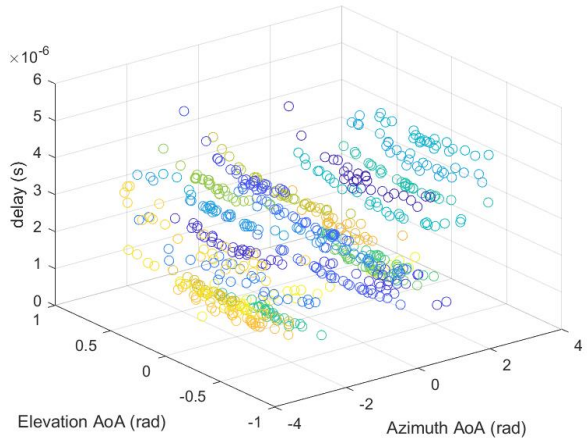

Fig. 1. Multipath components (MPCs) generated by the COST 2100 channel model where a group of MPCs coded with the same color is classified as a multipath cluster.

Different clustering techniques that automatically determine the clusters have been introduced to overcome these concerns over the years. Among them is K Power Means (KPM), which uses K-means in clustering the multipaths [11]. The powers of the multipaths are included, and the distance between cluster centroids is minimized to determine the number of clusters. KPM, however, needs the

initial number of clusters as a priori. Kurtosis Measure (KuM) overcomes the sensitivity of KPM to the input settings by detecting the time of arrival of the multipaths and partitioning them into clusters [12]. KuM is independent of the channel and is still applicable even without prior knowledge of the impact of the environment on the CIR. Ant Colony Clustering (ACC) combines the decaying amplitude and the time of arrivals of MPCs [13]. Clusters are identified based on the population and the positive-feedback collaboration of the evolution of the ant agents. Automatic Cluster Identification (ACId) improved the mean cluster distance of K-means by iteratively assigning MPCs to a cluster as long as the cluster distance is within a threshold [14]. The cluster centroid position is dynamically updated and reassigns MPCs that might be closer to existing clusters.

The Sparsity-Based Method (SBM) is built on the Saleh-Valenzuela (SV) model feature, that with increasing delay, the power of the MPCs exponentially decreases [15]. SBM does not need prior knowledge of clusters, such as the number and initial cluster locations, because it incorporates the expected behavior of clusters into the clustering framework. Kernel Power Density (KPD) utilizes the kernel density and power of multipaths to identify the local density variations of MPCs [16]. A heuristic approach to cluster merging is used to improve the performance of the clustering approach. The Gaussian Mixture Model (GMM) relates the covariance structure with the mean information of the multipaths to reveal their similarity [17]. A compact index validates the close relationship between the GMM clustering mechanism and the multipath propagation characteristics.

The clustering approaches mentioned above give only the number of clusters of MPCs and do not consider the accuracy of the cluster membership. Thus, the number of clusters may be correct, but it does not necessarily mean the correct members are in the clusters. This problem can be solved by simultaneously determining the number of clusters and the membership of clusters. Simultaneous identification of the number of clusters and the membership of clusters is made to solve the problem of giving just the number of clusters. Simultaneous clustering and model selection matrix affinity (SCAMSMA) [18], three-constraint affinity matrix SCAMSMA (3CAM-SCAMSMA) [19], Spectral Clustering (SC) [20], and 3CAM-SC [21] can be used to address the issue.

The significance of the study is it solves the number of multipath clusters and the membership of multipath clusters simultaneously. This is a new technique of presenting clustering accuracy as it shows at the same time the correctness of the number of clusters and the contents of the clusters. Some results were from previous works [19, 21]. The paper compares their performance variation, which was not done previously [13, 15–17], and looks if there is statistical significance. Thus, the research gap in finding the statistical significance is addressed by this work, in particular, working with multipath datasets with apriori multipath cluster membership. As discussed and cited above, many contributions to multipath clustering fail to communicate their statistical significance, which the authors labored to express in this present work. The impact of sharing the statistical significance is that researches in the field can clearly comprehend and understand the performance of the clustering approaches and identify which of them gives the best results. The clustering approach with outstanding performance can then be used best in clustering multipaths. The main contributions of this work are as follows:

- Applying SCAMSMA, 3CAM-SCAMSMA, SC, and 3CAM-SC to cluster wireless multipaths which adds to the literature of multipath clustering four more clustering approaches that can be used for the task of clustering multipaths,
- Adopting 3CAM to improve the clustering accuracy of SCAMSMA and SC which can also be used in conjunction with other clustering approachess to further improve their accuracy,
- Conducting a performance evaluation of the clustering approaches to show which of them is the best and most robust for the task of clustering multipaths.

The paper is organized in the following way: Section II discusses the methodology. Section III presents the results of the clustering approaches and elaborates on the findings, and Section IV concludes the work.

II. Method

Fig. 2 outlines the methodology of the study. The COST 2100 Channel Model (C2CM) [9] generates the multipath components and clusters, serving as the reference dataset.

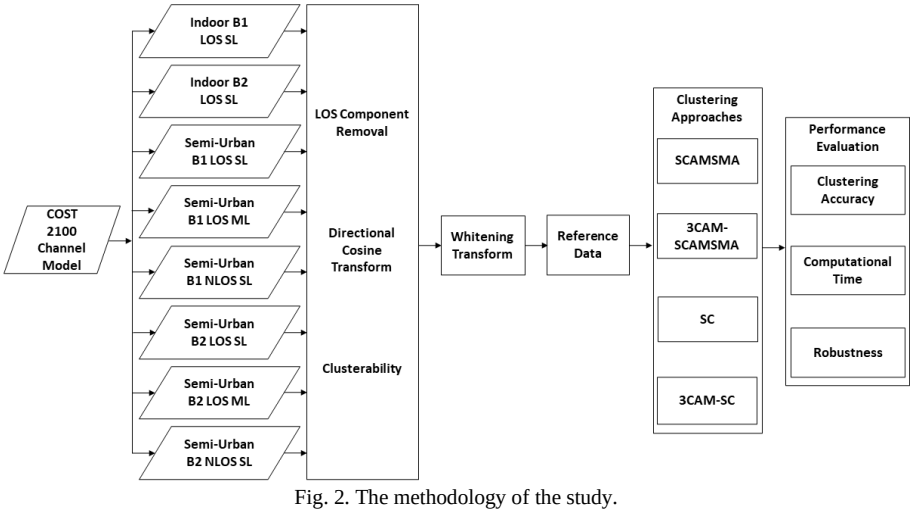

Fig. 2. The methodology of the study.

Indoor and semi-urban comprise the environment of the channel model, while Band 1 (B1) and Band 2 (B2) compose the frequency bands. The transmission of the wireless signals can be chosen as a Line-of-Sight (LOS) or Non-Line-of-Sight (NLOS), while the communication link between the transmitter and the receiver can be set to single or multiple. The COST 2100 dataset is preprocessed to become the reference data and clustered using four shallow learning clustering approaches. The performance of the clustering approaches is then compared using three evaluation criteria. The discussion in Section II-A to Section II-C is needed to be done to extract the labeled multipaths from [9]. Each label also referred to as identification or ID, indicates which cluster a multipath belongs to. Such labels enable the checking for the correctness of the clustering approach taken.

A. Generation of COST 2100 Channel Wireless Multipaths

C2CM can reproduce the stochastic properties of MIMO propagation channels. It is a Geometry-Based Stochastic Channel Model (GSCM) that is generic and flexible in its approach. It is suitable to model multi-user MIMO scenarios [22]. Multipath clusters characterize C2CM, and groups of MPCs with similar delays and angles comprise a multipath cluster. An MPC is classified based on the delay, angle of departure (Azimuth of Departure (AoD), Elevation of Departure (EoD)), angle of arrival (Azimuth of Arrival (AoA), and Elevation of Arrival (EoA)).

A CIR that changes with time (designated by *t*) is the group of MPCs from all the multipath clusters according to the location of the Mobile Station (MS) with the Base Station (BS). The CIR is based on the delay and direction domain and is given as

$$h\left(t,\tau,\boldsymbol{\Theta}^{\mathrm{BS}},\boldsymbol{\Theta}^{\mathrm{MS}}\right)= \sum_{n=1}^{K}\sum_{p}\alpha_{n,p}\delta\left(\tau-\tau_{n,p}\right)\delta\left(\boldsymbol{\Theta}^{\mathrm{BS}}-\boldsymbol{\Theta}_{n,p}^{\mathrm{BS}}\right)\delta\left(\boldsymbol{\Theta}^{\mathrm{MS}}-\boldsymbol{\Theta}_{n,p}^{\mathrm{MS}}\right) \quad (1)$$

where K is the set of visible cluster indexes, $\alpha_{n,p}$ is the complex amplitude of the *p*th MPC in the *n*th cluster, $\boldsymbol{\Theta}_{n,p}^{\mathrm{BS}}$ is the Direction of Departure (AoD, EoD), and $\boldsymbol{\Theta}_{n,p}^{\mathrm{MS}}$ is the Direction of Arrival (AoA, EoA) of the MPC.

Eight different channel scenarios generate the multipaths that serve as the input data for preprocessing. The eight channels are as follows:

- indoor, Band 1 (B1), Line-of-Sight (LOS), Single Link (SL),
- indoor, Band 2 (B2), line-of-sight, single link,
- Semi-Urban (SU), band 1, line-of-sight, single link,
- semi-urban, band 2, line-of-sight, single link,
- semi-urban, band 1, Non-Line-of-Sight (NLOS), single link,
- semi-urban, band 2, non-line-of-sight, single link,
- semi-urban, band 1, line-of-sight, Multiple Links (ML), and
- semi-urban, band 2, line-of-sight, multiple links.

Thirty trials were selected to represent a more extensive set based on the central limit theorem [23, 24]. Each trial has different multipaths and multipath clusters representing standard propagation settings in a wireless communications system. The study uses the MATLAB implementation of the COST 2100 channel [25, 26]. The generation of the eight-channel scenarios has the following initializations:

- the network characteristics are based on the parameterization of C2CM [27, 28],
- the BS location is at the geometric reference point (0, 0, 0),
- the MS position is randomized at a given distance from BS with a maximum distance of up to $\frac{\sqrt{2}}{2}\times \text{cell radius}$ to ensure that the cluster measurements are nontrivial (greater than 2) and that the MS position is within the cell radius of the network,
- the MS elevation is randomized for the indoor and semi-urban channel scenarios with a random height difference for BS of up to 15 meters for the semi-urban environment and 9 meters for an indoor environment and
- the MS velocity is randomized to be either standing still or at the average walking speed of 1.1 m/s in any random direction.

The randomized BS-MS distances were developed using the random generator of 1×2 vector of random numbers drawn from the uniform distribution in the interval (0,1). On the other hand, the randomized MS height was generated using the random integer generator drawn from the discrete uniform distribution on the interval 0 to 6 for the indoor environment. In contrast, the random scalar generator drawn from the uniform distribution in the interval (0,1) was used for the semi-urban environment [25]. The MS velocity was calculated using the pseudorandom generator Mersenne Twister. The MS antenna configuration is omnidirectional, while the BS antenna is single input single output (SISO) omnidirectional for both indoor and semi-urban single link scenarios. Moreover, the BS antenna is a 2×2 MIMO omnidirectional antenna for semi-urban multiple links scenarios [29].

B. Extraction of Wireless Channel Multipaths

The clustering procedure begins with feature selection [30]. For a double-directional radio channel [31], the parameters τ, ϕ_{AOD}, θ_{AOD}, ϕ_{AOA}, and θ_{AOA} are extracted and generated using MATLAB to serve as the raw data, which can be expressed as

$$X_{\mathrm{RAW}}{}' = [\tau \quad \phi_{\mathrm{AOD}} \quad \theta_{\mathrm{AOD}} \quad \phi_{\mathrm{AOA}} \quad \theta_{\mathrm{AOA}}] \quad (2)$$

The extraction process concurs with C2CM. Each snapshot generates $X_{\mathrm{RAW}}{}'$ with a dimension of 5. There are thirty sets of $X_{\mathrm{RAW}}{}'$ data per channel scenario for clustering. The parameters obtained are representations of each multipath pre-assigned to a particular cluster. The multipaths are filtered to get only those visible in a single snapshot. The LOS component with the highest amplitude and the least delay is removed as it does not constitute multipaths.

C. Transformation of Input Data

The input data from C2CM is transformed using the Directional Cosine Transform (DCT) and the Whitening Transform (WT). The problem with the circular nature of the angular domain is solved by the directional cosine Cartesian equivalents. The result is the transformation of Eq. (2) from 5 dimensions to 7 dimensions, which can be expressed as

$$X_{\mathrm{RAW}} = [\tau \quad x_{\mathrm{AOD}} \quad y_{\mathrm{AOD}} \quad z_{\mathrm{AOD}} \quad x_{\mathrm{AOA}} \quad y_{\mathrm{AOA}} \quad z_{\mathrm{AOA}}] \quad (3)$$

There are two additional columns generated but not required in clustering the multipaths. Column eight power component is removed as it is not needed in the clustering process. Column nine cluster Identifications (IDs) are also eliminated in the clustering process as they are only used as reference IDs to compare with the calculated IDs. Dip-dist examines the cluster ability of the transformed data where data with two or more clusters can be clustered while data with only one cluster cannot be clustered [32]. WT follows to standardize the data since they have different units from the dimensions, angle, and delay. WT eliminates unwanted noise resulting in a more efficient clustering of data. The whitened data, X_{WT}, is then normalized [0, 1] using

$$X_{\mathrm{NORM}} = \left(X_{\mathrm{WT}} - X_{\min}\right) \cdot \left(X_{\max} - X_{\min}\right) \quad (4)$$

where X_{NORM} is the normalized value of the whitened data, X_{WT} is the whitened data, $X_{\max}$ is the maximum value of each column, $X_{\min}$ is the minimum value of each column, and $\cdot$ is the Hadamard product. X_{NORM} is the input to the clustering approaches and serves as the reference data in calculating the Jaccard index. It is easier to see the statistical significance of methods among one another if one dataset is the focus. Their cluster labels were regarded as ground truth for the simulated datasets used in the paper. Measurements do not have their multipath cluster labels. So even if the clusters in the measurement were determined, the labels might not be the ground truth. If the cluster labels of the measurement are ground truth, then that would be future work.

D. Clustering of the Input Data

X_{NORM} is clustered using SCAMSMA, 3CAM-SCAMSMA, SC, and 3CAM-SC. SCAMSMA represents the data as the product of the data and an affinity matrix to solve the number and membership of multipath clusters simultaneously. An ideal affinity matrix is introduced, which can be factorized by an indicator matrix whose rows indicate to which cluster a point belongs. The accuracy of SCAMSMA depends on the correct formulation of the affinity matrix so that a 0-1 block diagonal is formed. SCAMSMA clusters the data accurately when the affinity matrix is formed correctly. 3CAM-SCAMSMA is a modified SCAMSMA that uses 3CAM to formulate the affinity matrix. 3CAM depends on three constraints: pairwise, binary, and proximity. The pairwise constraint is based on the absolute distance between the corresponding data pair for all dimensions. The binary constraint takes on the sum of the values of the pairwise constraints of all dimensions. It returns a value of one (same cluster) if the sum is greater than or equal to a predefined value or otherwise zero (not on the same cluster). The proximity constraint combines all the data points to form clusters around the main diagonal, which form a 0-1 block diagonal of the similarity matrix. The rest of the procedure for SCAMSMA follows to calculate the output clusters.

SC is a data analysis technique that reduces complex multidimensional datasets into clusters with fewer dimensions. The goal is to cluster the data based on their similarity. SC accepts the similarity matrix $S \in R_{n \times n}$ with k clusters as input. The similarity graph is constructed with the weighted adjacency matrix W. The normalized Laplacian L is computed, followed by the k eigenvectors. The points are then clustered using K-means to give the clusters as the output. 3CAM-SC is a modified SC that calculates the similarity matrix using 3CAM. The clustering algorithms used by the authors focus on the computational aspect over the physics aspect, though a joint method would undoubtedly be helpful, but that requires computationally demanding resources. Also, they need cluster labels which measurements do not have, as shown in Table 1. The labels are not directly available from the output of the channel model but must be obtained in the code before channel snapshot generation.

Table 1. Presence of cluster labels in the dataset where the membership ID or cluster label of the simulated data was extracted from the COST 2100 channel model

Dataset	Cluster Label	Example
Measured data	Without	Channel impulse response and channel frequency response
Simulated data	With	COST 2100, IMT-2020, and QuaDRiGa

The clustered data serves as the calculated data in computing the Jaccard index. Using common data for the clustering approaches standardizes the comparison of their clustering performance. The Jaccard index, which serves as the similarity measure, is calculated as

$$\eta = \frac{\left|C_{11}\right|}{\left|C_{11} + C_{10} + C_{01}\right|} \quad \in \quad [0,1] \quad (5)$$

where $|\cdot|$ refers to cardinality, $C_k \in C$, $K = |C|$ is the number of multipath clusters, C_{11} is the number of clusters that are present in the calculated clusters that are also present in the reference clusters, C_{10} is the number of clusters that are present in the calculated clusters but not present in the reference clusters, and C_{01} is the number of clusters that are present in the reference clusters but not present in the calculated clusters. For the membership of the clusters, the Jaccard index is calculated as

$$\eta = \frac{M_{11}}{M_{11} + M_{01} + M_{10}} \quad \in \quad [0,1] \quad (6)$$

where M_{11} is the number of members that are present in the calculated clusters that are also present in the reference clusters, M_{10} is the number of members that are present in the calculated clusters but not present in the reference clusters, and M_{01} is the number of members that are present in the reference clusters but not present in the calculated clusters. A Jaccard index of one means that the calculated multipath clusters are the same as the reference multipath clusters or the membership of the calculated multipath clusters is the same as the membership of the reference multipath clusters. In contrast, a zero Jaccard index means no calculated multipath clusters equal to the reference multipath clusters or no membership of the calculated multipath clusters equal to the membership of the reference multipath clusters.

E. Clustering Performance Evaluation

The performance of the clustering approaches in clustering the multipaths is evaluated through clustering accuracy, computational time, and robustness. Performance analyses on these areas show the strengths and weaknesses of the clustering algorithms quantitatively.

The clustering accuracy of the clustering approaches is evaluated using the Jaccard index. Thirty sets of data, each with seven dimensions, are generated and clustered. The indices are assessed with each other and compared with the results of the state-of-the-art clustering approaches.

Computational time is measured when an algorithm clusters the data from the press of the start button until the results are displayed. The mean serves as the basis since there are thirty sets of data for each algorithm. A short duration means the algorithm is straightforward to compute, while a more prolonged period means the algorithm is more likely to be computationally complex.

Robustness is based on the performance of the clustering algorithms on the eight-channel scenarios. A clustering algorithm can be robust when it performs consistently well for all channels. Robustness is assessed objectively by the standard deviation of the Jaccard indices. Analysis of Variance (ANOVA) is also applied to evaluate the consistency of the performance of the clustering algorithm. If the F-statistic p-value is smaller than the significance level (0.05), then the test rejects the null hypothesis that all group means are equal and concludes that at least one of the group means is different from the others.

III. Results and Discussions

The clustering results are presented and analyzed. The performance of the clustering approaches is compared based on their clustering accuracy, computational time, and robustness. Results show that 3CAM-SC has the best clustering performance.

A. Clustering Accuracy

The Jaccard index is used as the validation metric in analyzing the accuracy of the clustering approaches in clustering the multipaths. The Jaccard index compares the similarity of the reference dataset, which serves as the input to the clustering approach, and the clustered dataset, the output of the clustering approach. The mean Jaccard indices of the number clusters of the four clustering approaches for all channel scenarios are presented in Table 2. In contrast, that of the membership of clusters can be found in Table 3 [19, 21]. For both tables, indoor channel scenarios have better accuracy due to the fewer multipaths and multipath clusters generated by the enclosed space where reflections of signals are limited. Also, semi-urban scenarios have lower accuracy due to the higher number of multipaths and multipath clusters generated by the broader surroundings where more interacting objects reflect the signals. SCAMSMA and SC registered Jaccard indices close to zero for the number of clusters in semi-urban scenarios. At the same time, 3CAM-SCAMSMA and 3CAM-SC recorded Jaccard indices close to 1 for the membership of clusters for all channel scenarios. Among all the clustering approaches, 3CAM-SC registered the best clustering accuracy.

Table 2. Number of clusters mean Jaccard indices of the four clustering approaches for the eight-channel scenarios

Channel Scenario	SCAMSMA	3CAM-SCAMSMA	SC	3CAM-SC
Indoor B1	0.6034	0.8226	0.5405	0.9027
Indoor B2	0.6487	0.8004	0.2847	0.8894
SU B1 LOS SL	0.0186	0.5525	0.0199	0.5533
SU B2 LOS SL	0.0159	0.6721	0.0176	0.6730
SU B1 NLOS SL	0.0052	0.4594	0.0126	0.4594
SU B2 NLOS SL	0.0108	0.5325	0.1940	0.5372
SU B1 LOS ML	0.0080	0.3229	0.0078	0.3227
SU B2 LOS ML	0.0084	0.5805	0.0087	0.5809

Table 3. Membership of clusters mean Jaccard indices of the four clustering approaches for the eight-channel scenarios

Channel Scenario	SCAMSMA	3CAM-SCAMSMA	SC	3CAM-SC
Indoor B1	0.7305	0.9641	0.7612	0.9640
Indoor B2	0.7582	0.9588	0.5936	0.9585
SU B1 LOS SL	0.1875	0.9761	0.2496	0.9761
SU B2 LOS SL	0.1818	0.9815	0.2459	0.9815
SU B1 NLOS SL	0.1597	0.9825	0.2344	0.9825
SU B2 NLOS SL	0.1505	0.9858	0.5344	0.9859
SU B1 LOS ML	0.1459	0.9600	0.1798	0.9598
SU B2 LOS ML	0.1436	0.9779	0.1771	0.9779

Table 4 shows the percentage increase in the mean Jaccard indices when 3CAM is used to calculate the affinity matrix of SCAMSMA [19] and the similarity matrix of SC. All the mean Jaccard indices of the number of clusters and the membership of clusters for all channel scenarios increased, most notably in the semi-urban scenarios of the number of clusters. For the indoor scenarios, the improvement is at most 212.40% for the number of clusters and 61.47% for the membership of clusters from SC to 3CAM-SC. However, in the semi-urban scenarios, the improvement reaches 8734.62%, as shown in the number of clusters of SU B1 NLOS SL and 580.99%, as manifested in the membership of clusters of SU B2 LOS ML both from SCAMSMA to 3CAM-SCAMSMA. The increase in the Jaccard indices conveys that 3CAM improves the clustering performance of SCAMSMA and SC.

Table 4. Percentage increase in the mean Jaccard indices due to 3CAM in formulating the affinity matrix of SCAMSMA and similarity matrix of SC

Channel Scenario	SCAMSMA to 3CAM-SCAMSMA Number of Clusters	SCAMSMA to 3CAM-SCAMSMA Membership of Clusters	SC to 3CAM-SC Number of Clusters	SC to 3CAM-SC Membership of Clusters
Indoor B1	36.33	31.98	67.01	26.64
Indoor B2	23.39	26.46	212.40	61.47
SU B1 LOS SL	2870.43	420.59	2680.40	291.07
SU B2 LOS SL	4127.04	439.88	3723.86	299.15
SU B1 NLOS SL	8734.62	515.22	3546.03	319.16
SU B2 NLOS SL	4830.56	555.02	176.91	84.49
SU B1 LOS ML	3936.25	557.98	4037.18	433.82
SU B2 LOS ML	6810.71	580.99	6577.01	452.17

The performance of the clustering approaches is assessed based on the cluster classification and channel scenario. The clustering approaches can be compared since all solved the number and membership of clusters simultaneously. Fig. 3 presents the performance comparison of the four clustering approaches for the number of clusters in indoor scenarios (blue) and semi-urban scenarios (red). 3CAM-SC has the highest clustering accuracy in indoor scenarios, while 3CAM-SCAMSMA and 3CAM-SC performed equally in the semi-urban scenarios. Fig. 4 shows the performance comparison of the clustering approaches for the membership of clusters in indoor scenarios (green) and semi-urban scenarios (yellow). 3CAM-SCAMSMA and 3CAM-SC have almost identical performances in indoor and semi-urban scenarios. The figures clearly show that 3CAM significantly improves the performance of SCAMSMA and SC in formulating the affinity matrix and similarity matrix, respectively.

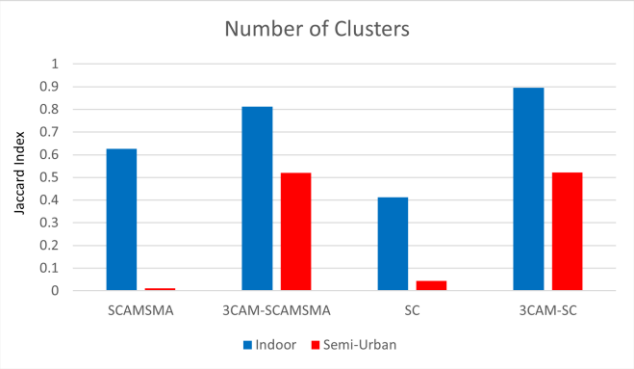

Fig. 3. Performance comparison of the clustering approaches for the number of clusters in indoor scenarios (blue) and semi-urban scenarios (red).

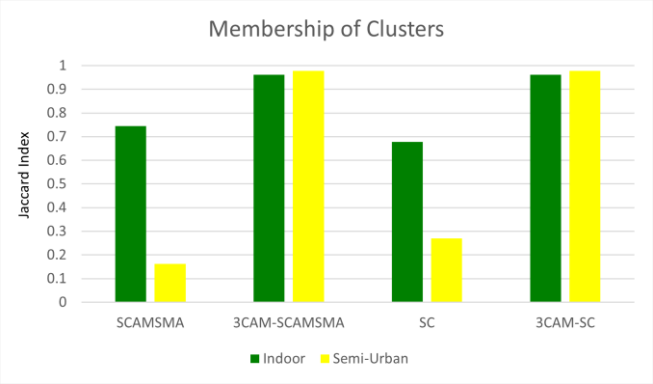

Fig. 4. Performance comparison of the clustering approaches for the membership of clusters in indoor scenarios (green) and semi-urban scenarios (yellow).

B. Computational Time

The mean computational time of the thirty sets of data per channel scenario using 3CAM-SCAMSMA, SC, and 3CAM-SC is presented in Table 5 [19, 21]. The simulations were done in MATLAB 2019a on a Dell 7730 mobile workstation with Windows 10 operating system, Intel Xeon E2186M 2.90 GHz CPU, and 64 GB memory. The computing time is based on a period counter function in MATLAB. The timer begins at the pressing of the start button and ends when the simulation stops. The computational duration depends on the number of multipath components and clusters. The value means that the higher the number of multipath components and clusters, the longer the computational duration. That is why the indoor scenarios have a short computational duration due to fewer multipath components and multipath clusters. In comparison, semi-urban line-of-sight single-link scenarios increased due to a higher number of multipath components and multipath clusters. 3CAM-SC has the least computational complexity for all channel scenarios, while 3CAM-SCAMSMA has the highest computational time.

Table 5. Mean computational duration (in seconds) of 3CAM-SCAMSMA, SC, and 3CAM-SC for the eight-channel scenarios

Channel Scenario	3CAM-SCAMSMA	SC	3CAM-SC
Indoor B1	2.96	0.94	0.93
Indoor B2	2.57	0.95	0.94
SU B1 LOS SL	330.92	6.29	3.33
SU B2 LOS SL	400.88	7.34	3.58
SU B1 NLOS SL	2780.95	26.67	16.08
SU B2 NLOS SL	2634.04	24.76	16.14
SU B1 LOS ML	4515.58	38.72	15.02
SU B2 LOS ML	6168.09	39.40	17.47

C. Robustness

The robustness of a clustering approach is based on its consistent performance in clustering multipaths in all channel scenarios. It is assessed by the standard deviations of the mean Jaccard indices and the box plots using the anova1 function of MATLAB. The clustering approach with the slightest variations is said to be robust.

1) Standard deviation

Table 6 shows the standard deviations of the mean Jaccard indices of the number of clusters of the four clustering approaches for all channel scenarios. In contrast, Table 7 shows the membership of clusters [19]. The standard deviation gives the variation of the Jaccard indices from the mean. A low standard deviation means that most of the Jaccard indices are close to the mean, while a high standard deviation indicates that the Jaccard indices are more spread out. For the number of clusters in indoor scenarios, SCAMSMA and SC have higher standard deviations than 3CAM-SCAMSMA and 3CAM-SC, even though they have lower mean Jaccard indices. It shows that those standard deviations and mean Jaccard indices are inversely related. On the other hand, the two are directly related to semi-urban scenarios since 3CAM-SCAMSMA and 3CAM-SC have higher standard deviations than SCAMSMA and SC due to their higher mean Jaccard indices.

Table 6. Standard deviations of the mean Jaccard indices of the number of clusters of the four clustering approaches for the eight-channel scenarios

Channel Scenario	SCAMSMA	3CAM-SCAMSMA	SC	3CAM-SC
Indoor B1	0.2435	0.1783	0.2655	0.0997
Indoor B2	0.3038	0.1949	0.2201	0.1116
SU B1 LOS SL	0.0100	0.3579	0.0079	0.3568
SU B2 LOS SL	0.0089	0.3860	0.0067	0.3850
SU B1 NLOS SL	0.0096	0.3716	0.0110	0.3716
SU B2 NLOS SL	0.0122	0.3506	0.0902	0.3471
SU B1 LOS ML	0.0042	0.2855	0.0061	0.2855
SU B2 LOS ML	0.0046	0.3221	0.0048	0.3215

For the membership of clusters, 3CAM-SCAMSMA and 3CAM-SC have almost the same standard deviations, showing that their Jaccard indices have similar variations

from the means. Their standard deviations are lower in indoor scenarios, even though they have high means, and this result indicates closer Jaccard indices from the means. For the semi-urban scenarios, 3CAM-SCAMSMA and 3CAM-SC have lower standard deviations except in scenarios with multiple links. This outcome signifies that their Jaccard indices are more stable around the mean despite higher Jaccard indices.

Table 7. Standard deviations of the mean Jaccard indices of the membership of clusters of the four clustering approaches for the eight-channel scenarios

Channel Scenario	SCAMSMA	3CAM-SCAMSMA	SC	3CAM-SC
Indoor B1	0.1808	0.0379	0.1670	0.0379
Indoor B2	0.2261	0.0435	0.1501	0.0439
SU B1 LOS SL	0.0283	0.0227	0.0280	0.0227
SU B2 LOS SL	0.0216	0.0236	0.0333	0.0236
SU B1 NLOS SL	0.0278	0.0148	0.0237	0.0148
SU B2 NLOS SL	0.0244	0.0128	0.0590	0.0126
SU B1 LOS ML	0.0123	0.0213	0.0179	0.0213
SU B2 LOS ML	0.0137	0.0192	0.0182	0.0192

3CAM-SCAMSMA and 3CAM-SC have almost identical standard deviations. However, when the two are compared closely, 3CAM-SC has lower values. Thus, it has the slightest Jaccard index variation and is the most robust among the four clustering approaches.

2) Analysis of Variance (ANOVA)

The box plots of the mean Jaccard indices of the number of clusters in indoor scenarios are shown in Fig. 5. The halfway mark (red line segment) indicates the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered are outliers. The box plots are generated using the one-way ANOVA of MATLAB. The purpose of one-way ANOVA is to determine whether data from several groups of a factor have a common mean. That is, one-way ANOVA can determine whether different groups of an independent variable have different effects on the response variable. The probability value (*p*-value), the box plots of the independent variable, and tests (the hypothesis that the samples in the independent variable are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same) are drawn from the utilized ANOVA tool. Values of $p < 0.05$ indicate that the means of the clustering approaches are significantly different. The *p*-value of Fig. 5 is 0.0217, which validates that the mean Jaccard indices of the clustering approaches are significantly different. 3CAM-SC has the best performance, as shown by the higher central mark, while SC has the worst accuracy.

The box plots of the mean Jaccard indices of the number of clusters in semi-urban scenarios are shown in Fig. 6. The *p*-value is 5.0224×10^{-10}, which indicates that the mean Jaccard indices differ significantly. 3CAM-SCAMSMA and 3CAM-SC have almost the same performance, as their box plots show. SCAMSMA and SC have mean Jaccard indices close to zero, and SC has an outlier Jaccard index, indicated by the red plus symbol.

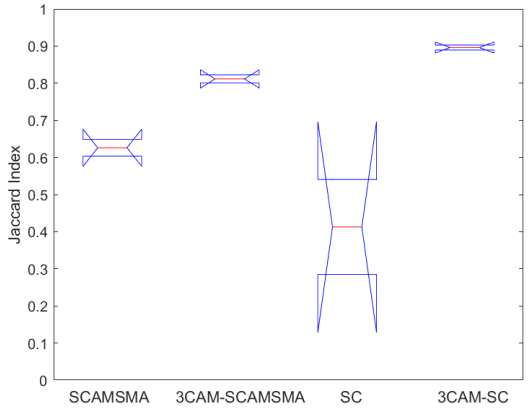

Fig. 5. Box plots of the indoor scenarios mean Jaccard indices of the number of clusters using the anova1 one-way approach of MATLAB with *p*-value = 0.0217.

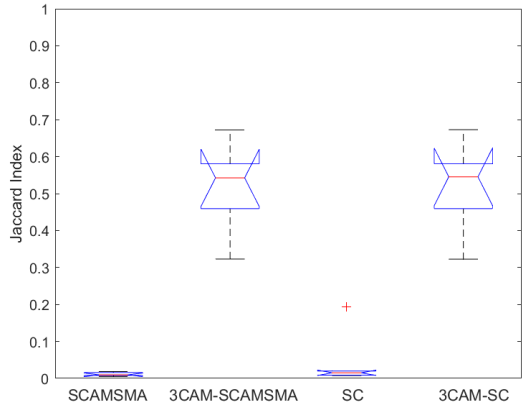

Fig. 6. Box plots of the semi-urban scenarios mean Jaccard indices of the number of clusters using the anova1 one-way approach of MATLAB with *p*-value = 5.0224×10^{-10}.

The box plots of the mean Jaccard indices of the membership of clusters in indoor scenarios are shown in Fig. 7. The p-value is 0.0182, indicating that the mean Jaccard indices differ significantly. 3CAM-SCAMSMA and 3CAM-SC have almost identical performances, as their box plots show. SC has the most varied mean Jaccard indices, as its more extended box plot shows.

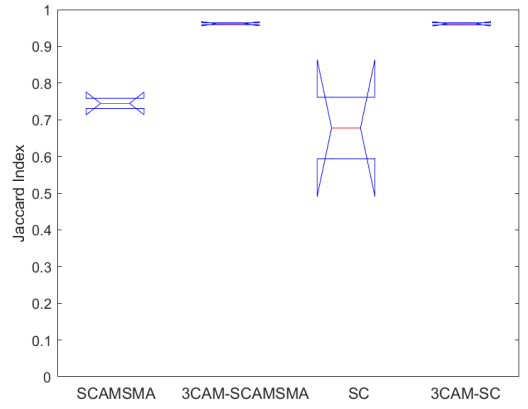

Fig. 7. Box plots of the indoor scenarios mean Jaccard indices of the membership of clusters using the anova1 one-way approach of MATLAB with *p*-value = 0.0182.

The box plots of the mean Jaccard indices of the membership of clusters in semi-urban scenarios are shown in Fig. 8. The *p*-value is 4.0754×10^{-16}, which indicates that the mean Jaccard indices differ significantly. 3CAM-SCAMSMA and 3CAM-SC have almost identical performances with mean Jaccard indices close to one, as shown by their box plots. SC has the most varied mean Jaccard indices, as shown by its more extended box plot. It also has an outlier mean Jaccard index, as indicated by the red plus symbol.

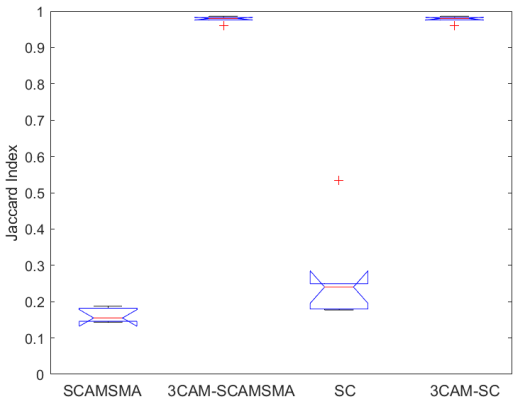

Fig. 8. Box plots of the semi-urban scenarios mean Jaccard indices of the membership of clusters using the anova1 one-way approach of MATLAB with *p*-value = 4.0754×10^{-16}.

3CAM-SCAMSMA and 3CAM-SC have almost the same box plots, as shown in Fig. 6 to Fig. 8. However, 3CAM-SC performs better, as shown in Fig. 5. Thus, 3CAM-SC is the most robust based on the box plots.

IV. Conclusion

The work presented the performance of SCAMSMA, 3CAM-SCAMSMA, SC, and 3CAM-SC in clustering wireless multipaths generated by C2CM. The clustering approaches solved the number and membership of multipath clusters concurrently. The clustering accuracy was assessed using the Jaccard index, the computational complexity was compared using computing time, and the robustness on standard deviation and box plots. Results show that SCAMSMA and SC fared well in indoor scenarios but had poor performance in semi-urban scenarios. 3CAM-SCAMSMA and 3CAM-SC had improved performance in all channel scenarios. The two have almost identical mean Jaccard indices. Nevertheless, with 3CAM-SC having a faster computational duration, lesser standard deviations, and higher box plots, it is the most robust among the clustering approaches.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

JB conducted the research; LM analyzed the data; JB wrote the paper; LM edited the paper; all authors approved the final version.

Acknowledgment

The authors wish to thank De La Salle University for the publication support.

References

[1] M. S. Haghighi, H. S. Yazdi, and A. Vahedian, "A hierarchical possibilistic clustering," *International Journal of Computer Theory and Engineering*, vol. 1, no. 4, pp. 465–472, 2009.

[2] S. Vijendra, S. Laxman, and K. Ashwini, "Mining clusters in data sets of data mining: an effective algorithm," *International Journal of Computer Theory and Engineering*, vol. 3, no. 1, pp. 171–177, 2011.

[3] A. Kamble, "Incremental clustering in data mining using genetic algorithm," *International Journal of Computer Theory and Engineering*, vol. 2, no. 3, pp. 326–328, 2010.

[4] S. Aranganayagi and K. Thangavel, "Clustering categorical data using bayesian concept," *International Journal of Computer Theory and Engineering*, vol. 1, no. 2, pp. 119–125, 2009.

[5] H. Q. Bao, L. V. Vinh, and T. Van Hoai, "A deep embedded clustering algorithm for the binning of metagenomic sequences," *IEEE Access*, vol. 10, pp. 348–357, 2022.

[6] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz, 3GPP TR 38.901 V17.0.0," Tech. Rep., 2022.

[7] ITU-R, "Series M guidelines for evaluation of radio interface technologies for IMT-2020, ITU-R M.2412-0," Tech. Rep., 2017.

[8] S. Jaeckel, L. Raschkowski, K. Börner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. on Antennas and Propagation*, vol. 62, no. 6, pp. 3242–3256, 2014.

[9] R. Verdone and A. Zanella, *Pervasive Mobile and Ambient Wireless Communications: COST Action 2100*, Springer London: Signals and Communication Technology, 2012.

[10] K. Yu, Q. Li, D. Cheung, and C. Prettie, "On the tap and cluster angular spreads of indoor WLAN channels," in *Proc. 2004 IEEE 59th Vehicular Technology Conf.*, 2004, pp. 218–222.

[11] N. Czink, P. Cera, J. Salo, E. Bonek, J. Nuutinen, and J. Ylitalo, "A framework for automatic clustering of parametric MIMO channel data including path powers," in *Proc. 2006 IEEE 64th Vehicular Technology Conf.*, 2006, pp. 1–5.

[12] C. Gentile, "Using the kurtosis measure to identify clusters in wireless channel impulse responses," *IEEE Trans. on Antennas and Propagation*, vol. 61, no. 6, pp. 3392–3395, 2013.

[13] B. Li, C. Zhao, H. Zhang, Z. Zhou, and A. Nallanathan, "Efficient and robust cluster identification for ultrawideband propagations inspired by biological ant colony clustering," *IEEE Trans. on Communications*, vol. 63, no. 1, pp. 286–300, 2015.

[14] S. Cheng, M.-T. Martinez-Ingles, D. Gaillot, M.-P. J-M., M. Lienard, and P. Degauque, "Performance of a novel automatic identification algorithm for the clustering of radio channel parameters," *IEEE Access*, vol. 3, pp. 2252–2259, 2015.

[15] R. He, W. Chen, B. Ai, A. Molisch, W. Wang, Z. Zhong, J. Yu, and A. Sangodoyin, "On the clustering of radio channel impulse responses using sparsity based methods," *IEEE Trans. on Antennas and Propagation*, vol. 64, no. 6, pp. 2465–2474, 2016.

[16] R. He, Q. Li, B. Ai, Y.-A. Geng, A. F. Molisch, V. Kristem, Z. Zhong, and J. Yu, "A kernel-power-density-based algorithm for channel multipath components clustering," *IEEE Transactions on Wireless Communications*, vol. 16, no. 11, pp. 7138–7151, 2017.

[17] Y. Li, J. Zhang, Z. Ma, and Y. Zhang, "Clustering analysis in the wireless propagation channel with a variational Gaussian mixture model," *IEEE Trans. on Big Data*, vol. 6, no. 2, pp. 232–238, 2020.

[18] Z. Li, L.-F. Cheong, S. Yang, and K.-C. Toh, "Simultaneous clustering and model selection: Algorithm, theory and applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1964–1978, 2018.

[19] J. Blanza and L. Materum, "Three-constraints affinity matrix on simultaneous identification of the clustering and cardinality of wireless propagation multipaths," *Journal of the Franklin Institute*, vol. 359, no. 5, pp. 2359–2376, 2022.

[20] A. Ng, Y. Weiss, and M. Jordan, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. 2001, vol. 14, pp. 849–856.

[21] J. Blanza, "Wireless propagation multipaths using spectral clustering and three-constraint affinity matrix spectral clustering," *Baghdad Science Journal*, vol. 18, no. 2 (Supplement June), pp. 1001–1011, 2021.

[22] L. Liu, C. Oestges, J. Poutanen, K. Haneda, P. Vainikainen, F. Quitin, F. Tufvesson, and P. Doncker, "The COST 2100 MIMO channel model," *IEEE Wireless Communications*, vol. 19, no. 6, pp. 92–99, 2012.

[23] P. Peebles, *Probability, Random Variables And Random Signal Principles*, McGraw-Hill Higher Education. McGraw-Hill Education (India) Pvt Limited, 2002.

[24] D. Stroock, *Probability Theory: An Analytic View*, Cambridge University Press, 2010.
[25] L. Liu. (2018). COST 2100 channel model. [Online]. Available: http://github.com/cost2100/cost2100/tree/master/matlab
[26] J. Blanza and L. Materum. (2022). Concurrent and spectral clustering of wireless waves. [Online]. Available: https://codeocean.com/capsule/1851626/tree
[27] 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. the 5th European Conf. on Antennas and Propagation*, 2011, pp. 3606–3610.
[28] 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. on Wireless Communications*, vol. 12, no. 2, pp. 888–897, 2013.
[29] 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 *Proc. the 2019 International Symposium on Multimedia and Communication Technology*, 2019, pp. 1–5.
[30] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Trans. on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.
[31] M. Steinbauer, A. Molisch, and E. Bonek, "The double directional radio channel," *IEEE Antennas and Propagation Magazine*, vol. 43, no. 4, pp. 51–63, 2001.
[32] M. Maechler. (2003). Hartigan's diptest for unimodality. [Online]. Available: https://github.com/mmaechler/diptest