IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 22, 2021, accepted February 5, 2022, date of publication February 8, 2022, date of current version February 17, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3150006

Xplaces: Segmenting Physical Space Through Wi-Fi Traces Using Eigendecomposition and X-Means

SUNSIKA CHAIKUL[1], SANTI PHITHAKKITNUKOON[1,2], AND CARLO RATTI[3]

[1]Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Chiang Mai 50200, Thailand
[2]Excellence Center in Infrastructure Technology and Transportation Engineering (ExCITE), Chiang Mai University, Chiang Mai 50200, Thailand
[3]SENSEable City Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Santi Phithakkitnukoon (santi@eng.cmu.ac.th)

ABSTRACT This study makes use of Wi-Fi connectivity data to understand how physical spaces are utilized and how it can be segmented, from which the insight gained can facilitate spatial planning and design. To carry out this study, we used a Wi-Fi connectivity data collected from a university network of 291,124 devices from 2,980 access points located across three campuses. For space segmentation, we've defined three features that characterize space utilization: crowdedness, mobility, and connectivity entropy. We've developed a new method called Xplaces that employs PCA to reduce high dimensionality of the features, eigendecomposition to extract behavioral signatures of the access points, and X-means to cluster access points without predefined number of clusters. Silhouette value was used to measure how well clusters were formed for our evaluation. Our method outperforms the state-of-the-art model i.e., eigenplaces. Our further investigation on the impact of area usage temporality on space segmentation shows that the Xplaces performs better with specific features for different temporal observation windows. For example, Xplaces works well with the crowdedness feature for the weekend's space segmentation. A set of recommended features for different temporal windows is thus also part of our study's contributions in addition to the development of the Xplaces.

INDEX TERMS Urban informatics, space segmentation, opportunistic sensing, Wi-Fi data analysis.

I. INTRODUCTION

Today, cities are growing at unprecedented rates [1]. Their forms, functions, and structures are being vastly transformed more rapidly than ever before. Urbanization brings with it other challenges as well. Many urban environments have been devastated by urbanization with increasing issues in pollution and transportation [2]. As urbanization accelerates across the globe, many cities have been preparing to be a smart city – urban area that is information and communication technologies (ICT)-driven to become 'smart' in servicing citizens in a sustainable way with better informed decisions and mechanisms. Several urban areas have started their smart city journeys by instrumenting their areas with sensing technologies such as CCTV, motion sensors, GPS units, and wireless sensing network to 'actively' sense information about their citizen behaviors, because one of the keys for making informed decision is data. Therefore, to become a smart city environment, data concerning citizens must be collected upon which the right innovations and insightful decisions as well as policies can be made.

Besides the aforementioned active sensing mechanism, another approach is called opportunistic sensing by which the data is collected for one purpose by it creates 'opportunity' for another purpose. For example, Wi-Fi connectivity data is originally collected for network activity monitoring, performance evaluation, and billing purposes, but it can be opportunistically analyzed to understand area crowdedness [3] and mobility [4], [5]. The opportunistic sensing has the edge over its counterpart approach as the data can be collected in a large scale with no or minimal user awareness or interaction with sensing activity. In addition, for the most part, it does not violate user privacy as most of the opportunistic sensed data is anonymized and hence identifying the individuals is not possible. Unlike location-aware sensors such as GPS tracking units that can track and collect a fine-grained trajectory

The associate editor coordinating the review of this manuscript and approving it for publication was Chenshu Wu.

data [6], however the privacy issues and regulations, e.g., EU GDPR (general data protection regulation), have largely limited this type of detailed mobility data to be available for a large-scale analysis. Recent attempts have produced data that are limited to specific type of tracked individuals, such as university students [7] and customers of a particular service provider where the data was obtained in exchange of some incentives [8], and urban cyclists [9].

Wi-Fi is a family of wireless network protocol IEEE 802.11 [10], which allows electronic devices such as smartphones, desktop computers, laptops, and tablets to ex-change data or connect to the internet using a wireless network, which are widely used in both public and private places. As several places provide Wi-Fi connection for free, people use it to connect to the internet. Collectively, these connectivity logs constitute a large-scale behavioral data which can be used to better understand human behavior in various perspectives.

Wi-Fi data has been used to analyze interesting aspects of its user behavior that are useful for built-environment design and planning as well as location-based services. For example, Calabrese *et al.* [11] examined Wi-Fi connectivity pattern on a campus to show its correlation to physical environment which was then used to cluster space according to their usage. Sevtsuk *et al.* [12] analyzed logs of Wi-Fi usage to understand how people use space, which reflected on occupancy and movements of its users. Occupancy detected from the Wi-Fi connectivity was shown as an advantageous alternative with a higher level of accuracy and a much lower cost compared to the use of CO2 sensors in the study done by Ouf *et al.* [13]. An analysis of campus Wi-Fi logs by Kim and Kotz [14] shows that influx and outflux of users between access points (APs) had a periodic pattern, which was then be used to model movements in terms of arrival rate and distribution that was closely related to the non-homogeneous Poisson processes. Although the AP usage was shown to be periodic, this behavior seemed to be independent of their geographical locations but may depend on the relative locations of nearby APs as observed in another study by Kim and Kotz [15]. At the individual level, Kang *et al.* [16] developed a method to detect the user's significant places from Wi-Fi traces based on their time-based clustering approach and discussed that the detected places information could be useful for location-aware services.

Motivated by the work of Calabrese *et al.* [11], this study aims at utilizing the Wi-Fi connectivity data to understand people's behaviors and how physical spaces are used as reflected by the Wi-Fi traces, so that more informed decisions can be made upon insights gained from the analysis, especially concerning spatial planning and design. As places and buildings are built for different purposes, the Wi-Fi connectivity of different places can reflect on how places are utilized and how behaviors of people are shaped by the physical environment.

II. METHODOLOGY

To carry out this study, we used a Wi-Fi connectivity data collected from the users of a campus Wi-Fi network provided by Chiang Mai University (CMU). There was a total of 2,980 access points (APs) across three campuses; Suan Sak, Suan Dok and Mae Hia, covering a combined area of 6.88 km^2, as shown in Fig. 1. Suan Sak is the main campus occupying 2.93 km^2 site that includes the university's administrative center, the science, engineering, humanities, and social science faculties, political science and public administration, law, the graduate school, all of the campus resource facilities and services and major sports facilities. Suan Dok campus is the health science complex occupying 0.45 km^2 site that includes faculties of medicine, associated medical sciences, nursing, dentistry, pharmacy, and university hospital, known locally as Suan Dok, the largest teaching hospital in northern Thailand. Mae Hia campus is about 5 km south of the main campus that occupies 3.50 km^2 site, which houses the faculties of veterinary medicine and agro-industry.

TABLE 1. Distribution of APs across six categories.

AP category	Suan Sak	Suan Dok	Mae Hia	Total
Residence	556	284	27	867
Academic building	731	399	74	1,204
Administrative building	76	40	0	116
Service center	311	27	9	347
Research institute	32	25	12	69
Other	61	316	0	377

Each data record consisted of the Wi-Fi network connectivity information including the connected device ID (hashed media access control (MAC) address), received signal strength indicator (RSSI) between the AP and the connected device, MAC address of the corresponding AP, and timestamp. Data sampling rate was 5 minutes. Logs were collected from January 9th – February 3rd, 2020, which included 133,754,260 records of connectivity by 291,124 unique device IDs. Geolocation of each AP was given in a separate lookup table based on the AP's ID (i.e., MAC address). Geographic coordinates (latitude and longitude) were provided with seven decimal points or about 1-cm precision level. APs were labeled into six categories according to their locations; residence, academic building, administrative building, service center, research institute, and other. The APs classified as 'other' were those located in areas that did not belong to any other five categories, such as museum, hall, and convention center. Distribution of APs in each of the six categories is shown in Table 1. Academic building has largest number of APs followed by the residence e.g., student dormitory, faculty and staff apartment, and university guest house, while the research institute has the fewest APs. This distribution is very much in line with the built environment of the university.

Our goal was to understand how physical spaces on campus were used differently through digital traces i.e., Wi-Fi connectivity, so that it can be used to segment the space from which the insight gained will facilitate spatial planning and design. To do so, we explored potential features, which could be extracted from the Wi-Fi connectivity data that contribute to characterization of space utilization profile. The overview of our methodology namely *Xplaces* is shown in Fig. 2, which includes data preprocessing, feature extraction, principal component analysis (PCA), eigendecomposion, and X-means clustering, to produce a set of clustered APs from the raw Wi-Fi connectivity data (i.e., AP usage logs).

Data preprocessing step mainly dealt with noise removal and AP labeling. Data with incorrect or missing AP locations were considered a noise and removed from the preprocessed dataset. Incorrect locations were those geographically positioned outside the campuses. We developed a simple tool for this specific preprocessing task with which AP locations were plotted on a map and those located off campus were removed. Since the raw data did not contain a complete information about the AP's belonging buildings or faculties, so we needed to geographically plot them on the map with our developed tool and labelled each AP according to their location to six categories (as listed in Table 1) for future reference. Therefore, a set of the preprocessed data of each i^{th} AP, denoted by D_i, for our further analysis can be defined as follow.

$$D_i = \Big\{id, lat, lon, \{d_j \left(timestamp, \left\{Dev_j^i\right\}\right) \times |j = 1, 2, 3, \ldots, z_i\}\Big\}, \quad (1)$$

where id is the AP's ID, lat and lon are the latitude and longitude coordinates of the AP, d_j is the connectivity log of j^{th} timestamp, Dev_j^i is the set of device IDs connected to the i^{th} AP at j^{th} timestamp, and z_i is the total number of unique timestamps.

According to the literature, utilization of physical space can be characterized by the density of people spending time engaged in activities within the space that produces some degrees of crowdedness [17], [18], and the movement of people within the space that creates dynamism and forms some levels of mobility [19], [20]. With our Wi-Fi data, we defined the degree of crowdedness for each AP as the maximum number of unique device IDs connected simultaneously during a time period T. Suppose that C_i is a set of crowdedness values of i^{th} AP over multiple time periods, C_i can be defined as follows.

$$C_i = \{c_1, c_2, c_3, \ldots, c_t, \ldots, c_N\}, \quad (2)$$

where c_t is the crowdedness of t^{th} time period and N is the total number of periods. In our analysis, with T set to 15 minutes, an average value of crowdedness was calculated for each quarter of the hour, for each hour of the day, and for each day of the week. Let C_i' denote a set of these average values that characterizes the crowdedness characteristics of the AP, as follows.

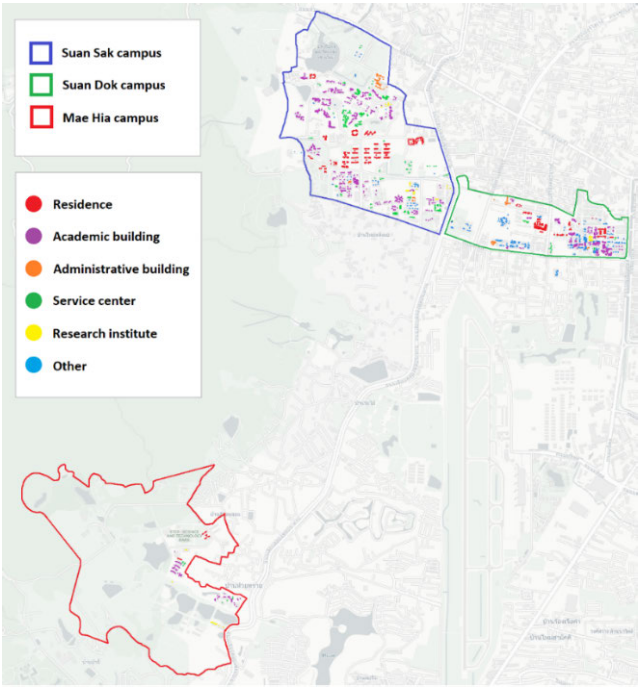

FIGURE 1. Locations of 2,987 Wi-Fi access points across the Chiang Mai University's three campuses of considered in this study; Suan Sak, Suan Dok, and Mae Hia. Geographic coordinates of this map's upper right and lower left corners are 18.811077, 98.978503 and 18.751595, 98.920256, respectively.

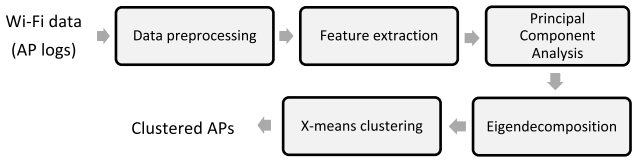

FIGURE 2. Overview of our proposed methodology, Xplaces.

$$C'_i = \{c'(q, h, d) | q \in Quarters, h \in Hours, d \in Days\}, \quad (3)$$

where $c'(q, h, d)$ is the average crowdedness during quarter q of hour h of day d, where $Quarters = \{1, 2, 3, 4\}$, $Hours = \{0, 1, 2, 3, \ldots, 23\}$, and $Days = \{\text{Monday, Tuesday, Wednesday, } \ldots, \text{Sunday}\}$. Thus, there are $4 \times 7 \times 24 = 672$ members of set C'_i from $c'(1, 0, Monday)$ to $c'(4, 23, Sunday)$. In the other words, the considered crowdedness feature has 672 dimensions.

Along the same lines, the mobility level was defined for each AP as the total number of connections and disconnections of device IDs during a time period T, as occurrences of AP connection and disconnection were considered a proxy for people moving in and out from the AP, hence it indicates mobility. Given that M_i is a set of mobility values of i^{th} AP over multiple time periods, M_i can be defined as follows.

$$M_i = \{m_1, m_2, m_3, \ldots, m_t, \ldots, m_N\}, \quad (4)$$

where m_t is the mobility of t^{th} period and N is the total number of periods. Similar to the crowdedness feature, with T

set to 15 minutes, the mobility feature was a set of the average mobility values over each quarter of the hour, each hour of the day, and each day of the week, as follows.

$$M'_i = \{m'(q, h, d) | q \in Quarters, h \in Hours, d \in Days\}, \tag{5}$$

where $m'(q, h, d)$ is the average mobility during the quarter q of hour h of day d, where $Quarters = \{1, 2, 3, 4\}$, $Hours = \{0, 1, 2, 3, \ldots, 23\}$, and $Days = \{\text{Monday, Tuesday, Wednesday, } \ldots, \text{ Sunday}\}$. Thus, there are $4 \times 7 \times 24 = 672$ members of set M'_i from $m'(1, 0, Monday)$ to $m'(4, 23, Sunday)$. So, the mobility feature of each AP has 672 dimensions.

Yet individual human behavior can appear almost random, typically there are repeating and easily identifiable routines. Collectively, these patterns become more apparent when contextualized temporally, and there are degrees of randomness associated with them. Shannon's entropy has been used to measure these degrees of uncertainty or randomness in human behavioral patterns [21], which is applicable in our case of Wi-Fi connectivity behavior. So, in addition to crowdedness and mobility, connectivity entropy was also extracted from the data and considered as another feature that characterizes physical space. Entropy was calculated to measure the degree of randomness associated with connectivity during each quarter of the hour of each hour of the day, and each day of the week. So, a set of connectivity entropy values (H'_i) of i^{th} AP can be defined as follows.

$$H'_i = \{H_i(X(q, h, d)) | q \in Quarters, h \in Hours, d \in Days\}, \tag{6}$$

where $H_i(X(q, h))$ can be calculated based on the Shannon's entropy [22] as following.

$$H_i(X(q, h, d)) = -\sum_{k=1}^{M} P\left(x_k\left(q, h, d\right)\right) \times \log_2 P\left(x_k(q, h, d)\right), \tag{7}$$

where M is the total number of connections occurred in the data during the quarter q (where $q = 1, 2, 3, 4$) of hour h (where $h = 0, 1, 2, 3, \ldots, 23$) of day d (where $d =$ Monday, Tuesday, Wednesday, $\ldots$, Sunday), $P\left(x_k(q, h, d)\right)$ is the probability of the connectivity k or $x_k(q, h, d)$, which is calculated as $x_k(q, h, d) / \sum_{k=1}^{M} x_k(q, h, d)$. For each AP, its connectivity was measured by the total number of connections made by any device IDs. Hence, the connectivity entropy feature has $4 \times 7 \times 24 = 672$ dimensions.

Examples of crowdedness, mobility, and connectivity entropy values of APs that were located in a female dorm, the University's main lecture building, and the University's main library, which were labelled as residence, academic building, and service center, respectively, are shown in Fig. 3. Intuitively, the crowdedness values are low during daytime but high at night at the dorm, while very low on Wednesdays as there were a very few lectures held on Wednesday at the lecture building and nearly no crowd at all on weekend when there was no lecture. Library, on the other hand, draws quite a consistent crowd throughout the week except for Sunday. Mobility values are naturally relative to the crowdedness and offers some insight on how much those crowds in motion. Connectivity entropy values at the dorm tend to be lower or in the other words there were more consistent connectivity on Tuesdays and Wednesdays, which may be due to the no-class-on-Wednesday schedule of most students who lived in the dorm. On other days of the week at the dorm, the connectivity is rather more random.

With the crowdedness, mobility, and connectivity entropy used as the features into the next process of the Xplaces, the PCA [23] was applied to reduce their high dimensionality by projecting the original high-dimensional feature vector onto the principal axes from which a lower dimensional feature vector called principal components are retained. Scree plot was used in deciding the number of principal components to retain based on the *elbow* criterion [24], which suggests to retain the principal components whose variance explained seem to level off from the point of the elbow of the graph. Each feature as well as their combination (i.e., $672 \times 3 = 2{,}016$ dimensions) were processed by the PCA from which their scree plots are shown in Fig. 4. There were 2, 2, 2, and 3 principal components retained for the crowdedness, mobility, connectivity entropy, and combined features, respectively. Figure 5 shows the principal components that capture the features' most significant aspects. Daily cycle is observed for each of the feature's principal components. For crowdedness and mobility, their first principal components show a daily cycle that peaks around the middle of the day, while the second principal components depict a cycle that rises later in the day (late afternoon into evening hours). On the other way, the entropy's second principal component shows the earlier peak around morning to noon, while the first principal component shows a daily cycle that rises in the afternoon.

Adapting a technique used in linear algebra, we applied eigendecomposition [25] to extract a behavioral signature of each AP based on its connectivity features as a time series whose dimensions were reduced by minimizing its information loss while maximizing its variance in forms of the principal components. We then represented these feature components of an AP over time as a vector and assembled the feature components from all APs into a covariance matrix. By applying the eigendecomposition, we factorized the matrix as a sum of the matrix's eigenvectors ($\boldsymbol{v_k}$) by a coefficient ($\boldsymbol{u_{i,k}}$) particular to that AP, which makes up its signature ($\boldsymbol{S_i}$) i.e.,

$$\boldsymbol{S_i} = \boldsymbol{u_{i,1}v_1} + \boldsymbol{u_{i,2}v_2} + \cdots + \boldsymbol{u_{i,w}v_w} \tag{8}$$

Hence, the signature $\boldsymbol{S_j}$ of AP j would be described by the same set of eigenvectors ($\boldsymbol{v_k}$'s) but with different coefficients ($\boldsymbol{u_{i,k}}$'s). These derived AP signatures can then be exploited to understand how physical spaces are used as reflected by the Wi-Fi traces.

Since all AP signatures are derived with the same principal, characterized by behavioral vector set, and so they're

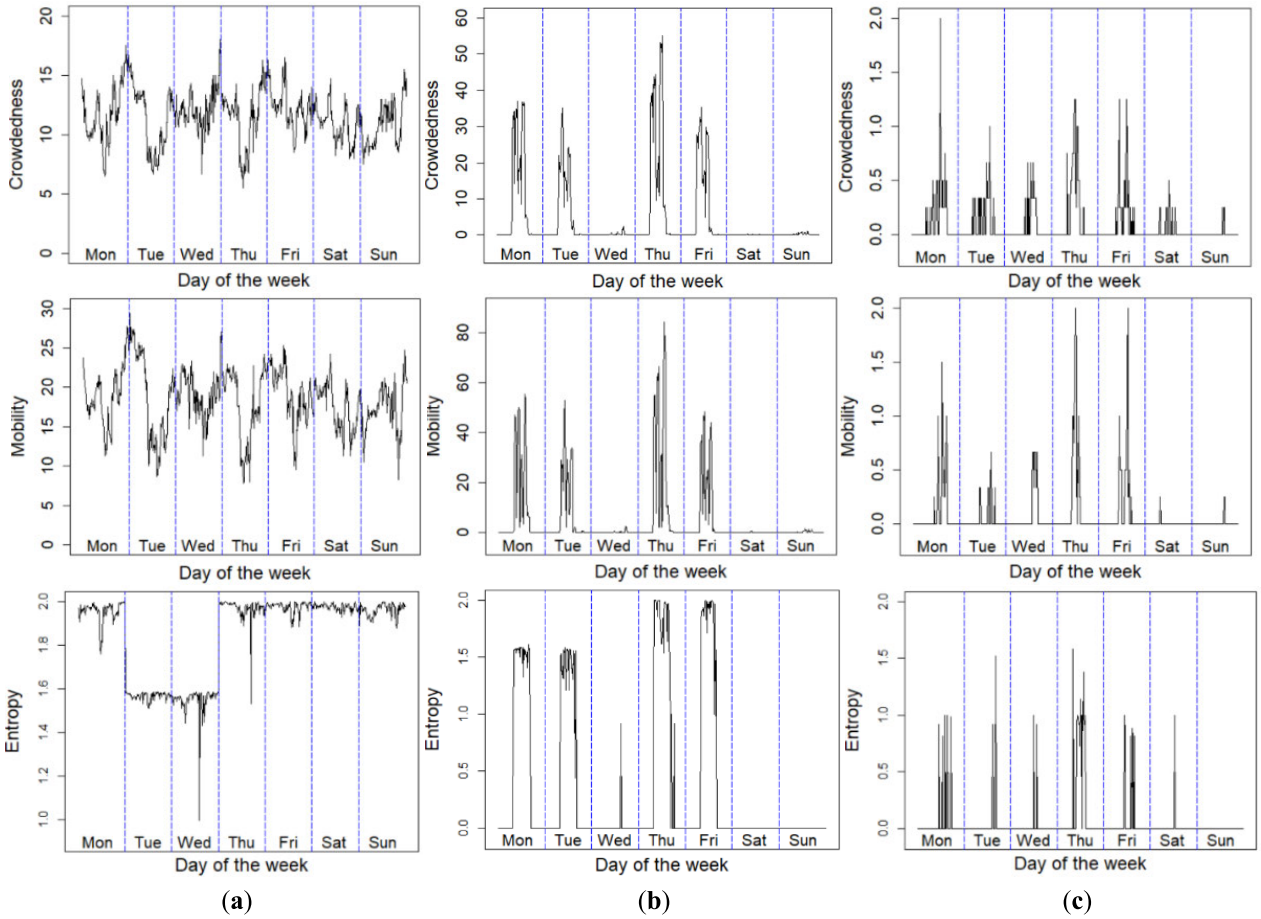

FIGURE 3. Crowdedness, mobility, and connectivity entropy values of APs located in (a) a female dorm (residence), (b) a main lecture building (academic), and (c) the main library (service center).

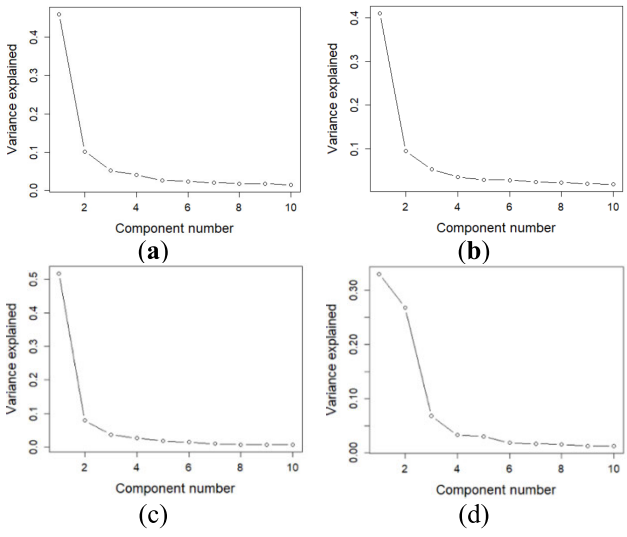

FIGURE 4. Scree plots of (a) crowdedness, (b) mobility, (c) entropy, and (d) their combination from which 2, 2, 2, and 3 principal components were retained based on the elbow criterion, respectively.

quantitatively comparable. We can cluster APs based solely on their coefficients, which are simple scalar, and then examine their grouping across the campuses. A simple clustering algorithm like k-means [26] cannot be utilized in our case here as the number of clusters, or k, must be known beforehand for the clustering to begin. In our scenario, the clusters should emerge inherently according to the similarities and differences in space utilization characterized by AP connectivity patterns in the vicinity. To address the k-means' shortcoming, we applied a technique called X-means clustering [27], which is an approach that can cluster data points without a predefined number of clusters by estimating the value k by making local decisions about which subset of the current centroids should be split to properly fit the data. Its splitting decision is based on the Bayesian Information Criteria (BIC) [28], so the key is an optimization the BIC value.

The X-means approach sets k to two (to initially create two clusters). The cluster centers are updated based on the renewed cluster mean, and each data point is repeatedly assigned to the nearest centroid. After then, the data points are redistributed and the cluster centers are updated once again. This procedure keeps repeating itself. Each centroid is split into two children for each value of k, who are then transported in opposing directions along a randomly determined

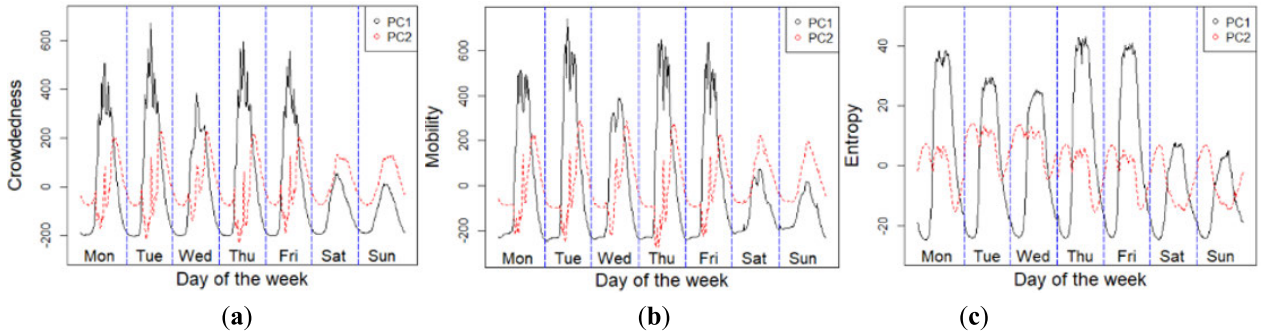

FIGURE 5. The primary principal components that capture the (a) crowdedness, (b) mobility, and (c) connectivity entropy features' most significant aspects.

vector for a distance proportional to the region's size. For each pair of children, a k-means algorithm is run locally in each parent region with k = 2. Within the parent region, data points are grouped to the children. The split decision is then made on the basis of the BIC value at the local level. Iteratively, the method continues until the global BIC value is optimized.

III. RESULTS

We implemented our Xplaces method using crowdedness, mobility, connectivity entropy, and their combination as four different sets of features in our experiment, as well as benchmarked our result against Calabrese *et al.*'s *eigenplaces* method [11], which is based on the number of connections over 15-minute interval as its feature, as well as eigendecomposition and *k*-means for its clustering process. For evaluation, the Silhouette value [29] was used to measure how well clusters were formed. Technically, it measures how similar a data point is to its own cluster compared to other clusters. The Silhouette value ranges from -1 to 1, where a greater value implies higher degree of similarity of data points to their own clusters than other ones.

As a result, silhouette plots of Xplaces based on crowdedness (C), mobility (M), connectivity entropy (E), and their combination (CME), as well as eigenplaces are shown in Fig. 6, where the overall average silhouette values are listed on Table 2. The Xplaces with E as its feature has the highest average silhouette value of 0.65, followed by M (0.64), CME (0.63), C (0.53), and lastly the eigenplaces method (0.33). This suggests that in general our Xplaces method performs better than the eigenplaces, which is the state of the art. Moreover, the Xplaces performs best with using connectivity entropy as its feature.

With the approach of Xplaces and the connectivity entropy employed as its feature, Figs. 7 – 9 show geolocations of the clustered APs as well as their corresponding building types across the three campuses. Interestingly, these results show that without seeking any reference data, our approach can accentuate important information that characterizes space utilization across the areas. At a glance, we can clearly see that most APs used in the residential areas are well clustered together, while the APs located in academic buildings tend to be also clustered together.

TABLE 2. Overall average silhouette values.

Method	Average silhouette value
Xplaces with crowdedness (C)	0.53
Xplaces with mobility (M)	0.64
Xplaces with entropy (E)	0.65
Xplaces with CME	0.63
Eigenplace	0.34

As spaces on the campuses were already designated to serve different purposes, the behavioral characteristics observed from the data do indeed reflect their real-word functionalities. Despite the difference in building types, some spaces in various buildings may be used similarly. For example, cafeterias that are located in academic buildings, dormitories, service center, and research institutes can draw comparable space usage patterns in those areas due to our regular food eating routines. Therefore, there is also a mixture of different clustered APs types within the same buildings.

Nonetheless, it is still intriguing to see how these user-generated maps reasonably resemble the real-word space utilization. As we observed that most residential APs were seemingly clustered well together, and likewise for the academic buildings, it is probably due to the fact that activities carried out in these two building types are relatively distinctive, as oppose to academic building versus research institute, for instance.

To further investigate how the APs were clustered from the perspective of building type and vice versa, Figs. 10 and 11 illustrate the cluster distribution and composition in terms of percentages of clustered APs based on the labelled building types, respectively. The precise numerical values of results depicted in Figs. 10 and 11 are listed on Table 3 and IV, respectively. APs labelled as residence were mostly clustered well within its own labelled building type with 72.43 percent or 628 APs of the cluster 4's members. A combine of clusters 3 and 4 accounts from over 93% of the residence APs. Intuitively, this is likely due to the nature of how people use

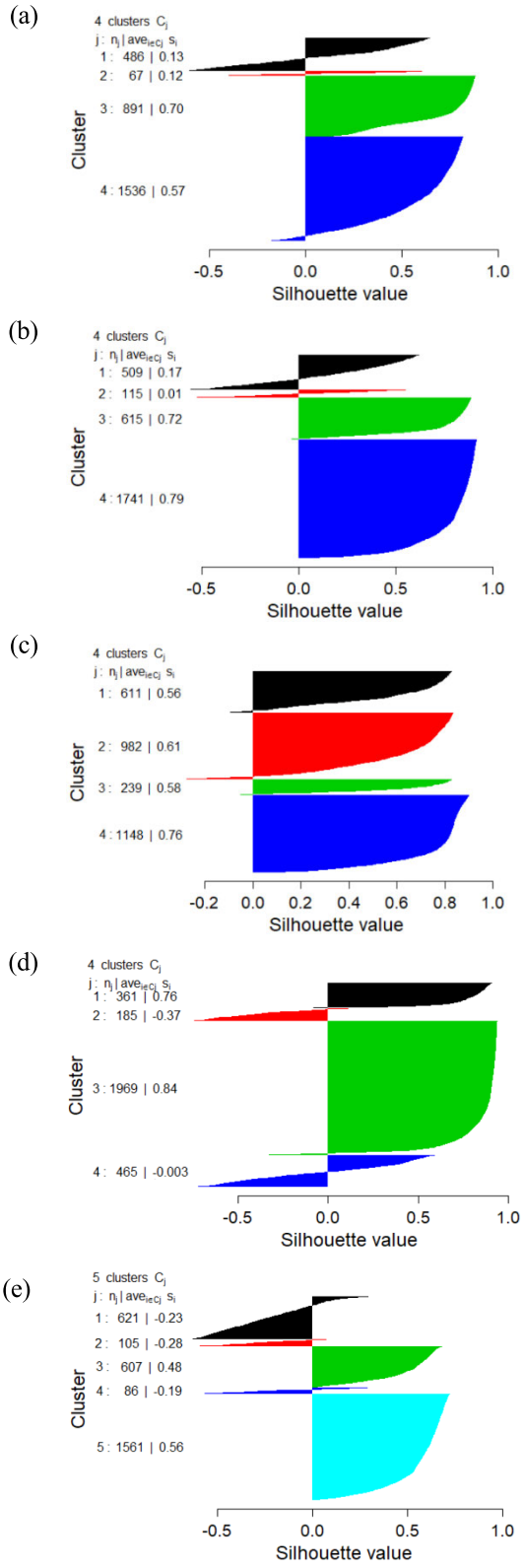

FIGURE 6. Silhouette plots of clustered APs based on the Xplaces with (a) crowdedness, (b) mobility, (c) connectivity entropy, (d) combined features, and (e) eigenplaces.

residential area, which is highly distinct from other building types. There were slightly over half of the academic

TABLE 3. Percentages and amount of clustered APs for each labelled building type.

Building Types	Cluster1	Cluster2	Cluster3	Cluster4
Residence	5.19% (45)	1.15% (10)	21.22% (184)	72.43% (628)
Academic building	26.91% (324)	50.42% (607)	2.82% (34)	19.85% (239)
Administrative building	12.93% (15)	54.31% (63)	2.59% (3)	30.17% (35)
Service center	38.33% (133)	29.11% (101)	1.73% (6)	30.84% (107)
Research institute	13.04% (9)	62.32% (43)	0.00% (0)	24.64% (17)
Other	22.55% (85)	41.91% (158)	3.18% (12)	32.36% (122)

TABLE 4. Percentages and amount of labelled building type APs for each cluster.

Building Types	Cluster1	Cluster2	Cluster3	Cluster4
Residence	7.36% (45)	1.02% (10)	76.99% (184)	54.70% (628)
Academic building	53.03% (324)	61.81% (607)	14.23% (34)	20.82% (239)
Administrative building	2.45% (15)	6.42% (63)	1.26% (3)	3.05% (35)
Service center	21.77% (133)	10.29% (101)	2.51% (6)	9.32% (107)
Research institute	1.47% (9)	4.38% (43)	0.00% (0)	1.48% (17)
Other	13.91% (85)	16.09% (158)	5.02% (12)	10.63% (122)

building APs clustered together. Similarly, over 50 percent of administrative building APs were clustered together. Likewise, there were over 60 percent of the research institute APs clustered together. Overall, there were three large groups: residence that was made up of clusters 3 and 4 (93.65 percent, 812 APs); academic building that was composed of clusters 1 and 2 (77.33 percent, 934 APs); and administrative building that consisted of clusters 2 and 4 (84.48 percent, 98 APs). When considered each cluster based on their building types (Fig. 11), the clusters 1 and 2 were dominated by academic building APs, while the clusters 3 and 4 were dominated residence.

As space usage pattern varies with time due to our regular working/studying schedules, which creates diurnal variations of population or rhythms [30], there might be a temporal variation in space utilization that affects its pattern and hence segmentation. To investigate the impact of this temporality of area usage on our space segmentation, we reran our experiment on data selected from particular times of observation window. Three different observation window schemes were considered, which includes 3-hour period, 6-hour period, and day of the week. Five different approaches for space segmentation were implemented, including Xplaces with C, M, E, CME, and eigenplaces for which an average silhouette value was calculated as its evaluation.

For the 3-hour period scheme, we intuitively chose the first time period to start from midnight (00:00) and hence

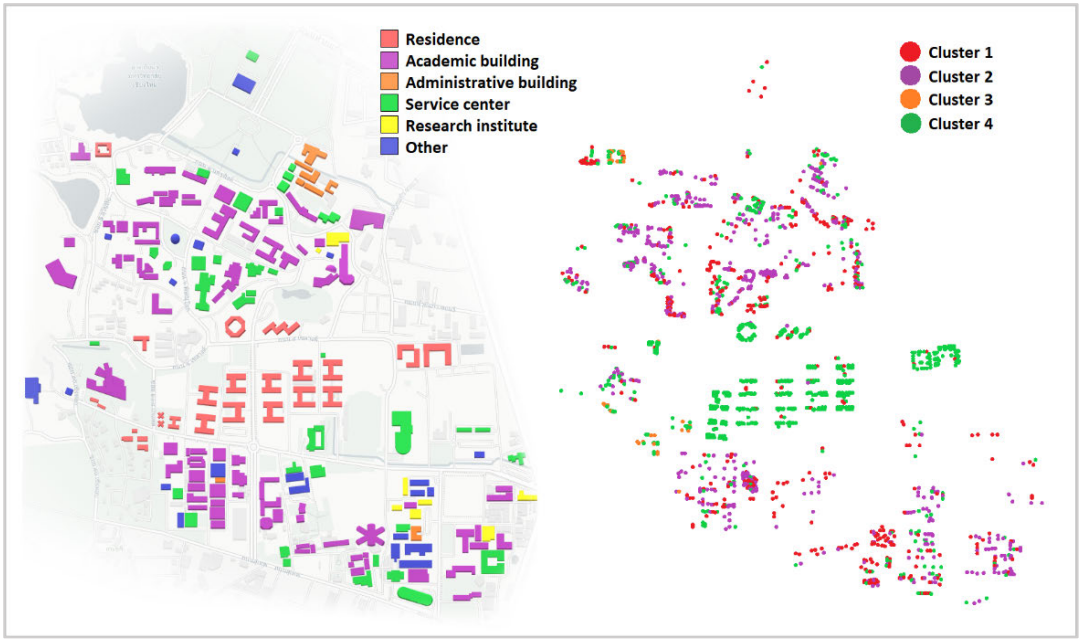

FIGURE 7. Xplaces resulting clustered APs and their corresponding building types on the Suan Sak campus (main campus).

TABLE 5. Silhouette values for 3-hour temporal window scheme in space usage segmentation.

Time	Xplaces-C	Xplaces-M	Xplaces-E	Xplaces-CME	Eigenplace
00:00 - 03:00	0.65	0.74	0.59	0.86	0.64
03:00 - 06:00	0.65	0.73	0.58	0.88	0.66
06:00 - 09:00	0.62	0.57	0.49	0.74	0.46
09:00 - 12:00	0.44	0.43	0.54	0.56	0.30
12:00 - 15:00	0.68	0.30	0.58	0.59	0.24
15:00 - 18:00	0.33	0.29	0.44	0.44	0.44
18:00 - 21:00	0.36	0.51	0.45	0.67	0.46
21:00 - 00:00	0.77	0.74	0.60	0.67	0.59

TABLE 6. Silhouette values for 6-hour temporal window scheme in space usage segmentation.

Time	Xplaces-C	Xplaces-M	Xplaces-E	Xplaces-CME	Eigenplace
00:00 - 06:00	0.72	0.80	0.78	0.84	0.64
06:00 - 12:00	0.54	0.66	0.60	0.65	0.34
12:00 - 18:00	0.46	0.49	0.49	0.35	0.21
18:00 - 00:00	0.72	0.70	0.64	0.71	0.60

the period series were 00:00 – 03:00, 03:00 – 06:00, …, 21:00 – 00:00, which accounted for eight periods in total. For each time period, the five approaches were implemented and silhouette values were calculated. The result is shown in Table 5, where the highest silhouette values for each time period are in bold for visibility. Xplaces-CME has the highest silhouette values for the first four periods (i.e., from 00:00 to 12:00), especially for the first three periods that cover the last night until morning hours where the silhouette values were very high (above 0.7). This is presumably due to the typical resting period of students. Xplaces-C performs better than other approaches for the 12:00 – 15:00 period, during which the crowdedness seems to be the key factor.

There appears to be some degrees of randomness in space usage between 15:00 and 18:00, which makes it difficult to characterize and distinguish among clusters as reflected by relatively low silhouette values where three approaches were tied for the highest segmentation performance, i.e., Xplaces-E, Xplaces-CME, and eigenplaces at silhouette value of 0.44. This observation is in line with the study by Horanont *et al.* [30] which also discovered that people's activity patterns are highly random during

FIGURE 8. Xplaces' resulting clustered APs and their corresponding building types on the Suan Dok campus (health science complex).

TABLE 7. Silhouette values when day of the week is used as temporal window in space usage segmentation.

Day	Xplaces-C	Xplaces-M	Xplaces-E	Xplaces-CME	Eigenplace
Monday	0.47	0.60	0.65	0.65	0.43
Tuesday	0.50	0.66	0.51	0.67	0.43
Wednesday	0.44	0.58	0.69	0.59	0.41
Thursday	0.51	0.67	0.66	0.62	0.47
Friday	0.50	0.65	0.65	0.65	0.46
Saturday	0.70	0.52	0.57	0.62	0.43
Sunday	0.72	0.61	0.56	0.65	0.55

15:00 – 18:00 period. For the period 18:00 – 21:00, when most academics and university staff are leaving the campuses while most students return to their dorms or go out for a dinner, Xplaces-CME performs better than other approaches with silhouette value of 0.67. Lastly, the Xplaces-C outperforms other methods for the period 21:00 – 00:00 with a high silhouette value of 0.77, during which most students are flowing back to their dorms or going to bed.

The 6-hour scheme also started from midnight, hence the period series were 00:00 - 6:00, 6:00 - 12:00, 12:00 - 18:00, and 18:00 - 00:00. As listed on Table 6, the result shows that the Xplaces-CME is the best performing method for space segmentation based on the usage happening late at night and toward morning (00:00 – 6:00) with a high silhouette value of 0.84, while the Xplaces-M is the top performer for space usage taking place in the morning and toward noon

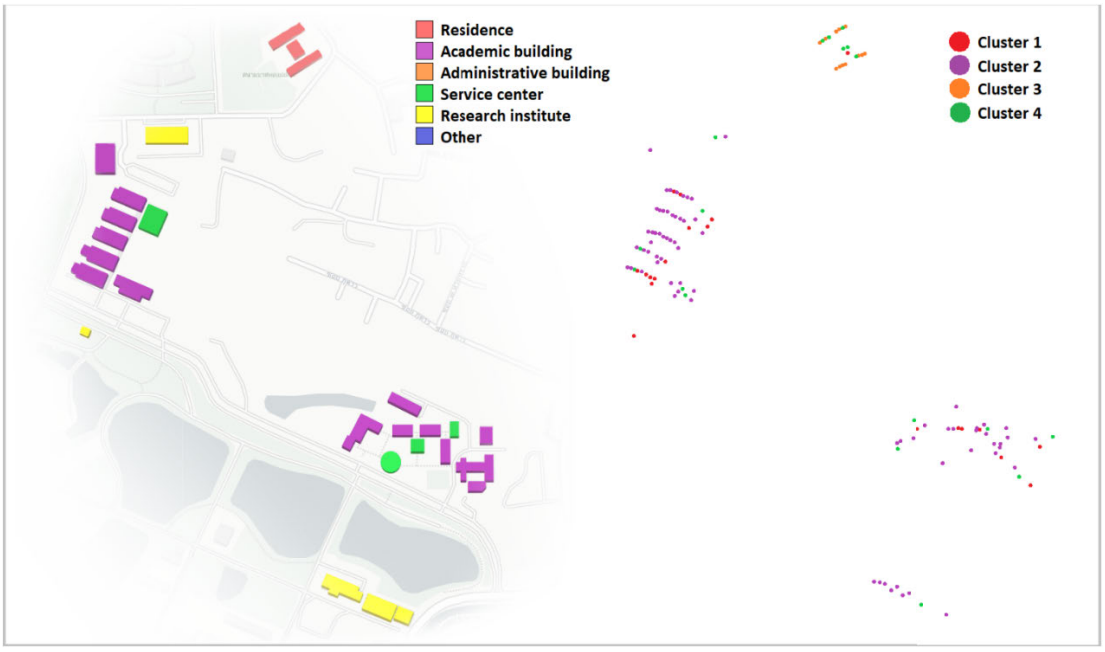

FIGURE 9. Xplaces' resulting clustered APs and their corresponding building types on the Mae Hia campus (veterinary medicine and agro-industry faculties).

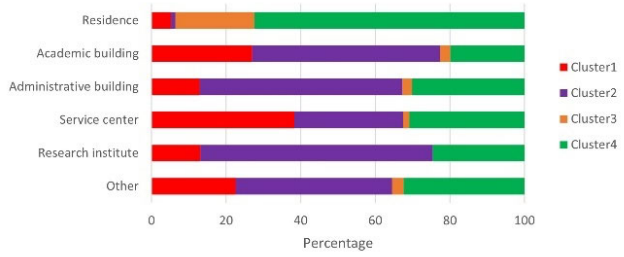

FIGURE 10. Cluster distribution based on building types.

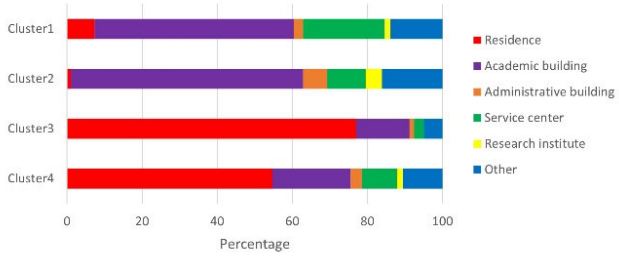

FIGURE 11. Cluster composition based on building types.

(6:00 – 12:00) with a silhouette value of 0.66. Then, the space utilization becomes more difficult to characterize from 12:00 – 18:00, as also observed previous in the 3-hour temporal window scheme, the Xplaces-M and Xplaces-E perform equally well as the best performing methods sharing the same silhouette value of 0.49. For the evening hours toward the midnight (18:00 – 00:00), the Xplaces-C outperforms other methods with a silhouette value of 0.72. The crowdedness is once again a decisive feature for late evening space utilization when most campus populations are returning to their residence, which subsequently triggers crowd shifting.

As most people go about their everyday activities around work/study schedules, so each day of the week thus affects how people utilize space differently. So, we continued to examine the impact of the area usage temporality on our space segmentation on a daily scale by looking at segmentation performance from the perspective of day of the week. As shown in Table 7, different models perform better than others for different days of the week. For Monday, the Xplaces-E and Xplaces-CME both are the best performing approaches sharing a silhouette value of 0.65. The Xplaces-CME is the best model for Tuesday with its silhouette value of 0.67, while Xplaces-E outperforms other methods for Wednesday with its highest silhouette value of 0.69. For Thursday, the Xplaces-M is the best model with a silhouette value of 0.67. Three models including Xplaces-M, Xplaces-E, and Xplaces-CME perform equally well with the best silhouette value of 0.65. Interestingly, the Xplaces-C does not perform well throughout the weekdays, however it becomes the best performing method for the weekend (Saturday and Sunday) with highest silhouette values of 0.70 and 0.72, respectively, which is intuitive due to a clear variation in the level of crowdedness on campuses during the weekend compared to the weekdays. So, the crowdedness subsequently emerges as a decisive feature for the weekend.

Overall, the Xplaces performs better than the state-of-the-art technique, i.e., eigenplaces, in all temporal window schemes. As reflected by the findings, the choice of

considered features of the Xplaces depends on the temporal window of consideration. For example, if it is desired that the space is segmented according to its spatial utilization during the weekend, then the Xplaces-C is the most suitable option. However, if the segmentation is to be done based on the characteristics of its spatial utilization during the time period 6:00 – 12:00, then the Xplaces-M is the recommended method.

IV. CONCLUSION

As we're narrowing the digital divide, the coverage of wireless networks such as Wi-Fi for the internet access increasingly expends. Especially during the pandemic, access to the internet has been even more essential. While people are connected to a Wi-Fi network, their connectivity is recorded for network monitoring. Collectively, over several access point locations, these connectivity logs can be analyzed opportunistically to reveal how people interact and utilize built environment and physical space. This paper presents a development of new method called Xplaces that segments physical space based on area utilization reflected by Wi-Fi connectivity, which can be useful for spatial design and planning. As a case study, we used a Wi-Fi data collected from a university network of 2,980 access points that serve 291,124 unique devices located across three campuses for our analysis. We've defined three features that characterize space utilization, i.e., crowdedness, mobility, and connectivity entropy, for space segmentation. Xplaces consists of three main procedures. It firstly reduces the feature's high dimensionality by employing the PCA. It then extracts a behavioral signature of each access point by applying eigendecomposition, a technique used in linear algebra. It finally uses these signatures for its segmentation by applying X-means clustering technique that does not require a predefined number of clusters. Such that, the resulting segmentation emerges from the actual characteristics of space utilization. For evaluation, silhouette value was used to measure how well clusters were formed. Our Xplaces outperforms the state-of-the-art model, i.e., eigenplaces with the silhouette value deficit of 0.31. We further investigated the impact of the temporality of area usage on our space segmentation by examining the area segmentation from three different temporal window schemes: 3-hour period, 6-hour period, and day of the week. This investigation shows that Xplaces performs well with particular features for different schemes, and thus yields a set of recommended features for area segmentation based on its utilization within a chosen temporal window of observation. For instance, we found that the Xplaces method works well with the crowdedness employed as its feature for the space segmentation that is based on weekend usage of the area.

Therefore, the main contributions of this work are the development of Xplaces and a set of recommended features for different temporal windows of observation of space utilization, which enable more informed spatial design and planning. Nonetheless, there are some limitations of our study. To begin with, we only introduced and examined three different features in this study. Clearly, there are other potential features that can potentially be extracted from the Wi-Fi data, which characterize area usage. Exploring other influential features is thus worth future research. Second, network connection issues may have caused some connecting and disconnecting events in the logs, and hence it may have affected our analysis, particularly for the mobility feature calculation. Since, these connection issues were marginal, the negative effect was thus believed to be minor. Nevertheless, future investigation may take this issue into consideration. Lastly, there was a lack of ground truth confirmation of our resulting area segmentations. There is still an open question of how to properly measure area usage. Approaches and methods for sensing and assessing how space is utilized are thus among those of potential future work.

ACKNOWLEDGMENT

The authors would like to thank the CMU's Computer Network Operation Center (CNOC) and Information Technology Service Center (ITSC) for providing the Wi-Fi connectivity data for their research.

REFERENCES

[1] S. Affairs, "World urbanization prospects," United Nations, New York, NY, USA, Tech. Rep. ST/ESA/SER.A/366, 2014, doi: 10.4054/DemRes.2005.12.9.
[2] H.-S. Cho and M. Choi, "Effects of compact urban development on air pollution: Empirical evidence from Korea," *Sustainability*, vol. 6, no. 9, pp. 5968–5982, Sep. 2014, doi: 10.3390/su6095968.
[3] A. O. Araico, "Detection of the crowdedness of a place sensing the devices in the area," Univ. Twente, Enschede, The Netherlands, Tech. Rep. 72436, 2017.
[4] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, "Tracking human mobility using WiFi signals," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0130824, doi: 10.1371/journal.pone.0130824.
[5] M. Uras, R. Cossu, E. Ferrara, A. Liotta, and L. Atzori, "PMA: A real-world system for people mobility monitoring and analysis based on Wi-Fi probes," *J. Clean. Prod.*, vol. 240, Oct. 2020, Art. no. 122084.
[6] L. Shen and P. R. Stopher, "Review of GPS travel survey and GPS data-processing methods," *Transp. Rev.*, vol. 34, no. 3, pp. 316–334, May 2014, doi: 10.1080/01441647.2014.903530.
[7] A. Cuttone, S. Lehmann, and M. C. González, "Understanding predictability and exploration in human mobility," *EPJ Data Sci.*, vol. 7, pp. 1–17, Dec. 2018, doi: 10.1140/epjds/s13688-017-0129-1.
[8] S. Phithakkitnukoon, T. Horanont, A. Witayangkurn, R. Siri, Y. Sekimoto, and R. Shibasaki, "Understanding tourist behavior using large-scale mobile sensing approach: A case study of mobile phone users in Japan," *Pervas. Mobile Comput.*, vol. 18, pp. 18–39, Apr. 2015, doi: 10.1016/j.pmcj.2014.07.003.
[9] F. Rupi, C. Poliziani, and J. Schweizer, "Data-driven bicycle network analysis based on traditional counting methods and GPS traces from smartphone," *ISPRS Int. J. Geo-Information*, vol. 8, no. 8, p. 322, Jul. 2019, doi: 10.3390/ijgi8080322.
[10] W. Lemstra, V. Hayes, and J. Groenewegen, *The Innovation Journey of Wi-Fi: The Road to Global Success*. Cambridge, U.K.: Cambridge Univ. Press, 2010. [Online]. Available: https://www.amazon.com/Innovation-Journey-Wi-Fi-Global-Success/dp/0521199719
[11] F. Calabrese, J. Reades, and C. Ratti, "Eigenplaces: Segmenting space through digital signatures," *IEEE Pervasive Comput.*, vol. 9, no. 1, pp. 78–84, Jan. 2010, doi: 10.1109/MPRV.2009.62.
[12] A. Sevtsuk, S. Huang, F. Calabrese, and C. Ratti, "Mapping the MIT campus in real time using WiFi," in *Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City*. Hershey, PA, USA: IGI Global, 2008.
[13] M. M. Ouf, M. H. Issa, A. Azzouz, and A.-M. Sadick, "Effectiveness of using WiFi technologies to detect and predict building occupancy," *Sustain. Buildings*, vol. 2, p. 7, Oct. 2017, doi: 10.1051/sbuild/2017005.

[14] M. Kim and D. Kotz, "Modeling users' mobility among WiFi access points," Dartmouth College, Hanover, NH, USA, Tech. Rep. 3344, 2005.
[15] M. Kim and D. Kotz, "Classifying the mobility of users and the popularity of access points," Dartmouth College, Hanover, NH, USA, Tech. Rep. 3320, 2005, doi: 10.1007/11426646_19.
[16] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 9, no. 3, pp. 58–68, Jul. 2005, doi: 10.1145/1094549.1094558.
[17] K. Qin, Y. Xu, C. Kang, S. Sobolevsky, and M. P. Kwan, "Modeling spatio-temporal evolution of urban crowd flows," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 12, pp. 1–22, 2019, doi: 10.3390/ijgi8120570.
[18] A. Binthaisong, J. Srichan, and S. Phithakkitnukoon, "Wi-crowd: Sensing and visualizing crowd on campus using Wi-Fi access point data," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2017, pp. 441–447, doi: 10.1145/3123024.3124413.
[19] R. Smarzaro, C. A. Davis, Jr., and J. A. Quintanilha, "Creation of a multimodal urban transportation network through spatial data integration from authoritative and crowdsourced data," *Int. J. Geo- Inf.*, vol. 10, no. 7, p. 31, 2021, doi: 10.3390/ijgi10070470.
[20] E. Graells-Garrido, F. Serra-Burriel, F. Rowe, F. M. Cucchietti, and P. Reyes, "A city of cities: Measuring how 15-minutes urban accessibility shapes human mobility in Barcelona," *PLoS ONE*, vol. 16, no. 5, 2021, Art. no. e0250080, doi: 10.1371/journal.pone.0250080.
[21] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Pers. Ubiquitous Comput.*, vol. 10, pp. 255–268, 2006, doi: 10.1007/s00779-005-0046-3.
[22] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. July, pp. 379–423, 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.
[23] R. Vidal, Y. Ma, and S. S. Sastry, "Generalized principal component analysis," in *Interdisciplinary Applied Mathematics*. New York, NY, USA: Springer, 2016, pp. 25–62.
[24] A. Dmitrienko, C. Chuang-Stein, and R. B. D'Agostino, *Pharmaceutical Statistics Using SAS: A Practical Guide*. Cary, NC, USA: SAS Institute, 2007.
[25] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.
[26] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. Hoboken, NJ, USA: Wiley, 1973. [Online]. Available: https://www.amazon.com/Pattern-Classification-Scene-Analysis-Richard/dp/0471223611
[27] D. Pelleg and A. W. Moore, "X-means: Extending K-means with efficient estimation of the number of clusters," in *Proc. 17th Int. Conf. Mach. Learn. Table Contents*, 2000, pp. 724–734, doi: 10.1007/3-540-44491-2_3.
[28] R. E. Kass and L. Wasserman, "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion," *J. Amer. Stat. Assoc.*, vol. 90, no. 431, pp. 928–934, 1995, doi: 10.1080/01621459.1995.10476592.
[29] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.
[30] T. Horanont, S. Phithakkitnukoon, T. W. Leong, Y. Sekimoto, and R. Shibasaki, "Weather effects on the patterns of people's everyday activities: A study using GPS traces of mobile phone users," *PLoS ONE*, vol. 8, no. 12, Dec. 2013, Art. no. e81153, doi: 10.1371/journal.pone.0081153.

SUNSIKA CHAIKUL received the B.S. degree in computer science from Nation University and the M.S. degree in computer engineering from Chiang Mai University, Thailand, where she is currently pursuing the Ph.D. degree with the Department of Computer Engineering. Her research interests include urban data science and visual analytics.

SANTI PHITHAKKITNUKOON received the B.S. and M.S. degrees in electrical engineering from Southern Methodist University, Dallas, USA, in 2003 and 2005, respectively, and the Ph.D. degree in computer science and engineering from the University of North Texas, USA. He is currently an Associate Professor with the Department of Computer Engineering, Faculty of Engineering, Chiang Mai University, Thailand. Before joining Chiang Mai University, he was a Lecturer in computing at The Open University, U.K.; a Research Associate at Newcastle University, U.K.; and a Postdoctoral Fellow with the SENSEable City Laboratory, Massachusetts Institute of Technology, USA. His research interest includes urban informatics.

CARLO RATTI received the Ph.D. degree in architecture from the University of Cambridge. He is an Architect and an Engineer, who practices architecture in Turin and teaches at the Massachusetts Institute of Technology (MIT), where he directs the SENSEable City Laboratory. His research interests include urban design, human–computer interfaces, electronic media, and the design of public spaces. He is a member of the Ordine degli Ingegneri della Provincia di Torino, the Architects Registration Board, U.K.; and the Association des Anciens Élèves de l'École Nationale des Ponts et Chaussées.