Clustering Validation for mmWave Multipath Components in Outdoor Transmissions

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

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

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

***Abstract*—Radio channel propagation models for the mmWave spectrum are of paramount importance for the design and planning of future 5G wireless communications systems. Since transmitted radio signals are received as clusters of multipath rays, the problem arise about how to identify them, which is functional to extract better spatial and temporal characteristics of the mmWave channel. This paper deals with the validation of the results produced by the clustering process. Specifically, we estimate the effectiveness of the k-means clustering algorithm in predicting the number of clusters by using cluster validity indices (CVIs) and score fusion techniques. We consider directive transmissions in outdoor scenarios and we show the importance of the correct estimation of the number of clusters for the mmWave radio channel simulated with a software ray-tracer tool. Our investigation shows that clustering is no trivial task because the optimal number of clusters is not always given by one or by a combination of more CVIs. In fact, a few of the CVIs used in our study were not capable to determine correct partitioning. However, using score fusion methods and additional techniques we find two solutions for the number of clusters based on power and time of arrival of the multipath rays or based on their angle of arrival.**

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

I. Introduction

Radio channel propagation models are at the forefront of the research in wireless communications because the performance of any real network implementation relies heavily on them. These models are obtained through extensive measurements (performed with steerable antennas and channel sounders), or via software ray-tracing simulators. Whether we measure or we simulate a radio channel for indoor or outdoor scenarios, the results show that the received *multipath components* (MPCs) are distributed in a number of clusters based on many factors, including transmission frequency, number and size of obstacles, frequency absorption and penetration, and weather [1]. A *cluster* is a set of paths with similar attributes or parameters including power, Time-of-Arrival (ToA), Angle-of-Arrival (AoA) and Angle-of-Departure (AoD).

Clustering analysis—a topic widely used in connection with machine learning and pattern recognition research—is applied for partitioning the MPCs arriving at the receiver with the aim of providing a better understanding of the relationship among their channel parameters. The next step is cluster validation. This step is equally important because it measures the quality of clustering results, i.e., how well the proposed partitioning fits the data. In addition, it helps to compare the results of different clustering algorithms that provide different numbers of clusters. In fact, the relationship between clustering procedures and their validation is very subtle because knowing how to define a good clustering criterion requires a good understanding of the input data, but, at the same time, clustering is an important tool that helps understand the data.

In this paper, we focus on the well known k-means clustering algorithm where we replace the usual Euclidean distance with the *multipath component distance* (MCD). We create a *multi-dimensional space* based on the Time-of-Arrival and the azimuth and elevation of the Angle-of-Arrival and Angle-of-Departure. This space is fed into the k-means clustering algorithm to provide the partitioning of all MPCs that arrive at the receiver. We quantify the goodness of the clustering algorithm through the use of five *cluster validity indices* (CVIs) and three *score fusion* techniques. Our results show that using only CVIs fails to find the optimal clustering number K because they might be able to only capture 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. To check the resulting clusters, we visualize them using polar plots of the AoA/AoD vs. ToA. The variance values of the parameters of the MPCs (power, ToA, AoA and AoD) are calculated, and a dual solution (5 vs. 8 clusters) is presented. One produces clusters with MPCs that arrive closer in time and with similar power values, the other one groups multipath rays with similar spatial characteristics.

Our investigation uses a professional software ray-tracer tool (Wireless InSite by Remcom), to produce the channel simulations for each receiver location in a mmWave urban scenario. The estimated channel parameters are then processed in Matlab, which we use to implement the clustering algorithm and to analyze the validity of its results.

The rest of the paper is organized as follows. Section II reviews various clustering concepts and algorithms applied to the

978-1-7281-0117-0/19/$31.00 ©2019 IEEE

partitioning of the MPCs generated in wireless transmissions. Section III provides a quick view of the cluster validation process and itemizes the CVIs and the score fusion methods used in our work. Section IV describes the outdoor simulation environment and presents the results of the k-means algorithm and of the score fusion techniques. Section V draws the conclusions regarding the optimal number of clusters and the effectiveness of the CVI/score fusion technique in estimating this number.

II. Clustering for Channel Modeling

Transmitted rays get reflected, diffracted or scattered due to various obstacles in their path to the receiver. Thus, the *multipath components* (MPCs) that arrive at the receiver are characterized by different delays and attenuation. Capturing and processing all possible rays at the receiver and grouping them in clusters is very important because the modeled root-mean-square (RMS) delay spread (that evaluates the time dispersive properties of the channel) can be affected if some of the clusters are not taken into account in the estimation [2].

A *cluster* is defined not only as a group of rays with similar attenuation, but also as arriving paths that have similar angular profile. Time-of-Arrival (ToA), Angle-of-Arrival (AoA) and Angle-of-Departure (AoD) can be extracted from our simulations because the ray-tracer provides the excess time delay for each MPC, as well as information about their directions of departure and arrival. This gives some insight about how *dense* or how *sparse* the channel is with respect to the number of MPCs. A basic method to identify clusters in the channel impulse response (CIR) is by *visual inspection* [3]. This method assumes that we can assign a certain separation time between clusters, to allow the partitioning of the CIR in different bins [4]. Visual inspection is possible for few simulations in which the maximum number of arriving rays considered for processing at the receiver is small enough. If the number of rays 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.

Some of the well known clustering algorithms used in conjunction with machine learning techniques are the *center-based* ones in which the input gets partitioned around few centroids or central points [5]. The most common algorithm (k-means) [6] and one of its variations (k-power-means) are applied in many studies [7], [8], [9], [10]. 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 (a guess) 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):

$$D = \sum_{l=1}^{L} d(x_l, c_{x_l}), \tag{1}$$

where 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 [11]. In this case, the distance described by (1) is replaced by the *multipath component distance* (MCD). The result is a sphere with a radius given by:

$$MCD_{ij} = \sqrt{||MCD_{AoA,ij}||^2 + ||MCD_{AoD,ij}||^2 + ||MCD_{\tau,ij}||^2}, \tag{2}$$

where i and j are any two estimated MPCs. First, all rays are sorted with respect to their delays. Then, the MCDs between the ray with the shortest delay and all the other rays are calculated. In this step, all rays that produce an MCD value smaller than a chosen threshold are grouped together with the ray with the shortest delay. The same procedure is applied again to the remaining rays until all rays are assigned to a cluster. The delays and angular characteristics for each cluster are considered as the ones of the ray with the shortest delay in that specific cluster.

k-means has been used heavily in research literature because of its simplicity. However, it has three main drawbacks. It does not scale well from computational point of view, the number of clusters K has to be supplied by the user, and running it with a fixed value for K vs. a dynamic one can result in worse local optima. The convergence to a local minimum that is not the global minimum is also due to an initialization error. Many studies try to solve the sensitivity-to-initialization problem [5], and to improve the quality of clustering and the speed-up of the convergence. One method is simply to run the clustering algorithm several times from different starting points (i.e., random restart) and to take the best solution [12]. Another way is to search for the best initialization possible (e.g., random, Forgy, MacQueen, Kaufman) [13], [14], [15]. Yet another procedure is to find improvements to the k-means algorithm through model selection (e.g., dynamic increase in the number of clusters K), and to run it once rather than looping over K with a fixed model [16]. Other center-based algorithms like *Gaussian expectation-maximization*, *fuzzy k-means*, and *k-harmonic means* also exist [5], but they will be the subject of our future research. For now, we focus on the k-means clustering algorithm.

III. Cluster Validity Indices

Clustering in itself is an *unsupervised* pattern classification method that partitions the elements in a data set into clusters. Thus, the goal is to group similar elements within a cluster by identifying similar values for various parameters that characterize these elements. In our case, we have MPCs arriving at the receiver with various values for their radio channel parameters (e.g., power levels, ToA, AoA, AoD).

Once the clustering algorithm has processed the input data set, an obvious question arises: How well does this partitioning

fit the input data? In other words, how accurate the number of clusters provided by the algorithm is. The question is relevant for many reasons. First, an optimal clustering algorithm does not exist (i.e., different algorithms produce different partitions and none of them prove to be the best in all situations). Second, many clustering algorithms must be supplied initially with a guess about the number of possible clusters K, but this information is difficult to estimate a priori. So, the usual approach is to run the algorithm several times with a different K value for each run, and then to evaluate all resulted partitions, to see which one best fits the input data (e.g., k-means algorithm).

Cluster validation is a difficult task, and the techniques used cannot be easily classified. Nevertheless, there is a clear distinction if we relate to the information available during the validation process. Some techniques belong to the *external validation* methods because they validate the clustering by comparing it with the *correct* partitioning; this makes sense in a controlled test environment when the exact value is known. On the other hand, *internal validation* methods validate the partitioning results by examining only the clustered data, measuring the *compactness* and *separation* of the clusters. This is the category that we apply in our paper by proposing few CVIs for our analysis: Calinski-Harabasz [17], Davies-Bouldin [18], generalized Dunn [19], [20], Xie-Benie [21] and PBM [22]. There is yet a third category labeled *relative validation* based on comparisons of partitions generated by the same clustering algorithm with different parameters or with different subsets of data.

For all CVIs described below, we use the MCD metric defined by (2) and the following notations. L is the number of all 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 while c is the position of the global centroid. 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 [23], [24] to clustering radio channel parameters [9], [25]. It is an index in which the compactness of a cluster is estimated based on the distances from the points in a cluster to its centroid. The separation of the clusters is based on the distance from the centroids to the global centroid:

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

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 $c = \frac{1}{L}\sum_{l=1}^{L} x_l$. The optimal K number is represented by the highest value of the ν_{CH} index.

Davies-Bouldin (DB) is yet 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: $d_{k_1,k_2} = MCD(c_{k_1}, c_{k_2})$. The overall DB index is then calculated as:

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

By starting with different K values, the optimal number of clusters is achieved when the index takes the smallest value: $\nu_{DB_{opt}} = \arg\min_K \{\nu_{DB}(K)\}$.

Generalized Dunn (GD) index was introduced in [19] to ameliorate 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, thus quantifying both the separation of clusters and their spread. All-together there are 18 forms 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. In our paper, we use 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: $\Delta_3 = \frac{2}{L_k}(\sum_{l=1}^{L_k} MCD(s_l, c_k))$. Their ratio provides the Generalized Dunn index:

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

in which we capture the worst case scenario (i.e., the smallest separation and the largest cluster). The optimal value for K is given by the maximum $\nu_{D_{53}}$ index.

Xie-Beni (XB) index was initially proposed for cluster validation on fuzzy partitions, but may be used on hard partitions as well [26], [22] (i.e., for crisp clustering where the CVIs are best for their lowest or highest values).

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

Based on this formula, more compact clusters (the nominator) and larger separations between clusters (the denominator) result in smaller values for this index. Therefore, the optimal value of the XB index is the one when it reaches its minimum value for a specific clustering solution.

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

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

Based on the data analyzed in [22], it performed better than the Davies-Bouldin, Dunn and Xie-Beni indices. Nevertheless, this is not a rule, so it remains to be discovered when the above indices validate the partitioning generated by k-means applied to the MPCs in our mmWave outdoor scenario.

A. Using multiple CVIs to compare clustering solutions

Finding the number of clusters present in any 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 devised for clustering algorithms to quantify various properties of the solution such as *compactness* and *separation* between clusters. The preferred clustering solution for a specific algorithm is obtained by finding the value of K (in a certain range) that provides the optimal (min or max) value of the CVI. Nevertheless, based on their formulas, CVIs might capture only specific aspects of the clustering solution (e.g., some clusters might not be considered compact just because they have an elongated shape). Thus, no CVI should be assumed a-priori better than its alternatives. Since the hypothesis is that no single CVI can capture correctly the validity of any clustering solution (i.e., work well with all data sets), [27] proposes a conciliation of multiple CVIs for this task. As such, the value of each CVI is captured in an *ensemble* that could represent a better predictor of the clustering quality than any of the CVIs taken separately. In this 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 (9), geometric (10) and harmonic mean (11). While there are many normalization methods (e.g., z-norm, global z-norm), according to [27], the best one is the *min-max*. All indices are first normalized, to produce values in the range $[0, 1]$. Since Xie-Beni and Davies-Bouldin CVIs select the optimal K with their *min* value, we subtract their normalized value from 1. We obtain normalized and biased sets of CVIs (columns in Table I). Each column points to the K number through its *max* value. Now, the three SF_x formulas can capture the CVIs on each row in the table.

$$SF_a = \frac{1}{M}\sum_{i=1}^{M}\nu_i \tag{9}$$

$$SF_g = \Big(\prod_{i=1}^{M}\nu_i\Big)^{\frac{1}{M}} \tag{10}$$

$$SF_h = M\Big(\sum_{i=1}^{M}\frac{1}{\nu_i}\Big)^{-1}. \tag{11}$$

IV. Simulation Results

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

We simulated 28 GHz transmissions between transmitter and receiver units using one of the urban scenarios (Rosslyn, VA) delivered with the ray-tracing tool. The advantage of using this professional electromagnetic simulation tool is that we input site-specific data for any scenario, and we evaluate the signal propagation characteristics by taking into consideration the effects of buildings, terrain and even weather. In addition to that, the tool generates rays with a very high angle resolution (0.2°). As a consequence, we 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), and we feed them to the clustering algorithm.

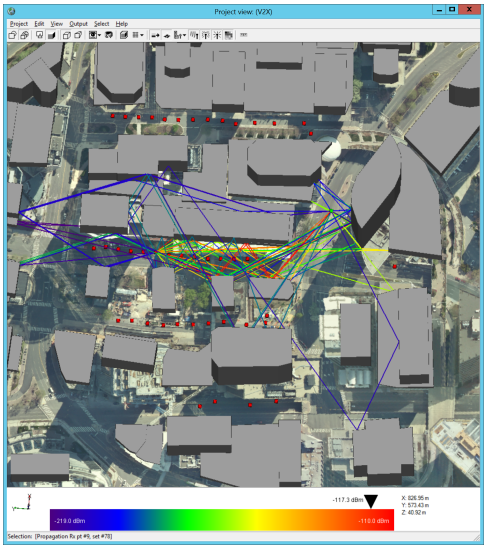

Fig. 1. 44 MPCs at the receiver Rx#9.

In our scenario, the Tx (base station) was located at a fixed site 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 was installed in a vehicle at approximately 1.5 m above ground (any of the red dots). The LOS transmission was 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 was 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 focused 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°) were used at both Tx and Rx locations in one experiment. The maximum power of the transmitted signal was 24 dBm. The ray-tracer followed a certain number of reflections (6) and diffractions (1) for each path from transmitter to receiver. Two beam alignment methods were always considered in all our studies. In the *no beam alignment* (Fig. 1), the Tx and Rx antennas were simply oriented with the street direction, whereas the *beam alignment* procedure implied that the bore-sight of the Rx antenna was 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 used Matlab to generate a random Rx point that was given to the ray-tracer for simulation. We captured the values of the *received power*,

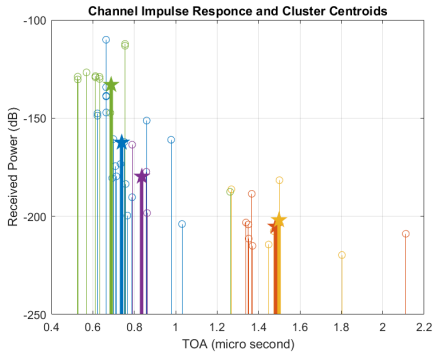

Fig. 2. Clustered CIR - each cluster has an average power and ToA.

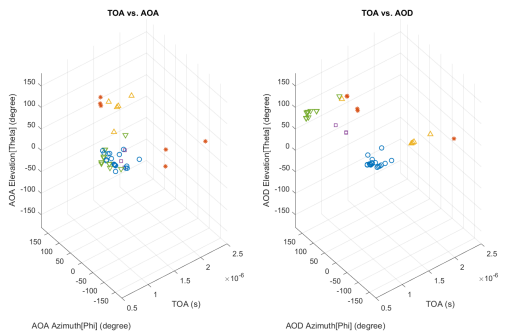

Fig. 3. Clustering via k-means algorithm - ToA vs. AoA, AoD.

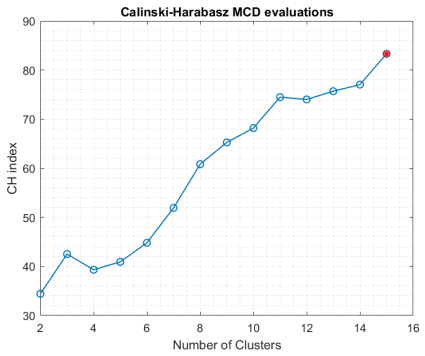

Fig. 4. CH index applied to clustering results for Rx#9.

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, to find how received power is distributed over time, how much the received rays are delayed in comparison to the LOS path, and finally, what preferred arrival directions are within each cluster. Alternatively, a multidimensional space (e.g., the MCD metric [11]) can be used to find a correlation among these parameters.

A. Clustering Algorithm Results

This section summarizes the results of the simulations for the k-means clustering algorithm with MCD metric. The urban scenario in Fig. 1 shows a number of 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 mark with a star the average power value and ToA of each cluster, both values calculated using the channel parameters of the MPCs in that cluster. Considering the large number of MPCs, it is impossible to apply a clustering procedure based on *visual inspection*. Using the k-means with MCD clustering algorithm, we obtain 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.

B. CVIs and Score Fusion Results

Once the clustering algorithm provides a partitioning for various input K values, each CVI described in Section III is applied in an attempt to find the optimal value K. As mentioned in Section III-A, this might not be the case (i.e., one or more CVIs might not be able to solve this task). Thus, combining more CVIs in a fusion classifier could potentially provide an easier way to find the optimal value of the number K of clusters. This section provides the results of the clustering validation process and of the score fusion methods described by equations (9), (10) and (11).

The CVI plots show how these indices performed on the clustered data. Receiver Rx#9 placed on a side street at approximately 150 m (Euclidean distance) from the transmitter (Fig. 1) is used for this analysis. With only 44 MPCs reaching this receiver, we can potentially consider a maximum number of 15 clusters. Thus, our k-means clustering algorithm uses as an initial guess input a number K in the range $[2, 15]$.

As mentioned, not all CVIs can find the optimal value K. For example, even if the CH index is supposed to identify this value when the index itself reaches its maximum value, we cannot say from Fig. 4 that we believe our findings; we know that practically we cannot have 15 clusters. We have a similar problem with the DB index (Fig. 5); we know that we cannot have only 2 clusters. The same small number of clusters is reported by the GD index (Fig. 6). Nevertheless, the other two indices XB (Fig. 7) and PBM (Fig. 8) behave as expected; XB shows 6 as the optimal number of clusters while PBM finds 5.

At this point, we have few CVIs that report a number of clusters hard to believe, and we also have a couple of CVIs that report different values for the K number. To check the theory about the better ensemble predictor, we plug the

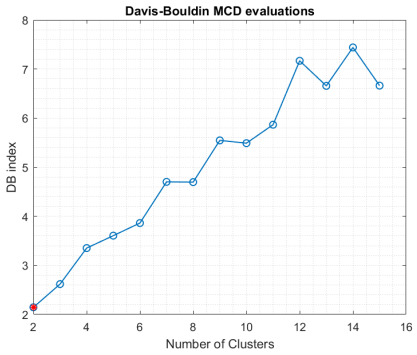

Fig. 5. DB index applied to clustering results for Rx#9.

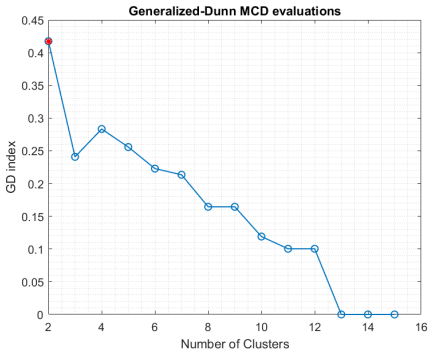

Fig. 6. GD index applied to clustering results for Rx#9.

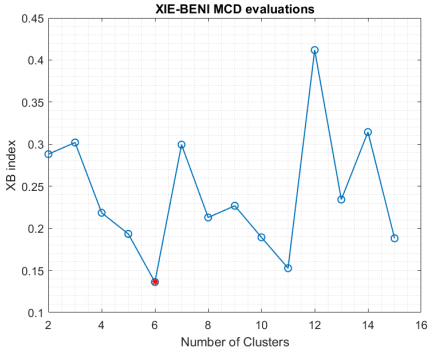

Fig. 7. XB index applied to clustering results for Rx#9.

normalized and biased CVI values obtained for Rx#9 (for each input value K) into the score fusion formulas (9), (10), and (11), as explained in Section III-A (see Table I). We notice that the optimum value K cannot be predicted using only the maximum value of the CVIs because not all CVIs have a max value on the same row. Thus, we turn to score fusion methods. The results are better in the way that at least two scores (SF_g and SF_h) agree with each other for receiver point Rx#9. As a final step (last column in Table I) we calculate the average of the three SF_x scores, and the result points to an optimal value of $K = 8$ clusters, which agrees with both geometric (SF_g) and harmonic (SF_h) mean-based scores.

We repeat this study for all 14 receivers installed on the

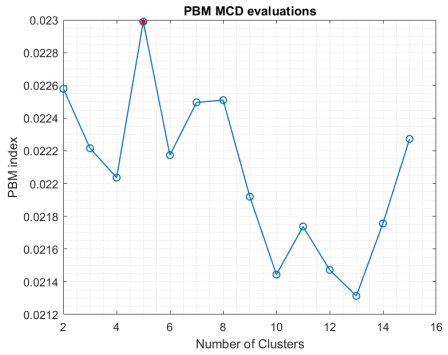

Fig. 8. PBM index applied to clustering results for Rx#9.

TABLE I
NORMALIZED AND BIASED CVIS AND SF VALUES FOR RX#9.

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

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

TABLE II
OPTIMAL VALUE K FOR EACH CVI AND SF METHOD FOR FEW RX.

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

points 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).

Going back to the solution provided by Table I, we ask ourselves if only two score fusion models or even only one had been enough to take a decision on the correct partitioning. Thus, we eliminate the score fusion based on the harmonic mean of the CVIs because its indication ($K = 8$) is the same as the one based on their geometric mean. The arithmetic mean-based score fusion points to a solution with 5 clusters while the geometric mean-based one indicates 8 clusters. However,

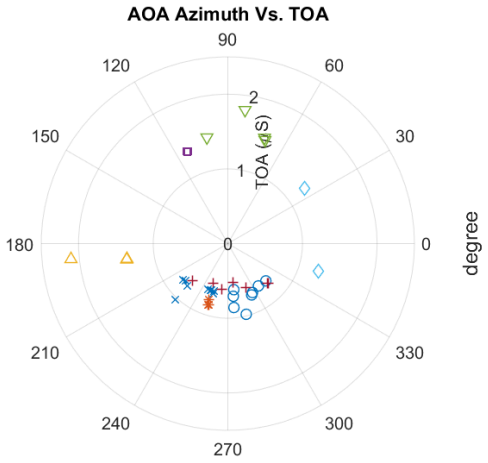

Fig. 9. Polar plot of azimuth AoA vs. ToA for K=8.

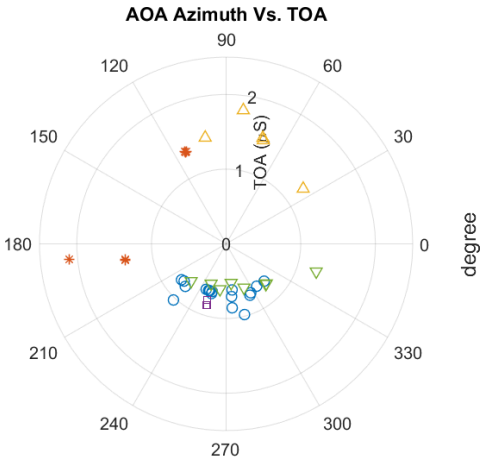

Fig. 10. Polar plot of azimuth AoA vs. ToA for K=5.

when we take the average value for the two score fusion factors (SF_a and SF_g), the optimal clustering solution is $K = 5$. To prove that one solution might be better than the other one, we decide to 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. Since the *elevation* component of the two angles shows little spatial variation, we focus on the *azimuth* component. For lack of space, we show only the polar plots of the AoA for both solutions (Fig. 9 and Fig. 10). 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. 11 and Fig. 12), in order to understand the advantage of each potential clustering. As Fig. 12 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. Using $K = 5$ as input to the k-means algorithm, we group the MPCs around their centroids (Fig. 3).

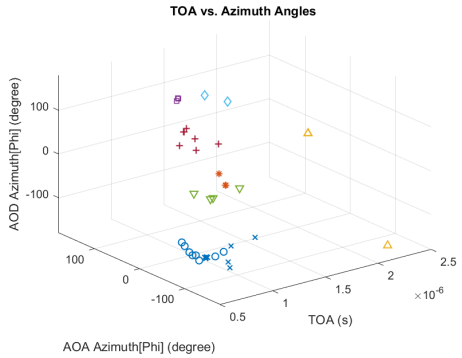

Fig. 11. 3D representation based on ToA and azimuth of AoA, AoD for K=8.

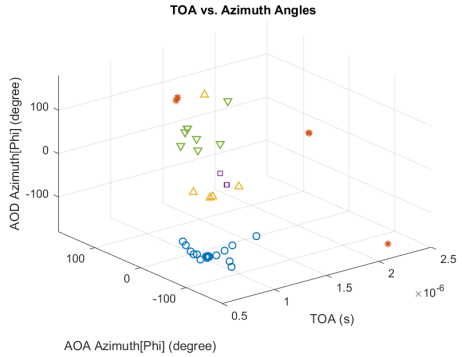

Fig. 12. 3D representation based on ToA and AoA, AoD for K=5.

Polar or 3D plots are definitely helpful when the number of estimated clusters is small and we can infer something from their visualization. However, when we deal with a very large number of MPCs, 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. We propose a *statistics-based* decision by calculating the *variance* of the values for various parameters of the MPCs in each cluster. Thus, 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 variance calculation implies obtaining first the mean value for each parameter in each cluster, so the total variance values can be compared, even though one solution has 5 clusters and the other one has 8. The results in Table III show that if we are

TABLE III
TOTAL VARIANCE OF MPC PARAMETERS FOR 5 AND 8 CLUSTERS.

MPC Parameter	Total Variance K=5	Total Variance 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. Many times, this is one important thing for a wireless network architect who wants to identify quickly the dominant clusters that account for most of the received power, and the most relevant directions of reception for orienting a single-antenna receiver. On the other hand, mmWave transmissions consider *directivity* as one of their dominant traits, so it is equally important to analyze the clusters purely 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. This solution will prevail in our future research when we plan to consider the influence of diffuse scattering and also of MIMO transmissions where grouping rays based on their angle-of-arrival is crucial.

V. Conclusions

mmWave communication systems will rely heavily on directive transmissions and MIMO antenna configurations for which accurate channel models are required. The MPCs arriving at the receiver occur in clusters, which, properly identified, influence the generation of these models. Our paper emphasized the importance of clustering algorithms and of their validation for predicting optimal partitioning of MPCs. 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 CVIs. We noticed that few of the CVIs used in our study were not capable to find the correct partitioning. However, using score fusion techniques allowed us to narrow down to only two options for the optimal value K. Further statistics-based decisions selected the most appropriate clustering solution.

Acknowledgments

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

References

[1] A. A. M. Saleh and R. Valenzuela, "A Statistical Model for Indoor Multipath Propagation," *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 2, pp. 128–137, February 1987.

[2] C. Gustafson, D. Bolin, and F. Tufvesson, "Modeling the cluster decay in mm-wave channels," in *Proceedings of the 8th European Conference on Antennas and Propagation (EuCAP)*, The Hague, Netherlands, April 6–11 2014, pp. 804–808.

[3] 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.

[4] M. K. Samimi and T. S. Rappaport, "3-D statistical channel model for millimeter-wave outdoor mobile broadband communications," in *Proceedings of the IEEE International Conference on Communications (ICC)*, London, UK, June 8–12 2015, pp. 2430–2436.

[5] 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.

[6] 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.

[7] M. T. Martinez-Ingles, D. P. Gaillot, J. Pascual-Garcia, J. M. Molina Garcia-Pardo, M. Lienard, J. V. Rodrguez, 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognition*, vol. 36, no. 2, pp. 451–461, February 2003.

[13] J. M. Pena, J. A. Lozano, and P. Larranaga, "An empirical comparison of four initialization methods for the K-Means algorithm," *Pattern Recognition Letters*, vol. 20, no. 10, pp. 1027–1040, October 1999.

[14] M. Meila and D. Heckerman, "An Experimental Comparison of Model-Based Clustering Methods," *Machine Learning*, vol. 42, no. 1–2, pp. 9–29, January 2001.

[15] P. S. Bradley and U. M. Fayyad, "Refining Initial Points for K-Means Clustering," in *Proceedings of the 15th International Conference on Machine Learning (ICML)*, Madison, WI, July 24–27 1998, pp. 91–99.

[16] D. Pelleg and A. W. Moore, "X-means: Extending K-means with Efficient Estimation of the Number of Clusters," in *Proceedings of the 17th International Conference on Machine Learning (ICML)*, Stanford, CA, June 29–July 2 2000, pp. 727–734.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.