Clustering Algorithms and Validation Indices for mmWave Radio Multipath Propagation

Miead Tehrani Moayyed
Institute for the Wireless IoT
Northeastern University
Boston, USA
tehranimoayyed.m@husky.neu.edu

Bogdan Antonescu
Institute for the Wireless IoT
Northeastern University
Boston, USA
antonescu.b@husky.neu.edu

Stefano Basagni
Institute for the Wireless IoT
Northeastern University
Boston, USA
basagni@ece.neu.edu

***Abstract*—Transmissions in the mmWave spectrum benefit from a-priori knowledge of radio channel propagation models. This paper is concerned with one important task that helps provide a more accurate channel model, namely, the clustering of all multipath components arriving at the receiver. Our work focuses on directive transmissions in urban outdoor scenarios and shows the importance of the correct estimation of the number of clusters for mmWave radio channels simulated with a software ray-tracer tool. We investigate the effectiveness of k-means and k-power-means clustering algorithms in predicting the number of clusters through the use of cluster validity indices (CVIs) and score fusion techniques. Our investigation shows that clustering is a difficult task because the optimal number of clusters is not always given by one or by a combination of more CVIs. However, using score fusion methods, we find the optimal partitioning for the k-means algorithm based on the power and time of arrival of the multipath rays or based on their angle of arrival. When the k-power-means algorithm is used, the power of each arriving ray is the most important clustering factor, making the dominant received paths pull the other ones around them, to form a cluster. Thus, the number of clusters is smaller and the decision based on CVIs or score fusion factors easier to be taken.**

***Index Terms*—mmWave, clustering algorithms, cluster validity indices, channel propagation models**

I. Introduction

5G wireless standards are a promising solution for many problems of current wireless networks, especially concerning high-speed data transfers and ubiquitous connectivity requiring very low latency responses. In this context, one option to achieve these goals is *spectrum extension* through the use of *millimeter-wave* (mmWave) band (30–300 GHz) with multiple GHz of unused bandwidth. Unfortunately, mmWave transmissions suffer from high propagation loss, sensitivity to blockage, atmospheric attenuation and diffraction loss, so implementing transmissions in these extremely high frequency bands brings in new challenges. Tackling them requires well-thought *radio channel propagation models* that are obtained through extensive measurements (via steerable antennas and channel sounders), or via software ray-tracing simulators.

In this paper we are concerned with one important task that leads to the generation of better radio channel models. We emphasize the role of *clustering algorithms* in grouping the incoming rays at the receiver site. They are paramount for the fast processing of the received rays, and thus for extracting channel parameters in an efficient manner when the volume of data generated through simulations is huge. We use two variants of the well known k-means clustering algorithm in which we replace the usual Euclidean distance metric with the multipath component distance (MCD). Thus, we create a multi-dimensional space that is defined by the channel parameters of the multipath components (MPCs). This space—based on the Time-of-Arrival (ToA), azimuth and elevation of the Angle-of-Arrival (AoA) and Angle-of-Departure (AoD)—is fed into the clustering algorithms to provide the partitioning of all MPCs. We also quantify the goodness of these algorithms through the use of five *cluster validity indices* (CVIs) and three *score fusion* techniques. Our results show that by only using CVIs we sometimes fail to find the optimal clustering number K because the indices might capture only specific aspects of a clustering solution. Thus, we combine all five CVIs in an *ensemble* to provide a predictor of clustering quality that is better than any of the CVIs taken separately. The solution is represented by few score fusion techniques. We check this solution by visualizing the resulted clusters using polar plots of the AoA/AoD vs. ToA and by calculating the variance of the parameters that characterize the MPCs (power, ToA, AoA and AoD).

Our investigation uses a professional software ray-tracer tool (Wireless InSite by Remcom), to produce the channel simulations for each receiver location considered in the mmWave urban scenario. The estimated channel parameters at those locations are then processed in MATLAB by applying clustering algorithms and analyzing the validity of their results.

The rest of the paper is organized as follows. Section II reviews clustering concepts and the algorithms applied to the partitioning of the MPCs generated in our mmWave outdoor scenario. Section III enumerates the cluster validity indices and the score fusion methods used in our research. Section IV describes the outdoor simulation environment and presents the results of the two variants of the k-means algorithm and the validation of their results. Section V draws the conclusions regarding the optimal number of clusters proposed by each clustering algorithm and the effectiveness of the CVI/score fusion techniques in confirming these numbers.

978-1-5386-8380-4/19/$31.00 ©2019 IEEE

II. Clustering for mmWave MPCs

A *cluster* is defined as a group of rays with similar attenuation and angular profile. Channel parameters like Time-of-Arrival (ToA), Angle-of-Arrival (AoA) and Angle-of-Departure (AoD) are reported in our simulations by the ray-tracer tool for each arriving ray at the receiver. As a first order of magnitude, *visual inspection* [1] can identify clusters in the channel impulse response (CIR) of the channel. Unfortunately, this method is possible for simulations in which the number of received rays is small. If this number increases or the number of simulations becomes orders of magnitude larger, more automated procedures and algorithms need to replace the visual inspection of the CIR.

In our paper, we consider some of the well known *center-based* clustering algorithms in which the input gets partitioned around few centroids or central points [2]. The most common algorithm (k-means) [3] and one of its variants (k-power-means) are applied in many studies [4], [5], [6], [7]. k-means groups the rays with similar features (e.g., ToA, AoA, AoD) into a number of K clusters based on an a-priori decision about their number. Each MPC is assigned to a specific cluster by calculating the distance to these centroids and choosing the minimum one (i.e., finding the closest centroid): $D = \sum_{l=1}^{L} d(x_l, c_{x_l})$, where L is the total number of MPCs, x_l is the parameter of the l-th MPC, c_{x_l} is the parameter of the cluster centroid closest to the l-th MPC, and $d(\cdot)$ denotes the *distance function* between any two points in the parameter space. In subsequent iterations, the algorithm tries to find the optimum location of the centroids in order to minimize the distance from each MPC to its centroid. While each of the distances for ToA, AoA, AoD can be calculated separately, and delay and angular domains can be searched sequentially, there is an improvement if we use them jointly. In this case, the Euclidean distance is replaced by the *multipath component distance* (MCD) [8]. The result is a hypersphere with a radius in the normalized multipath parameter distance space:

$$MCD_{ij} = \sqrt{||MCD_{AoA,ij}||^2 + ||MCD_{AoD,ij}||^2 + ||MCD_{\tau,ij}||^2}, \quad (1)$$

where i and j are any two estimated MPCs.

As a variation, in the k-power-means algorithm [6], the same distance metric MCD is applied, but it is weighted by the power values P_l of the MPCs:

$$D = \sum_{l=1}^{L} P_l \cdot MCD(x_l, c_{I_l^{(i)}}), \quad (2)$$

where index $I_l^{(i)}$ is the cluster number for the l-th MPC in the i-th iteration. The idea of including power values into the distance function "forces" the centroids towards the points with strong powers. This lines up with the receiver's usual desire of finding and latching on the strongest rays in the transmitted spectrum.

III. Cluster Validity Indices

Clustering is an *unsupervised* pattern classification method that partitions the elements in a data set into clusters. Grouping elements within a cluster requires the identification of similar values for the parameters that characterize these elements. The MPCs arriving at the receiver have various values for their radio channel parameters (e.g., power levels, ToA, AoA, AoD).

Once the clustering algorithm finishes processing the input data set, an indicator is required to prove how accurate the number of clusters is. *Cluster validation* is a difficult task, so the techniques used are not easy to be classified. However, we can group them based on the information available during the validation process. *External validation* methods validate the clustering result by comparing it with the *correct* partitioning; it makes sense when the exact value is known (i.e., in a controlled test environment). *Internal validation* methods validate the partitioning results by examining only the clustered data, measuring the *compactness* and *separation* of the clusters. This category is applied in our paper, using the following CVIs: Calinski-Harabasz [9], Davies-Bouldin [10], generalized Dunn [11], [12], Xie-Benie [13] and PBM [14]. Yet a third category labeled *relative validation* compares partitions generated by the same clustering algorithm with different parameters or with different subsets of data.

All CVIs described in this section use the MCD metric defined by (1) and the following notations. L is the total number of MPCs arriving at the receiver while L_k is the number of MPCs in cluster k. c_k is the position of the centroid of cluster k, c is the position of the global centroid, and s_l is the data of subpath l in cluster k.

Calinski-Harabasz (CH) is one the most used CVIs in research, from pattern recognition papers [15], [16] to clustering radio channel parameters [6], [17]. The index estimates the compactness of a cluster based on the distances from the points in the cluster to its centroid. The separation of the clusters is measured as the distance from the centroids to the global centroid:

$$\nu_{CH} = \frac{\frac{\sum_{k=1}^{K} L_k (MCD(c_k, c))^2}{K-1}}{\frac{\sum_{k=1}^{K}\sum_{l=1}^{L_k} L_k (MCD(s_l, c))^2}{L-K}}\ , \quad (3)$$

where the location of the centroid of cluster k is calculated as $c_k = \frac{1}{L_k}\sum_{l=1}^{L_k} x_l$ while the one of the global centroid is computed as $c = \frac{1}{L}\sum_{l=1}^{L} x_l$. If k-power-means is used, then the position of the global centroid becomes $c = \frac{\sum_{l=1}^{L} P_l x_l}{\sum_{l=1}^{L} P_l}$ while the position of the k-th centroid (c_k) is given by a similar formula in which L is replaced by L_k. The optimal K number is represented by the highest value of the ν_{CH} index.

Davies-Bouldin (DB) is another index widely used in CVI comparative studies. The compactness is computed as the average distance of all patterns for the points in the cluster to its centroid $S_k = \frac{1}{L_k}\sum_{l=1}^{L_k} MCD(s_l, c_k)$ while the separation is based on the distance between centroids

$d_{k_1,k_2} = MCD(c_{k_1}, c_{k_2})$. Then, the DB index is calculated as:

$$\nu_{DB} = \frac{1}{K}\sum_{k=1}^{K} R_k, \quad R_k = \max_{k_1,k_2} \frac{S_{k_1} + S_{k_2}}{d_{k_1,k_2}}. \tag{4}$$

Trying different input K values, the optimal number of clusters is achieved for the smallest value of the index: $\nu_{DB_{opt}} = \arg\min_K\{\nu_{DB}(K)\}$.

Generalized Dunn (GD) index [11] was meant to improve the sensitivity of Dunn's index to noisy points (i.e., outliers and inliers to the cluster structure). The initial *Dunn index* was the ratio of two distances, the minimum distance between two points belonging to different clusters to the maximum distance between any two points selected from the same cluster; hence, it was quantifying both the separation of clusters and their spread. 18 forms are known for the generalized index based on 6 formulas for the calculation of the distance δ between clusters and 3 formulas for the diameter Δ of the cluster. Our paper uses two of the most researched forms that define the D_{53} index. The *distance* δ between two clusters depends on *all points* in each cluster, so averaging reduces the effect of adding/deleting points to/from any two clusters:

$$\delta_5 = \frac{1}{L_{k1} + L_{k2}}\Big(\sum_{l=1}^{L_{k1}} MCD(s_l, c_{k1}) + \sum_{m=1}^{L_{k2}} MCD(s_l, c_{k2})\Big). \tag{5}$$

The *diameter* of each cluster is also based on all points in the cluster: $\Delta_3 = \frac{2}{L_k}(\sum_{l=1}^{L_k} MCD(s_l, c_k))$. The Generalized Dunn index is given by the ratio:

$$\nu_{D_{53}} = \frac{\min_{k1,k2} \delta_5(k1, k2)}{\max_k \Delta_3(k)}. \tag{6}$$

The worst case scenario is captured as the smallest cluster separation and the largest cluster. The optimal value for K is given by the maximum value of the $\nu_{D_{53}}$ index.

Xie-Beni (XB) index was initially proposed for cluster validation on fuzzy partitions, but may be used on hard partitions as well [18], [14] (i.e., for crisp clustering where the CVIs are best for their lowest or highest values).

$$\nu_{XB} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L_k}(MCD(s_l, c_k))^2}{L \times [\min_{k1,k2}(MCD(c_{k1}, c_{k2}))^2]} . \tag{7}$$

More compact clusters (the numerator) and larger separations between clusters (the denominator) result in smaller values for this index. Thus, the optimal clustering solution is the one for which the XB index has the minimum value.

The last index accounted in our analysis is the ***PBM*** index.

$$\nu_{PBM} = \Big(\frac{1}{K} \times \frac{\max_{k1,k2}(MCD(c_{k1}, c_{k2}))}{\sum_{k=1}^{K}\sum_{l=1}^{L_k} MCD(s_l, c_k)}\Big)^2. \tag{8}$$

According to the authors of [14], it performed better than Davies-Bouldin, Dunn and Xie-Beni indices for their specific data. However, this is not a rule.

A. Using multiple CVIs to compare clustering solutions

Finding the correct number of clusters in an analyzed data set has no theoretically optimal method. Existing algorithms include other methods than CVIs (e.g., stability-based methods, model-fitting-based algorithms). CVIs are meant to quantify various properties of the clustering solution such as *compactness* and *separation* between clusters. The optimal clustering solution K is pointed out by the min or max value of the CVI. Nevertheless, their formulas might capture only specific aspects of the clustering solution, so an elongated shape cluster might not be considered compact. Therefore, no CVI should be assumed a-priori better than its alternatives. Considering that no single CVI can capture correctly the validity of any clustering solution (i.e., work well with all data sets), [19] proposes that the value of each CVI be captured in an *ensemble* that could represent a better predictor of the clustering quality than any of the CVIs taken separately. Therefore, in our paper, the solution is represented by few *score fusion-based* techniques. A combined score SF_x is computed using M normalized CVIs. Three such examples shown below are based on the arithmetic, geometric and harmonic mean (9).

$$SF_a = \frac{1}{M}\sum_{i=1}^{M}\nu_i; SF_g = \Big(\prod_{i=1}^{M}\nu_i\Big)^{\frac{1}{M}}; SF_h = M\Big(\sum_{i=1}^{M}\frac{1}{\nu_i}\Big)^{-1} \tag{9}$$

From the many normalization methods (e.g., z-norm, global z-norm), [19] claims *min-max* to be the best. First, all indices are normalized, to produce values in the range $[0, 1]$. Then, to capture in all SF_x formulas only CVI values that point the optimal K with their *max* value, we subtract their normalized values from 1 for the indices that actually show this optimal value with their *min* value (e.g., Xie-Beni, Davies-Bouldin).

IV. Simulation Results

This section describes our ray-tracer simulations, the results of both clustering algorithms and the decision on the optimal number K of clusters based on the CVIs and score fusion techniques.

We simulated 28 GHz transmissions using one urban scenarios (Rosslyn, VA) delivered with the ray-tracing tool. The advantage of using this professional electromagnetic simulation tool is the input of site-specific data for any scenario, and the evaluation of the signal propagation characteristics by taking into consideration the effects of buildings, terrain and even weather. The tool generates rays with a very high angle resolution (0.2°), allowing us to collect very accurate channel parameters at a fraction of the time required to measure them with dedicated hardware (e.g., channel sounders and horn antennas). The estimated values are then fed to the clustering algorithms.

Our scenario has the Tx (base station) located on a light/traffic pole (with a height of 8 m) in the North part of Fig. 1 (the green dot) while the Rx point is installed in a vehicle at approximately 1.5 m above ground (any of the

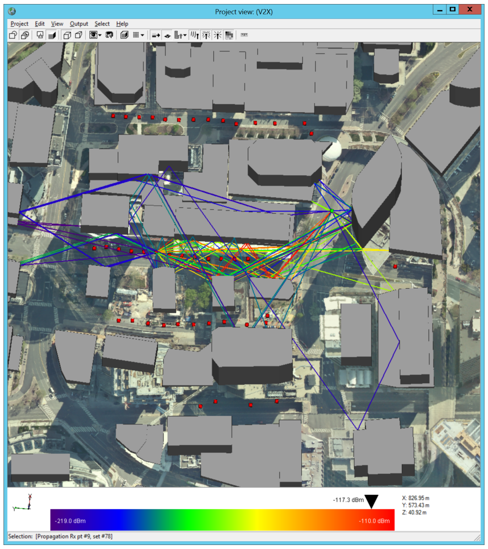

Fig. 1. 44 MPCs at receiver Rx#9.

red dots). The LOS transmission is simulated in the North-South direction by placing the vehicle along the wide-open boulevard at different locations up to 150 m in front of the transmitter. The NLOS reception mode is simulated in the East-West orientation in Fig. 1 by moving the vehicle at distances 70 to 150 m from Tx, on a side street behind very tall buildings. Since NLOS is a much more challenging scenario, we focus our simulations primarily on this case. We set the ray-tracer to use two horn antenna models with different half-power beamwidth (HPBW) and gain (7°/25 dBi and 22°/15 dBi). In all simulations described in this paper, the same antennas (7° or 22°) are used at both Tx and Rx locations in one experiment. The maximum power of the transmitted signal is 24 dBm. The ray-tracer follows a certain number of reflections (6) and diffractions (1) for each path from transmitter to receiver. Two methods are always considered in all our studies. In the *no beam alignment* (Fig. 1), the Tx and Rx antennas are simply oriented with the street direction, whereas the *beam alignment* procedure implies that the boresight of the Rx antenna is oriented with the direction of the strongest reception path, at that specific location. To take less time for running the simulations, in this paper, we applied only the no beam alignment procedure. At each Tx-Rx separation distance, we use MATLAB to generate a random Rx point that is given to the ray-tracer for simulation. We capture the values of the received power, excess delay, angle-of-arrival and angle-of-departure of all MPCs arriving at each randomly placed Rx point. Thus, each of these channel parameters is an array with L values due to the L MPCs. The clustering algorithm can be applied to each parameter, or a multi-dimensional space (e.g., the MCD metric [8]) can be used to find a correlation among these parameters.

A. Clustering Algorithm Results

This section summarizes the clustering results obtained when the two variants of the k-means algorithm are applied to the MPCs collected for our simulations. The urban scenario in Fig. 1 shows 44 MPCs at a specific receiver point (Rx#9) placed on one of the side streets. Each path/MPC has its own received power level, AoA, AoD, and comes with a certain excess delay (ToA). The real part of the complex impulse response (CIR) for this one-time channel realization (Fig. 2) shows the relationship between received power levels of various MPCs and their ToA. Using different colors, we show the average power value of each cluster and its average ToA, as marked by stars. Both values are calculated using the channel parameters of the MPCs in each cluster; the partitioning is performed with the k-means with MCD algorithm. Considering the large number of MPCs, it is impossible to apply a clustering procedure based on *visual inspection*. The same clustering algorithm gives us the 3D result in Fig. 3, in which MPCs are grouped in different clusters based on their temporal and spatial characteristics (i.e., delay spread and azimuth & elevation values of their AoA/AoD). The results show that capturing all five parameters of the MPCs (azimuth & elevation for AoA and AoD, and excess delay) allows us to correlate the *temporal* and *spatial* characteristics of the radio channel and to provide a better clustering solution.

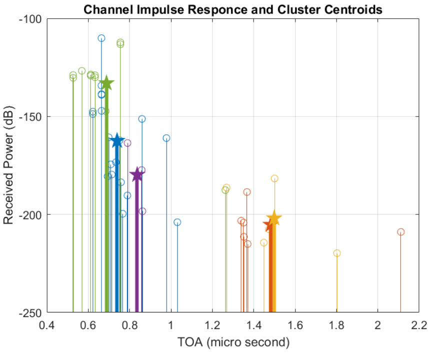

Fig. 2. Clustered CIR at Rx#9 based on k-means with MCD.

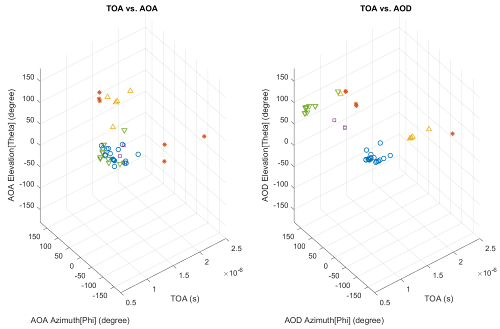

Fig. 3. Clustering via k-means algorithm—ToA vs. AoA, AoD.

Using the other variant of the clustering algorithm (k-power-means with MCD), we obtain different CIR (Fig. 4) and ToA vs. AoA/AoD clustering pictures (Fig. 5). We can notice that the average values of the Rx power and ToA in each cluster (marked with a star in Fig. 4) are very close to each other,

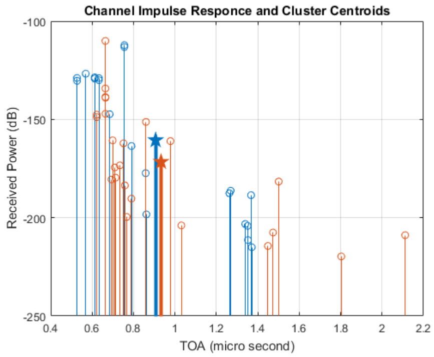

Fig. 4. Clustered CIR at Rx#9 based on k-power-means with MCD.

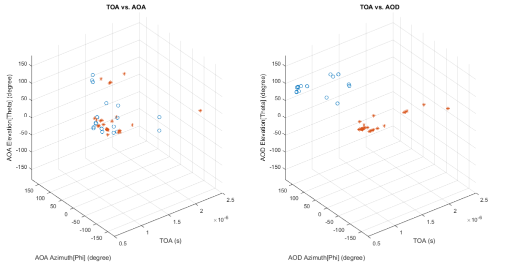

Fig. 5. Clustering via k-power-means algorithm—ToA vs. AoA, AoD.

even though the MPCs in each cluster can be dispersed in time. Comparing Fig. 2 and Fig. 4, we can see that the 5-dimensional space that we had initially is now totally biased (in the latter picture) by the received power. In this case, partitioning around the most representative MPCs (power-wise) is the important factor that reduces the number of clusters to a minimum.

B. CVIs and Score Fusion Results

Once the clustering phase is finished for various input K values, the CVIs described in Section III are applied, to find the optimal K value. As mentioned in Section III-A, one or more CVIs might not be able to solve this task, but combining CVIs in a fusion classifier could potentially provide a better way to find the optimal value of the number of clusters K. This section provides the results of the clustering validation process and of the score fusion methods described by equations (9).

For this analysis, we use receiver Rx#9 placed on a side street at approximately 150 m (Euclidean distance) from the transmitter (Fig. 1). With only 44 MPCs reaching this receiver and considering only 3 rays per cluster, we could have a maximum of 15 clusters. This assumption sets the initial K input of the clustering algorithm in the range $[2, 15]$.

As mentioned, not all CVIs can find the optimal K value. For example, when we apply k-means with MCD algorithm, indices CH, DB and GD cannot find this number correctly. Nevertheless, the other two indices XB and PBM find a number of clusters of 6 and 5, respectively (Table I). The conclusion is that few CVIs report a number of clusters hard to believe, and a couple of CVIs report different values for the K number. Using the ensemble predictor, we plug the normalized and biased CVI values obtained for Rx#9 (for each input value K) into the score fusion formulas (9), as explained in Section III-A (Table I). While the optimum value K cannot be predicted using only CVIs because not all CVIs have their maximum value on the same row, by using score fusion methods, we find that two scores (SF_g and SF_h) agree with each other. If we calculate the average of the three scores (last column in Table I), the maximum value points to an optimal value of $K = 8$ clusters, which agrees with both geometric (SF_g) and harmonic (SF_h) mean-based scores.

TABLE I
NORMALIZED AND BIASED CVIS AND SF VALUES FOR RX#9—k-MEANS ALGORITHM.

K	CH	XB	PBM	DB	GD	SF_a	SF_g	SF_h	M_{SF}
2	0.000	0.449	0.755	**1.000**	**1.000**	0.641	0.000	0.000	0.214
3	0.166	0.399	0.539	0.911	0.577	0.518	0.451	0.378	0.449
4	0.100	0.702	0.431	0.772	0.679	0.537	0.437	0.303	0.426
5	0.133	0.793	**1.000**	0.724	0.613	**0.653**	0.542	0.391	0.529
6	0.213	**1.000**	0.513	0.676	0.534	0.587	0.524	0.454	0.522
7	0.358	0.408	0.706	0.517	0.511	0.500	0.486	0.474	0.487
8	0.540	0.722	0.714	0.518	0.394	0.578	**0.564**	**0.549**	**0.563**
9	0.631	0.671	0.362	0.358	0.394	0.483	0.464	0.448	0.465
10	0.691	0.807	0.078	0.368	0.285	0.446	0.340	0.230	0.339
11	0.819	0.940	0.253	0.297	0.240	0.510	0.425	0.363	0.433
12	0.810	0.000	0.095	0.051	0.240	0.239	0.000	0.000	0.080
13	0.844	0.644	0.000	0.148	0.000	0.327	0.000	0.000	0.109
14	0.872	0.353	0.264	0.000	0.000	0.298	0.000	0.000	0.099
15	**1.000**	0.811	0.572	0.147	0.000	0.506	0.000	0.000	0.169

We repeat this study for all 14 receivers installed on the side street where Rx#9 is located. For lack of space, we show in Table II the optimal K clustering values only for three receivers, including Rx#9. We notice that for other receiver locations on the same street the score fusion factors and their average value all agree on the same optimal K value ($K = 3$ for Rx#5 and $K = 4$ for Rx#13).

TABLE II
OPTIMAL K VALUE FOR EACH CVI AND SF METHOD FOR FEW RX—k-MEANS ALGORITHM.

Rx	CH	XB	PBM	DB	GD	SF_a	SF_g	SF_h	M_{SF}
5	21	17	2	2	2	3	3	3	3
9	15	6	5	2	2	5	8	8	8
13	17	8	4	2	2	4	4	4	4

The second part of our analysis is the k-power-means variant of the clustering algorithm. The distance metric used in the clustering algorithm (2) and in the local and global centroids computed for the CVI formulas in Section III is weighted by the power of each MPC. The clustering results are validated by the same five CVIs and three score fusion factors for the same set of MPCs received at Rx#9. In this case, the optimal K number is 2. As with the first algorithm, we repeat the study for all 14 receivers located on the same street

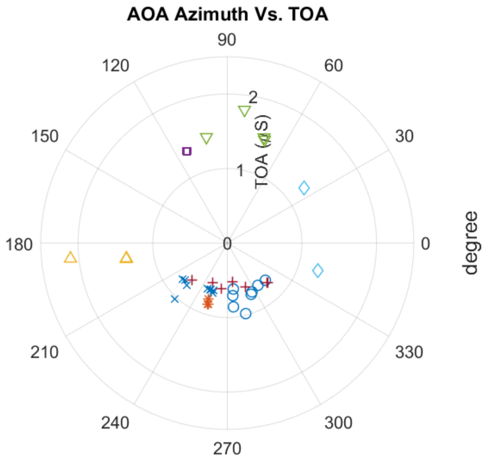

Fig. 6. Polar plot of azimuth AoA vs. ToA for k-means with K=8.

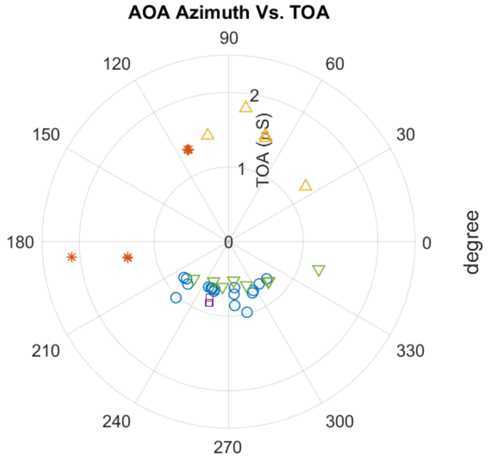

Fig. 7. Polar plot of azimuth AoA vs. ToA for k-means with K=5.

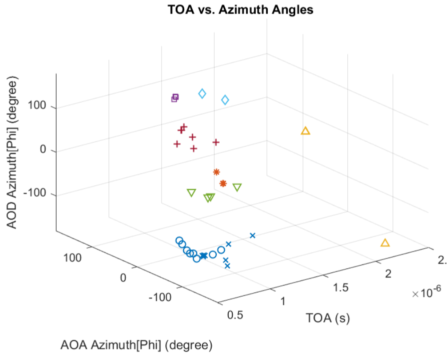

Fig. 8. 3D representation based on ToA and azimuth of AoA, AoD for K=8.

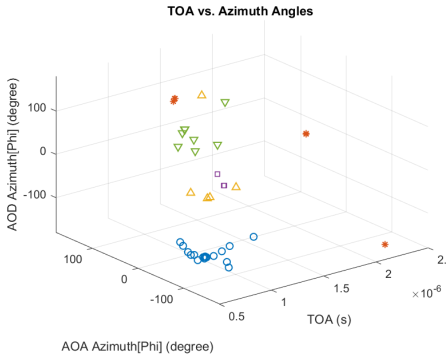

Fig. 9. 3D representation based on ToA and AoA, AoD for K=5.

with Rx#9. For the other two receivers (Rx#5 and Rx#13) mentioned in Table II, the optimal K values are 2 and 3.

Going back to the values in Table I, we question the option of providing a decision on the correct partitioning if only two score fusion models are used. The arithmetic mean-based score fusion points to a solution with 5 clusters while the geometric mean-based one indicates 8 clusters. However, when we take the average value for SF_a and SF_g, the optimal clustering solution is $K = 5$. To investigate both solutions even more, we use the *polar plots* of the AoA and AoD vs. ToA for all MPCs when they are grouped in either 5 or 8 clusters. The advantage of this method is that it considers the cyclic feature of the angles and becomes easier to find how close the MPCs are in comparison with the 3D visualization. The *elevation* component of the two angles shows little spatial variation. Thus, we focus only on the *azimuth* component, and we show the polar plots of the AoA for both solutions (Fig. 6 and Fig. 7). Based on the azimuth component information for both AoA and AoD, we build a 3D plot in which the third dimension is the ToA of each MPC (Fig. 8 and Fig. 9), in order to understand the advantage of each potential clustering. As Fig. 9 shows, this solution is able to gather more MPCs in at least one cluster and to merge two other clusters with a low number of multipaths. Thus, a partitioning with only 5 clusters might be more realistic (Fig. 3).

Polar or 3D plots are helpful when the number of estimated clusters is small and we can infer something from their visualization. However, when the number of MPCs is large, and the decision based on CVIs or SF scores is in between two clustering solutions, we want a more analytical method, to find the K number. A *statistics-based* decision using the *variance* of the values for various parameters of the MPCs in each cluster could be a solution. To choose one partitioning against the other, we find which one produces more *compact* clusters, i.e., with the smallest variance for received power, ToA, and AoA (Table III). The results show that if we are

TABLE III
TOTAL VARIANCE OF MPC PARAMETERS

MPC Parameter	k-means K=5	k-means K=8
Rx_Power	1687.63	2095.64
ToA	0.19	0.25
$Elevation\_AoA$	811.78	1090.89
$Azimuth\_AoA$	4166.87	2808.96
$Elevation\_AoD$	51.12	52.92
$Azimuth\_AoD$	4361.98	1997.11

interested in clusters that group more MPCs, a solution with 5 clusters would be better. The total variance values of the Rx power and ToA are smaller, so this solution produces clusters with rays coming closer in time to each other and with power values closer to the average value in each cluster. On the other hand, mmWave transmissions consider *directivity* as one of their dominant traits, so it is equally important to analyze the clusters predominantly from the AoA of their constituent MPCs. In that case, the solution with 8 clusters gives a better result since it groups MPCs based on their spatial parameters rather than temporal and power ones.

A similar analysis consisting of polar plots for AoA/AoD vs. ToA and variance values for the channel parameters of the received MPCs can be applied to the k-power-means algorithm. Fortunately, in this case, the validation of the clustering results is easier. All score fusion factors point to the same optimal K value for Rx#9. Moreover, this consensus applies for all receivers placed on that street. Since power was the major weight factor, in addition to the number of clusters, we are also interested in the received power levels of the dominant path that defines each cluster. These values are getting smaller (i.e., from -104 dB for Rx#5 to -137 dB for Rx#13) as we move on the street from East to West towards the edge of the cell.

V. Conclusions

Our paper compared two variants of the well known k-means clustering algorithm from the point of validating and predicting the optimal partitioning of the MPCs generated in our simulations. Our results show that clustering is not a trivial task because finding the optimal number K of clusters is not always given by one or more cluster validity indices. For the k-means algorithm, we noticed that few of the CVIs used in our study were not able to find the correct partitioning. Nevertheless, score fusion techniques and further statistics-based decisions allowed us to choose the optimal value for K. When the k-power-means algorithm was applied to the same set of MPCs (at the same receiver), the Rx power factor weighted more and "pulled" many MPCs around the dominant path in each cluster. It also generated a much smaller number of clusters at many locations on the street. The end result was an easier decision about the correctness of the clustering solution based on both CVIs and score fusion factors. In the future we will analyze the effect of diffuse scattering to the partitioning solution, and then we will quantify the influence of clustering to the generation of mmWave channel models.

Acknowledgments

This work is supported in part by MathWorks under Research Grant "Cross-layer Approach to 5G: Models and Protocols."

References

[1] D. Shutin, "Cluster analysis of wireless channel impulse responses," in *Proceedings of the International Zurich Seminar on Communications*, Zurich, Switzerland, February 18–20 2004, pp. 124–127.

[2] G. Hamerly and C. Elkan, "Alternatives to the k-means algorithm that find better clusterings," in *Proceedings of the eleventh International Conference on Information and Knowledge Management (CIKM)*, McLean, VA, November 4–9 2002, pp. 600–607.

[3] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York, NY, USA: Springer, 2nd edition, 2016.

[4] M. T. Martinez-Ingles, D. P. Gaillot, J. Pascual-Garcia, J. M. Molina Garcia-Pardo, M. Lienard, J. V. Rodríguez, and L. Juan-Llacer, "Impact of clustering at mmW band frequencies," in *Proceedings of the IEEE International Symposium on Antennas and Propagation & USNC/URSI National Radio Science Meeting*, Vancouver, BC, Canada, July 19–24 2015, pp. 1009–1010.

[5] C. Gustafson, K. Haneda, S. Wyne, and F. Tufvesson, "On mm-Wave Multipath Clustering and Channel Modeling," *IEEE Transactions on Antennas and Propagation*, vol. 62, no. 3, pp. 1445–1455, March 2014.

[6] N. Czink, P. Cera, J. Salo, E. Bonek, J. P. Nuutinen, and J. Ylitalo, "A Framework for Automatic Clustering of Parametric MIMO Channel Data Including Path Powers," in *Proceedings of the IEEE 64th Vehicular Technology Conference (VTC Fall)*, Montreal, Quebec, Canada, September 25–28 2006, pp. 1–5.

[7] N. Czink, R. Tian, S. Wyne, F. Tufvesson, J. P. Nuutinen, J. Ylitalo, E. Bonek, and A. F. Molisch, "Tracking Time-Variant Cluster Parameters in MIMO Channel Measurements," in *Proceedings of the Second International Conference on Communications and Networking, CHINACOM*, Shanghai, China, August 22–24 2007, pp. 1147–1151.

[8] N. Czink, P. Cera, J. Salo, E. Bonek, J. P. Nuutinen, and J. Ylitalo, "Improving clustering performance using multipath component distance," *Electronics Letters*, vol. 42, no. 1, pp. 33–45, January 2006.

[9] T. Caliński and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, January 1974.

[10] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, April 1979.

[11] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 28, no. 3, pp. 301–315, June 1998.

[12] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, September 1973.

[13] X. L. Xie and G. Beni, "A validity measure for fuzzy clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 8, pp. 841–847, August 1991.

[14] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern Recognition*, vol. 37, no. 3, pp. 487–501, March 2004.

[15] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, June 1985.

[16] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Perez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, January 2013.

[17] S. Mota, F. Perez-Fontan, and A. Rocha, "Estimation of the Number of Clusters in Multipath Radio Channel Data Sets," *IEEE Transactions on Antennas and Propagation*, vol. 61, no. 5, pp. 2879–2883, May 2013.

[18] U. Maulik and S. Bandyopadhyay, "Performance Evaluation of Some Clustering Algorithms and Validity Indices," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 12, pp. 1650–1654, January 2002.

[19] K. Kryszczuk and P. Hurley, "Estimation of the Number of Clusters Using Multiple Clustering Validity Indices," in *Proceedings of the 9th International Workshop on Multiple Classifier Systems, Cairo, Egypt*, Springer, Berlin, Heidelberg, April 7–9 2010, pp. 114–123.