Computers, Environment and Urban Systems 67 (2018) 55–67

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Activity patterns mining in Wi-Fi access point logs

Guilhem Poucin[a], Bilal Farooq[b,*], Zachary Patterson[c]

[a] Laboratory of Innovations in Transportation (LITrans), Département de Génies Civil, Géologique et des Mines, École Polytechnique Montréal, 2500 Ch. Polytechnique Montréal, H3T 1J4 Montréal, Canada
[b] Laboratory of Innovations in Transportation (LITrans), Department of Civil Engineering, Ryerson University, 350 Victoria Street, M5B 2K3 Toronto, Canada
[c] Transportation Research for Integrated Planning (TRIP) Laboratory, Geography, Planning and Environment Department, Concordia University, Montreal, Canada

ARTICLE INFO

Article history:
Received 21 June 2016
Received in revised form 12 September 2017
Accepted 13 September 2017
Available online 20 September 2017

Keywords:
Information mining
Activity pattern recognition
Wi-Fi networks
Clustering
Facility usage

ABSTRACT

This article proposes a methodology to mine valuable information about the usage of a facility (e.g. building, open public spaces, etc.), based *only* on Wi-Fi network connection history. Data are collected at Concordia University in Montréal, Canada. Using the Wi-Fi access log data, we characterize activities taking place within a building without any additional knowledge of the building itself. The methodology is based on identification and generation of pertinent variables derived by Principal Component Analysis (PCA) for clustering (i.e. PCA-guided clustering) and time-space activity identification. K-means clustering algorithm is then used to identify 7 activity types associated with buildings in the context of a campus. Based on the activity clusters' centroids, a search algorithm is proposed to associate activities of the same types over multiple days. The spatial distribution of the computed activities and building plans are then compared, which shows a more than 85% match for the weekdays.

1. Introduction

Most traditional efforts to collect data on mobility and human behavior involve surveys (e.g. regional origin-destination surveys) are based on sampling a small proportion of the known population (see e.g. Ortúzar and Willumsen (2011)). Such data collection methods are expensive in terms of direct costs and time required. They may not represent actual behavior of users due to sampling biases (Richardson, Ampt, & Meyburg, 1995; Zmud, Lee-Gosselin, Munizaga, & Carrasco, 2013) and the reliability of data reported by respondents may suffer due to difficulty in recalling the activities (Frick & Grabka, 2005). Moreover, due to their predominantly cross-sectional nature, traditional mobility data are not able to observe the evolution of an individual's behavior over time. The development of free Wi-Fi networks in cities (e.g. Smart Sidewalks in the UK) and the spread of smartphones represent an opportunity to capture a larger sample of the population continuously at very low cost. While such passive collection technologies and longitudinal behavioral data represent tremendous opportunities, methodologies and tools to exploit these new sources are in their infancy.

Using data from pervasive and ubiquitous networks for mobility studies is an emerging area of research (Cao et al., 2015). Munizaga and Palma (2012), Kusakabe and Asakura (2014), Long and Thill (2015) and other recent studies have used smart-card transaction data from Automated Fare Systems (AFS) to study urban mobility patterns. Iqbal, Choudhury, Wang, and González (2014) used cellular phone network data to develop origin-destination matrices in an urban area. Meneses and Moreira (2012) used Wi-Fi network data for localization and routing on a university campus. The main challenge faced while using these datasets is the lack of information concerning users, which is caused by the common necessity of anonymizing the data. Recent studies have tried to incorporate other data sources (e.g. land use data, travel diary surveys, time tables etc.) to overcome such issues (Calabrese, Lorenzo, & Ratti, 2010; Danalet, Farooq, & Bierlaire, 2014; Grapperon, Farooq, & Trépanier, 2016). However in many cases, it is very difficult to access such data, especially at a very disaggregate level.

The Media Access Control (MAC) address is a unique identifier associated with each network interface and is used as a unique address in a Wi-Fi network. This address is fixed for a Wi-Fi enabled device and remains the same throughout the life of a device. Wi-Fi networks are composed of sets of Access Points (AP) to which a device can connect using its MAC address. APs provide Wi-Fi services, i.e. a connection to the internet. APs are spatially distributed covering large areas (e.g. campus, shopping center, etc.) and collectively comprise a Local Area Network (LAN). Our methodology only uses

[*] Corresponding author.
E-mail addresses: guilhem.poucin@polymtl.ca (G. Poucin), bilal.farooq@ryerson.ca (B. Farooq), zachary.patterson@concordia.ca (Z. Patterson).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.09.004
0198-9715/© 2017 Elsevier Ltd. All rights reserved.

communication between devices and APs over the LAN to develop the traces of people over time and space. We advance the current state of research by proposing a PCA-guided K-means clustering to associate to each Wi-Fi AP the dominant activity performed at its location over time. The methodology is applied at a large scale in terms of space (i.e. a high-rise building) and on a very disaggregate time scale (i.e. day level), without any previous spatial knowledge of the infrastructure. To confirm the consistency of mined activities, we extend our analysis to a period of an entire week. Furthermore, to indirectly test the accuracy of our methodology, we compare our results to the designated usage of spaces in the building.

The rest of the article is structured as follows: a review of current literature on the use of ubiquitous networks, especially Wi-Fi networks is followed by a description of the case study location, and the dataset used in the analysis. This leads us to a section describing the methodology adopted and results related to the classification of access points in terms of their surrounding activities. The next section compares the activity inference results with designated usage of the space on building plans. In the end we discuss our conclusions, limitation, and possible applications.

2. Literature review

The study of pervasive systems such as cellular networks, Global Positioning Systems (GPS) or Wi-Fi networks, has received growing attention from researchers during the last decade. Indeed, the development and growth of these ubiquitous networks, combined with the improvement of data collection processes, the spread of smartphones, and the emergence of data science have opened promising perspectives towards the understanding and characterization of human behavior. This interest has resulted in applications related to network optimization, urban modeling and even transportation policy. In the following section, we inventory few of the studies analyzing the relationship between human activities and infrastructure space. We then explore the potential of network trace data (especially Wi-Fi) to study human behavior, proposing an overview of the literature from the different data collection processes, to the challenges encountered and the different methodologies proposed to analyze the data.

2.1. Relationship between human activity and space

Recent studies on the relationship between activities/behavior and spatial data have benefited from the large improvement of the datasets available. These studies show a high regularity in human trajectories (Gonzalez, Hidalgo, & Barabasi, 2008) and daily routine (Song, Qu, Blumm, & Barabási, 2010). Wang, Pedreschi, Song, Giannotti, and Barabasi (2011) study the relationship between human mobility and their social network connections.

However, privacy issues surrounding such data reduce the accuracy of the analysis–especially with respect to the socio-demographic variables driving human behavior. Some studies are based on geographic data, for instance Eagle and Pentland (2009). In parallel, other studies, such as (Jiang, Ferreira, & González, 2012), base themselves on more conventional survey data (travel diary survey), benefiting from the richness of these datasets, but limiting their large scale applicability due to cost.

2.2. Network traces

As suggested in Aschenbruck, Munjal, and Camp (2011), user traces can be acquired through three different methods: monitoring location, communications or contacts.

The monitoring of location involves collecting successive positions of a user's device and is mostly done with GPS. Using a network of 72 satellites, this technology can furnish a user's position within a few meters. In the past, Liu, Andris, and Ratti (2010) used GPS traces to study the mobility behavior of taxi drivers. Patterson and Fitzsimmons (2016) analyzed trace data collected through the smartphone travel survey application DataMobile. However, GPS data show limited application in indoor and dense urban environments, where obstacles can create a shadowing effect. This problem can be overcome by coupling the information from GSM and Wi-Fi networks (Aschenbruck et al., 2011), or in some cases with additional data sources like GTFS, as Zahabi, Ajzachi, and Patterson (in press) did to infer transit itineraries from smartphone data. However, as mentioned in J.~Su, Chin, Popivanova, Goel, and De~Lara (2004), some users can be reluctant to share the history of their positions. Since this method is device-centric, it needs a user's cooperation by accepting the burden of an additional device (e.g. GPS unit) or an energy consuming application on their own device (e.g. smartphone). We refer to this kind of location monitoring (i.e. location monitoring by a user's device) as device-centered monitoring. This is distinguished from network-centered monitoring when device information is collected passively and automatically by Wi-Fi or GSM networks (Nguyen-Vuong, Agoulmine, & Ghamri-Doudane, 2007), which we divide into two broad categories.

The first type of network-centered monitoring relies on the monitoring of communication and it uses interactions between devices and a communication system (cellular or Wi-Fi) to recreate a user's mobility history. This information is regularly collected by network operators and represents a low-cost (and low-burden on the user) source of locational information. The pervasive nature of these networks allows the capture of a large sample of the population; even if the characterization of this sample is still a challenge (Calabrese, Diao, Di~Lorenzo, Ferreira, & Ratti, 2013). The improvement of the relatively low accuracy of the locational data obtained is considered in Mao, Fidan, and Anderson (2007) and Wymeersch, Lien, and Win (2009). Usually this leads to work on symbolic spaces rather than geographic spaces, as described in Meneses and Moreira (2012). The use of signal strength can improve location accuracy and can be further improved with other information fusion processes (Aschenbruck et al., 2011). Cellular (GSM) data, which can provide locational accuracy at the size of neighborhoods in cities, are discussed in Calabrese et al. (2013). Wi-Fi data have been used at finer scales in locations such as campuses, offices or festivals. (More detail on this literature is discussed in the next section.)

The second type of network-centered monitoring of locational data is done by monitoring contact between users through technologies such as Bluetooth or Wi-Fi. Monitoring of contacts can allow the characterization of social networks existing in closed environments. This topic has received growing attention thanks to developments in multi-hop networks (Conti & Giordano, 2007). Monitoring of contacts involves unloading a part of the Wi-Fi network and making information transit through a chain of user devices. In J.~Su et al. (2004) and J.~Su, Goel, and De~Lara (2006), a sample of students are monitoring their contacts with other users within a campus to explore the feasibility of such a technology, and such an approach has been used to highlight the social behavior of the students. Another use of this process is proposed in Naini, Dousse, Thiran, and Vetterli (2011), where phone Bluetooth activity at a festival allowed the estimation of the size of the entire population of festival goers using a statistical algorithm derived from biology.

2.3. Challenges to using network-centric Wi-Fi data

While the information from a phone's connection history gives us the advantage of collecting data on a large sample of individuals at low cost, important challenges to using this data have been highlighted in the literature. The first challenge results from the need for anonymization of user data, essential to guaranteeing user privacy (Meneses & Moreira, 2012), which naturally removes socio-demographic information. A related issue is the absence of a distinction between different types of devices–smartphones, tablets or laptops appear in the same way in the database and can even represent the same users. Yoon, Noble, Liu, and Kim (2006) and others have tried to overcome these problems by endeavoring to identify differences between the behavior of different devices: e.g., laptops are connected longer but to fewer APs. A detailed analysis of subtle behavioral differences remains an open question.

In addition to this, the reliability of Wi-Fi data is dependent upon the status of the antenna on a user's device. If the antenna is turned off, the user is invisible on the network. At the same time, the geographically limited range of these networks means that devices easily go outside the range of any given network, thus creating gaps in a user's location history (Meneses & Moreira, 2012).

Another important issue that network-centered Wi-Fi data processing encounters is the so-called Ping-Pong effect. This arises when the user is connected and disconnected to different APs while the user is not mobile (Danalet et al., 2014). If this happens, non-existent trips can be created within datasets. A number of solutions have been proposed to deal with this issue (Aschenbruck et al., 2011; Yoon et al., 2006).

2.4. Previous research on the analysis of Wi-Fi network data

A large amount of work has considered how to analyze Wi-Fi usage for a given network, such as Henderson, Kotz, and Abyzov (2008). Through the aggregation of connections and statistical tools, Henderson et al. (2008) present the main parameters of network usage, user behavior and the mobility of users. The goal of this research is to develop analysis tools for the design and efficiency of Wi-Fi networks in highly frequented places. In the list of possible optimization problems, we find the optimization of hand off latency (Ramani & Savage, 2005), anticipation of resource allocation (Katsaros et al., 2003), improvement of routing protocols (W.~Su, Lee, & Gerla, 2001) and the development of energy-efficient localization (You et al., 2006). Work in this area can be found at different scales, such as a university campus (Henderson et al., 2008), buildings of a corporation (Balazinska & Castro, 2003), hospitals (Prentow, Ruiz-Ruiz, Blunck, Stisen, & Kjærgaard, 2015), or even cities (Afanasyev, Chen, Voelker, & Snoeren, 2010).

The necessary pre-processing of data to address the challenges raised in the previous section leads to a diversity in the assumptions concerning user behavior and the spatial distribution of infrastructure in literature. Meneses and Moreira (2012), for example, use connection rates to compute the most important corridors within a campus, aggregating APs of the same building (movements within buildings are thereby not considered).

From a theoretical perspective, there have been attempts to move away from a strict geographical concept of location (i.e. coordinates in space) to include in the notion of locations, the activities that are conducted there and as a result, to emphasize the concept of *place* in the description of a wireless environment (Kang, Welbourne, Stewart, & Borriello, 2005). From this perspective, it is thought that users are more interested in the type of activities taking place in a location rather simply its spatial location.

Related to this emphasis on activities surrounding APs, Calabrese, Reades, and Ratti (2010) introduce the possibility of identifying the type of activities taking place around APs, adopting an eigenvector analysis of temporal signatures in the connection data. They use an eigen decomposition to characterize the number of connected devices through time at each access point with four variables. These values are then clustered to generate 5 families of access points associated with a specific type of activity. The identification of the activity performed is mainly done by analyzing the time of day or day of week, when the access points are the most used. The characterization of access points is developed over 7 days. Behavioral differences observed between the different days (especially between weekdays and weekend) are an input data for the clustering. Note that here a week level analysis is not possible. Also, the information on actual use of these locations are included in their clustering initialization, making such information necessary to applying their methodology.

In this paper, we propose a method to analyze infrastructure usage–especially the activities undertaken around APs. Calabrese, Reades, and Ratti (2010) propose such a process through the analysis of the peak number of connections during a week at a campus scale. We build on this by developing a methodology that can be applied to periodic events by first studying one day at a finer scale. Instead of clustering the number of connections as a function of time, we generate explanatory variables characterizing the dynamics of connections and time spent by users. We then use clustering methods to extract the main activities performed in a building. These activities are generated for each day of the week. We also compare the spatial distribution of activities with the designated usage of space in the building–unlike Calabrese, Reades, and Ratti (2010) where this information was used to calibrate the model.

This work aims to enrich the Wifi databases bringing an additional layer of information concerning the infrastructure usage. We believe it could be used for infrastructure monitoring, furnishing an understandable analysis of the network usage allowing a comparison between the designated use of the spaces in the building and the actual use made of it. Then, it could help building preliminary analysis on Wi-Fi sensors based data collection performed on special events e.g. festivals (Farooq, Beaulieu, Ragab, & Dang~Ba, 2015).

3. Concordia University Wi-Fi access point log dataset

This study is based on AP log data from Concordia University in Montreal, Canada. Concordia is made up of 57 buildings across two campuses (one in Montreal's downtown core, and the other in suburban Montreal). It counts around 47,000 students, faculty and staff with 36,000 undergraduate students. Our initial dataset included all connection data from all APs of the campus during the week of February 2 to 8, 2015. The data include the connection history of all devices connected to the Concordia network over this period. In order to ensure anonymity, each device initially identified by its MAC had a new randomly generated identifier associated to it in the database. Each connection record included the access point ID, user device ID as well as connection and disconnection time. In the raw data, connections with duration of less than 5 min were not recorded. Altogether, over the course of the week there were 1.7 million connections by 60,500 devices to 1048 Access Points. The minimum connection time was 5 min, and the maximum 5 days and 20 min. The latter was surely a fixed device such as desktop computer with a very long connection time. As mentioned above, there are many buildings on the two university campuses. Instead of considering all APs across all buildings, we concentrated on the APs of only one building. As such, in this article, we focus on the set of data concerning one of the largest buildings at Concordia University (for security reasons we are not permitted to reveal the name or exact location of the building).

It is necessary to highlight three potential limitations of this data and our proposed methodology. The first relates to the representativeness of observed users compared to the total population. We do not have access to any data concerning the smartphone and laptop ownership by Concordia University users. However recent studies have indicated the penetration rate of between 35 to 60% is observed for smartphone with Wi-Fi functionality being turned on (Farooq et al., 2015). Then as mentioned in the literature, only connected users appear in the AP logs. Thus, users exploring the building while staying offline are not visible. However, due to significantly large size of the sample, representativeness is assured (Farooq, Bierlaire, Hurtubia, & Flötteröd, 2013).

Second, these data correspond to devices (nodes) and not users. The diversity of devices available (smartphones, tablets and laptops) can make one user appear multiple times under different IDs in the data. This can be a problem when it comes to user mobility, but not in our case, since we are characterizing the differences between categories of access points and the number of unique users is not required. Moreover, the fact that the use of particular devices (e.g. laptops) is correlated to particular types of activities (e.g. working) could, in fact, amplify the differences between different activity types.

Finally, our goal is to associate to each access point a main type of activity. The methodology assumes that a unique type of dominant activity is performed at each location which may not be the case. This assertion is even amplified by the large range of access points, covering different rooms and spaces at the same time.

3.1. Complementary data

In addition to the connection log database, maps of facilities with the location of each AP were available. While available, we decided not to use the map information when developing the methodology. Instead we used the maps uniquely for the validation of our inferred activities. This is useful, because while Wi-Fi log data are easily accessible, facility characteristics can be harder to obtain for security reasons (this was the case in our study). As a result, this approach could also be applied to data collected by monitoring Wi-Fi connected devices in public areas like in (Farooq et al., 2015). In this work, we explore the possibilities offered by these limited data, and compare our inferred activities with those that can be derived from the supply (map) data.

4. Methodology

We propose a methodology to infer the main activity performed around each access point using unsupervised machine learning with the K-means algorithm (Jain, 2010). We start by creating indicators that characterize the different aspects of user connections of each AP considering the number of connections and connection duration. The clustering algorithm organizes APs into families considering the similarities in the computed variables. A principal component analysis (PCA) was used as a space reduction technique, to decrease the correlation between the computed indicators. This process improves the efficiency of the K-means algorithm and decreases the probability of the algorithm getting stuck in local minima. After the generation of clusters, we associate semantic meaning to the different families of APs based on the values of indicators. Finally, we propose an algorithm tracing the cluster found through the data collection period and then observe the variation of clusters found across the days.

4.1. Activity indicators

We characterize the mobility of users at each AP by generating indicators based on the main components of human behavior. The use of signal decomposition to generate the clustering variable as in (Calabrese, Reades, & Ratti, 2010) leads to a better characterization of the dynamics in the number of connections. However, such a process aggregates the set of users at a place without taking into account individual user characteristics. Indicators (Table 1) provide a way to understand the characteristics of clusters once generated. While the values obtained through eigen decomposition, describe accurately the time signature of the connections, our indicators give interpretative descriptions of the usage of the AP in a cluster through the day. We make the hypothesis here that density/attendance of the place and time spent by users have an impact on the human behavior. These indicators are mainly generated by computing a given variable through time over the course of the day, and then characterizing the distribution of values obtained through their mean and standard deviation.

Table 1
Indicators interpretation.

Indicators	Code
Number of connections	Number of connections performed
Number of devices	Number of people who went in the place
Primary derivative SD	Intensity of the variation of connected users
Primary derivative amplitude	Scale of the variation of connected users
Secondary derivative SD	Intensity of the speed of the variation of connected users
Secondary derivative amplitude	Scale of the speed of the variation of connected users
Average connection duration	Average connected time of users
SD of connection duration	Standard deviation of the connected time
Maximum connection duration	Maximal time users spend in a place

In consequence, one parameter to choose for this indicator is the time period of aggregation throughout the day. Time period of aggregation used in time series analysis depends on the phenomenon being studied.

Before clustering the indicators described in this section, we first proceed to a variable space reduction technique. While the described indicators are representative of important aspects of individual behavior, they may also be correlated. Thus they may result in biasing the clustering process. In Fig. B.9 we can actually observe these high correlations. To address this problem we perform a reduction of the space through a Principal Component Analysis (PCA). In the following section, we'll show the results obtained for one day of the week. A comparison between different days of the week follows in Section 4.4.

4.2. Principal component analysis

Principal component analysis is a multivariate analytic tool applied to data containing inter-correlated quantitative variables. The first purpose of this method is to extract the important elements of the data and represent it through a new set of orthogonal (uncorrelated) variables called principal components. The second purpose is the compression of data by reducing the space to only the most representative components. Finally, PCA simplifies description of the dataset and allows examination of the structure of data (Abdi & Williams, 2010)

As mentioned in Xu, Ding, Liu, and Luo (2015), the K-means algorithm is prone to falling into local minima. As this phenomenon tends to become more likely with the increase in the number of dimensions, it is recommended to perform the clustering on a reduced space generated by a PCA.

Ding and He (2004) show that because of the close relationship existing between PCA and K-Means algorithm, applying the K-means algorithm in the subspace of the principal component can improve the quality of the clustering. Example of the application of PCA-based clustering on the human activity clustering can be found in Jiang et al. (2012).

We apply PCA on the computed indicator to extract a subset of uncorrelated variable. We then select the subset of principal component ensuring it represent 95% of the global variance, and use this space for the clustering.

4.3. Clustering

Izakian, Mesgari, and Abraham (2016) pointed out that clustering is one of the most powerful technique to reveal hidden patterns and structures in the data. Among many clustering algorithms, unsupervised machine learning algorithms such as the K-means algorithm, help to avoid the transposition of expectations on clustering results (Jain, 2010). However, the K-means algorithm requires that one specify the number of clusters as a prerequisite. The appropriate number of clusters can be extrapolated from previous knowledge of classes expected or through the iterative analysis of the evolution of error. We minimize the sum of square of distance between the points within a cluster, and maximize the distance between the clusters.

The first step is to identify extreme clusters containing few elements and outlying values; these points represent specific elements of the infrastructure. We then associate each cluster to a characteristic type of activity and campus location. This part cannot be computed automatically and requires human analysis to interpret and give semantic meaning to the centroid indicator values of each cluster. While PCA subspace is indeed useful to increase the chance of finding an optimal solution of the K-means algorithm, the description of clusters is far easier in the original space of the indicators rather than in principal component space.

4.4. Analysis of clusters over the course of the data collection period

We develop the methodology at the scale of a day; this way, the resulting daily clusters were independent of each other's centroid positions and spread over the collection period. Applying clustering on a whole collection period would have missed the specific day level variations.

By applying the K-mean clustering on each day independently with the same number of clusters r, these daily clusters are not linked across different days of the data collection period. Our hypothesis is that clusters are the same along the collection and only their properties (e.g. centroid position, spread, size, etc.) change slightly. We have to find a way to associate them to each other. To do so, we compute all the possible combinations of clusters and use an optimization method based on the minimization of within sum of square (as the K-means). We define the vectors for coordinates of the centroids as follows:

$C_{i,j} \in \Re^k$

with r number of clusters, t the number of day in the data collection period, $i \in [1,t]$ days in the collection period, and $j \in [1,r]$ cluster number

$C_{(i,j)} = \left\{C^1_{(i,j)}, \ldots, C^d_{(i,j)}\right\}$ with d dimension of the space S

$C_i = \{C_{(i,1)}, \ldots, C_{(i,r)}\}$

We define a category of clusters as:

$V_q = \left\{V^1_q, \ldots, V^t_q\right\} \in C_1 \mathrm{x} C_2 \mathrm{x} \ldots \mathrm{x} C_t$

A solution to our combinatorial problem can be expressed as:

$V = \{V_1, \ldots, V_t\} \,/\, V^p_n \neq V^p_m \forall p \in [1,t], \forall (m,n) \in [1,t]\mathrm{x}[1,r]$

We are looking for the solution:

$$min_{V \in V_T}\left(\sum_{q=1}^{t} \sum_{(i,j)\in[1,t]*[1,r]} d\left(V^i_q, V^j_q\right)\right)$$

with $d(,)$ the Euclidian distance between two points in the space S

At the end of the process, we have r families of clusters, containing one cluster for each day and representing the same dominant activity. The result of this methodology allows us to study the evolution of clusters (size and elements) over the data collection period.

To conclude, in this section, we presented a methodology to extract detailed information about facility usage with only Wi-Fi Access Point logs. Such an approach is applicable in environments where a facility's purposes are not well defined in advance e.g. public places for events or festivals. It would be possible to use this method as a tool to help assign the appropriate usage of spaces within such a facility based on how they are actually used–a method to crowd source appropriate use. Definition of semantic meaning clearly is still up to human judgment and cannot be automated.

5. Results

5.1. Activity indicators

Here we focus on activities that can be from *transitions between places*, which are of few seconds or minutes, to *work sessions*, which can have a duration in hours. As a consequence, we use a period of aggregation close to the shortest activity duration we want to describe thus choosing 5 minutes.

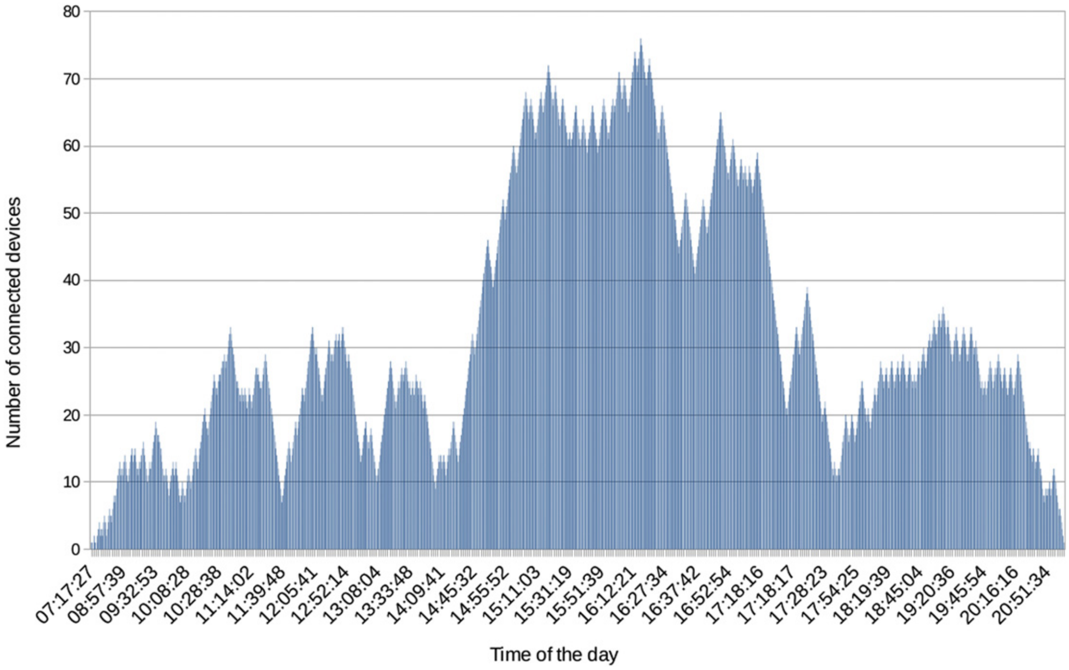

Fig. 1. Connection profile of a representative AP during one day.

Table 2
Importance of the components.

Components	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Portion of the variance	0.53	0.21	0.11	0.08	0.03	0.02	0.01	0.00	0.00	0.00	0.00
Cumulative variance	0.52	0.74	0.86	0.94	0.97	0.99	0.99	0.99	0.99	0.99	1.00

In the first family of indicators, at each AP, we want to characterize traffic or flow as it is referred to in the literature (Meneses & Moreira, 2012). To describe the density of users at one place, we compute the number of users connected to each access point as presented in Fig. 1. To describe the flow/variation of users at one place, we generate the primary and secondary derivatives of the number of connections through time. For each of the three signals generated, we use the mean, standard deviation, and amplitude as an indicator of attendance. Finally, we add the total number of connections and unique devices through the day.

Time is an essential component of human behavior and is highly correlated to the type of activity performed. We found the average connection time of users at the access point to be a decisive component of activity behavior. Detailed descriptive analysis of the used indicators can be found in Table B.7. A possible interpretation of the selected indicators is proposed in Table 1.

5.2. Principal component analysis

We summarize the results obtained by applying the algorithm for 1 day of the week in Table 2. The space reduction does not follow specific rules and has to be chosen after several iterations and adjustments. We observe that more than 95% of the variance is captured by the first 5 components. Thus, we used these 5 for the clustering. Our initial data is then expressed in the new space using the eigen vector generated by the algorithm.

The eigen vector Table 3 allows us to explore the structure of the space generated. We observe that the first component is linked to the number of connections and variation in the number of connections representing 50% of the variance, while the second component is linked to variation in connection times, representing 22% of the variance.

5.3. Clustering

The computational time being relatively small, we are able to compute the within group sum of squared error for each number of clusters r as shown Fig. 2. We had to find a compromise between the minimization of inner distance and reasonable number of clusters. The analysis of evolution of clusters' centroids along the week shows that 7 clusters allow having the best stability of characteristics over the week. It is also reflected in Fig. 2 where at $r = 7$, the slope begins to level out.

Fig. 3 shows APs, classified by the clusters obtained from the K-means analysis as a function of the first three components in the PCA subspace. At first glance, one can observe the presence of an extreme value far from the rest of the distribution, leading to a cluster containing a unique access point. This access point is characterized by a very high connection number and very short connection times. This cluster containing only one AP, represent a unique type of behavior within the infrastructure and is in fact the place presenting the highest flow of people. The comparison with the maps in Section 6 shows that this location is, in fact, the entrance of the building. Proximity of some clusters shows the limit of the approach: even if the locations show some characteristics and patterns allowing them to be classified as a "main activity", the gap between different types of APs can be small. Indeed, we considered in our study, that each place/access point would be associated with a unique main activity, which is a simplification of reality. However, the semantic meaning that we associate with these unique activities is generic and realistic enough to incorporate more disaggregate activities e.g. a *sharing place* activity at an access point can incorporate eating, chatting, waiting and other such detailed activities.

Table 3
Principal components contribution.

Indicators	PC1	PC2	PC3	PC4	PC5
Number of devices	−0.38	0.00	−0.11	0.23	−0.53
Number of connections	−0.38	0.03	−0.08	0.19	−0.49
Primary derivative standard deviation	−0.40	0.08	0.13	−0.10	−0.08
Primary derivative amplitude	−0.38	0.08	0.21	−0.20	0.28
Secondary derivative standard deviation	−0.40	0.11	0.17	−0.02	0.04
Secondary derivative amplitude	−0.38	0.08	0.23	−0.20	0.34
Average connection duration	0.04	0.51	−0.20	−0.56	−0.24
standard deviation of Connection duration	0.10	0.59	−0.25	−0.09	−0.02
Maximum connection duration	0.01	0.53	−0.05	0.56	0.29

Table 4 shows centroids in terms of indicators value for each cluster. As mentioned before, the variation of centroids through the week gave us 1 stable cluster containing a unique access point, four stable clusters representing more than 85% of the access points, and two unstable clusters. While the stable clusters are defined throughout the week, unstable clusters have to be analyzed for each day.

Cluster 1: *Entrance* This is an extreme point with an large flow of devices connecting to it briefly.
Cluster 2: *Offices* They have a very low number of connections, but are very regular in their behavior with low variation in the number of connections. The average connection time is the highest of the 7 clusters.
Cluster 3: *Classroom* These clusters are characterized by very strong and punctual variations (explaining the high standard deviation of the derivative). At the same time, the connection

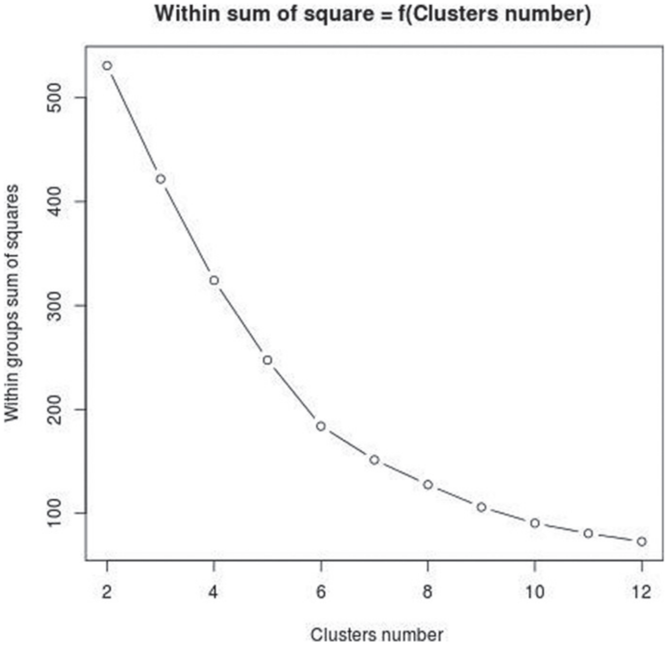

Fig. 2. Within sum of square of error versus number of clusters.

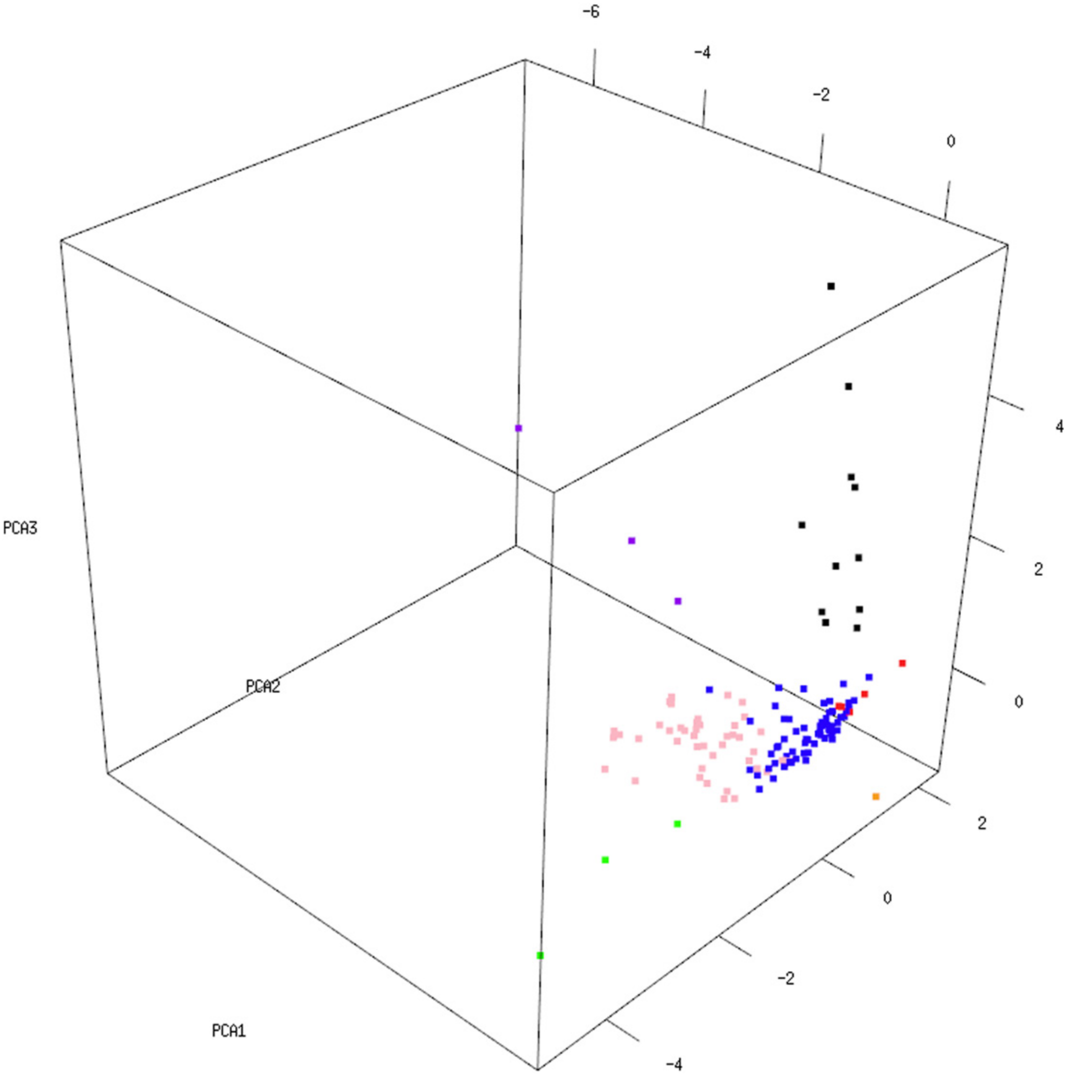

Fig. 3. Representation of the clusters as a function of first three principal components.

number stays relatively constant during the peaks: number of users vary before and after the class but not during the activity time interval.

Cluster 4: *Sharing place* These locations are public places shared by all users of the building such as cafeterias or common working spaces. They often have an increase in connection number during lunch times. The number of connections tends to vary in a smoothly as these locations fill and empty progressively.

Cluster 5: *Corridors* These access points are passing places, they are characterized by short connection times.

Cluster 6: *High frequency corridor (unstable)* This cluster looks similar to the corridors one with a higher number of people connected to it, and high standard deviations.

Cluster 7: *High-frequency sharing places(unstable)* This cluster is a hybrid between sharing places and classrooms.

5.4. Analysis of clusters over the course of a week

We present in Fig. 4 the evolution of the number of APs in the 4 stable clusters (excluding the *entrance* cluster, as behaviorally there

Table 4
Characteristics of each clusters. Values represent the centroid of each cluster.

Cluster	1	2	3	4	5	6	7
Semantic meaning	Entrance	Office	Classroom	Sharing places	Corridors	High corridors	High sharing places
Device number	1.00	0.04	0.09	0.03	0.02	0.35	0.06
Connection number	1.00	0.05	0.12	0.05	0.03	0.33	0.09
Primary derivative standard deviation	1.00	0.10	0.19	0.07	0.06	0.36	0.43
Primary derivative amplitude	1.00	0.09	0.19	0.07	0.06	0.36	0.43
Secondary derivative standard deviation	1.00	0.11	0.21	0.08	0.07	0.48	0.43
Secondary derivative amplitude	0.95	0.10	0.20	0.07	0.06	0.35	0.53
Average connection duration	0.12	0.55	0.31	0.36	0.22	0.21	0.38
Connection duration standard deviation	0.12	0.46	0.27	0.47	0.18	0.20	0.31
Maximum connection duration	0.36	0.43	0.38	0.70	0.19	0.32	0.32

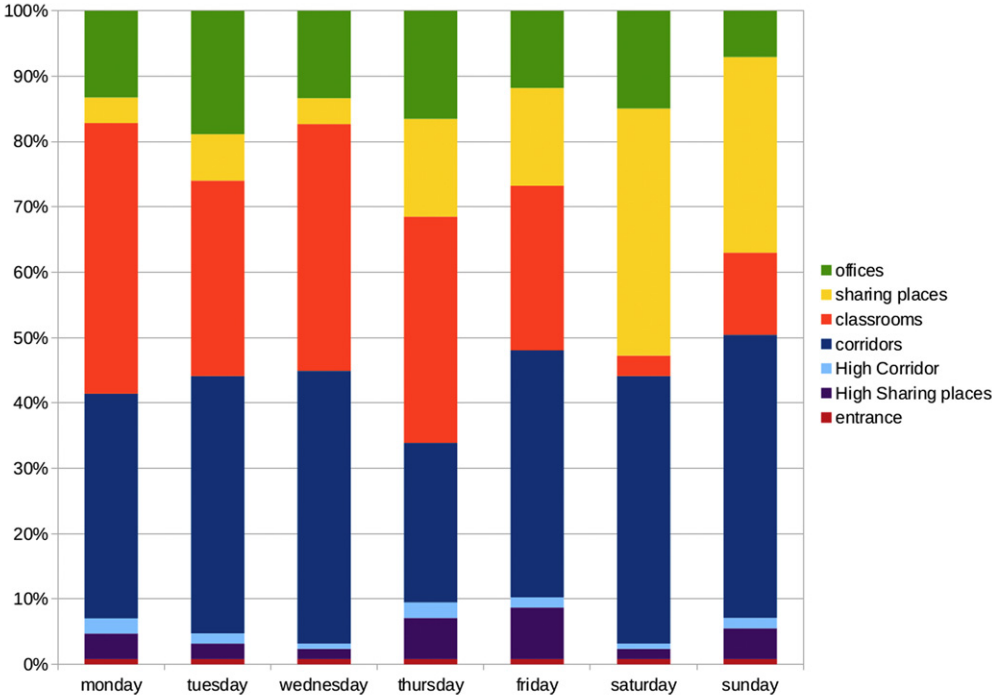

Fig. 4. Size of the clusters over the week.

is no interesting activity going on) over the week, using the semantic meaning associated with them in the previous section. We first observe the relative continuity of the cluster sizes over the week. It is also interesting to notice the difference in behavior between the weekdays and weekend. We observe that the number of classroom access points are decreasing drastically, while the portion of sharing places increases. This is representative of the real behavior of the students who often have classes during weekdays and work on their own during the weekend. The number of office clusters decreases while the number of corridor clusters increases.

While these results do not systematically ensure the validity of our results, they are representative of the general behavior of university population through the weekdays.

To conclude, in this section, we demonstrated the application of a methodology to extract detailed information about facility usage with only Wi-Fi Access Point logs available from Concordia University. In the next section, we'll bring space functionality related data to put the results obtained in perspective.

6. Comparison with the campus maps and limitations

In this section we compare the location of activities we inferred with the designated usage derived from the architectural plans of the building. In essence, we indirectly attempt to evaluate the accuracy of our methodology. Detailed spatial data at building level may not always be easy to access for obvious security reasons. As such, we developed our methodology so that such information would not be required. We did, however, want to be able to evaluate how good our methodology was at inferring activities by comparing them with activities that we could infer from the architectural plans of the building.

6.1. Actual activities inferred from building plans

The architectural plans of the university's buildings, allow us to locate access points on the campus and to then associate them to each part of the building using semantic labels. The plans were sufficiently accurate to allow us to identify four main space types: corridors, open public spaces, offices and meeting rooms. Examples of the geographic and semantic description of the spaces inferred are shown in Fig. 5.

The spatial location of access points allows us to analyze the environment surrounding them. Using the connection range for access points, we create buffers around each AP and compute the ratio of activity type occupation on each floor. Fig. 6 shows the summary results of this analysis. We can notice that classes are located on the 10th floor, while a big proportion of the offices are located above it. Corridors and sharing places are relatively equally distributed. We observe the peak of sharing places on the 10th floor. It seems that the building has two different parts in term of activities.

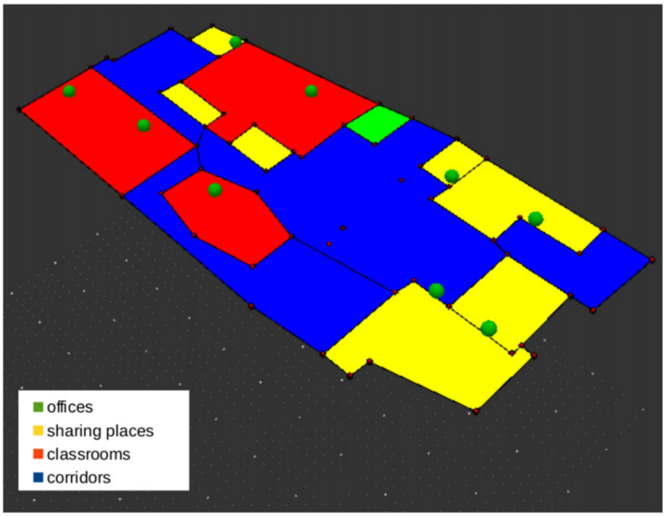

Fig. 5. Codification of the maps.

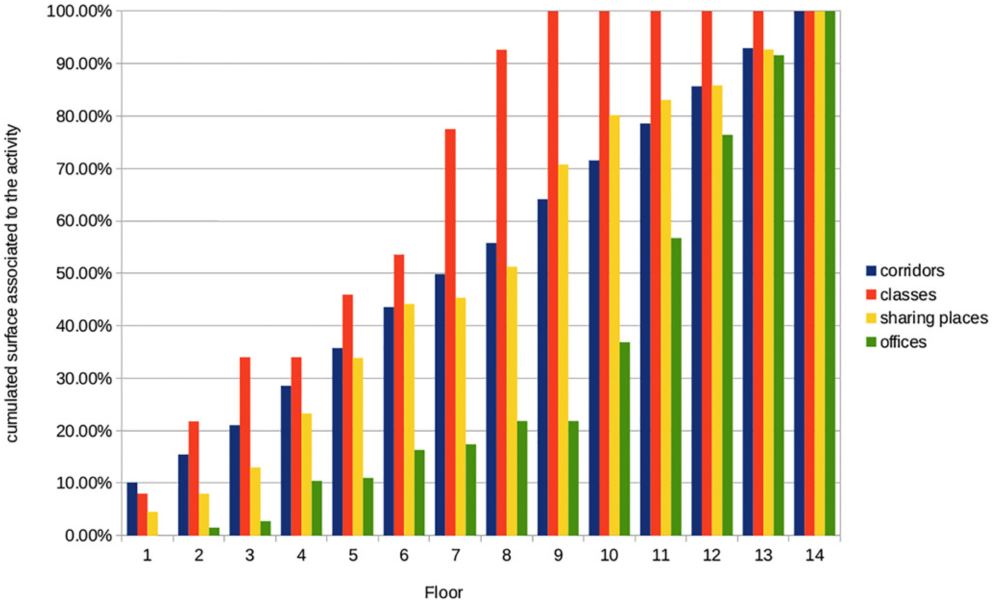

Fig. 6. Cumulated surface of the building.

It is important to note that the supply data we used here are not necessarily representative of the realized activities of users, but they do give us some idea on the possible distribution of actual activities. Indeed, the density of users performing an activity is not necessarily proportional to the space allocated for that given activity. We could use the example of classroom presenting a far higher density of devices than offices.

6.2. Comparison between supply and activity demand

The first approach to test the robustness of our results is to consider the accuracy of the distribution of activities in the building. Using clustering results shown in the previous section, we compare the inferred activities with those suggested by the architectural plans, aggregating both at the level of the floor. A summary of these results appears in Table 5. This level of aggregation is sufficiently fine to allow comparison between floor occupation and activities. In the table, we computed the correct or incorrect inference of the presence of each activity type, leading to a confusion matrix. Relatively good results obtained are encouraging concerning our ability to detect the different type of activities performed in the building. Results show two incorrect predictions for corridors and sharing places type, while the presence of classes and offices were correctly computed for all the floors.

The second approach is a comparison between the ratio of floor occupation and inferred access point activities. This level of aggregation leads to differentiation of the supply (space reserved for an activity) and the demand (activities performed by users). The results obtained for one day in the week and one in the weekend is presented in Fig. 7.

On Friday, as shown with the confusion matrix in the previous section, most of the activities on all the floors are detected correctly. However, we see a difference between the portion of space associated with an activity and the portion of activities inferred by the access point log analysis. Classroom activity tends to have a larger portion than the space associated to them around the access point. On that day, the ratio of the activities computed is relatively similar to the ratio of floor occupation.

On Sunday, the distribution of activities through the building is very different from the one observed during the week. Some parts of the building are not used at all, thus creating non observable places in our data. The portion of class activities has heavily decreased to let place to sharing places. Considering that there is no class on Sunday, we suppose that there are clusters of users working in the building, dense enough to reproduce the connection pattern of classrooms. The main part of the building hosts some low-density activities in sharing places or short connection times in corridors. These behaviors seem coherent with weekend activities and show how differently the infrastructure are used over time during the week.

We now present a disaggregated comparison between the activities computed for each APs and the environment surrounding them for the Friday. These environments are characterized here by the distribution of the type of spaces available around the assess point: the results are presented in Table 6. A first observation is the dominance of corridors around the APs which is explained by the global occupation of the floors presented in Fig. 7. We observe that the *High attendance corridors* and the *Entrance* are surrounded almost exclusively by corridors and sharing spaces. While the *High attendance sharing places* are mainly found around classes and corridors.

Table 5
Confusion matrix of the floors guesses.

		Real				
		Corridor	No corridor		Classes	No classes
Computed	Corridor	12	0	Classes	8	0
	No corridor	2	0	No classes	0	4
		Sharing places	No sharing places		Offices	No offices
	Sharing places	12	0	Office	12	0
	No sharing places	2	0	No office	0	2

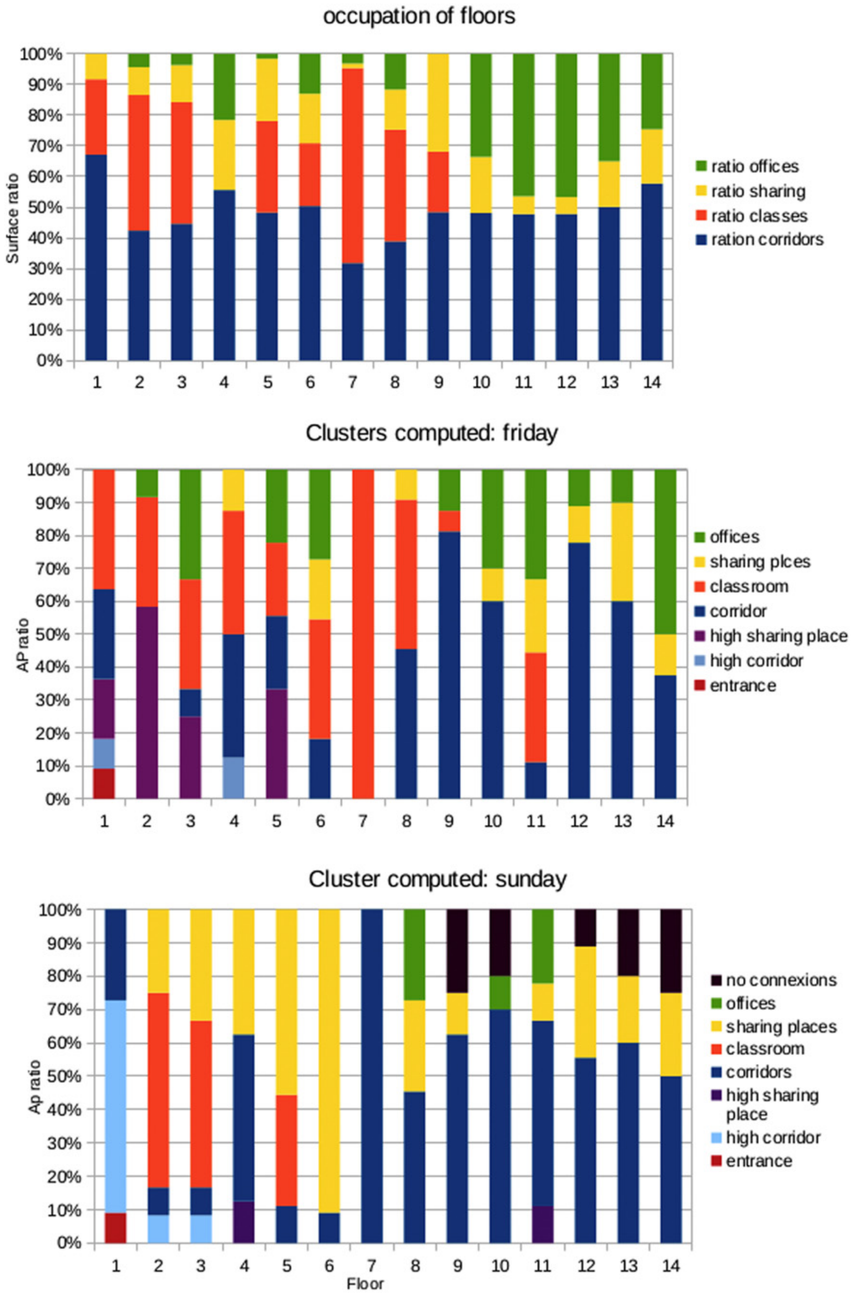

Fig. 7. Comparison between supply and demand.

These three cluster shows a clear relationship between the activity computed and their environment, which is also due to the low number of AP they contain. The four next clusters show a more heterogeneous environment. Classrooms and offices show a dominance of their respective associated space (making abstraction of the corridors). Corridors and Sharing places shows similar ratios of classrooms, offices and sharing spaces. This level of accuracy highlight a limitation brought by the computation of a single main activity per AP, as it doesn't take into account the heterogeneity of the activities in the places.

6.3. *Limitations and further work*

The indicators we used in our methodology allow a better identification of the clusters once computed, and as a result, a better understanding of the human behaviors around the APs. These indicators are based on the flow and temporal dimensions as they are observable within the raw data. It could be pertinent to add

Table 6
Average surface ratio around the APs.

Cluster of activity	Ratio of corridors	Ratio of sharing spaces	Ratio of classrooms	Ratio of offices
Entrance	72.83%	14.26%	12.92%	0.00%
High attendance corridor	56.57%	43.43%	0.00%	0.00%
High attendance sharing places	36.90%	3.07%	58.69%	1.35%
Corridor	44.37%	17.96%	15.95%	21.72%
Sharing place	42.77%	20.40%	16.18%	20.65%
Classes	45.00%	15.56%	31.03%	8.41%
Offices	48.72%	13.27%	4.90%	33.11%

new dimensions, as long as we could ensure they were not correlated with the previous one, or adjust for their correlation using techniques like PCA. Some possible variables to include in future work would be the indicator representing the importance of the place for the users, or the strength of the social connection existing between the people attending the same place.

The availability of validation data, describing the activities of a sample of users would allow robust validation of the results found. If the comparison with actual supply of the facility show sensible results, they cannot ensure the actual accuracy of the computed activity.

Thus, the results obtained here apply to a building in a campus, hosting defined types of activities. This context is very similar to the office environment, where our methodology could be applied. However, this process would not be appropriate for application to an open urban environment, proposing a huge set of activities, and a very limited knowledge of the users purpose and behavior.

7. Conclusion

In this paper we have proposed a methodology for identifying the main activities performed by users around wireless access points. The methodology is developed at a disaggregate scale both in terms of space (i.e. within building, and across floors) and time (i.e. over a day and along the week). The time period studied allows us to iterate the methodology over a week and compare the evolution of the distribution of activities over the days of the week. The clustering variables are computed from the connection log data for each location. These variables exhibit varying degrees of correlation among each other. For instance, these correlations are very high in the case of primary and secondary derivatives, while there is low correlation between the number of connections and average connection duration. We use a PCA-guided K-means clustering process to decrease the correlation between variables and improve the solution found by avoiding local minima. The results obtained allow us to classify the access points according the main activities taking place around them for each day independently. An optimization algorithm is then proposed for the week level analysis allowing us to see the evolution of different activity clusters between days. Analysis of the weekly results and the comparison to actual occupation of the facilities reveals various types of activities taking place around the access points and allows us to perform infrastructure usage analysis. We observed that the match between spatial distribution of activities computed and of building space usage vary depending on the day; for instance, the differences observed between Friday and Sunday. To be able to accurately validate our results, the use of building activity schedule seems necessary.

The methodology is based on a few restricting hypotheses, which may limit the validity and accuracy of our results. First, the association of a unique activity to each AP does not take into account the mixture of activities likely performed at a given location. A possible solution could be to create latent clusters of users within the population of people connected to an access point. Another restricting hypothesis is the time independence of activities through the day. In reality this is not the case, e.g. there is a typical time for eating and there is a difference between classes offered during the day and those offered at night. This leads to insensitivity of our current clustering to time constraints of the activities. Then, variables chosen to explain user behavior could be improved by adding social components (e.g. relationship between users) or the attachment to a place, for compensating the absence of socio-demographic data.

The development of the methodology in this paper allows us to take raw, rich, and readily available Big Data and develop an in-depth analysis of activity behavior. This therefore represents a great opportunity to furnish detailed and low-cost analytics. However, our methodology shows its limits when it comes to the accuracy of the activity computed: while Wi-Fi data furnished a better view of the global population of activities, surveys still furnish a more detailed description of the activities undertaken.

In recent years, there has been a major push from the Information and Communication Technology (ICT) industry to provide city level Wi-Fi services. For instance, Sidewalk Labs is installing a Wi-Fi network across New York City. In the UK, similar services are being offered by Smart Sidewalks. The growth of these urban large-scale ubiquitous networks presents us with great potential for the automatic monitoring of urban infrastructure usage and activity pattern analysis. While our work has not demonstrated the ability to infer activity information at the urban scale, it would still be able to furnish us with the ability to monitor activities at a more local and a finer grained local scale.

Acknowledgments

This research would not have been possible without the help of a number of people and institutions. First, we would like to thank Mike Babin, Myung Hong and Lichuan Yang of IITS at Concordia University for extracting the data from Wi-Fi network. Special thanks to Adam Taka, a Ph.D. student at LITrans, Polytechnique for developing the MAC address anonymization script for the preprocessing of Wi-Fi network logs. This research has been funded by FRQNT-New Researcher grant.

Appendix A. Buildings map

We provide the space distribution of the entire building in Fig. A.8

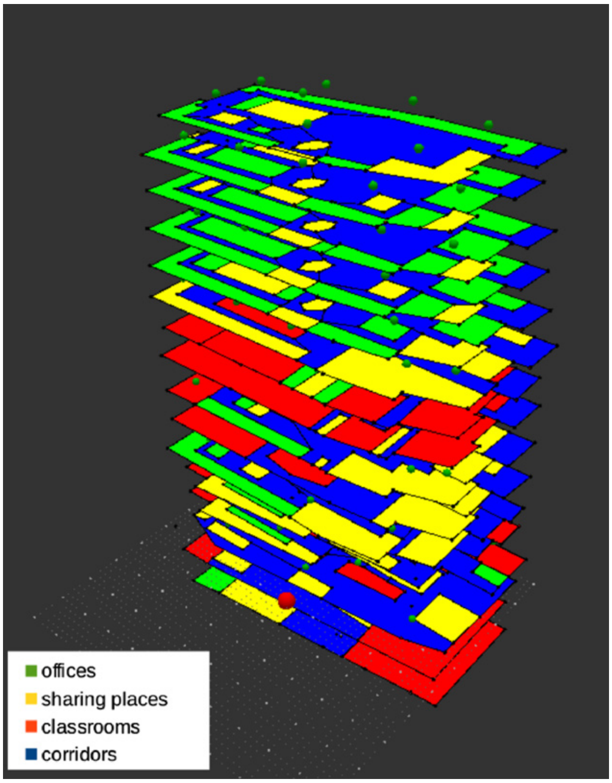

Fig. A.8. Building space distribution map.

Appendix B. Statistics

In Table B.7 we present the descriptive analysis of the used variables. While in Fig. B.9 we demonstrate the existence of high correlation among the variables.

Table B.7
Indicators statistics

Indicators	Code	Mean	Median	Standard deviation	Minimum	Maximum
Number of connections	X1	155.833	90.000	225.805	2.000	2090.000
Number of devices	X2	89.442	43.500	166.004	2.000	1553.000
Primary derivative SD	X4	0.358	0.234	0.382	0.000	2.400
Primary derivative amplitude	X5	3.942	2.000	4.626	0.000	25.200
Secondary derivative SD	X7	0.102	0.069	0.101	0.000	0.629
Secondary derivative amplitude	X8	1.088	0.600	1.250	0.000	6.720
Average connection duration	X9	27.682	26.125	13.648	5.017	84.400
SD of connection duration	X10	40.694	39.383	22.469	0.017	140.983
Maximum connection duration	X11	217.399	201.500	120.457	5.000	644.000

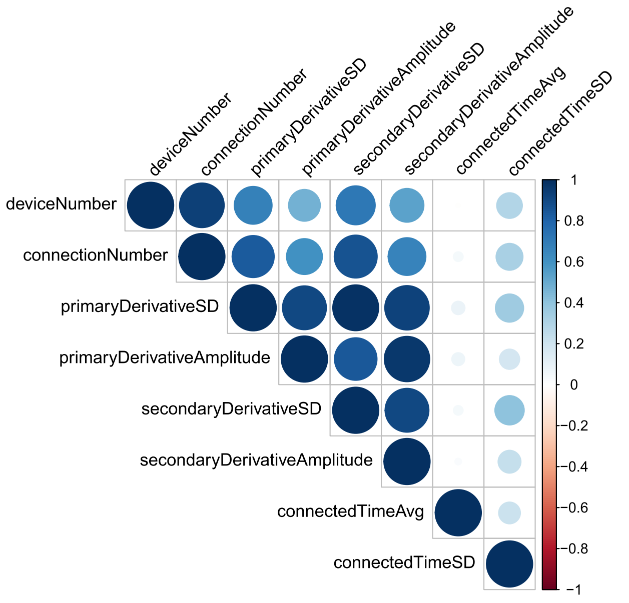

Fig. B.9. Correlation matrix.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433− 459.

Afanasyev, M., Chen, T., Voelker, G. M., & Snoeren, A. C. (2010). Usage patterns in an urban wifi network. *Networking, IEEE/ACM Transactions on*, *18*(5), 1359–1372.

Aschenbruck, N., Munjal, A., & Camp, T. (2011). Trace-based mobility modeling for multi-hop wireless networks. *Computer Communications*, *34*(6), 704–714.

Balazinska, M., & Castro, P. (2003). Characterizing mobility and network usage in a corporate wireless local-area network. *Proceedings of the 1st international conference on mobile systems, applications and services*. (pp. 303–316). ACM.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., & Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation research part C: emerging technologies*, *26*, 301 −313.

Calabrese, F., Lorenzo, G. D., & Ratti, C. (2010). Human mobility prediction based on individual and collective geographical preferences. *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE conference on. IEEE*. (pp. 312–317).

Calabrese, F., Reades, J., & Ratti, C. (2010). Eigenplaces: Segmenting space through digital signatures. *Pervasive Computing, IEEE*, *9*(1), 78–84.

Cao, G., Wang, S., Hwang, M., Padmanabhan, A., Zhang, Z., & Soltani, K. (2015). A scalable framework for spatiotemporal analysis of location-based social media data. *Computers, Environment and Urban Systems*, *51*, 70–82. http://www.sciencedirect.com/science/article/pii/S0198971515000162.

Conti, M., & Giordano, S. (2007). Multihop ad hoc networking: The theory. *Communications Magazine, IEEE*, *45*(4), 78–86.

Danalet, A., Farooq, B., & Bierlaire, M. (2014). A bayesian approach to detect pedestrian destination-sequences from wifi signatures. *Transportation Research Part C: Emerging Technologies*, *44*, 146 −170.

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on machine learning*. (pp. p.~29.). ACM.

Eagle, N., & Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, *63*(7), 10 57−1066.

Dang~Ba, V.Farooq, B., Beaulieu, A., & Ragab, M. (2015). Ubiquitous monitoring of pedestrian dynamics: Exploring wireless ad hoc network of multi-sensor technologies. *Sensors, 2015 IEEE*. (pp. 1–4). IEEE.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 2 43−263.

Frick, J. R., & Grabka, M. M. (2005). Item nonresponse on income questions in panel surveys: Incidence, imputation and the impact on inequality and mobility. *Allgemeines Statistisches Archiv*, *89*(1), 49–61.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782.

Grapperon, A., Farooq, B., & Trépanier, M. (2016). Activity based approach to estimation of dynamic origin-destination matrix using smartcard data. *TRISTAN IX*. (pp. 1–4).

Henderson, T., Kotz, D., & Abyzov, I. (2008). The changing usage of a mature campus-wide wireless network. *Computer Networks*, *52*(14), 2690–2712.

Iqbal, M. S., Choudhury, C. F., Wang, P., & González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, *40*, 63–74.

Izakian, Z., Mesgari, M. S., & Abraham, A. (2016). Automated clustering of trajectory data using a particle swarm optimization. *Computers, Environment and Urban Systems*, *55*, 55–65. http://www.sciencedirect.com/science/article/pii/S0198971515300302.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, *31*(8), 651−66 6.

Jiang, S., Ferreira, J., & González, M. C. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, *25*(3), 478− 510.

Kang, J. H., Welbourne, W., Stewart, B., & Borriello, G. (2005). Extracting places from traces of locations. *ACM SIGMOBILE Mobile Computing and Communications Review*, *9*(3), 58–68.

Katsaros, D., Nanopoulos, A., Karakaya, M., Yavas, G., Ulusoy, Ö., & Manolopoulos, Y. (2003). Clustering mobile trajectories for resource allocation in mobile environments. *Advances in intelligent data analysis v*. (pp. 319–329). Springer.

Kusakabe, T., & Asakura, Y. (2014). Behavioural data mining of transit smart card data: A data fusion approach. *Transportation Research Part C: Emerging Technologies*, *46*, 179 −191. http://www.sciencedirect.com/science/article/pii/S0968090X14001612.

Liu, L., Andris, C., & Ratti, C. (2010). Uncovering cabdrivers behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, *34*(6), 541−54 8. GeoVisualization and the Digital CitySpecial issue of the International Cartographic Association Commission on GeoVisualization. http://www.sciencedirect.com/science/article/pii/S0198971510000773.

Long, Y., & Thill, J.-C. (2015). Combining smart card data and household travel survey to analyze jobs-housing relationships in Beijing. *Computers, Environment and Urban Systems*, *53*, 19–35. Special Issue on Volunteered Geographic Information. http://www.sciencedirect.com/science/article/pii/S0198971515000356.

Mao, G., Fidan, B., & Anderson, B. D. (2007). Wireless sensor network localization techniques. *Computer networks*, *51*(10), 2529–2553.

Meneses, F., & Moreira, A. (2012). Large scale movement analysis from wifi based location data. *Indoor Positioning and Indoor Navigation (IPIN), 2012 International Conference on*. (pp. 1–9). IEEE.

Munizaga, M. A., & Palma, C. (2012). Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile. *Transportation Research Part C: Emerging Technologies*, *24*, 9− 18.

Naini, F. M., Dousse, O., Thiran, P., & Vetterli, M. (2011). Population size estimation using a few individuals as agents. *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*. (pp. 2499–2503). IEEE.

Nguyen-Vuong, Q.-T., Agoulmine, N., & Ghamri-Doudane, Y. (2007). Terminal-controlled mobility management in heterogeneous wireless networks. *Communications Magazine, IEEE*, *45*(4), 122–129.

Ortúzar, J. d. D. S., & Willumsen, L. G. (2011). *Modelling transport*. John Wiley & Sons..

Patterson, Z., & Fitzsimmons, K. (2016). DataMobile: A smartphone travel survey experiment. *Transportation Research Record*, *2594*, 35–43.

Prentow, T. S., Ruiz-Ruiz, A. J., Blunck, H., Stisen, A., & Kjærgaard, M. B. (2015). Spatio-temporal facility utilization analysis from exhaustive wifi monitoring. *Pervasive and Mobile Computing*, *16*, 305−3 16.

Ramani, I., & Savage, S. (2005). Syncscan: Practical fast handoff for 802.11 infrastructure networks. *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE. Vol. 1*. (pp. 675–684). IEEE.

Richardson, A. J., Ampt, E. S., & Meyburg, A. H. (1995). *Survey methods for transport planning*. Eucalyptus Press Melbourne..

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018–1021.

De~Lara, E.Su, J., Chin, A., Popivanova, A., & Goel, A. (2004). User mobility for opportunistic ad-hoc networking. *Mobile Computing Systems and Applications, 2004. WMCSA 2004. Sixth IEEE Workshop on*. (pp. 41–50). IEEE.

De~Lara, E.Su, J., & Goel, A. (2006). An empirical evaluation of the student-net delay tolerant network. *Mobile and Ubiquitous Systems: Networking & Services, 2006 Third Annual International Conference on*. (pp. 1–10). IEEE.

Su, W., Lee, S.-J., & Gerla, M. (2001). Mobility prediction and routing in ad hoc wireless networks. *International Journal of Network Management*, *11*(1), 3—3 0.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A.-L. (2011). Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. (pp. 1100–1108). ACM.

Wymeersch, H., Lien, J., & Win, M. Z. (2009). Cooperative localization in wireless networks. *Proceedings of the IEEE*, *97*(2), 427–450.

Xu, Q., Ding, C., Liu, J., & Luo, B. (2015). Pca-guided search for k-means. *Pattern Recognition Letters*, *54*, 50–55. http://www.sciencedirect.com/science/article/pii/S0167865514003675.

Yoon, J., Noble, B. D., Liu, M., & Kim, M. (2006). Building realistic mobility models from coarse-grained traces. *Proceedings of the 4th international conference on Mobile systems, applications and services*. (pp. 177–190). ACM.

You, C.-w., Chen, Y.-C., Chiang, J.-R., Huang, P.-Y., Chu, H.-h., & Lau, S.-Y. (2006). Sensor-enhanced mobility prediction for energy-efficient localization. *Sensor and Ad Hoc Communications and Networks, 2006. SECON'06. 2006 3rd Annual IEEE Communications Society on*. *Vol. 2*. (pp. 565–574). IEEE.

Zahabi, A., Ajzachi, A., & Patterson, Z. (in press). Transit trip itinerary inference with general transit feed specification and smartphone data. Accepted in Transportation Research Record.

Zmud, J., Lee-Gosselin, M., Munizaga, M., & Carrasco, J. A. (2013). *Transport survey methods: Best practice for decision making*. Emerald Group Publishing Limited..