TELKOMNIKA Telecommunication Computing Electronics and Control
Vol. 23, No. 4, August 2025, pp. 871~881
ISSN: 1693-6930, DOI: 10.12928/TELKOMNIKA.v23i4.26439

Outlier detection and clustering of fifth-generation wireless channel model datasets

Jojo Blanza[1], John Bernard Cipriano[2]
[1]Department of Electronics Engineering, Faculty of Engineering, University of Santo Tomas, Manila, Philippines
[2]Bank of the Philippine Islands, Makati, Philippines

Article Info

Article history:

Received June 24, 2024
Revised Mar 25, 2025
Accepted May 10, 2025

Keywords:

5G
Channel model
Clustering
Multipaths
Outlier

ABSTRACT

The fifth-generation (5G) wireless communications system offers faster data rates, lower latency, and more interconnecting devices. Various 5G channel models were developed to study its stochastic characteristics before implementation. These channel models generate multipath components that are grouped into clusters. The multipath clusters serve as datasets in multipath clustering. The clustering results are then used to examine the propagation properties of the 5G system. However, datasets are prone to outliers. They tend to affect clustering accuracy. Hence, this study clusters the datasets generated by the channel models using five clustering approaches, removes the outliers using mean-shift outlier detection, and clusters the datasets free of outliers again using the same clustering algorithms. Outlier detection shows that 5G channel model datasets contain noise, and outlier removal improves the modeling characteristics, as demonstrated by enhanced clustering accuracy. Results show that most of the outliers are detected in the 2×SD threshold. The removal of the outliers using the said threshold increased the clustering accuracy of K-means and AC-Single in Semi-Urban B1 LOS multiple links by 78.85% and 55%, respectively, and DBSCAN in Semi-Urban B2 LOS multiple links by 57.14%. Outlier detection and removal also work well with 5G channel model datasets.

Corresponding Author:

Jojo Blanza
Department of Electronics Engineering, Faculty of Engineering, University of Santo Tomas
España, Manila 1015, Philippines
Email: jfblanza@ust.edu.ph

1. INTRODUCTION

The fifth-generation (5G) wireless system increased bandwidth, shortened latency, and permitted more interconnecting devices. The physical implementation of 5G is costly; hence, correct modeling of the 5G system is necessary. 5G channel models such as the European Cooperation in Science and Technology (COST 2100) [1], International Mobile Tecommunications-2020 (IMT-2020) [2], Quasi-Deterministic Radio Channel Generator (QuaDRiGa) [3], and Wireless World Initiative New Radio II (WINNER II) [4] are used to study the stochastic properties of 5G before putting up the physical system. These 5G channel models generate wireless multipath components (MPCs) that form multipath clusters (MCs) when they have a similar delay, angle of departure, and angle of arrival. The MPCs and MCs serve as datasets used in multipath clustering to study the propagation characteristics of 5G. Understanding the characteristics of 5G channels helps designers compare and deploy the most appropriate wireless technologies.

A previous study [5] on multipath clustering has shown that clustering accuracy is low due to outliers. However, clustering accuracy can be improved when outliers are removed from the dataset.

Examples are the synthetic and map datasets using an outlier removal clustering algorithm [6], the Forest Cov dataset using a scalable and robust clustering algorithm [7], Real-World datasets using an improved cuckoo search-based K-means [8], UCI datasets using local density and natural neighbor-based outlier detection [9], synthetic and real datasets using angle-based outlier factor [10], and 2-D datasets using mean-shift outlier detection and filtering [11].

Even multipath datasets contain outliers. A previous study [12] tried to detect and remove the outliers of the COST 2100 dataset involving just indoor scenarios. It used mean-shift outlier detection [11] to identify and remove outliers and simultaneous clustering and model selection matrix affinity (SCAMSMA) [13] for the clustering. Results were positive, as clustering accuracy improved by an average of 2.2%. It was the first study to implement outlier detection on multipath datasets, but it only used one 5G channel model (COST 2100) and two indoor channel scenarios. The current study expands the previous research by considering four 5G channel models (COST 2100, IMT-2020, QuaDRiGa, and WINNER II) and their complete channel scenarios (total of thirty-three). This paper contributes to the characterization of multipath datasets by detecting their outliers. Furthermore, this research addresses the problem of improving the accuracy of clustering multipaths by removing outliers. Lastly, by pruning the outliers, the users can optimize the parameters of the channel model for a more accurate implementation of 5G technology.

2. METHOD

The flowchart for the methodology is shown in Figure 1. The datasets generated by the channel models are clustered to get the clustering accuracy. The outliers in the datasets are removed to create new datasets free of outliers. They are again clustered to get the new accuracy. The original clustering accuracy from the datasets with outliers is then compared to the new clustering accuracy from those without outliers.

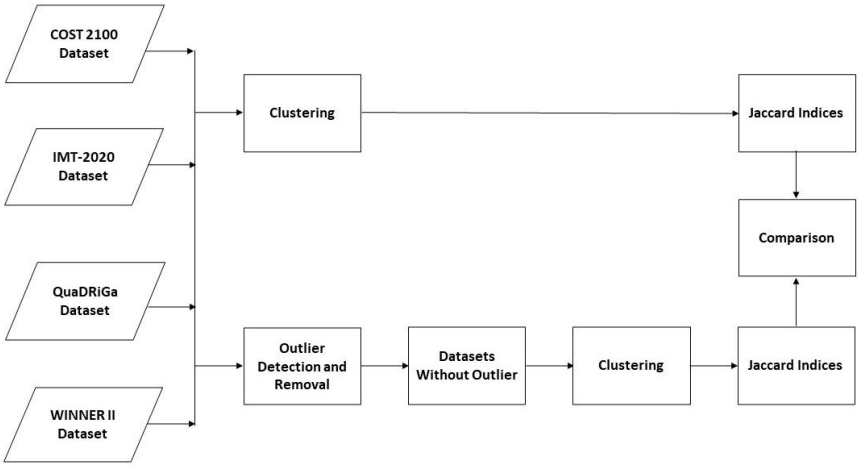

Figure 1. The flowchart of method

2.1. Datasets for outlier detection and clustering

The datasets consist of MPCs and MCs generated by the 5G channel models. They serve as reference data in multipath clustering. They were also used in the process of outlier detection and removal. The COST 2100 dataset was taken from [14], whereas the IMT-2020, QuaDRiGa, and WINNER II datasets were lifted from [15].

Table 1 shows the summary of the four datasets. Each channel scenario has thirty sheets of Excel file data. COST 2100 has eight channel scenarios. The number of clusters and multipaths varies for each channel scenario. The numbers given pertain to the maximum number of clusters and multipaths per channel scenario. The IMT-2020 dataset has eleven channel scenarios with the same number of clusters and multipaths per channel scenario. It is the most among the four datasets. The QuaDRiGa dataset has the same number of clusters and multipaths per channel scenario. However, it has only eight channel scenarios. Lastly, the WINNER II dataset has six channel scenarios, the least among the four. Like IMT-2020 and QuaDRiGa, WINNER II has the same number of clusters and multipaths per channel scenario.

Table 1. Number of clusters and number of multipaths per cluster for each channel scenario

Channel model	Channel scenario	Number of clusters	Number of multipaths
	Indoor B1 LOS single link	27	81
	Indoor B2 LOS single link	26	78
	Semi-Urban B1 LOS single link	35	945
COST 2100	Semi-Urban B2 LOS single link	38	1,026
	Semi-Urban B1 NLOS single link	34	1,632
	Semi-Urban B2 NLOS single link	33	1,584
	Semi-Urban B1 LOS multiple links	63	1,701
	Semi-Urban B2 LOS multiple links	66	1,782
	InH A LOS	15	1,425
	InH A NLOS	19	3,895
	RMa A LOS	11	31,768
	RMA A NLOS	10	26,600
	RMa A O2I	10	58,520
IMT-2020	UMa A LOS	12	11,172
	UMa A NLOS	20	77,140
	UMa A O2I	12	216,144
	UMi A LOS	12	12,084
	UMi A NLOS	19	24,187
	UMi A O2I	12	109,440
	BERLIN UMa LOS	15	18,000
	BERLIN UMa NLOS	25	30,000
	BERLIN UMi Campus LOS	12	7,200
QuaDRiGa	BERLIN UMi Campus NLOS	20	12,000
	BERLIN UMi Square LOS	12	7,200
	BERLIN UMi Square NLOS	20	12,000
	Industrial LOS	25	7,500
	Industrial NLOS	26	7,800
	Indoor A1 LOS	12	3,600
	Indoor A1 NLOS	16	4,800
WINNER II	UMa C2 LOS	8	9,600
	UMa C2 NLOS	20	24,000
	UMi B1 LOS	8	4,800
	UMi B1 NLOS	16	9,600

2.2. Clustering approaches

The datasets with outliers were clustered using K-means [16], K-medoids [17], agglomerative hierarchical clustering (AC-Single) [18], density-based spatial clustering of applications with noise (DBSCAN) [19], and spectral clustering (SC) [20]. Their MATLAB implementations can be found in [21]. The datasets without outliers were again clustered using the above mentioned five clustering approaches. The clustering process is shown in Figure 2. The clustering of datasets with outliers is illustrated in Figure 2(a), while the clustering of datasets without outliers is presented in Figure 2(b).

Figure 2. Clustering of datasets: (a) with outliers and (b) without outliers [11]

2.3. Outlier detection and removal

The outliers were detected using the mean-shift outlier detector [11]. The detector uses the mean-shift technique, which replaces every object by the mean of its K-nearest neighbors (KNN). The method forces an object to move towards a dense area. Outliers are then determined by the movement of the objects using an outlier score. An object has a higher chance of being an outlier if it has a more significant movement.

The mean-shift outlier detector analyzes the distribution of the calculated outlier scores to detect the outliers. The standard deviation (SD) of all the scores is used as the global threshold. The study uses three popular outlier thresholds: 2×SD, 2.5×SD, and 3×SD [22]. Any object with an outlier score higher than the outlier threshold is considered an outlier.

2.4. Clustering accuracy

The Jaccard index (η) is used to evaluate the clustering accuracy of the five clustering approaches. It represents the intersection over the union of the reference data and the clustered data. Its numeric values can range from 0 to 1, with one being the highest. A Jaccard index of 1 means a perfect match between the

reference data and the clustered data, while an index of 0 means no match between the two datasets. The Jaccard index objectively shows the match and mismatches between the reference and clustered data. The clustering metric can be computed as follows:

$$\eta = \frac{m_{11}}{m_{11}+m_{10}+m_{01}} \quad (1)$$

where:

m_{11} is the number of pairs that are correctly classified.
m_{01} is the number of pairs that are not correctly classified.
m_{10} is the number of pairs that are incorrectly classified when they are not supposed to.

2.5. Comparison of clustering accuracy

The Jaccard indices of the clustered datasets with outliers and datasets without outliers are compared using SD and box plots. The SD of the Jaccard indices of the thirty sheets of Excel data per channel scenario is calculated using the Excel function STDEV.S. The Excel formula estimates the SD of a sample, ignoring logical values and text in the sample. The SDs of the Jaccard indices are then compared to determine the compactness of clusters in datasets with outliers and datasets without outliers. A smaller SD indicates a more compact MC. Moreover, the Jaccard indices can be used to identify which of the clustering approaches is robust. Robustness depends on the clustering performance for all channel scenarios. A clustering approach is robust when it performs well for all channels.

Analysis of variance (ANOVA) is also used to determine the consistency of clustering performance. The one-way ANOVA of MATLAB (anova1) [23] is used to generate the box plot [24], as shown in Figure 3. The red mark indicates the median, the bottom edge of the blue box is the 25th percentile, and the top edge is the 75th percentile. The whiskers extend to the most extreme data points not considered outliers. The outliers are marked individually by the red '+' symbol. One-way ANOVA determines whether data from several groups have a common mean. The ANOVA1 tool gives the F-statistic probability value (*p*-value) and the box plots of the independent variable. It tests the hypothesis that the samples in the independent variable are drawn from populations with the same mean against the alternative hypothesis that the population means are not all the same. If the *p*-value is smaller than the significance level of 0.05, the test rejects the null hypothesis that all group means are equal and concludes that at least one group means differs from the others. Using box plots, two medians are significantly different at the 5% significance level if their intervals do not overlap.

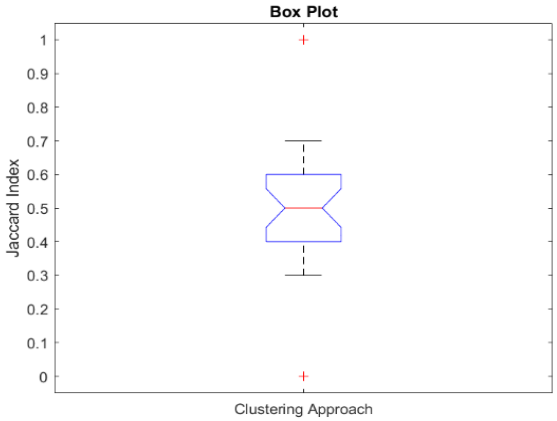

Figure 3. Box plot of Jaccard indices showing the performance of a clustering approach

3. RESULTS AND DISCUSSION

Simulations were done using a laptop with an Intel 11th-generation processor, 2.4 GHz speed, and 16 GB of RAM. Spyder of Anaconda 3 [25] was used for the outlier detection and removal, while MATLAB 2023a was used to cluster the datasets, compute the Jaccard index, and generate anova1 boxplots. The clustering accuracy of the datasets with outliers is discussed first, followed by outlier detection and removal. Furthermore, the clustering performance of the datasets without outliers is elaborated. Lastly, the clustering results of the datasets with and without outliers are compared.

3.1. Clustering of datasets with outliers

The four datasets were clustered using the MATLAB implementations of the five clustering approaches. The mean Jaccard indices of the 30 sheets of Excel data per channel scenario are shown in Table 2. The blank entries indicate that the computational requirements exceed the 16 GB RAM of the computer. The clustering performance for all channel scenarios is shown in Figure 4. The channel scenarios are arranged chronologically as presented in Table 2, with channel scenario 1 about indoor B1 LOS single link, channel scenario 2 indoor B2 LOS single link, and so on up to channel scenario 33 UMi B1 NLOS. The five clustering approaches work well with channel Scenarios 1 and 2, the two indoor scenarios of COST 2100, with K-medoids as the most accurate. This is due to the small number of MPCs per MC. The accuracy of the five clustering approaches ranges from 0 to 0.1 for channel scenarios 3 to 33. The low accuracy is due to the fact that there are more MPCs per MC for the scenarios. The clustering approaches have almost the same accuracies in the semi-urban scenarios of COST 2100 (channel scenarios 3 to 8). Also, nearly the same values are evident for IMT-2020 channel scenarios 9 to 19, except channel scenario 13, when SC has no data, and channel scenarios 15, 16, and 19, when both AC-Single and SC have no data.

Table 2. Mean Jaccard indices of datasets with outliers. Blank entries have higher computational memory requirements which exceed the 16 GB RAM of the computer

Channel model	Channel scenario	K-means	K-medoids	DBSCAN	AC-Single	SC
COST 2100	Indoor B1 LOS single link	0.5391	0.5660	0.2952	0.4141	0.2352
	Indoor B2 LOS single link	0.4323	0.4719	0.1754	0.4096	0.2246
	Semi-Urban B1 LOS single link	0.0228	0.0257	0.0219	0.0251	0.0241
	Semi-Urban B2 LOS single link	0.0242	0.0201	0.0215	0.0231	0.0253
	Semi-Urban B1 NLOS single link	0.0217	0.0218	0.0234	0.0226	0.0240
	Semi-Urban B2 NLOS single link	0.0217	0.0256	0.0227	0.0221	0.0224
	Semi-Urban B1 LOS multiple links	0.0104	0.0116	0.0106	0.0120	0.0149
	Semi-Urban B2 LOS multiple links	0.0133	0.0130	0.0112	0.0120	0.0136
IMT-2020	InH A LOS	0.0464	0.0487	0.0345	0.0361	0.0460
	InH A NLOS	0.0332	0.0333	0.0270	0.0274	0.0318
	RMa A LOS	0.0558	0.0566	0.0476	0.0476	0.0570
	RMA A NLOS	0.0530	0.0530	0.0526	0.0527	0.0529
	RMa A O2I	0.0524	0.0526	0.0526	0.0526	-
	UMa A LOS	0.0528	0.0531	0.0435	0.0436	0.0506
	UMa A NLOS	0.0320	0.0317	0.0256	-	-
	UMa A O2I	0.0473	0.0474	0.0435	-	-
	UMi A LOS	0.0500	0.0506	0.0435	0.0435	0.0499
	UMi A NLOS	0.0296	0.0298	0.0270	0.0270	0.0287
	UMi A O2I	0.0479	0.0478	0.0435	-	-
QuaDRiGa	BERLIN UMa LOS	0.0634	0.0625	0.0359	0.0345	0.0418
	BERLIN UMa NLOS	0.0366	0.0354	0.0308	0.0205	0.0251
	BERLIN UMi Campus LOS	0.0825	0.0815	0.0815	0.0436	0.0553
	BERLIN UMi Campus NLOS	0.0487	0.0487	0.0441	0.0258	0.0327
	BERLIN UMi Square LOS	0.0777	0.0772	0.0740	0.0436	0.0582
	BERLIN UMi Square NLOS	0.0467	0.0463	0.0383	0.0257	0.0322
	Industrial LOS	0.0272	0.0266	0.0283	0.0207	0.0241
	Industrial NLOS	0.0248	0.0240	0.0279	0.0198	0.0224
	Indoor A1 LOS	0.0756	0.0754	0.0782	0.0440	0.0563
	Indoor A1 NLOS	0.0608	0.0612	0.0541	0.0325	0.0452
WINNER II	UMa C2 LOS	0.1135	0.1117	0.0667	0.0668	0.0796
	UMa C2 NLOS	0.0442	0.0441	0.0344	0.0257	0.0302
	UMi B1 LOS	0.0927	0.0910	0.0864	0.0669	0.0769
	UMi B1 NLOS	0.0446	0.0444	0.0367	0.0324	0.0373

The industrial scenarios of QuaDRiGa (channel scenarios 26 and 27) have almost the same Jaccard indices of 0.02. However, the clustering approaches have varying indices for the rest of QuaDRiga (channel scenarios 20 to 25) and WINNER II (channel scenarios 28 to 33). K-medoids is the most consistent clustering approach, registering the highest clustering accuracy in most channel scenarios. On the other hand, AC-Single is the least consistent of the clustering approaches as it gives the least clustering accuracy in most channel scenarios.

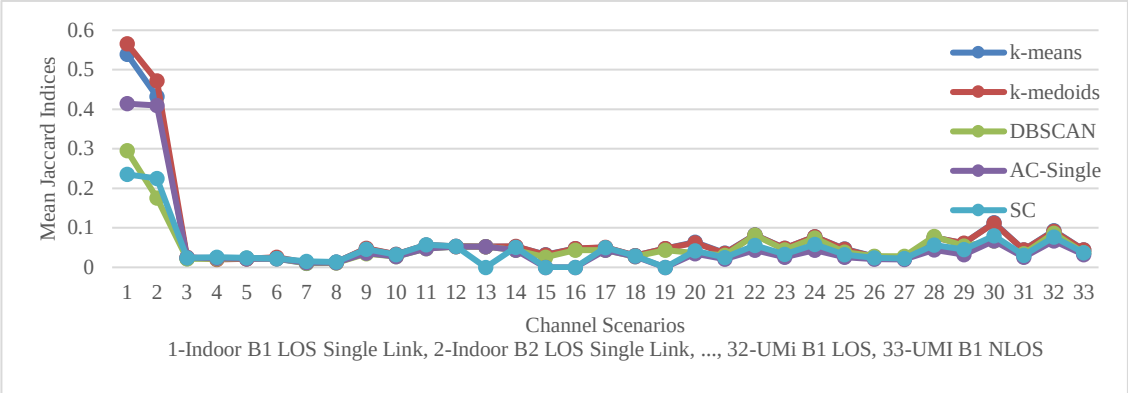

Figure 4. Clustering performance for all channel scenarios

3.2. Outlier detection and removal

The outliers were detected and removed using the mean-shift outlier detection. The method is implemented in Python using Spyder in Anaconda 3. The outliers are based on the outlier thresholds 2×SD, 2.5×SD, and 3×SD. A datapoint outside of the threshold is considered an outlier. The mean of its KNN replaces it. Table 3 gives the number of outliers for the 30 sheets of Excel data per channel scenario based on the outlier threshold. The 2×SD outlier threshold gives the most outliers since a smaller region is considered for normal datapoints. On the other hand, the 3×SD outlier threshold results in the least number of outliers due to a larger region being considered for normal datapoints.

Most outliers are generated by the IMT-2020 RMa A O2I channel scenario using the 2×SD outlier threshold, as shown in Table 3. This is due to a large propagation area with many interacting objects (IO), such as houses, buildings, and trees. Also, MPCs experience attenuation, scattering, diffraction, or reflection when traveling outdoors to indoors.

Table 3. Number of outliers for the 30 sheets of data per channel scenario based on the outlier score threshold

Channel model	Channel scenario	2×SD	2.5×SD	3×SD
COST 2100	Indoor B1 LOS single link	63	43	31
	Indoor B2 LOS single link	60	45	37
	Semi-Urban B1 LOS single link	1,072	571	319
	Semi-Urban B2 LOS single link	1,167	640	334
	Semi-Urban B1 NLOS single link	1,978	1,104	597
	Semi-Urban B2 NLOS single link	1,961	1,077	604
	Semi-Urban B1 LOS multiple links	2,217	1,378	852
	Semi-Urban B2 LOS multiple links	2,200	1,320	793
IMT-2020	InH A LOS	410	14	0
	InH A NLOS	923	36	0
	RMa A LOS	34,395	8,173	1,742
	RMA A NLOS	42,935	23,006	9,458
	RMa A O2I	92,906	10,480	0
	UMa A LOS	1,284	523	267
	UMa A NLOS	12,436	4,192	2,530
	UMa A O2I	16,973	8,669	3,971
	UMi A LOS	4,101	1,205	535
	UMi A NLOS	21,847	5,033	24
	UMi A O2I	806	291	100
QuaDRiGa	BERLIN UMa LOS	28,911	16,307	7,719
	BERLIN UMa NLOS	44,374	25,262	14,077
	BERLIN UMi Campus LOS	9,960	3,394	1,102
	BERLIN UMi Campus NLOS	19,097	7,927	2,847
	BERLIN UMi Square LOS	11,800	6,517	2,897
	BERLIN UMi Square NLOS	20,256	11,892	6,195
	Industrial LOS	10,738	6,666	4,188
	Industrial NLOS	11,199	6,920	4,232
WINNER II	Indoor A1 LOS	4,120	1,279	129
	Indoor A1 NLOS	6,643	3,715	2,052
	UMa C2 LOS	15,299	6,792	1,924
	UMa C2 NLOS	35,538	20,908	12,148
	UMi B1 LOS	8,464	5,352	3,188
	UMi B1 NLOS	11,479	6,773	4,268

Considering the 2×SD threshold, in general, the indoor scenarios (indoor B1/B2 LOS single link of COST 2100, InH A LOS/NLOS of IMT-2020, Industrial LOS/NLOS of QuaDRiGa, and indoor A1 LOS/NLOS of WINNER II) of the four channel models generate the least number of outliers relative to their non-indoor counterparts in the same channel model. Indoor scenarios have a smaller area for the propagation of signals, lesser IO, and a lesser number of MPCs and MCs. InH A LOS, InH A NLOS, and RMa A O2I have zero outliers using the 3×SD threshold because all the multipaths are within the radius for normal datapoints.

A new dataset without outliers was generated when the outliers were replaced by the means of their knn. The new dataset still consists of 30 Excel sheets of data per channel scenario, the same number of MPCs, and the same number of MCs. The outliers were highlighted in red and replaced by the mean of the knn. The new dataset is uploaded to the IEEE Dataport [26].

3.3. Clustering of datasets without outliers

The datasets without outliers were again clustered using the MATLAB implementation of the five clustering approaches. Table 4 shows the clustering accuracies using the 2×SD outlier threshold, Table 5 for the 2.5×SD threshold, and Table 6 for the 3×SD threshold. No general trend exists that outlier detection and removal increases clustering accuracy, nor does a smaller outlier threshold improve clustering performance. It shows that the outlier detection method is not accurate in detecting and removing the outliers of the clusters.

For K-means, the most significant increase in clustering accuracy is in the Semi-Urban B1 LOS multiple links 2×SD threshold with a value of 78.85%, while the highest decrease is in indoor B1 LOS single link 3×SD threshold with a value of 19.53%. For K-medoids, the most significant gain of 50.86% is in the Semi-Urban B1 LOS multiple links 3×SD threshold, while the most substantial drop of 11.28% is in the Industrial LOS 2×SD threshold. For DBSCAN, Semi-Urban B2 LOS multiple links 2×SD registers the highest increase of 57.14%, while indoor A1 LOS 3×SD gives the most significant decrease of 43.86%.

Table 4. Mean Jaccard indices of datasets without outliers with outlier score threshold 2×SD

Channel model	Channel scenario	K-means	K-medoids	DBSCAN	AC-Single	SC
COST 2100	Indoor B1 LOS single link	0.4644	0.6461	0.3491	0.5499	0.2731
	Indoor B2 LOS single link	0.4275	0.5672	0.2042	0.4424	0.2186
	Semi-Urban B1 LOS single link	0.0224	0.0239	0.0210	0.0267	0.0264
	Semi-Urban B2 LOS single link	0.0238	0.0231	0.0216	0.0249	0.0223
	Semi-Urban B1 NLOS single link	0.0231	0.0217	0.0237	0.0228	0.0249
	Semi-Urban B2 NLOS single link	0.0227	0.0240	0.0228	0.0232	0.0236
	Semi-Urban B1 LOS multiple links	0.0186	0.0162	0.0156	0.0186	0.0200
	Semi-Urban B2 LOS multiple links	0.0167	0.0183	0.0176	0.0179	0.0225
IMT-2020	InH A LOS	0.0472	0.0486	0.0345	0.0360	0.0456
	InH A NLOS	0.0337	0.0329	0.0270	0.0275	0.0322
	RMa A LOS	0.0566	0.0565	0.0476	0.0481	0.0607
	RMA A NLOS	0.0528	0.0527	0.0526	0.0527	0.0550
	RMa A O2I	0.0527	0.0529	0.0526	0.0528	-
	UMa A LOS	0.0531	0.0536	0.0435	0.0436	0.0502
	UMa A NLOS	0.0319	0.0318	0.0256	-	-
	UMa A O2I	0.0470	0.0472	0.0435	-	-
	UMi A LOS	0.0498	0.0500	0.0435	0.0437	0.0516
	UMi A NLOS	0.0292	0.0291	0.0270	0.0271	0.0312
	UMi A O2I	0.0478	0.0477	0.0435	-	-
QuaDRiGa	BERLIN UMa LOS	0.0631	0.0628	0.0422	0.0346	0.0463
	BERLIN UMa NLOS	0.0353	0.0357	0.0306	0.0205	0.0282
	BERLIN UMi Campus LOS	0.0823	0.0817	0.0809	0.0437	0.0615
	BERLIN UMi Campus NLOS	0.0484	0.0488	0.0440	0.0258	0.0364
	BERLIN UMi Square LOS	0.0774	0.0764	0.0728	0.0437	0.0554
	BERLIN UMi Square NLOS	0.0476	0.0470	0.0380	0.0258	0.0332
	Industrial LOS	0.0254	0.0236	0.0280	0.0206	0.0244
	Industrial NLOS	0.0234	0.0230	0.0278	0.0198	0.0241
WINNER II	Indoor A1 LOS	0.0741	0.0743	0.0777	0.0441	0.0575
	Indoor A1 NLOS	0.0606	0.0606	0.0531	0.0327	0.0452
	UMa C2 LOS	0.1118	0.1114	0.0982	0.0668	0.0859
	UMa C2 NLOS	0.0443	0.0444	0.0354	0.0257	0.0326
	UMi B1 LOS	0.0969	0.0955	0.0903	0.0669	0.0780
	UMi B1 NLOS	0.0412	0.0409	0.0363	0.0324	0.0387

Table 5. Mean Jaccard indices of datasets without outliers with outlier score threshold 2.5×SD

Channel model	Channel scenario	K-means	K-medoids	DBSCAN	AC-Single	SC
	Indoor B1 LOS single link	0.5317	0.6559	0.3556	0.5314	0.2409
	Indoor B2 LOS single link	0.4673	0.5530	0.2042	0.4550	0.2587
	Semi-Urban B1 LOS single link	0.0271	0.0249	0.0218	0.0250	0.0287
COST 2100	Semi-Urban B2 LOS single link	0.0206	0.0223	0.0217	0.0251	0.0255
	Semi-Urban B1 NLOS single link	0.0237	0.0195	0.0235	0.0225	0.0230
	Semi-Urban B2 NLOS single link	0.0210	0.0231	0.0226	0.0225	0.0249
	Semi-Urban B1 LOS multiple links	0.0153	0.0169	0.0153	0.0173	0.0211
	Semi-Urban B2 LOS multiple links	0.0210	0.0176	0.0173	0.0176	0.0228
	InH A LOS	0.0465	0.0476	0.0345	0.0362	0.0452
	InH A NLOS	0.0335	0.0338	0.0270	0.0273	0.0315
	RMa A LOS	0.0563	0.0560	0.0476	0.0477	0.0586
	RMA A NLOS	0.0527	0.0529	0.0526	0.0527	0.0542
	RMa A O2I	0.0525	0.0527	0.0526	0.0526	-
IMT-2020	UMa A LOS	0.0537	0.0534	0.0435	0.0436	0.0493
	UMa A NLOS	0.0320	0.0321	0.0256	-	-
	UMa A O2I	0.0469	0.0476	0.0435	-	-
	UMi A LOS	0.0498	0.0505	0.0435	0.0437	0.0511
	UMi A NLOS	0.0296	0.0294	0.0270	0.0271	0.0299
	UMi A O2I	0.0479	0.0483	0.0435	-	-
	BERLIN UMa LOS	0.0626	0.0633	0.0393	0.0346	0.0485
	BERLIN UMa NLOS	0.0356	0.0358	0.0308	0.0205	0.0262
QuaDRiGa	BERLIN UMi Campus LOS	0.0821	0.0809	0.0813	0.0437	0.0569
	BERLIN UMi Campus NLOS	0.0482	0.0484	0.0441	0.0258	0.0335
	BERLIN UMi Square LOS	0.0772	0.0771	0.0734	0.0437	0.0558
	BERLIN UMi Square NLOS	0.0472	0.0479	0.0381	0.0258	0.0327
	Industrial LOS	0.0253	0.0255	0.0282	0.0207	0.0240
	Industrial NLOS	0.0227	0.0227	0.0281	0.0199	0.0233
	Indoor A1 LOS	0.0728	0.0737	0.0780	0.0440	0.0590
	Indoor A1 NLOS	0.0611	0.0611	0.0535	0.0326	0.0447
WINNER II	UMa C2 LOS	0.1119	0.1112	0.0667	0.0668	0.0777
	UMa C2 NLOS	0.0443	0.0426	0.0356	0.0257	0.0312
	UMi B1 LOS	0.0976	0.0958	0.0914	0.0670	0.0777
	UMi B1 NLOS	0.0428	0.0421	0.0365	0.0324	0.0383

Table 6. Mean Jaccard indices of datasets without outliers with outlier score threshold 3×SD

Channel model	Channel scenario	K-means	K-medoids	DBSCAN	AC-Single	SC
	Indoor B1 LOS single link	0.4338	0.6347	0.3413	0.4866	0.2780
	Indoor B2 LOS single link	0.4154	0.5611	0.1720	0.4483	0.2141
	Semi-Urban B1 LOS single link	0.0225	0.0253	0.0217	0.0252	0.0261
COST 2100	Semi-Urban B2 LOS single link	0.0199	0.0201	0.0217	0.0240	0.0234
	Semi-Urban B1 NLOS single link	0.0229	0.0227	0.0236	0.0230	0.0244
	Semi-Urban B2 NLOS single link	0.0221	0.0240	0.0226	0.0228	0.0253
	Semi-Urban B1 LOS multiple links	0.0161	0.0175	0.0153	0.0170	0.0191
	Semi-Urban B2 LOS multiple links	0.0196	0.0189	0.0170	0.0173	0.0215
	InH A LOS	0.0464	0.0487	0.0345	0.0361	0.0460
	InH A NLOS	0.0332	0.0333	0.0270	0.0274	0.0318
	RMa A LOS	0.0566	0.0558	0.0476	0.0477	0.0576
	RMA A NLOS	0.0528	0.0528	0.0526	0.0527	0.0538
	RMa A O2I	0.0524	0.0526	0.0526	0.0526	-
IMT-2020	UMa A LOS	0.0531	0.0530	0.0435	0.0437	0.0497
	UMa A NLOS	0.0317	0.0319	0.0256	-	-
	UMa A O2I	0.0469	0.0479	0.0435	-	-
	UMi A LOS	0.0497	0.0502	0.0435	0.0436	0.0504
	UMi A NLOS	0.0297	0.0295	0.0270	0.0271	0.0288
	UMi A O2I	0.0468	0.0477	0.0435	-	-
	BERLIN UMa LOS	0.0638	0.0628	0.0361	0.0346	0.0457
	BERLIN UMa NLOS	0.0355	0.0350	0.0308	0.0205	0.0255
QuaDRiGa	BERLIN UMi Campus LOS	0.0817	0.0816	0.0815	0.0436	0.0556
	BERLIN UMi Campus NLOS	0.0482	0.0491	0.0441	0.0258	0.0322
	BERLIN UMi Square LOS	0.0766	0.0763	0.0739	0.0437	0.0575
	BERLIN UMi Square NLOS	0.0467	0.0467	0.0382	0.0258	0.0325
	Industrial LOS	0.0263	0.0256	0.0282	0.0207	0.0243
	Industrial NLOS	0.0245	0.0225	0.0284	0.0199	0.0235
	Indoor A1 LOS	0.0737	0.0730	0.0439	0.0439	0.0577
	Indoor A1 NLOS	0.0612	0.0609	0.0538	0.0325	0.0445
WINNER II	UMa C2 LOS	0.1129	0.1132	0.0667	0.0668	0.0808
	UMa C2 NLOS	0.0446	0.0437	0.0356	0.0257	0.0303
	UMi B1 LOS	0.0937	0.0933	0.0897	0.0669	0.0755
	UMi B1 NLOS	0.0438	0.0433	0.0366	0.0324	0.0380

For AC-Single, the most significant gain in clustering accuracy is registered by the Semi-Urban B1 LOS multiple links 2×SD threshold with a value of 55%. The Industrial LOS 2×SD threshold gives the most remarkable drop of 0.48%. For SC, the most significant increase of 67.65% is posted by the Semi-Urban B2 LOS multiple links 2.5×SD threshold, while the highest decrease is recorded by the Semi-Urban B2 LOS single link 2×SD threshold with a value of 11.86%. Since InH A LOS, InH A NLOS, and RMa A O2I have zero outliers in the 3×SD threshold, their clustering accuracies remain the same. Among the clustering approaches used, K-medoids is the most robust, registering the highest mean Jaccard indices for all channel scenarios with or without outliers and for all outlier thresholds.

The means of standard deviations of the clustering approaches using different outlier thresholds are shown in Table 7. The outlier threshold 2×SD has the least means for all clustering approaches. This indicates that the clusters are more compact due to the removal of the outliers. However, the other outlier thresholds have higher means for all clustering approaches. It shows that outliers were not removed correctly in these thresholds.

Table 7. Means of standard deviations

Outlier threshold	K-means	K-medoids	DBSCAN	AC-Single	SC
No Outlier	0.0241	0.0260	0.0171	0.0224	0.0172
2×SD	0.0209	0.0235	0.0166	0.0215	0.0155
2.5×SD	0.0306	0.0353	0.0239	0.0353	0.0242
3×SD	0.0335	0.0382	0.0220	0.0350	0.0205

The box plots of the mean Jaccard indices of the five clustering approaches are shown in Figure 5. For the Jaccard indices in clustering the datasets with no outliers, it is shown in Figure 5(a). The p-value is 0.9810. For the 2×SD outlier threshold, it is presented in Figure 5(b). The p-value is 0.9824. Figure 5(c) shows the boxplot of the 2.5×SD with a p-value of 0.9006. The box plot of the 3×SD outlier threshold is shown in Figure 5(d). The p-value is 0.7606. Since the p-values are all greater than 0.05, there is no significant difference in the mean Jaccard indices of the clustering approaches.

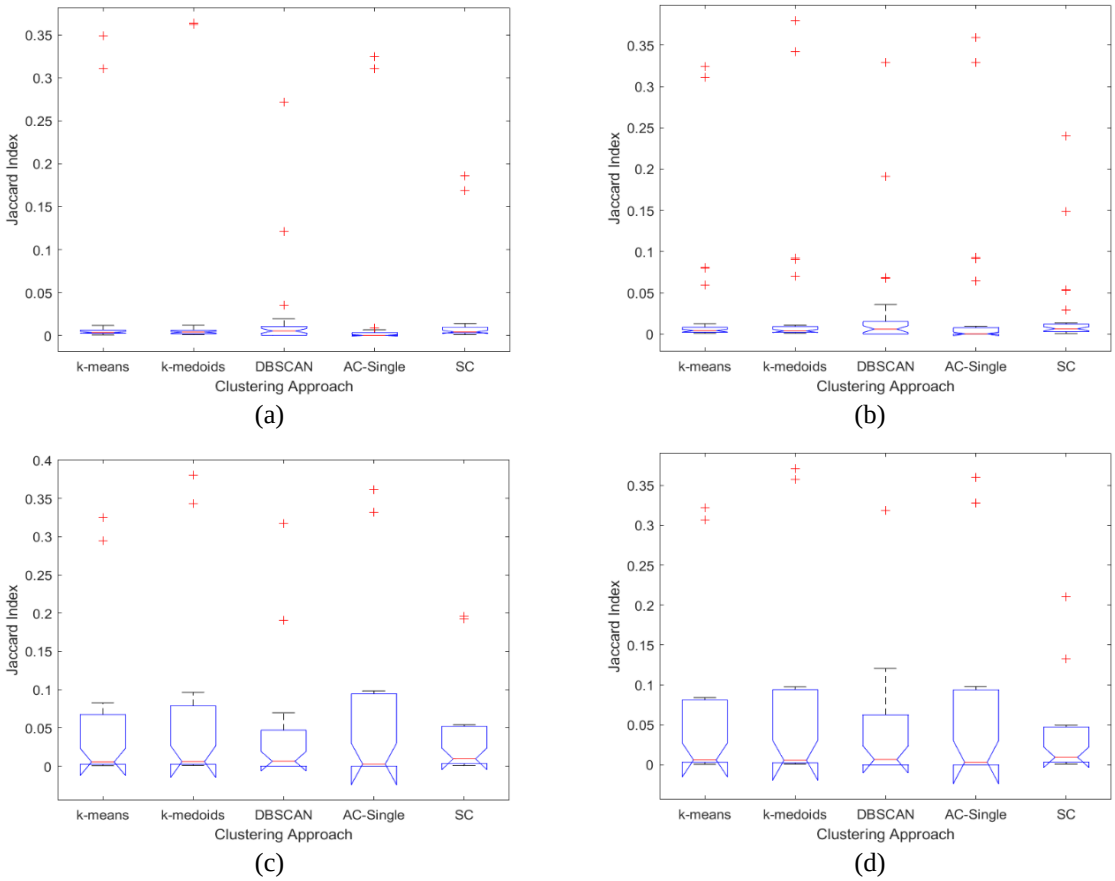

Figure 5. Box plots: (a) no outlier, (b) 2×SD, (c) 2.5×SD, and (d) 3×SD

4. CONCLUSION

Wireless multipath datasets generated by 5G channel models were clustered. The outliers were removed, and the datasets without outliers were again clustered. Results show that most outliers were identified using the 2×SD outlier threshold. Also, the threshold gave the highest clustering accuracy improvement. The clustering accuracy of K-means and AC-Single in Semi-Urban B1 LOS multiple links is increased by 78.85% and 55%, respectively, and DBSCAN in Semi-Urban B2 LOS multiple links by 57.14%. Furthermore, the study shows that outlier detection and removal work well with channel model datasets. The characterization of the multipath datasets was achieved by detecting their outliers. Finally, the results can be used to analyze the propagation characteristics of 5G further.

The implications of the study show that datasets generated by 5G channel models contain outliers and removing them greatly improves clustering accuracy. The study is limited to the channel models examined and cannot be used to generalize 5G systems. As for future work, the use of other outlier detection techniques can be considered to check the outliers of the 5G channel model datasets.

ACKNOWLEDGMENTS

Authors thank the Research Center for the Natural and Applied Sciences and the Office of the Vice-Rector for Research and Innovation of the University of Santo Tomas (UST) for supporting this work under the Memorandum of Agreement UST: SO21-00-FO04.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Jojo Blanza	✓	✓			✓	✓	✓		✓			✓	✓	✓
John Bernard Cipriano		✓	✓	✓		✓		✓		✓	✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in IEEE DataPort at http://doi.org/10.21227/q88v-xm96.

REFERENCES

[1] R. Verdone and A. Zanella, *Pervasive Mobile and Ambient Wireless Communications: COST Action 2100*. London: Spring, 2012. [Online]. Available: http://books.google.de/books?id=VVAmcZxn9akC
[2] M. Series, "Guidelines for evaluation of radio interface technologies for imt-2020," 2020.
[3] F. Burkhardt, S. Jaeckel, E. Eberlein, and R. Prieto-Cerdeira, "QuaDRiGa: A MIMO channel model for land mobile satellite," in *The 8th European Conference on Antennas and Propagation (EuCAP 2014)*, IEEE, Apr. 2014, pp. 1274–1278, doi: 10.1109/EuCAP.2014.6902008.
[4] P. Kyösti, "WINNER II Channel Models," *Radio Technologies and Concepts for IMT-Advanced*, pp. 39–92, 2009, doi: 10.1002/9780470748077.ch3.
[5] A. Teologo Jr., "Comparative Study of KPower Means, Ant Colony Optimization, Kernel Power Density-based Estimation, and Gaussian Mixture Model for Wireless Propagation Multipath Clustering," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 7, pp. 3942–3950, Jul. 2020, doi: 10.30534/ijeter/2020/164872020.

[6] V. Hautamäki, S. Cherednichenko, I. Kärkkäinen, T. Kinnunen, and P. Fränti, "Improving K-means by Outlier Removal BT - Image Analysis," H. Kalviainen, J. Parkkinen, and A. Kaarna, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 978–987.
[7] R. L. Adusumilli and M. Shashi, "Effective Outlier Detection using Scalable and Robust Clustering (SRC) Algorithm," in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, IEEE, Feb. 2022, pp. 1186–1193, doi: 10.1109/ICAIS53314.2022.9742956.
[8] Z. Han, B. Cheng, C. Wang, F. Feng, and Y. Wang, "An Improved Cuckoo Search Based K-means for Outlier Detection," in *2023 IEEE International Conference on Electrical, Automation and Computer Engineering (ICEACE)*, IEEE, Dec. 2023, pp. 447–450, doi: 10.1109/ICEACE60673.2023.10442856.
[9] J. Yang, L. Yang, W. Wang, and R. Pu, "An Outlier Detection Algorithm based on Local Density and Natural Neighbors," in *2023 2nd International Conference on Cloud Computing, Big Data Application and Software Engineering (CBASE)*, IEEE, Nov. 2023, pp. 51–56, doi: 10.1109/CBASE60015.2023.10439072.
[10] C. Wang and J. Ju, "Geodesic Affinity Propagation Clustering Based on Angle-Based Outlier Factor," *IEEE Access*, vol. 11, pp. 43619–43629, 2023, doi: 10.1109/ACCESS.2023.3271996.
[11] J. Yang, S. Rahardja, and P. Fränti, "Mean-shift outlier detection and filtering," *Pattern Recognition*, vol. 115, p. 107874, Jul. 2021, doi: 10.1016/j.patcog.2021.107874.
[12] J. Blanza, X. E. Cabasal, J. B. Cipriano, G. A. Guerrero, R. Y. Pescador, and E. V Rivera, "Indoor Wireless Multipaths Outlier Detection and Clustering," *Journal of Physics: Conference Series*, vol. 2356, no. 1, p. 012037, Oct. 2022, doi: 10.1088/1742-6596/2356/1/012037.
[13] Z. Li, L.-F. Cheong, S. Yang, and K.-C. Toh, "Simultaneous Clustering and Model Selection: Algorithm, Theory and Applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1964–1978, Aug. 2018, doi: 10.1109/TPAMI.2017.2739147.
[14] J. F. Blanza, A. T. Teologo, and L. Materum, "Datasets for Multipath Clustering at 285 MHz and 5.3 GHz Bands Based on COST 2100 MIMO Channel Model," in *2019 International Symposium on Multimedia and Communication Technology (ISMAC)*, IEEE, Aug. 2019, pp. 1–5, doi: 10.1109/ISMAC.2019.8836143.
[15] J. Blanza, E. Trinidad, and L. Materum, "Scedasticity descriptor of terrestrial wireless communications channels for multipath clustering datasets," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 13, no. 6, pp. 6547–6557, Dec. 2023, doi: 10.11591/ijece.v13i6.pp6547-6557.
[16] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982, doi: 10.1109/TIT.1982.1056489.
[17] L. Kaufman and P. J. Rousseeuw, "Partitioning Around Medoids (Program PAM)," in *Finding Groups in Data: An Introduction to Cluster Analysis*, Hoboken, New Jersey, United States: John Wiley & Sons, Inc., 1990, ch. 2, pp. 68–125, doi: 10.1002/9780470316801.ch2.
[18] R. Sibson, "SLINK: An optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16, no. 1, pp. 30–34, Jan. 1973, doi: 10.1093/comjnl/16.1.30.
[19] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.
[20] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., MIT Press, 2001.
[21] MATLAB, "Cluster Analysis and Anomaly Detection." Accessed: Mar. 25, 2025. [Online]. Available: https://www.mathworks.com/help/stats/cluster-analysis.html?s_tid=CRUX_lftnav
[22] J. Yang, S. Rahardja, and P. Fränti, "Outlier detection," in *Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing*, New York, NY, USA: ACM, Dec. 2019, pp. 1–6, doi: 10.1145/3371425.3371427.
[23] MATLAB, "anova1." Accessed: Mar. 25, 2025. [Online]. Available: https://www.mathworks.com/help/stats/anova1.html
[24] MATLAB, "boxplot." Accessed: Mar. 25, 2025. [Online]. Available: https://www.mathworks.com/help/stats/boxplot.html
[25] "Spyder," *MIT License*. Accessed: April 15, 2024. [Online]. Available: https://www.spyder-ide.org/
[26] J. Blanza and J. B. Cipriano, "Fifth-generation wireless channels outlier detection and clustering," *IEEE Dataport*, 2024, doi: 10.21227/q88v-xm96.

BIOGRAPHIES OF AUTHORS

Jojo Blanza received the B.Sc. degree in Electronics and Communications Engineering from the University of Santo Tomas, Manila, Philippines, in 1999, and the Master's degree and the Ph.D. degree in Electronics and Communications Engineering from the De La Salle University, Manila, Philippines, in 2005 and 2020, respectively. He is an Assistant Professor at the Department of Electronics Engineering, University of Santo Tomas. He is also a resident researcher at the Research Center for the Natural and Applied Sciences, University of Santo Tomas. His research interests include wireless communications, radio wave propagation, channel modeling, and multipath clustering. He can be contacted at email: jfblanza@ust.edu.ph.

John Bernard Cipriano received the B.Sc. degree in Electronics Engineering from the University of Santo Tomas, Manila, Philippines, in 2022. He is currently a Programmer at the Bank of the Philippine Islands. His current interest is data preprocessing. He can be contacted at email: jbcipriano6@gmail.com.