A Novel Clustering Method Based on Density Peaks and Its Validation by Channel Measurement

Hanqing Xue[1], Lei Tian[1], Yuxiang Zhang[1], Chao Wang[1], Jianhua Zhang[2], Wei Li[3], Hewen Wei[4]
[1]Key Lab of Universal Wireless Communications, Ministry of Education,
[2]State Key Lab of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, P. O. Box 92, Beijing 100876, China
[3]Department of Electrical Engineering Northern Illinois University, DeKalb, IL 60115
[4]National key laboratory of science and technology on blind signal processing, Chengdu, Sichuan 610041, China
Email: {xuehanqing, tianlbupt, jhzhang}@bupt.edu.cn; weili@ieee.org; weihewen@tsinghua.org.cn

***Abstract*—In this paper, cluster centers are defined as local maxima in the density which is called density peaks. We use an algorithm based on density peaks to cluster multipath components (MPCs) whose parameters are estimated by the space-alternating generalized expectation-maximization (SAGE) algorithm. To validate its performance, we apply it to the data from a channel measurement in an indoor scenario at 28 GHz and compare it to the widely used KPowerMeans algorithm with the same distance metric of multiple component distance (MCD). Through the results, we find it can cluster MPCs more accurately validated by Calinski-Harabasz (CH) and Davies-Bouldin (DB) indices. More importantly, it overcomes the shortcoming of setting the number of clusters manually because we can determine the cluster centers automatically. Besides, the efficiency of DP clustering is far superior to KPowerMeans because what it needs is only a distance matrix and we only need to compute it once. At the end, statistical analysis of clusters are presented to provide insights in channel modeling of millimeter wave.**

***Index Terms*—clustering algorithm, channel measurement, channel statistical characteristics**

I. Introduction

Facing the tremendous increase in the volume of mobile data and the high data transmission rate in the coming fifth generation (5G) wireless communication [1], an accurate but not so complex channel model is required in 5G system design. Because the cluster based channel modeling can maintain accuracy while reducing complexity, it has been playing an important role in the development of channel modeling. Therefore, many standardized channel models like ITU-R M.2135 [2] have adopted the cluster based modeling. Besides, many measurement results present that the multipath components (MPCs) are distributing as clusters [3], which are defined as a group of MPCs with similar parameters, e.g. azimuth angle of arrival (AOA), azimuth angle of departure (AOD), elevation angle of departure (EOD), elevation angle of arrival (EOA), and delay. So finding clusters of MPCs accurately and efficiently is a very attractive research topic now.

In recent years, many clustering algorithms were proposed and the KPowerMeans algorithm is one of the most representatives of them. In [4], the KPowerMeans algorithm is proposed which considers the influence of MPC power while computing cluster centers. Specially, in KPowerMeans algorithm, the multipath component distance (MCD) is used to define the distances between MPCs in [5]. However, because a MPC is always assigned to the nearest center, the KPowerMeans methods scale poorly with respect to the time taken to complete each iteration. An inignorable shortcoming is that the number of clusters k has to be supplied by the user. So we use an alternative approach proposed in [6] to cluster MPCs. It defines cluster centers as density peaks which are characterized by a higher density than their neighbors and a relatively large distance to higher densities. It has its basis only on the distances between data points similar to the KPowerMeans and can find the correct number of clusters automatically.

In this paper, we use the algorithm based on density peaks (DP clustering) to determine the number of clusters automatically and cluster the MPCs efficiently. To reveal the performance and advantages of the algorithm, it is validated using the data from the channel measurement in an indoor scenario at 28 GHz and a comparison is made between it and the widely used KPowerMeans algorithm. We also present statistical analysis of clusters in this scenario to provide reference about cluster propagation.

The rest of the paper is organized as follows. The distance metric and DP clustering algorithm are described in Section II. Section III introduces the measurement facilities and scenario. Section IV shows the results including a comparison between the algorithm and KPowerMeans and presents the statistical analysis of clusters. Section V concludes this paper.

II. Clustering Algorithm

A. MCD

The space-alternating generalized expectation-maximization (SAGE) algorithm [7] is used to extract channel parameters from the measured data. The MPCs to be clustered are described in the form of $[\tau, \theta_{EOA}, \theta_{EOD}, \varphi_{AOA}, \varphi_{AOD}]$. θ usually represents the elevation angle and φ stands for azimuth angle. And τ is the time delay. The distances between individual MPCs are calculated by MCD.

The delay distance between the ith and jth MPC is

$$MCD_{\tau,ij} = \alpha \cdot \frac{|\tau_i - \tau_j|}{\Delta\tau_{max}} \cdot \frac{\tau_{std}}{\Delta\tau_{max}} \tag{1}$$

In this equation, $\Delta\tau_{max}$ is the maximum excess delay and so $\Delta\tau_{max} = max_{i,j}\left|\tau_i - \tau_j\right|$. τ_{std} is the standard deviation of the delay values. α is a scaling factor to give the delay suitable "importance" in the final distance function.

The angle distance is given as

$$MCD_{T/R,ij} = \frac{1}{2}\left|\begin{pmatrix}\sin\theta_i\cos\varphi_i\\ \sin\theta_i\sin\varphi_i\\ \cos\theta_i\end{pmatrix} - \begin{pmatrix}\sin\theta_j\cos\varphi_j\\ \sin\theta_j\sin\varphi_j\\ \cos\theta_j\end{pmatrix}\right| \quad (2)$$

The angle distance is calculated in the spherical coordinate system for AOA and AOD likewise. The T and R represent Tx and Rx respectively.

Finally, the total distance measure is defined as

$$MCD_{ij} = \sqrt{\parallel MCD_{T,ij} \parallel^2 + \parallel MCD_{R,ij} \parallel^2 + {MCD_{\tau,ij}}^2} \quad (3)$$

We can see that the MCD integrates the effects of angle of arrival, angle of departure and delay together.

B. Two key quantities of DP clustering

The DP clustering is proposed on the basis of the assumptions that every cluster center is surrounded by MPCs with smaller local density and they are relatively far from any MPCs with a larger local density. For each MPC, we need to compute two important quantities: the local density ρ and the distance δ from the MPC to higher density.

We have two method to compute ρ, the first one is Cut-off kernel:

$$\rho_i = \sum_{j=1}^{N} f\left(d_{ij} - d_c\right) \quad (4)$$

where $f\left(x\right) = \begin{cases}1 & \text{if } x < 0\\ 0 & \text{if } x \geq 0\end{cases}$, N is the number of MPCs and d_c is named cutoff distance which is needed to be assigned. An advice to choose d_c is making the average number of neighbors being around 1% to 2% of the total number of MPCs. This percentage is a parameter to be set manually and we use $t \in (0,1)$ to represent it. So obviously, the ρ_i here equals to the number of MPCs whose distance to the ith subpath x_i is smaller than d_c.

The second one is named Gaussian kernel:

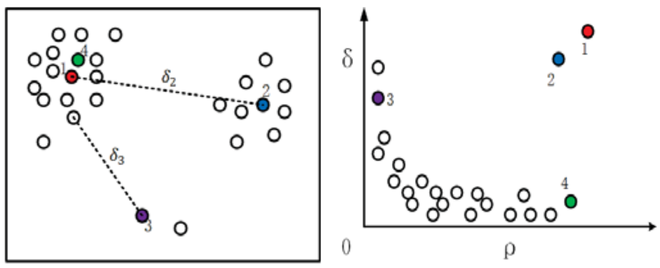
$$\rho_i = \sum_{j=1}^{N} e^{-\left(\frac{d_{ij}}{d_c}\right)^2} \quad (5)$$

We can see that the result is a discrete value when we use Cut-off kernel and it is easy that different MPCs have the same ρ value. So we choose to use Gaussian kernel because the result is a continuous value when using it.

δ_i is the minimum distance from the ith MPC to any other MPC with higher density:

$$\delta_i = \min_{j:\rho_j > \rho_i} d_{ij} \quad (6)$$

For the MPC with the highest density, we usually make $\delta_i = \max_j d_{ij}$.



(a) Point distribution (b) Decision graph for the data

Fig. 1. An example in two dimensions.

Here is an example shown in Fig. 1. Fig. 1 (a) shows several points in a 2D space. Obviously the density maxima are at the Point 1 and Point 2. δ_1 equals to the biggest distance from Point 1 because it has the largest local density. For Point 2, δ_2 is the distance from it to Point1, the nearest point whose local density is larger than its. The outliers like Point 3 are featured by large δ but small ρ. Point 4 has a large ρ, but its δ value is small with the cause of being close to a more dense area. So only Point 1 and 2, who are cluster centers, have large ρ and δ simultaneously and pop up from other points. Then we can know how many clusters there is. So the (ρ_i, δ_i) graph as shown in Fig. 1(b) is called Decision Graph because of its important role in determining the cluster centers.

In addition, because ρ_i and δ_i are usually on different orders of magnitude, we add the operation of normalization when searching for cluster centers. To find cluster centers automatically, we can also use another parameter $\gamma = \rho \cdot \delta$ with an appropriate threshold. There will be an obvious skip from cluster centers to those are not cluster centers.

C. DP Clustering

1) Initialization and preprocessing: In this step, we mainly work out the two key values, ρ and δ. To provide convenience for the following clustering, we also need $\left\{q_i\right\}_{i=1}^{N}$, the descending order of $\left\{\rho_i\right\}_{i=1}^{N}$, and $\left\{n_i\right\}_{i=1}^{N}$ which is the index of the nearest MPC whose ρ value is larger than the ith one.

Algorithm 1 Initialization and preprocessing

```
compute d_ij = MCD(x_i, x_j);
compute ρ_i using Gaussian kernel;
sort {ρ_i}_{i=1}^N in descending order to get {q_i}_{i=1}^N;
n_i=0, i ∈ I_S;
for each i ∈ [2, N] do
    δ_{q_i} = d_max;
    for each j ∈ [1, i−1] do
        if d_{q_i,q_j} < δ_{q_i} then
            δ_{q_i} = d_{q_i,q_j}
            n_{q_i} = q_j
        end if
    end for
end for
δ_{q_1} = max_{j≥2} δ_j
```

2) Determine the cluster centers: We can find cluster centers according to Decision Graph and the graph of γ as mentioned above.

3) Cluster the MPCs that are not cluster centers: We cluster the MPCs that are not cluster centers to the same center with the nearest MPC whose ρ value is larger than it. So if we traversal the MPCs with the order of q_i, we only need to cluster it to the center of n_{q_i}.

D. KPowerMeans algorithm

The steps of KPowerMeans algorithm are given briefly as followed [4]. In this algorithm, a range $[K_{min}, K_{max}]$ should be set at first. L is the number of MPCs. P_l is the power of the lth subpath and the meaning of x is mentioned before.

Algorithm 2 KPowerMeans clustering algorithm

Randomly initialize K centroid positions $c_1^{(0)}, ..., c_K^{(0)}$;
for $i = 1$ to MaxIterations **do**
  Assign MPCs to cluster centroids and store indices:
  $I_l^{(i)} = arg\min \left\{ P_l \cdot MCD\left(x_l, c_k^{(i-1)}\right)\right\}$
  Recalculate cluster centroids: $c_k^{(i)} = \frac{\sum_{j\in c_k^{(i)}}(P_j \cdot x_i)}{\sum_{j\in c_k^{(i)}} P_j}$
  if $c_k^{(i)} = c_k^{(i-1)}$ for all $k = 1, ..., K$ **then**
    break;
  end if
end for
Return $R_K = \left[I_{(i)}, c_k^{(i)}\right]$

III. Measurement Scenario and Facilities

To provide data for clustering algorithm in previous section and validate its performance, an indoor measurement of millimeter wave is performed.

A. Measurement Scenario

The measurement campaign was conducted in an indoor LOS scenario of Beijing University of Posts and Telecommunications (BUPT). As shown in Fig. 2, the geometric size of the office is equal to $10.97 \times 6.62 \times 2.40 m^3$. The transmitter (Tx) is located in a fixed position at the northeast corner and there are 16 receiver (Rx) locations distributing in the office.

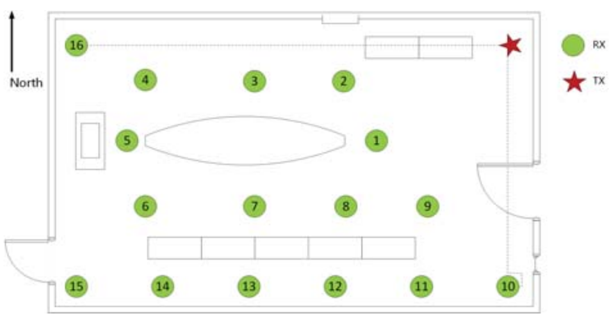


Fig. 2. Layout of measurement environment

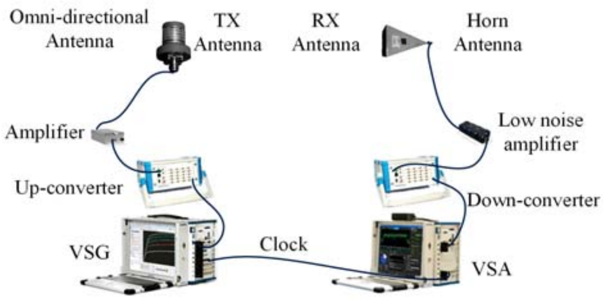


Fig. 3. Measure equipments

B. Measurement Facilities

The measurement campaign was conducted using a broadband correlator channel sounder at 28 GHz with 400 MHz bandwidth. The Tx and Rx antennas are at the same height of 1.55 m. In order to conduct a virtual Multiple-Input Multiple-Output (MIMO) measurement, there is a horn antenna at Rx being rotated around with 5° each step in azimuth domain and 10° each step in elevation domain to receive the channel characteristics. An omni-directional antenna is fixed at Tx. The receiver sensitivity is -70 dBm, so we use the amplifier and low noise amplifier (LNA) at both Tx and Rx to expand the dynamic range in the link budget. A common reference clock source is used to synchronize Tx and Rx. What should be specially noted is that the pseudo-random (PN) sequence with the length of 511 is generated with the chip rate of 400 MHz and the signal received is collected with a sampling rate of 1.2 GHz, because the effective bandwidth of the band-pass filter is only 400 MHz. Table I shows the basic parameters of our measurement.

TABLE I
MEASUREMENT CONFIGURATION PARAMETERS

Parameter	Value
Carrier Frequency	28 GHz
PN Sequence Length	511
BandWidth	400 MHz
Tx Antenna	Omni-directional
Rx Antenna	Horn
Tx Ant.AZ HPBW	360°
Rx Ant.AZ HPBW	10°
Tx Ant.Gain	2.93 dBi
Rx Ant.Gain	25 dBi
Ant.Polarization	V-to-V

IV. Results

A. Decision Graph

According to (5), we can compute the local density ρ of every MPC as shown in Fig. 4. The colors of the points and the color bar indicate the level of ρ. We can see that there are 9 areas having a MPC who has larger ρ than its neighbors. They are very likely to be cluster centers. Then we can plot the Decision Graph and the graph of γ as shown in Fig. 5 to further determine the cluster centers.

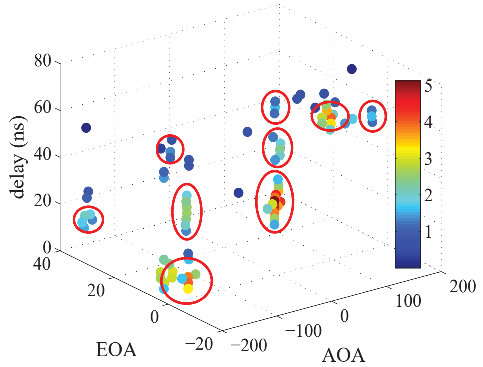
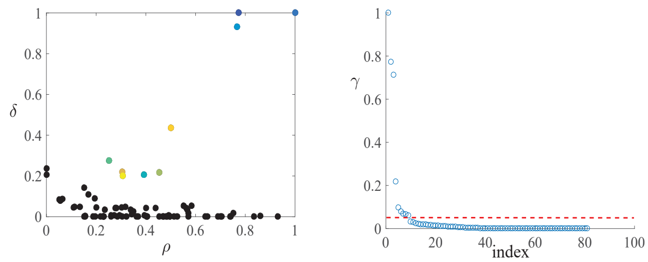


Fig. 4. Graph of local density



(a) Decision Graph

(b) Graph of γ

Fig. 5. Graph of finding cluster centers

We can find that there are 9 MPCs with relatively large ρ and δ values simultaneously. Their γ values also step up from all the others. Then we can set them to be cluster centers and cluster all the MPCs with the DP clustering algorithm mentioned above. The clustering result is showed in Fig. 6. The clusters are plotted with different colors. Comparing Fig. 4 and Fig. 6, we can see that clusters are identified with the center of MPCs whose local density ρ is relatively larger than its neighbors.

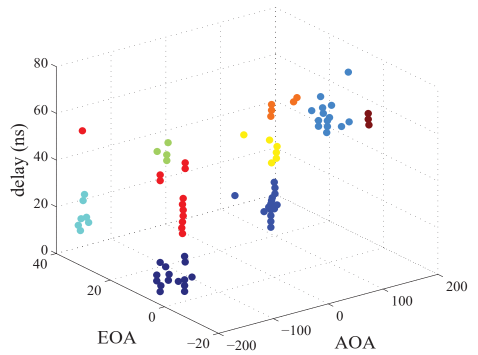


Fig. 6. The result of clustering by density peaks

B. Comparison with KPowerMeans algorithm

In channel modeling before, we often use KPowerMeans algorithm to cluster MPCs. Here we make a comparison between the two algorithms. Because the KPowerMeans algorithm cannot determine the number of clusters k automatically, we set it from 2 to 20 artificially and then select a good one according to the validity indices, Calinski-Harabaz (CH) and Davies-Bouldin (DB) [9]. Fig. 7 shows the CH and DB index values of different k using KPowerMeans and the same two values from the result of DP clustering algorithm. The clustering result of KPowerMeans is showed below in Fig. 8.

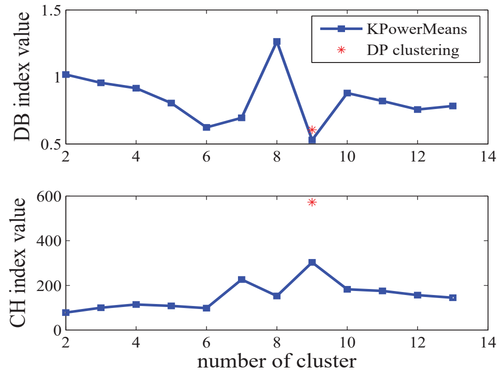


Fig. 7. CH and DB index value

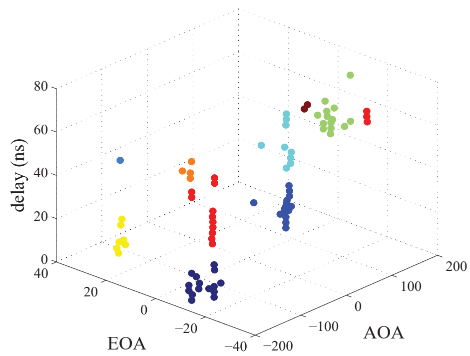


Fig. 8. Clustering result using KPowerMeans

According to the CH and DB values shown in Fig. 7, there are 9 clusters at this spot because higher CH and lower DB index values means that the clusters are more intra-compact and inter-separate [8]. Intuitively, the DP clustering and KPowerMeans algorithm can get the similar result using the same distance metric. Besides, the result using the DP clustering algorithm has a larger CH value and a close DB value than KPowerMeans. But if we use KPowerMeans algorithm, we might not be able to set an appropriate range of k when the scenario is complex or get a high computational complexity with a large range of k, while we can get the number of clusters obviously and cluster very efficiently for the only once computation of distance matrix using DP clustering algorithm.

So the algorithm provides a method to determine the number of clusters automaticlly instead of setting manually and to cluster MPCs efficiently.

C. Statistical analysis of clusters

Based on the clustering result, we get intra-cluster parameters, including DS, ASA and ESA.

1) Delay Spread: DS of kth cluster can be caculated by :

$$\mu_{\tau,k} = \left(\sum_{l=1}^{L_k} \tau_l \cdot P_l \right) / \sum_{l=1}^{L_k} P_l \tag{7}$$

$$\sigma_{\tau,k} = \sqrt{\left(\sum_{l=1}^{L_k} (\tau_l - \mu_{\tau,k})^2 \cdot P_l \right) / \sum_{l=1}^{L_k} P_l} \tag{8}$$

τ_l and P_l are the delay and normalized power of the lth path respectively. So $\mu_{\tau,k}$ means the weighted average of delays in kth cluster.

2) Angular Spread: We will get ASA and ESA of the kth cluster after changing the τ in (7) and (8) with the corresponding angle. The method proposed in [9] is used to solve angle ambiguity.

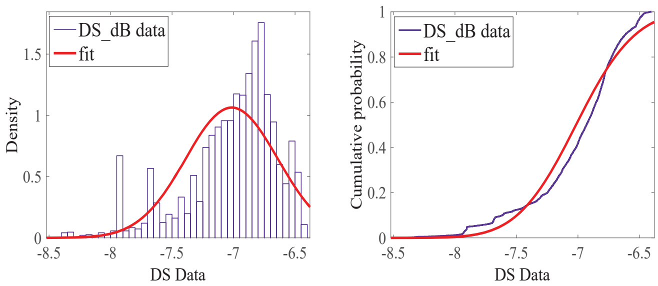


Fig. 9. Distribution of delay spread within a cluster

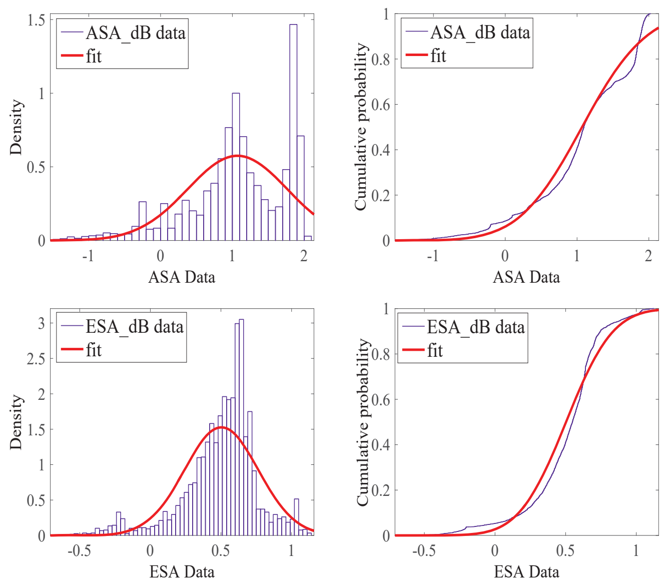


Fig. 10. Distribution of ASA and ESA within a cluster

We take the logarithms of DS, ASA and ESA and fit them using normal distribution with mean values 6.97ns, $8.08°$ and $3.18°$ respectively. That can be seen in Fig. 9 and Fig. 10.

V. Conclusion

In this paper, we use the clustering algorithm based on density peaks, with MCD as the distance metric, to cluster the MPCs successfully. As the results display, compared to the commonly used KPowerMeans algorithm, the DP clustering algorithm can get the same number of clusters and better performance seen from CH and DB values as validity indices. What is more, the algorithm overcomes the shortcoming of being difficult to determine the number of clusters and it can cluster the MPCs more efficiently. Besides, the intra cluster parameters including DS, ASA and ESA are obtained and their logarithms are all fitted well by normal distribution. These results will be very useful for indoor channel modeling of millimeter wave.

Acknowledgment

The research is supported by National Science and Technology Major Project of the Ministry of Science and Technology and project name is Development, Standardization and Verification Research on 5G Integrated Networking of High and Low Frequency Bands with 2017ZX03001012-003 and by Exploratory Project of State Key Lab of Networking and Switching Technology and project name is Research on millimeter-wave Channel Characteristics and Modeling with NST20170205.

References

[1] J. Zhang, "The interdisciplinary research of big data and wireless channel: A cluster-nuclei based channel model," *China Communications*, vol. 13, pp. 14–26, N 2016.

[2] M. Series, "Guidelines for evaluation of radio interface technologies for IMT-Advanced," *Report ITU*, no. 2135-1, 2009.

[3] A. F. Molisch and F. Tufvesson, "Propagation channel models for next-generation wireless communications systems," *IEICE Transactions on Communications*, vol. 97, no. 10, pp. 2022–2034, 2014.

[4] 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 *IEEE Vehicular Technology Conference*, pp. 1–5, Sept 2006.

[5] 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, pp. 33–5, Jan 2006.

[6] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[7] T. Jiang, L. Tian, P. Tang, Z. Hu, and J. Zhang, "Basestation 3-dimensional spatial propagation characteristics in urban microcell at 28 GHz," in *2017 11th European Conference on Antennas and Propagation (EUCAP)*, pp. 3167–3171, March 2017.

[8] C. Huang, J. Zhang, X. Nie, and Y. Zhang, "Cluster characteristics of wideband MIMO channel in indoor hotspot scenario at 2.35 GHz," in *2009 IEEE 70th Vehicular Technology Conference Fall*, pp. 1–5, Sept 2009.

[9] W. Dong, J. Zhang, X. Gao, P. Zhang, and Y. Wu, "Cluster identification and properties of outdoor wideband MIMO channel," in *2007 IEEE 66th Vehicular Technology Conference*, pp. 829–833, Sept 2007.