International Journal of Advanced Research in Science, Engineering and Technology

Vol. 12, Issue 7, July 2025

ISSN: 2350-0328

IJARSET IMPACT FACTOR 7.150

Density Based Optimal Spatial Clustering of Applications with Noise Using Uniform Manifold Approximation and Projection (UMAP) and T-Stochastic Neighbour Embedding (T-SNE) for 5G Propagation Scenarios

Sriram Yadav, Dr. Prajakta P Shirke

A.P., SOCSE, Sandip University Madhubani, India
A.P., SOCSE, Sandip University Nashik, India

ABSTRACT: The multipath dataset is reduced into 2-dimensions (2D) for visualization and clustering applying Uniform Manifold Approximation and Projection (UMAP). Clustering approach density-based spatial clustering of applications with noise (DBSCAN) is used and the performance of variant search radius epsilon ε is evaluated. The COST2100 channel model's (C2CM) semi-urban scenarios are used to cluster for proposed approach, which has more than one multipath component (MPC). Comparing the clustering results to the ground truth and computing the Adjusted Rand Index (ARI) and the cluster-wise Jaccard index η the approach is validated. Due to the overlapping nature of cluster, a median below 0.5 in the multiple link scenarios is achieved by lowering the search radius up to 0.2 as the results suggest. In spite of that, the single-link scenarios indicate the robustness of the approach if the median values for the ARI and Jaccard index η, respectively goes above 0.6 and 0.7

KEYWORDS: channel modelling, dimensionality reduction, multipath clustering.

I. INTRODUCTION

The development of wireless communications relies on characterizing the physical wireless channel. Utilizing Multiple-Input Multiple-Output (MIMO) antennas at both the transmitter and receiver enables the wireless systems to achieve fourth-generation (4G) and fifth-generation (5G) mobile communication standards. Channel models have been used to test algorithms and approaches before developing the actual system, using the characteristics of the complicated propagation environment; thus, channel modeling plays a vital role in wireless system design. Recently, the use of geometry-based stochastic channel models (GBSCM) has gained popularity and developed the double-directional channel models [1]. The GBSCM also uses cluster concepts, attributing clusters to scatterers in the propagation environment. The presence of multipath clusters and their exploitation opens up the benefits of spatial multiplexing, diversity, and beam forming. The scatterers produce multipath components (MPCs), which diffuse and scatter the signal propagated, causing delays and different paths called multipaths. Generating a channel model for different scenarios and accurately clustering the MPCs in the angular and time domain can simplify obtaining the channel impulse response (CIR). Accurately modeling the propagation space can produce better channel models for different scenarios that technically advance the wireless system's design and present achievable reliability, data rates, and latency. Hence, finding optimal ways to cluster MPCs is crucial in channel modeling.

Recently, the use of automatic algorithms gained attraction in clustering the MPCs. Different methods have been proposed to cluster MPCs in different scenarios. The first framework for applying an automatic algorithm was reported by [2], [3], where the *K*-means algorithm was utilized, and the Multipath Component Distance (MCD) was used. The *K*-means algorithm was also used in an urban scenario reported by [4]. A spectral-based power-weighted algorithm was proposed by [5] and applied to measured data in a hall environment. The automatic and manual approach combined to produce a middle-ground technique is reported in [6], in an urban macro-cell where it was stated that human interaction in clustering should not be ignored.

Furthermore, the K-means is extended to include the power, also known as the K-power means (KPM) algorithm, and tracking the multipaths was reported in [7]. The Variational Gaussian Mixture Model (GMM) was proposed to cluster the outdoor-to-indoor propagation scenario [8]. Additionally, a comparative study of different algorithms and their performance is presented in [9]. The use of Simultaneous Clustering and Model Selection (SCAMS) is proposed in [10,11] to cluster the C2CM generated datasets. Finally, a visualization tool using 3D point cloud data to locate small interacting objects in a microcell was proposed to include the visualization process in the wireless channel characterization [12]. Different approaches are presented in clustering the MPCs due to the difference in propagation environment, frequency band, and scenario being studied and modeled.

Big data analysis gained research interest due to the massive data available today. Visualization techniques have been one of the focuses of research in exploratory data analysis. Data's high dimensional nature gave birth to techniques to reduce the dimensionality of data to address the “curse of dimensionality.” Dimensionality reduction (DR) has benefitted other fields of science, especially for data with many features that need to be reduced, visualized, and clustered. DR techniques can preserve the global or local structure of the data while reducing the features that can be used to visualize in lower dimensions. The mutation dataset of the SARS-CoV-2 was visualized using t-distributed Stochastic Neighbor Embedding (t- SNE) and UMAP, followed by the K-means algorithm for clustering [13]. DR has gained popularity, especially in genome sequencing. In [14], DBSCAN is used after applying t-SNE and UMAP to find repeating patterns in the biological signaling of single-cell calcium spiking.

Furthermore, DR techniques have been used to improve the performance of different machine learning algorithms for intrusion detection systems [15]. Furthermore, DR techniques before automatic clustering have not been fully utilized in clustering the MPCs. The main contribution of this paper is to obtain accurate number of clusters and their membership using UMAP to reduce the dimension into 2 (for visualization) and cluster the latent space using DBSCAN. This paper applies the DR technique to the C2CM data to visualize the MPCs using Principal Component Analysis (PCA) and UMAP, followed by DBSCAN to cluster the reduced data.

II.METHODS

The methodology and techniques in this study are discussed. Figure 1 illustrates the procedures implemented from the dataset used, the PCA and UMAP techniques, and the MCD distance metric. The dataset is prepared and read using MATLAB software. Each scenario has 30 snapshots that are read per sheet in an Excel file imported to the MATLAB interface. The PCA is applied to rotate the data, capturing the maximum variance without reducing the dimension. The UMAP is modified using the MCD metric to suit the angular nature of data to capture the actual distance and avoid the circular nature of data. The computation of MCD is done using a developed script in MATLAB. The UMAP then computes a low-dimensional representation in 2D that can be visualized and utilized for clustering.

Additionally, the default DBSCAN implementation in MATLAB was used, and the ε was varied to improve the groupings of the UMAP output. Finally, the ARI and Jaccard index are computed to validate the approach in the number of clusters and their membership. This section discusses each of the methods used in this paper.

A. COST 2100 DATASET

The European Cooperation in Science and Technology (COST) 2100 channel model is based on the cluster of MPCs which has similar delays and angular parameters [16]. The C2CM generates the double-directional channel containing the large-scale and small-scale parameters. The dataset used in this work is found in the Institute of Electrical and Electronics Engineers (IEEE) Data Port [17], which has eight scenarios. Only the semi-urban scenarios were used in this work since the DBSCAN performance relies on the density of points and the neighbors.

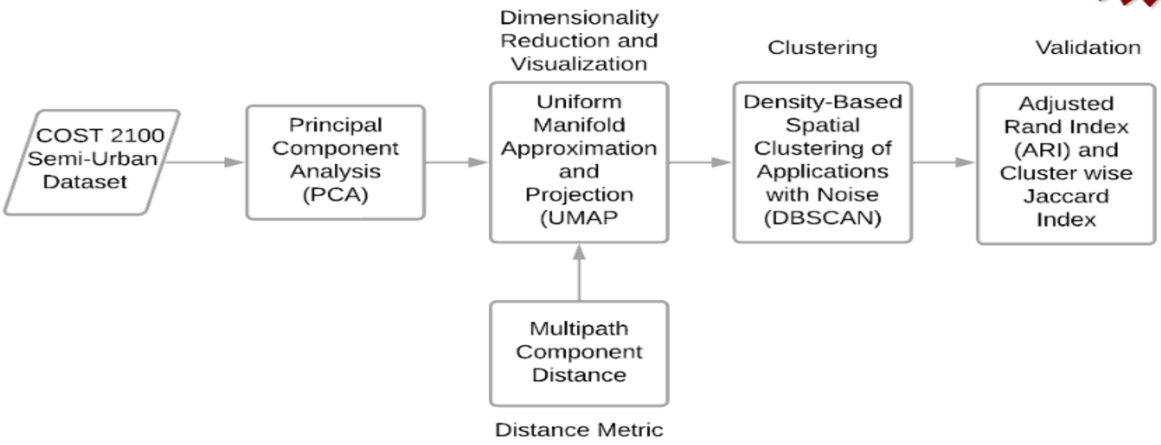

Dimensionality Reduction and Visualization
Clustering
Validation
COST 2100 Semi-Urban Dataset
Principal Component Analysis (PCA)
Uniform Manifold Approximation and Projection (UMAP
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Adjusted Rand Index (ARI) and Cluster wise Jaccard Index
Multipath Component Distance
Distance Metric

Fig. 1: Methodology of the study.

The settings are as follows:

1. Band 1 Semi-urban NLOS Single Link
2. Band 1 Semi-urban LOS Single Link
3. Band 1 Semi-urban LOS Multiple Links
4. Band 2 Semi-urban NLOS Single Link
5. Band 2 Semi-urban LOS Single Link
6. Band 2 Semi-urban LOS Multiple Links

Band 1 refers to the narrowband, while Band 2 is for the wideband. Additionally, the LOS refers to line-of-sight and NLOS refers to non-line-of-sight. A vector represents the MPCs

$\boldsymbol{x}_\ell = [\theta_{\boldsymbol{\ell},\text{AOA}}, \phi_{\ell,\text{AOA}}, \theta_{\boldsymbol{\ell},\text{AOD}}, \phi_{\ell,\text{AOD}}, \tau_{\boldsymbol{\ell}}, \alpha_\ell]$ where x_l represents the ℓ^{th} MPC which has the delay parameter τ, the Angle of Arrival (AOA), Angle of Departure (AOD), ϕ is the azimuth angle, θ as the elevation angle, and the power represented by α. The MPCs are clustered based on the similarity of these parameters, excluding the power. Each dataset contains 30 snapshots, where one snapshot is represented by a data matrix $\boldsymbol{X}$ that contains ℓ number of MPCs.

The C2CM is part of the COST action project and the COST family of channel models. A 200 MHz bandwidth is supported by this model for frequencies below 6 GHz. The model assumes that there is only one terminal fixed for the BS, which limits the dual mobility conditions. Dynamic modeling, multilink, spherical, and spatial consistency are the advantages of the C2CM as compared to other channel models.

B. DIMENSIONALITY REDUCTION

PCA is considered a DR technique that aims to direct the components to maximum variance using orthogonal axes. PCA produces uncorrelated variables, which can then be reduced to a number of principal components. The first principal component explains most of the variance in the data, followed by the second principal component. PCA can be achieved using the singular value decomposition (SVD) or the covariance matrix and its eigenvectors. In this work, all the components are retained; thus, PCA is used for the decorrelation of the variables, which are then fed to the UMAP algorithm.

UMAP is a relatively new technique for embedding high-dimensional data in low dimensions based on topological data analysis and graph theory [18]. The algorithm begins by constructing a K-neighbor graph G using a specific distance metric d which computes the distance in high dimension. The 'G' low dimensional graph is constructed using Laplacian eigenmaps. The cross-entropy between the two graphs is minimized, producing an optimized layout in the low dimension [19]. UMAP has hyperparameter n-neighbors, which in this work is set to the square root of N, where N is the number of paths in each snapshot. The reason for the square root is to deflate the varying number of MPC in each snapshot, and selecting √N is widely accepted when using KNN [20]. The MATLAB implementation of UMAP [21] is used in this work. The UMAP preserves both local and global structure of data. Another manifold learning technique that uses the same principle of optimizing the low dimensional embeddings is the t-SNE. In contrast, UMAP utilizes graph theory, while t-SNE is based on the student-t distribution in finding the neighboring points. As mentioned earlier, the PCA is a DR technique as well, but falls under the category of linear techniques and is employed alongside t-SNE or UMAP.

The parameters of the MPCs are in terms of angle and time domain. In clustering and neighborhood embedding, the similarity of the distance is the Euclidean distance which is the default distance metric of UMAP and other dimensionality reduction algorithms. It measures the distance between points in a linear manner. A more generalized distance metric is the Minkowski distance where p represents the norm where $p = 2$ for the Euclidean, and $p = 1$ for the Manhattan distance. However, because of the angular nature of the multipath components, the Euclidean distance should be modified to address the angular ambiguity and separation between the MPCs. Considering also the fact that the delay feature τ is a dimension of time, the Euclidean distance cannot measure the spatio-temporal difference of MPCs. The MCD is a measure to quantify the separations between multipaths i, introduced in [22] given by Eq. 1:

$$\mathrm{MCD}_{ij} = \sqrt{\|\mathrm{MCD}_{\mathrm{AoA},ij}\|^2 + \|\mathrm{MCD}_{\mathrm{AoD},ij}\|^2 + \mathrm{MCD}_{\tau,ij}^2} \tag{1}$$

$$\mathrm{MCD}_{\mathrm{AoA/AoD},ij} = \frac{1}{2}\left|\begin{pmatrix}\sin(\theta_i)\cos(\phi_i)\\ \sin(\theta_i)\sin(\phi_i)\\ \cos(\theta_i)\end{pmatrix} - \begin{pmatrix}\sin(\theta_j)\cos(\phi_j)\\ \sin(\theta_j)\sin(\phi_j)\\ \cos(\theta_j)\end{pmatrix}\right| \tag{2}$$

where $\mathrm{MCD}_{\mathrm{AoA/AoD},ij}$ computes the distance of the angle of arrival or departure; Equation 2 is used to compute the angular distance between the i^{th} and j^{th} multipath. Equation 3 quantifies the separation between delays given by $\mathrm{MCD}_{\tau,}$, where ζ is the scaling factor and the standard deviation of the delays denoted by τ_{std}.

$$\mathrm{MCD}_{\tau,ij} = \zeta \cdot \frac{|\tau_i - \tau_j|}{\Delta\tau_{max}} \cdot \frac{\tau_{std}}{\Delta\tau_{max}} \tag{3}$$

The MCD provides a metric integrated into the UMAP algorithm to measure the probability of neighboring points in the high-dimensional space. Using PCA and UMAP, where MCD is the distance metric, each snapshot is projected into 2D space for visualization, and the 2D data is clustered using the DBSCAN algorithm.

C.DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

The DBSCAN algorithm relies on three parameters, the core points, boundary points, and noise points [14]. The core points can be determined by the radius of the search represented by epsilon ε and the minimum number of neighbors (minPts). Furthermore, core points are defined as the reference of the data points in the center of the group with the least minPts connected within the distance ε. The number of core points can also be treated as the number of clusters since it defines how DBSCAN can find many core points. The DBSCAN treats the neighbors to be connected to a core point, and the non-core points are treated as noise. One advantage of DBSCAN from K-means is that DBSCAN does not require the number of clusters K in advance but produces clusters based on the density of points around the core point that the radius ε and the minPts specifies. In this paper, ε was varied from 1, 0.8, 0.5, and 0.3, while the minPts was set to the default value of 5. The reason for varying the ε in decreasing order is to separate highly dense points from the projection of UMAP. The abstract algorithm of DBSCAN is first to identify the core points, followed by assigning core points, and for non-core points, the border points are assigned, adding the neighboring points to a core point and finally assigning the noise points. The original DBSCAN paper was proposed in [23] and was recently criticized due to the misuse of distance metrics. However, the suitable use of its parameters ε and the distance metric are discussed and highlighted in [24] and still encourages its usage.

D. VALIDATION METRICS

The availability of the true cluster number and cluster membership from the C2CM data, external clustering validity indices are used. The Adjusted Rand Index (ARI), which compares the cluster members to the reference cluster membership, and the Jaccard index to assess the accuracy in identifying the number of clusters. The ARI is given in equation 4 [25].

$$\mathrm{ARI} = \frac{\boldsymbol{M}_{11} + \boldsymbol{M}_{00}}{\boldsymbol{M}_{00} + \boldsymbol{M}_{01} + \boldsymbol{M}_{10} + \boldsymbol{M}_{11}} = \frac{\boldsymbol{M}_{11} + \boldsymbol{M}_{00}}{(\mathrm{M}_{2})} \in [0,1] \tag{4}$$

The value of the ARI ranges from 1 if the clustering matches perfectly. In equation 4, $\boldsymbol{M}_{11}$ pertains to the number of pairs that exist in the same cluster while $\boldsymbol{M}_{00}$ are the instances that exist in a different cluster. The number of pairs that exist in the reference cluster but not in the clustering output is denoted by $\boldsymbol{M}_{01}$while $\boldsymbol{M}_{10}$ is for the opposite.

On the other hand, the number of clusters is evaluated using the Jaccard index, which ranges from 0 to 1, where 0 indicates void similarity and 1 for a perfect match. The clusterwise Jaccard index is computed using equation 5, where $|\cdot|$ denotes the cardinality, $\boldsymbol{C}_{\mathrm{ref}}$ is the reference multipath cluster, and $\boldsymbol{C}_{\mathrm{cal}}$ indicates the calculated clusters.

$$\eta_{\mathrm{Jac}} = \frac{|\boldsymbol{C}_{\mathrm{ref}} \cap \boldsymbol{C}_{\mathrm{cal}}|}{|\boldsymbol{C}_{\mathrm{ref}} \cup \boldsymbol{C}_{\mathrm{cal}}|} \in [0,1] \tag{5}$$

The results of cluster labels can sometimes change depending on the clustering discovery of DBSCAN. Consequently, the use of ARI to measure the similarity of cluster membership as compared to the Adjusted Mutual Index (AMI) is due to the balanced nature of clusters generated by C2CM. Since the dataset generated has equal number of MPC per cluster, the dataset is said to be balanced and the ARI is used toward a balanced clustering solution.

III. RESULTS AND DISCUSSION

The results of using PCA+UMAP and MCD are first projected with their reference clusters to validate the embedding quality. The DBSCAN is then used on the projected data to visualize and assess the performance by computing the ARI. Figure 2 illustrates the ground truth projection and the DBSCAN clustering on the UMAP results. The projected data is from one snapshot of the B1 NLOS scene with 911 MPCs. Varying the ε value shows a difference in the clusters in the middle being separated as the value of the search radius decreases.

Consequently, when the value of ε is small, the clusters are separated from the core points and are treated as noise, as illustrated in the last plot of Fig. 2, where the $\varepsilon = 0.1$ achieves an ARI of only 0.0814. The search radius ε was lowered to the value of 0.3; further lowering this value, the ARI and Jaccard index decreased. With these trials, the PCA+UMAP and MCD are applied to reduce the parameters into new variables UMAP1 and UMAP2 and are projected in 2D space for visualization and clustering. This process was applied to thirty snapshots for each semi-urban channel scenario.

Table 1 shows the mean ARI for each semi-urban scenario corresponding to the search radius values of 0.9, 0.7, 0.4, and 0.2. The number of clusters was evaluated using the Jaccard index η_{Jac} Furthermore, the mean computational durations using the proposed approach were recorded. The simulations were carried out using MATLAB R2017a on a desktop with AMD Ryzen 5 3600 6-Core Processor with 16 GB of installed RAM, and the duration was computed using the timer functions of MATLAB.

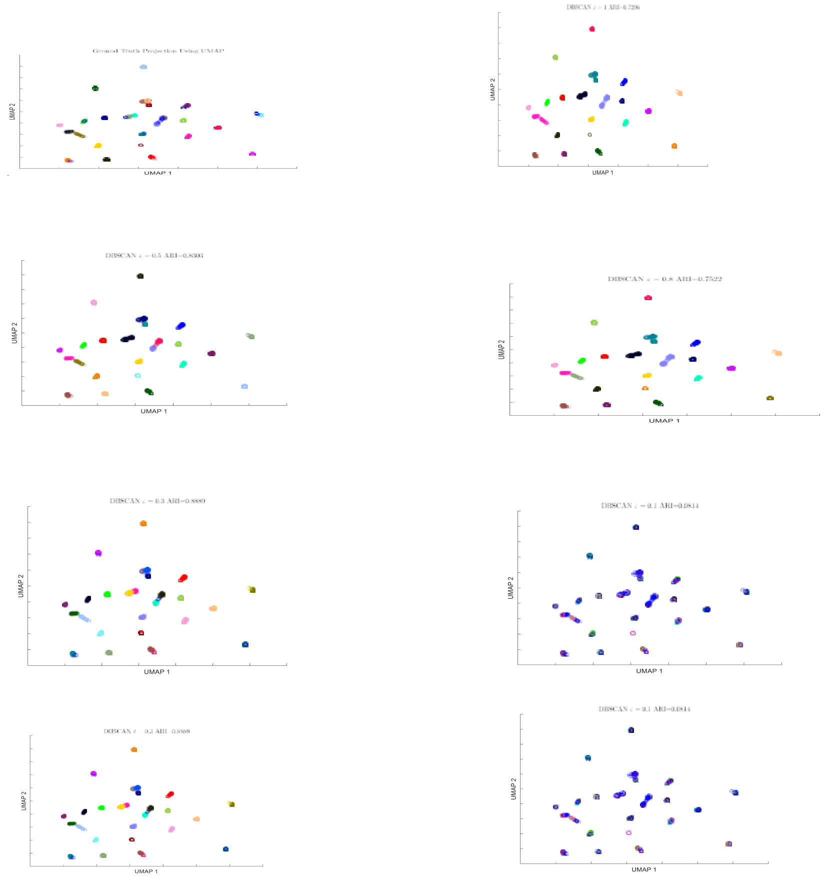


Fig. 2: UMAP of Ground-truth and DBSCAN results with varying ε.

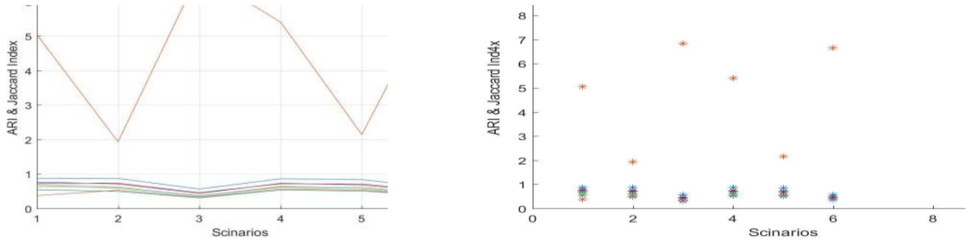
The applied approach had an excellent performance in the scenario B1 semi-urban LOS single link with the highest ARI and η_{Jac} of 0.7359 and 0.8726, respectively. The same scenario follows this result in the B2 with 0.7116 and 0.8390. On the other hand, the lowest scores are in the multiple links scenarios; this is due to the overlapping clusters in one dense point, and the UMAP cannot separate the clusters clearly, mainly due to the number of links that has the same parameters which only achieves only half of the number of clusters. This limitation also affects the memberships of the clusters resulting in lower ARI. Furthermore, in the multiple link scenarios, the highest mean computational duration of approximately 7 seconds can be attributed to the high number of paths. Finally, the overall scores show that at $\varepsilon = 3$ achieves the highest ARI and η_{Jac} as opposed to using $\varepsilon = 1$ where a significant difference in the scores is observed.

Table 1: Mean ARI, Jaccard index, and Computational Duration of UMAP and DBSCAN

Scenario	**ARI**				**ηJac**				**Computational Duration (seconds)**
	$\varepsilon = 0.9$	$\varepsilon = 0.7$	$\varepsilon = 0.4$	$\varepsilon = 0.2$	$\varepsilon = 0.9$	$\varepsilon = 0.7$	$\varepsilon = 0.4$	$\varepsilon = 0.2$	
B1 Semi-urban NLOS Single Link	0.5452	0.3731	0.6341	0.7298	0.6685	0.7046	0.7655	0.8736	5.0469
B1 Semi-urban LOS Single Link	0.4979	0.5306	0.6138	0.7358	0.5874	0.6257	0.7126	0.8726	1.9335
B1 Semi-urban LOS	0.3127	0.3244	0.3779	0.4611	0.3562	0.3704	0.4377	0.5637	6.8409

Multiple Links									
B2 Semi-urban NLOS Single Link	0.5361	0.5658	0.6276	0.7165	0.6150	0.6435	0.7301	0.8617	5.4050
B2 Semi-urban LOS Single Link	0.5096	0.5339	0.6113	0.7115	0.5739	0.6054	0.6850	0.8390	2.1443
B2 Semi-urban LOS Multiple Links	0.3535	0.3732	0.4304	0.4982	0.3726	0.3893	0.4579	0.5724	6.6561

Figure 3a illustrates the Empirical Cumulative Distribution Function (ECDF) of ARI for all 180 snapshots of channel scenarios and their corresponding ARI. The figure shows a significant difference and improved scores at $\varepsilon = 0.3$. The 10th percentile, median, and 90th percentile of $\varepsilon = 0.3$ are graphed in Figure 3a. The percentiles indicate that at less than 10% of the data, the performance is 0.4486; at less than 50%, an ARI of 0.6763; and at less than 90% of the ARI is 0.8018. The graph summarizes the performance in terms of the ARI.



(a) Grid plot (b) Star plot

Fig. 3: ECDF of ARI and η_{Jac} for all channel scenarios

Furthermore, Fig. 3b provides the ECDF of the cluster-wise Jaccard index for all scenarios. The ECDF shows a steeper curve in the ECDF of $\varepsilon = 0.3$ where at the median, less than 50% of the Jaccard indices are less than 0.8214, indicating a good approximation of the actual number of clusters. Consequently, the lower percentile indicates that at less than 10%, the Jaccard index of the number of clusters is 0.5193, which can be associated with the approach's performance in the multiple-links scenarios.

IV. CONCLUSION

This paper presents the performance results of applying UMAP to MPC for visualization prior to clustering. The UMAP reveals the clustering tendencies in the data utilizing PCA and the distance metric MCD. Visualizing the data can aid the clustering process, especially for the increasing number of paths. In addition, the clustering performance of DBSCAN to the membership and number of clusters, along with the varying values of the search radius ε are evaluated. The selection of the ε at 0.3 shows promising results in clustering the projected 2D transformed data of the MPCs. A median of 0.6763 and 0.8214 of the ECDF for the ARI and Jaccard index, respectively, suggests the accuracy of the approach. Incorporating DR techniques to visualize and cluster the MPCs proves to be optimal and can be used as an alternative to cluster the MPC of the C2CM semi-urban scenarios.

REFERENCES

[1] Bonek E. (2013) MIMO propagation channel modeling. 7th European Conference on Antennas and Propagation (EuCAP), pp. 2488-2492.

[2] Czink N, Cera P, Salo J, Bonek E, Nuutinen J, Ylitalo J. (2006) A framework for automatic clustering of parametric mimo channel data including path powers. In Proceedings IEEE Vehicular Technology Conference, pp. 1-5. https://doi.org/10.1109/VTCF.2006.35.

[3] Czink N, Cera P, Salo J, Bonek E, Nuutinen J, Ylitalo J. (2006) Improving clustering performance using multipath component distance. Electronics Letters. 42(1): 33-45. https://doi.org/10.1049/el:20063917.

[4] Moayyed MT, Antonescu B, Basagni S. (2019) Clustering algorithms and validation indices for mmwave radio multipath propagation. In Proceedings Wireless Telecommunications Symposium (WTS), pp. 1-7. https://doi.org/10.1109/WTS.2019.8715540.

[5] Hu M, Ye Y, He R, Ai B, Huang C, Zhong Z. (2020) A novel power weighted multipath component clustering algorithm based on spectral clustering. In Proceedings IEEE 91st Vehicular Technology Conference(VTC2020-Spring), pp.1-5. https://doi.org/10.1109/VTC2020-Spring48590.2020.9129206.

[6] Materum M, Takada J, Ida I, Oishi Y. (2009) Mobile station spatio-temporal multipath clustering of an estimated wideband MIMO double-directional channel of a small urban 4.5 GHz macrocell. EURASIP Journal on Wireless Communications and Networking, 2009: 1-16. https://doi.org/10.1155/2009/804021

[7] Hanpinitsak P, Saito K, Takada J, Kim M, Materum L. (2017) Multipath clustering and cluster tracking for geometry-based stochastic channel modeling. IEEE Transactions on Antennas and Propagation, 65(11): 6015-6028. https://doi.org/10.1109/TAP.2017.2754417.

[8] Li Y, Zhang J, Ma Z, Zhang Y. (2020) Clustering analysis in the wireless propagation channel with a variational gaussian mixture model. IEEE Transactions on Big Data. 6(2): 223-232. https://doi.org/10.1109/TBDATA.2018.2840696.

[9] Teologo A, Materum L, Blanza J, Hirano T. (2020) Comparative study of k-power means, ant colony optimization, kernel power density-based estimation, and gaussian mixture model for wireless propagation multipath clustering. International Journal of Emerging Trends in Engineering Research. 8(7). https://doi.org/10.30534/ijeter/2020/164872020