UNIVERSITÉ DE MONTRÉAL

URBAN ACTIVITY PATTERNS MINING IN WI-FI ACCESS POINT LOGS

GUILHEM POUCIN
DÉPARTEMENT DES GÉNIES CIVILS, GÉOLOGIQUE ET DES MINES
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE CIVIL)
MARS 2017

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

URBAN ACTIVITY PATTERNS MINING IN WI-FI ACCESS POINT LOGS

présenté par : POUCIN Guilhem
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. BILODEAU Guillaume-Alexandre, Ph. D., président
M. FAROOQ Bilal, Ph. D., membre et directeur de recherche
M. PATTERSON Zachary, Ph. D., membre et codirecteur de recherche
M. YU Jia Yuan, Ph. D., membre

DEDICATION

A toutes les personnes dont j'ai croisé le chemin,
A ma famille et mes amis aux quatres coins du monde,
A la princesse slave et au boulanger philosophe,
C'était cool de vous voir..

AKNWOLEDGMENTS

I want to thank my supervisor Bilal Farooq and Zachary Patterson for their support during this experience. Their patience and knowledge allowed me to complete this modest work and learn on a personal point of vue.

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.

Finally, I want to thank my friends and family for their unconditional support during this experience far from my native home.

RÉSUMÉ

Aujourd'hui la grande majorité des données sont basée sur des enquêtes ou des études appliquées à des échantillons définis de la population. De plus les méthodes traditionnelles de collecte de données en termes de coûts ainsi que de temps tout en ne garantissant pas la représentativité des observations du fait du biais d'échantillonages et de la relative fiabilité des répondants.

La disponibilité grandissantes de bases de données collectées passivements couplé à la forte pénétration des smartphones ont ouvert des perspectives intéressantes concernant la collecte et le traitement automatisé de données de mobilité.

L'utilisation de données provenant de réseaux omniprésents tels que les réseaux de communication téléphoniques ou Wi-Fi est un domain en cours de développement. Des cas d'étude ont été déeveloppés à differentes éechelles, qu'il s'agisse d'un batîment, d'un campus ou d'une ville. Des études récentes tentent de fusionner ces données avec d'autres sources dìnformation, ce processus se révélant compliqué, spécialement dans les cas désagrégés.

La Media adress Control (adresse MAC) est un identifiant unique propre à chaque appareil permettant de les identifier au sein dùn réseau. Cette adresse reste fixe à vie pour chaque appareil est permet ainsi de retracer leurs historiques de connexion au sein des bases de données Wi-FI. En associant les positions des bornes d'accées Wi-FI, généralement fixes, ont peut donc retracer les positions successives de l'utilisatuer a lorsque son appareil est connecté au réseau.

Nous proposons au travers de ce travail de classifier les activitées réalisées par les utilisateurs du réseau Wi-Fi, à une petite échelle (batîment) sans conaissance préalable de la géographie des infrastructures (localisation spatiale des points d'accés). Pour cela , nous nous appuyons sur des algorithmes d'extraction de données en utilisant la méthodologie des K-moyennes guidées par analyse en composantes principales. Les résultats obtenus, générés l'échelle temporelle d'un jours sont ensuite mis en perspective à l'échelle de la semaine. Nous utilisons finalement les données géospatiales afin de valider les résultats obtenus.

Travaux antérieurs

L'études des données issues des données de communications tels que les réseaux téléphniques, le GPS ou les réseaux de Wi-Fi à reçu une attention grandissante durant les dernières décénies. L'étude de la mobilité des utilisatuer se fait au travers de l'aquisition de leurs trace

qui peuvent être obtenues de trois manières différentes : la surveillance de leurs localisation, de leurs communications ou encore de leurs contact. LA surveillance de la localisation, majoritairement effectuées au travers u GPS consiste à enregistrer les positions successives de l'appareil grâce à son propre outils de géolocalisation : cette méthode est considérée comme centrées sur l'appareil. La surveillance des communications consite à reconstituer l'historiques des connectiosn effectiuées par l'appareil et des points d'accés concernés : cette méthode est considérée comme centrées sur le réseau.Enfin, la surveillance des contacts, encore en cours de développement consiste à enregistrer l'ensemble des appareils rencontrées par l'utilisateur.

Si les données collectées au travers de la surveillance des communications des appareils présente l'avantage de fournir une base de données contenant un écahntillon important de la population à moindre coût, certains probles doivent être surmontés. Le première obstacle découle de la nécessité d'anonimisation de ces bases de données pour des questions de sécurité de la vie privée : la base de données ne contient généralement aucune données socio-démographiques concernant les utilisateurs. Il devient alors difficile de caractériser l'échantillon de population étudié, notamment dû a la variation du nombre d'appareil possédés par chaques utilisateurs. Ensuite, une partie des compotements des utilisateurs n'est pas observable, à partir du moment ou l'utilisateur n'est plus connecté au réseau : différencier l'absence du batiment de l'extinction de l'appareil est compliqué. Enfin, l'absence de mesure de position géographique peut creer l'apparition d'erreur ou bruit connu sous le non d'effet Ping Pong, avec l'apparitions d'enregistrement non représentatif d'événement réels.

Un travail considérable a été effectué dans le domaine du traitement des données de connections Wi-FI. A l'origine destiné l'optimisation du design et de l'efficacité des réseaux de Wi-FI dans les lieux à forte fréquentation, le doamine d'application s'éetend aujourd'hui à l'étude de la mobilité des usagers. Divers process de pré traitement de ces données ont été proposés, notamment concernant la diminution de l'impact de l'effet Ping Pong, au travers de processus d'aggrégation.

Un travail a été effectué, proposant de classer les différents types d'activité réalisés autours des points d'accés WI-FI. En décomposant en élément propres la matrice de l'évolution du nombre d'usagers connectés en fonction du temps, puis en applicant l'algorithme des K moyennes, 5 familles d'activités sont générées. Nous proposons dans notre travail une alternative à la décomposition en éléments propre, permettant en autre d'incorporer les notions de temps de connections pour les utilisatuers. De plus, à l'instar de l'étude mentionnée, nous n'utilisons aucun données géospatiales lors de la calibration de l'algorithme, afin de limiter les données nécessaires.

Données et et traitement des erreurs

Notre méthodologie s'appuie sur le cas d'étude de l'université de Concordia (Montréal, Canada). La base de donnée concerne les connections au réseaux Wi-Fi de l'université sur une semaine, enregistée en février 2015. La base comporte environs 1.7 millions de connections effectuées par plus de 60.5000 appareils intelligents. Les appareils sont identifiés à l'aide de leurs MAC adress, permettant ansi de retracer leurs historique de connections au cours de la semaine.

La base de données comporte des erreurs qui doivent être corrigées avant tout traitement.

- L'erreur de mesure la plus courrante est le champ manquant de certains enregistremet au sein de la base de donnée. Dans notre cas 0.08% des enregistrement ne sont associés à aucune borne d'accès Wi-Fi. Considérant la faible portion de points d'accés concernés, nous remplissons les champ manquant avec une valeur générique "inconnnue".
- Une seconde limitation de notre base de donnée est l'aggrégation temporelles de certaines connections d'une durée inférieure à cinq minutes. Ce biais de mesure ne demande pas de correction mais necessite d'être pris en compte lors de l'analyse de données. De plus, ce phénomène rend difficile l'analyse des trajectoires des usagers au sein des infrastructures, le chemin emprunté (succession de bornes Wi-Fi) étant incomplet.
- Enfin, une erreur souvent mentionnée dans la litterature concernant les bases de données de connection Wi-Fi et "l'effet Ping-Pong" décrit prcédemment. Nous diminuonss ce bruit au sein de la base de donnée en détectant et aggrégeant les connections de durrées minimales et répétitives.

Méthodologie

Nous proposons une méthodologie s'appuyant sur la générations de variables caracteristiques des comportemments des usagers autours des points d'accès durant la journée, et le clustering de ces dernières afin de déterminer les groupes de comportement similaires. Les variables utilisée sont regroupées en deux familles principales : les variables liées aux fluctuations des connexions et celles liées au temps.

Afin d'améliorer les performaneces de l'algorithme de clustering, nous utilisons une analyse en composantes principales afin de réduire la dimension de l'espace des variables étudiées. Du fait du caractère arbitraire des variables générées, l'analyse des composantes principales afin de déterminer un sous espace de variables montrant une corélation moindre. Ces composantes sont ensuite utilisées pour catégoriser les bornes Wi-Fi, à l'aide de l'algorithme des K-moyennes. Le nombre de clusters utilisé dans l'algorithme est déterminé de manière empirique afin d'optimiser la solution trouvée tout en fournissant un nombre de groupe représentatif de la diversité des activitées réalisées au sein de l'infrastructure. La qualité de définie quantitativement par le calcul de du rapport entre distances internes des clusters et les distances externes qui doit etre minimisé.

Une fois la calibrations de l'algorithme de partitionement réalisée et les clusters générés, nous associons une activitée à chaque famille de points d'accès en nous appuyant sur les indicatuers précédemment générés. Cette association n'est pas automatisée et s'appuie sur la connaissance des activitées potentiellement réalisées au sein de l'infrastructure étudiée, à savoir une université pour notre cas d'étude. Le choix d'appliquer l'algorithme de partitionement à l'échelle d'une journée plutôt que d'une semaine permet d'observer l'évolution des activitées générées au cours de la semaine. Afin de pouvoir mettre en parrallèle les clusters générés pour chaque jour, nous utilisons un algorithme d'optimisation de distance afin de trouver les cluster similaires durant la semaine.

Résultats et comparaison aux infrastructures

L'analyse en composantes principale permet de réduire le nombre de variables étudiées au cinq composantes principales représentant 95% de la variance globale pour chaque jour de la semaine. L'algorithme de clustering est ensuite appliqués sur ces composantes afin de généré nos clusters d'activité. Les itérations successives en changeant en faisant varier le nombre de clusters en paramètre donnes des résultats optimals pour sept clusters.

Ces sept clusters sont associés a des activités en utilisant la connaissance préalable des activités réalisées au sein des infrastructures. Parmis ces activités, on trouvera un cluster ne comportant qu'un unique point d'accés associé à l'entrée dans le bâtiment. On lister ensuite quatre familles majeures représentant 90% des points d'accés, associées aux activitées de classes, passage, attente et bureaux. Les deux dernières familles correpondent enfin à des lieux de passage et de classes très fréquentés. L'analyse de l'évolution des clusters durant la semaine permet d'observer une relative stabilitée des résultats obtenues entre les jours de semaine avec une nette différentiation pour les jours de week end.

Afin de tester la validité des résultats générés au cours de notre méthodologie, nous comparons les clusters d'activité calculés avec l'occupation des sols des infrastructures du campus étudié. Si une telle comparaison permet de vérifier la cohérences des résultats obtenus, il ne permettent pas une validation rigoureuse du modèle, qui necéssiterait des données d'activitées réelles. Cependent ils permettent de comparer l'tilisation des lieux avec leurs fonctions.

Une première comparaison agrégée par étage nous permet d'évaluer la capacité de notre modèle a identifier l'amenagement du bâtiment. Nous comparons les types d'activités trouvées pour chaque étage avec les espaces disponibles en utilisant les ratios de points d'accès comparés aux ratios de surface disponible. On remarquera que les répartitions des activités et des surfaces aménagées sont relativement similaires en semaine, mais divergent durant les jours de week end.

Une comparaison désagrégée est ensuite proposée, mettant en parrallèle les types de locaux environant chaque point d'accès avec l'activité trouvée. Si les résultats trouvés sont cohents, ils montrent les limites de notre mode ainsi que les limites de l'assotiation d'une unique activité à un un espace acceuillant des activitées hétérogènes.

Approfondissement et limitations

Si notre travail présente des résultats cohérents, les résultats obtenus souffrent certaines limitations dues à la qualité des données ainsi qu'à certaines hypothèses inhérentes à la méthodologie.

La première limitation de notre travail est due à la qualité e la base de données utilisée. La présence de d'erreurs au sein de la base sous forme de champs manquant implique la non prise en compte de l'ensemble des phénomènes mesurés. L'impact de cette erreur reste toutefois limité du fait de la faible portions d'enregistrements concernés. Ensuite, l'aggrégation de l'ensemble des connections ayant une durée inférieure à cinq minutes crée aussi un biais au sein des donnée, créant de mauvaises durées de connections ainsi que des connections intermédiares manquantes.

Une seconde limitation est liée à la nature des données de Wi-Fi, créant un biais entre les phénomènes physiques observés (déplacements d'individus et réalisation d'activitées) et la mesure. Les données de connections, puisqu'utilisant la MAC adresse des appareilles afin d'dentifier la connection, sont liés à des appareils et non à des individus. De ce fait, la variabilité du nombre d'appareils connectés dépendamment des individus rend difficil l'évaluation précise des déplacements réels des individus. A cela s'ajoute le bruit induit par l'effet Ping Pong au sein de la base de donnée, créant potentiellement des déplacements fictifs.

Ensuite, le fait d'associer une unique activité à chaque point d'accès limite la précision des résultats. En effet, les bornes Wi-Fi couvrent de larges surfaces présentant des zones d'activités différentes (couloirs, salles de cours, bureaux). A cela s'ajoute le caractère hétérogène des activitées réalisées au sein d'un espace. L'aggrégation à l'échelle de la journée est aussi une simplification des comportements réels observés au sein d'une université.

ABSTRACT

This thesis proposes a methodology to mine valuable information about the usage of a facility (e.g. building), based only on Wi-Fi network connection history. Data are collected at Concordia University in Montreal, Canada, during one week in Febuary 2015. Using the Wi-Fi access log data, we characterize activities taking place within a building without any additional knowledge of the building itself. Such information can be used to monitor the use of a facility automatically, to study human mobility or as an input information for mobility models.

The methodology is based on the identification and generation of pertinent variables linked to the principal components of users mobility. Principal Component Analysis (PCA) based clustering (PCA-guided clustering) is used on these variables for the identification of main activities in space-time. A K-means clustering algorithm is then used to construct 7 activity types associated with a campus context. 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 are then compared (based on building plans) to the actual uses of the building. We found the computed distribution of the activities to be relatively similar to the distribution of dedicated spaces within the building during the week. We give an example with the variation in the building's usage during week-end, when activities differ more from the primary purpose of the places, e.g. classroom used as working/waiting places.

This work differentiates itself from the existing body of literature at various levels. We study a smaller scale area, on a smaller time period, increasing the need for accuracy in the clustering. In our definition of the influence variables of activity, we take into account other parameters than what has been used in the literature, e.g. the fluctuation in the connection number through time, as connections duration and users habits. Finally, we decided to exclude geo-location data from the input parameter, to obtain a methodology as general as possible. This way, the activities can be mined solely from the Wi-Fi signals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFS	Automated fare system
AP	Access point
CAD	Computer-aided design
CEMDAP	Comprehensive Econometric Model of Daily Activity Pattern
DM	Data mining
GIS	Geographic information system
GPS	Global positioning system
GSM	Global system for mobile communication
ICT	Communication Technology
ID	Identifier
IEEE	Institute of electrical and electronics engineers
KDD	Knowledge discovery
LAN	Local area network
MAC	Media access control
MB	Mega byte
ML	Machine learning
OD	Origin destination
PCA	Principal component analysis
PCATS	Prism-constrained activity simulator
SQL	Structured Query Language
TRB	Transportation research board
UK	United Kingdoms
Wi-Fi	Wireless fidelity
WLAN	Wireless local area network
WSS	Within sum of square

APPENDICES

CHAPTER 1 INTRODUCTION

An important challenge of transportation study is the characterization of population mobility through the treatment and analysis of different datasets (e.g. census, travel demand survey, GPS trajectories, smart-cards, etc.). The knowledge of this mobility behaviors can then be used for planning, design, and operations of infrastructure. A large amount of traditional methodology for collecting data on mobility involves the study of a small sample of a characterized population (e.g. OD survey conducted by Montreal every five years is a 5% sample). These surveys are usually expensive in terms of time and direct costs, and are based on the hypothesis that the sample is representative of the whole population. Opposite to these methodologies are the study based on the large ubiquitous sets of data such as the smart-card data, cell phone or Wi-Fi communications. However, their treatment to transformm raw data into useful informations concerning mobility is still a challenge.

1.1 Definition and Concepts

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 et al. (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 survey, time tables etc.) to overcome such issues (Grapperon et al., 2016; Danalet et al., 2014; Calabrese et al., 2010a). However in many cases, it is very difficult to access such data, especially at a very disaggregate level.

1.1.1 The Wi-Fi log data

The IEEE 802.11 standard or Wi-Fi (wireless fidelity) is the most popular protocol for local area network (LAN) today (Femijemilohun and Walker, 2013). It is composed of a set of standard allowing to implement a wireless local area network (WLAN) using a physical layer, to handle the transmission of data, and a medium access control layer, to keep order in the usage and distribution data (Cafarelli and Yildiz, 2004).

The Media Access Control (MAC) address is a unique identifier associated with each network interface of a device (e.g. smart phone, laptop etc.) 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 network interface. 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, for instance, a connection to the Internet. APs are spatially distributed covering large areas (e.g. campus, shopping centers, etc.) and collectively comprise a Local Area Network. Our methodology only uses communication between devices and APs over the LAN to develop traces of people over time and space.

Following the IEEE 802.11 standard, each connected device scan periodically its environment to inventory the available AP. The device sends a probe request and wait, listening the eventual probe responses from the APs around, informing it on the quality of the connections available. The device can then send a request to associate itself with one of the AP, or to re-associate itself to a new one if it is already connected to the network. If the device authenticates successfully, a connection bridge is set between the device and the AP. The re-association process is performed when an available AP presents a better connection than the current one. The device can send a disassociation request to the AP when the signal is too weak or the device owner disable the WI-Fi connectivity. These communication processes are recorded in databases of APs, informing us on the time and location, we refer to them in the following as "connections".

The only spatial information in these records is the name of the AP to which the device was connected to, that we refer further as the "location" as show Table 1.1. However, this spatial information is relatively weak at the scale of a building, allowing us to associate the connection to an area rather than to a spatial point. Contrarily to our case, some Wi-Fi data sets contain the signal strength, improving the potential spatial location of the users (Farooq et al., 2015). Working with a less extensive dataset increase the transferability of our methodology.

When it comes to the description of the AP areas, we considered their range zone to be fixed circular buffers as shown in Figure 1.1. This hypothesis doesn't take into account the non-uniform signal attenuation brought by the walls within the floor. This hypothesis is a

Table 1.1 Format of the input data

Fields	ID	Association time	Location	Duration	Disassociation time
Type	Integer	Time stamp	String	Time	Time stamp

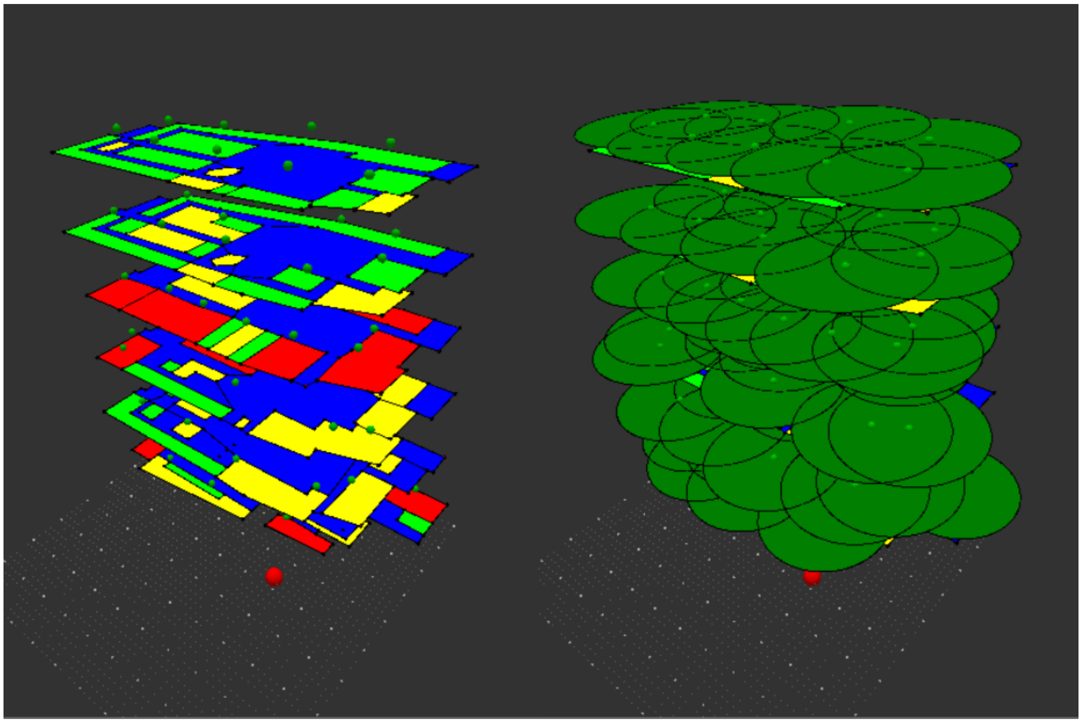

Figure 1.1 Access Points range area

simplification of reality as the devices usualy connect ot the closest AP, but facilitated the calcualtions. Using voronoi Poligons instead of circular buffers could increase the acuuracy of the model. We consider that the floor thickness forbid a connection to another floor's AP.

1.1.2 Demand : activity-based behavior

While the concept of mobility is extremely extensive, here we focus on the notion of activity performed by the users. In fact, activity-based behavior study is based on the assumption that the choice of a trip destination is mainly driven by the activity performed at this location. Our methodology is network centric, focusing on the infrastructure use description rather than the individual mobility behavior. In consequence, even if passing through a location is not the purpose of one's trip, it represents an activity performed and recorded in our database that we consider pertinent to characterize the behavior of users in a given area.

1.1.3 Supply : Concordia university infrastructures

In the study, we test our methodology using the case Concordia University in Montreal, Canada. To facilitate the treatment and the description, we choose one building as a representative case study. The previous knowledge of the building shows a pertinent diversity in the type of places found in the building : entrance hall, corridors auditoriums, cafeteria offices etc. This selected infrastructure also shows one of the highest attendance among the 57 buildings of the campus. While we can show some fuzzy maps and visualization of the building, the name and information concerning the building are hidden for security purposes.

1.1.4 Knowledge discovery methodology

The improvements performed in the last decades in computers' storage capacity and automated data collection processes have increased drastically the databases' sizes. These new sizes made difficult the extraction of useful information and insufficient the manual data analysis. This need in automated data treatment lead to the important development of the knowledge discovery (KDD) field, grouping methodologies from various disciplines such as statistics, machine learning (ML) or visualization (Kononenko and Kukar, 2007).

Definitions

The KDD process aims to extract useful information from data. In other terms, KDD generate high-level information (understandable, summarized and global description) called knowledge from low-level data (large amount of very specific measures or values) that could not decently be analyzed manually. Furthermore, KDD field also includes the pre- and post-processing of the data e.g. storage and access of the data, scale of the algorithms for large databases and interpretation or visualization of the results (Kononenko and Kukar, 2007).

ML is the field investigating the methods to make computers learn or increase their performances using data. A main component of the ML is to allow computers to perform their learning process automatically from the data. Despite this automation, a certain degree of supervision can exist to improve the performances of the algorithms : this degree is used to classify the methodologies (supervised, unsupervised, semi-supervised and active learning). An exhaustive discussion about these concepts can be found in Han et al. (2011).

Data mining (DM) is a sub-field of KDD using ML algorithm to extract useful information and patterns from data. In fact, DM is one of the steps of the KDD process as show Figure 1.2. However, the term DM is sometimes used to refer to the KDD field (Kononenko and Kukar, 2007).

KDD process

KDD is an iterative process composed of few steps, each one involving decisions and analysis from the user to calibrate the methodology. A possible decomposition of the process is shown Figure 1.2, identifying five main steps as proposed in (Fayyad et al., 1996).

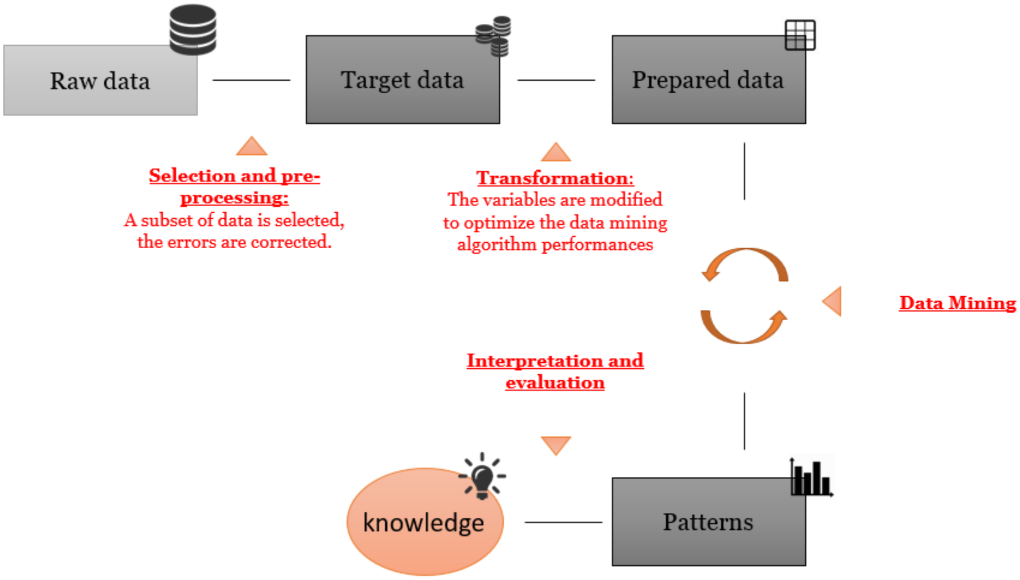
Raw data
Target data
Prepared data
Selection and pre-processing:
A subset of data is selected, the errors are corrected.
Transformation:
The variables are modified to optimize the data mining algorithm performances
Data Mining
Interpretation and evaluation
knowledge
Patterns

Figure 1.2 KDD process

- **Selection :** a subset of the data is selected to perform the knowledge discovery process.
- **Preprocessing :** strategies are chosen to handle the potential inner errors of the dataset : correction of the noise and handling of the missing fields.
- **Transformation :** exploration of the different ways to represent the data depending on the target of the KDD process. A space reduction or transformation can be performed to decrease the number of variables and optimize the DM process.
- **Datamining :** A DM algorithm is then chosen, parametrized and applied, depending on the transformed data obtained and the goal of the process.
- **Interpretation/evaluation :** Finally, the result of the algorithm processing has to be analyzed and interpreted considering their field of application, to furnish understandable knowledge.

All of these steps can be repeated to improve the calibration of the process.

KDD purposes

The field of ML propose different types of algorithm applying on defined type of input data and serving specific purposes. The choice in the algorithm is mainly function of the data analysis problematic to answer, of the input data available, and the specific characteristic of the background knowledge of the field studied. We propose a non-exhaustive overview of the different families of existing algorithm that can be found in Kononenko and Kukar (2007) for more details.

- **Classification :** these algorithms implement classifiers to assign each object of the database to a class (among a finite set of classes), depending of these objects' characteristics. These methodologies can be used for objects/people labeling, given a set of discrete attributes (e.g. medical diagnostic).
- **Regression :** they consist in predicting the value of an unobserved variable of an object in function of its others attributes values.
- **Logical relations :** they can be considered a generalization of discrete functions in the case the object possesses only discrete attributes.
- **Equations :** they calibrate a set of equation to model a real-world problem observable through the data.
- **Clustering :** these algorithms are different from the previous one because of their unsupervised nature. They group the object in a finite number of subsets called cluster. This grouping is performed optimizing the similarities between objects from a same cluster, and the disimilarities between object from different clusters. An important part of these algorithms is the choice of the metric used to evaluate (dis)similarity.

1.1.5 Principal component analysis and K-means algorithm

In the following project, we base our methodology on two algorithms to proceed to the clustering process : The principal component analysis (PCA) and the K-means algorithm.

As mentioned in Section 1.1.4, a space reduction can be performed during a KDD process to decrease the number of variables taken into consideration. PCA is a multivariate analytic tool applied on the 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 (non correlated) variables called principal components. The second purpose is the compression of data by reducing the space to only the most representative components. Finally, the PCA simplifies description of the dataset and allows examination of the structure of data (Abdi and Williams, 2010).

The K-means algorithm is part of the clustering process studied in ML, and is consequently unsupervised as mention in the previous section. The number of clusters have to be specified as a parameter of the algorithm, qualifying it as partial clustering, opposed to hierarchical clustering. If this additional parameter is one of the main problems of partial clustering, hierarchical clustering methods are limited for large datasets because of the computational cost of the dendograms construction (Kononenko and Kukar, 2007). The K-means minimize the within sum of squares (WSS) representing the sum of the square distance between each point and its cluster's centroid.

1.2 Elements of the problematic

The study aims to extract valuable information, especially performed activities, from Wi-Fi connection database using DM algorithms. One of the main characteristics of our project is to have a methodology not using the spatial characteristics of the building. This point is motivated by the sensitive nature of these information and the difficulty to obtain them. Being able to get rid of this limitation would facilitate the deployment of studies concerning mobility behavior through Wi-Fi. The study although present the general problematic of knowledge discovery work mentioned previously.

1.2.1 Selection and preprocessing : evaluation and correction of the errors

The first objective is to evaluate the errors within the data furnished and decide for a solution to clean them. These errors are mostly missing fields and errors in the measurement. Then a statistical analysis has to be performed on the data to evaluate and ensure the fit of the data studied with the physical reality. Such an analysis can lead to the highlight of hidden errors in the measurement as noise. Here this perturbation takes the form of the so called "Ping Pong effect" generating records non representative of any real event. By repetively switching their connection AP, devices will record events in the database that are not representative of any movement or action of the user. This error has to be reduced to improve the accuracy of the further results.

1.2.2 Transforation : generating variables without spatial knowledge of the infrastructure

Assuming we work on cleaned data, the variables have to be selected and transformed to optimize the performance of the DM algorithm. A detailed analysis of the variables generated can help to determine their pertinence for knowledge discovery. Once these variables are selected and that their eventual physical meaning is described, a spatial reduction can be performed. Such a process can increase the performance of DM algorithms and decrease the correlation between the variables in the case of a PCA. An important issue is to be able to identify the physical meaning of the groups of activities identified using their characteristics. This issue raise the importance of using understandable variables efficient for pattern identification and allowing real world analysis.

1.2.3 DM and Interpretation : implementation and Evaluation of the model using supply and demand

Once the descriptive variables have been computed and analyzed, a DM algorithm can be applied to classify the AP. The choice and calibration of a clustering algorithm is an iterative process based on the knowledge of the field of study and the empirical knowledge from trials. Such a process should generate a classification of the AP regarding the main activity performed around the location : we call the relationship AP/activity the demand.

The evaluation of the performance of a model is an essential step in every scientific process. While ideally, this step would require the accurate knowledge of the activities actually performed by the users, we do not possess such a data set. The closest set of data available are the supply corresponding to the architecture of the buildings. While comparing the activity performed with the purpose of a location doesn't ensure an accurate model, it allows to evaluate the realism of the result computed.

1.3 Objectives of the Research

The objective of this research is to propose a methodology to classify the types of activities performed in infrastructures, based on the Wi-Fi log data, without any spatial knowledge of the area. To compensate for this lack of spatial data, we generate indicators quantifying few aspects of the infrastructure's users' mobility behavior around the AP. Using indicators rather than the function of the number of connected users depending on time facilitate the identification and labeling of the activity categories found. We then use the evolution of the activity classification through the week to evaluate the robustness of the methodology.

1.4 Organization of the Thesis

In this thesis, we first give an overview of the WI-Fi dataset used, through statistics to evaluate the possibilities offered by such data. We then describe the different steps of the methodology proposed : pretreatment, transformation and space reduction, and clustering. After a discussion of the daily results obtained, we propose an algorithm to links the different daily results throughout the week. The last part of the article concerns the validation of the model using the comparison between supply and demand.

The global methodology is presented through the methodology section and through an article presented at the TRB 2016, currently under review in the Journal of Computers, Environment, an Urban Systems. The final version of the our process is the solution proposed in the discussion. In fact, the development of this work was done through few iterations, each one improving the accuracy or the robustness of the methodology.

CHAPTER 2 LITERATURE REVIEW

The study of pervasive systems such as cellular network, GPS traces or Wi-Fi logs, has received a growing attention during the last decades. Indeed, the development and growth of these ubiquitous networks combined with the improvement of data collection processes, data storage capacity, and DM methods opened promising perspectives towards the understanding and characterization of human mobility. These results show an application in network optimization, urban modeling or even transportation policies. As suggested in Aschenbruck et al. (2011), the user traces can be acquired through three different processes : monitoring the location, the communications or the contacts.

The monitoring of the location consists in collecting the successive positions of a user with his device and is mostly done with the Global Positioning System (GPS) data. Using satellites, this technology furnishes the position of the users with a margin of few meters. It, however, shows limited application when coming to indoor environments where obstacles can create a shadowing effect. This problem can be overcome coupling these data with GSM and Wi-Fi data (Aschenbruck et al., 2011). However, as mentioned in Su et al. (2004), some users can be reluctant to share the history of their position. That method being device-centric, it needs the users to cooperate accepting the burden of a device or an energy consuming application.

The monitoring of the communication use the interactions between the devices and a communication system (cellular or Wi-Fi) to recreate the users' mobility history. This information is regularly collected by the networks operators and represents a low-cost information. The pervasive nature of these networks allows capturing a large sample of the population even if the characterization of this sample is still a challenge (Calabrese et al., 2013). The data produced have a relatively low location accuracy. The simple information of the AP is very limiting and usually lead to work on symbolic spaces (space without geographic localization) rather than geographic one as shown in Meneses and Moreira (2012). The use of the signal strength can improve the location accuracy and can, therefore, be improved with data fusion process (Aschenbruck et al., 2011). Cellular data, which allow working at a large scale (city), are discussed in the following work (Calabrese et al., 2013). Wi-Fi works at on smaller scales on the specific environment as campus, offices or festivals.

The last process is the monitoring of the contact between users using technologies as bluetooth or Wi-Fi. Such a process allows characterizing the social network existing in closed environments. This topic has received a growing attention with the developments link with the multi-hop networks (Conti and Giordano, 2007). The global concept lies in unloading a part of the Wi-Fi network, making information transit through a chain of users' devices. In Su et al. (2004) and Su et al. (2006), a sample of students are monitoring their contact with other users within a campus to explore the feasibility of such a technology, and in the process are able to highlight some social behavior of the students. Another use of this process is proposed in Naini et al. (2011), where devices' Blue-tooth in a Festival allowed to estimate the size of the entire population using statistical algorithm derived from the biology field.

2.1 Wi-Fi network-centric trace data

While the data obtained monitoring an infrastructure show great promise to track a large sample of individuals for a low cost, a certain number of challenges are highlighted in the literature.

A first challenge follows the need of anonymization of users, essential to guaranty users privacy (Meneses and Moreira, 2012). This leads to an absence any users socio-demographic characteristics. This issue is even more problematic considering the fact that the characteristics of the sample of the population observed through the network are rarely characterized accurately.

A linked issue lies in the absence of distinction between the different type of devices, smartphones tablets or laptops appear in the same way in the database and can represent the same users. However, Yoon et al. (2006) shows differences between these devices' behavior : laptops are connected longer but to fewer APs.

Then, the data is highly dependent on the connection status of users' devices : non connected users are invisible in the network. Parallel to that is the geographically limited range of the network, devices getting out of range of the network are lost which create holes in the node's history (Meneses and Moreira, 2012).

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

2.2 Analysis of the network

A large amount of work has been done on analyzing the Wi-Fi usage of infrastructures as Henderson et al. (2008). Through the aggregation of the connections and statistical tools, the article presents main trends concerning the network usage, users' behavior and mobility of users. The end goal is to furnish tools to analyze the design and efficiency of highly frequented places. In the list of possible applications, we find the optimization of hand off latency (Ramani and Savage, 2005), anticipation of the resource allocation (Katsaros et al., 2003), improvement of routing protocols (Su et al., 2001) or development of energy-efficient location (You et al., 2006). Such works can be found at different scales as campus (Henderson et al., 2008), offices (Balazinska and Castro, 2003), hospital (Prentow et al., 2015) or even city (Afanasyev et al., 2010).

The necessary pre-treatment of the data to answer the challenges raised in the previous section lead to diversity in the hypothesis in the literature.Meneses and Moreira (2012), for example, use connection rates to compute the most important corridors within a campus aggregating the AP of the same building (the within movement are thus 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 et al., 2005). From this perspective, it is thought that users are more interested in the type of activities taking place in a location rather than its spatial location. Such considerations are taken into account into activity-based models discussed next section.

Related to this emphasis on activities surrounding APs, Calabrese et al. (2010b) introduce the possibility of identifying the type of activities taking place around Wi-Fi APs, adopting an eigenvector analysis of signals in connection patterns. The use of signal theory methods, such as signal decomposition and unsupervised machine learning on Wi-Fi connection data allows them to classify APs into 5 clusters associated with different types of activity.

In parallel with this descriptive analysis, an important part of the literature concerning Wi-Fi data deals with the implementation, calibration, and validation of mobility models as shown in the next section.

2.3 Mobility models

Mobility is an essential component in the optimization of the networks and therefore the understanding and simulation of human mobility. A huge amount of mobility models have been proposed during the last decades, the first level of classification is their level of aggregation. Microscopic models describe the evolution of the position and the speed of a single vehicle as the car-following model while mesoscopic models consider groups of vehicles. They have been largely developed and discussed in the literature (Bettstetter, 2001), but are not the focus of our work. Macroscopic models describe more aggregates data as the flow, density or activity. We present an overview of few approaches concerning macroscopic mobility models.

Location-based model attempt to predict the next destination of users given their previous locations in the day : these models are often based on Markov process (Francois, 2007). However, a part of them do not take into consideration another component in the user's destination choice as the time of the day, duration, or social groups. Some models go further segmenting the time of the day (Liu et al., 1998), or creating a hierarchy within the APs (Jain et al., 2005). These methods have trouble to identify points out of the norm and differentiate groups behaviors (Wanalertlak et al., 2011).

Social networks based model consider that the destination choice is driven by the social relationship existing between the different users. Then users would move to a place to encounter other people, this phenomenon being represented by social links bounding users. Based on the research in the social science topic as Watts (1999); Albert and Barabási (2002), some early models were developed as Herrmann (2003). In this work, users are clustered in groups based on their social network, a movement dynamic is then associated with each group. More sophisticated models were then proposed as Musolesi and Mascolo (2007) creating nodes networks creating social attraction each another, and Boldrini et al. (2008) who added location attraction to it.

Finally, activity-based models are built on the assumption previously mentioned, that users are more interested in the activity than the location where it is performed (Danalet et al., 2013). Time is fundamental variable in activity modeling (Yamamoto et al., 2000), and can be continuous, tour based or activity chains based. In general, activity-scheduled modeling can be divided into two tasks : activity generating which create the set of activity proposed to the user in his choice, and the activity scheduling which take into account the spatial-temporal constraints of the activity. The first approaches for activity generation were discrete choice models generating a choice set for activity patterns (Adler and Ben-Akiva, 1979). Hybrid simulations developed proposed a sequential constructing of activity patterns creating small choice sets through heuristics. Later models proposed to create a hierarchy within the activity type ordering them according to their importance in the activity chain structure. In that approach, The activity generation is divided into few tours associated with the activity importance, choices being based on people's utility maximization. Illustration of sequential construction of activity patterns can be found through the PCATS (Prism-constrained activity simulator) or the CEMDAP (Comprehensive Econometric Model of Daily Activity Pattern) (Danalet et al., 2013).

2.4 Directions for our work

We propose here to go further in the analysis of infrastructure usage, especially on the major activity realized at one place by clustering the APs. While Calabrese et al. (2010b) proposed such a process through the analysis of the connections number during a week at a campus scale, we develop a methodology applying for non periodic event studying one day at a smaller scale. We develop indicator taking into consideration the different component of people mobility choices highlighted in the mobility model literature. We then use clustering methods to extract the main activities performed in a building.

CHAPTER 3 METHODOLOGY

In this section, we present an overview of the adopted methodology in compelemt to the article presented in Section 4. The project follows the same steps as the global KDD process presented in the Section 1.1.4, especially Figure 1.2. As mentioned earlier, the objective of the project is to be able to identify the activities taking place in the different locations using the WI-Fi connection behavior of the users. To do so, we organize and correct the input data, generate pertinent indicators for each location and then cluster these values to generate families of location showing similar types of behaviors.

The data were furnished to us within 6 csv files representing 150 MB of data. These raw data correspond to the connections of different users locating them through time and places with the time and the associated AP. Considering the size of data, a first step is to export it in an SQL database to improve the calculation times of the queries. Then a first pre-processing is necessary to identify and correct the eventual errors existing in the database. When data have been corrected and organized in a database, we generate indicators representative of the users behaviors within the place through the day. These indicators aim to characterize the level of visits to the place and the time spent by users at these locations. We compute for each day and each AP a set of variables supposed to encapsulate the observable differences of behavior in our data. These variables were chosen to be a compromise between the pertinence of the data represented and the interpretability of it for the semantic meaning association coming later. A correlation analysis of the indicators generated show existing links between the variables. Clustering a correlated set of variable could not be pertinent or consistent. To overcome this issue, we used a space reduction algorithm, the principal component analysis to extract a subspace of the data in which we could perform a consistent clustering. The space generated is supposed to have a lower number of orthogonal dimensions solving the correlation issue. We choose to use the K-means algorithm to cluster the set of data, this method present the advantage to be unsupervised. The number of clusters in the parameters is chosen considering the number of activities expected and the behavior of the sensitivity of the clusters found. The clustering is then applied independently to each day creating independant dayly solutions. We use an optimization algorithm to associate the different cluster found during the week allowing to track the evolution of distribution of the activities through the week. We finally involve the map of the facilities furnished to compare the activities found through our methodology with the actual use of the facility.

The methodology was coded in Java and divided into 4 modules as presented Figure 3.1.

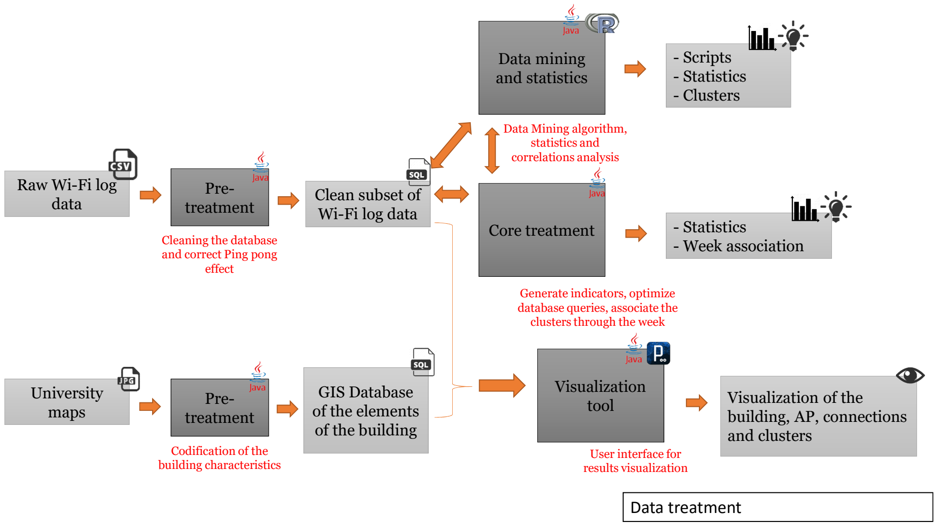
Data mining and statistics
- Scripts
- Statistics
- Clusters
Data Mining algorithm, statistics and correlations analysis
Raw Wi-Fi log data
Pre-treatment
Cleaning the database and correct Ping pong effect
Clean subset of Wi-Fi log data
Core treatment
- Statistics
- Week association
Generate indicators, optimize database queries, associate the clusters through the week
University maps
Pre-treatment
Codification of the building characteristics
GIS Database of the elements of the building
Visualization tool
User interface for results visualization
Visualization of the building, AP, connections and clusters
Data treatment

Figure 3.1 Work structure

- **Pretreatment module :** it exports the data from csv files to a postgre SQL database. This first step is also the opportunity to generate some global statistics concerning the datas and identify the potential errors existing within the database.
- **Main treatment module :** it is the main part of the program corresponding to the processing of raw records of connections and results of the clustering. Indicators are generated for a given day and location. This module also contains the post clustering methods identifying the evolution of the clusters through the week.
- **DM module :** it is Java based code communicating with a R server to proceed to the DM algorithm allowing the automatization of the process. The R functions coded correspond to descriptive statistics scripts and DM methods as customized versions of the PCA and K-means algorithm.
- **Visualization module :** it is a visualization interfaces using the Processing libraries allowing to input graphic data as maps and visualize the results generated by the clustering methods.

Table 3.1 Format of the input data

Fields	ID	Association time	Location	Duration	Disassociation time
Type	Integer	Time stamp	String	Time	Time stamp

Table 3.2 Databases created

Connection DB	User ID	Connection ID	Building	Room	Association time	Disassociation time		
Type	Integer	Integer	String	String	Time stamp	Time stamp		
Transition DB	User ID	Connection ID	O. Building	O. Room	D. Building	D. Room	Departure	Arrival
Type	Integer	Integer	String	String	String	String	Time stamp	Time stamp

3.1 Data furnished

3.1.1 Global statistics

The data used in the research were provided by Concordia University in Montreal, Canada. Concordia Concordia University is made up of 57 buildings across two campuses (one in downtown, and the other in suburban, Montreal). The university counts of 47,000 students, faculty and staff with 36,000 undergraduate students. In this study, we analyzed connection data from all APs of the campus for one week (02/02/15-08/02/15). The data identifies each connected individuals through a unique code associated to the device's MAC address in a confidential database to provide users anonymity. Each connection record includes an AP's ID, user device ID, connection and disconnection time as shown Table 3.1.

We expect that this sample of device connections will be representative of the activity on the Concordia campus, despite the fact that individuals who don't own a connected device are invisible while users owning more than one (e.g. smartphone & laptop & tablet) are over represented. Past studies and our experience in previous usage of similar data in public spaces (e.g. outdoor street festival, train station), suggest that the sample from Wi-Fi connection data represents 35% to 65% of the population, especially in case of a younger population, which is the case here, the usage percentage is near 65% (Farooq et al., 2015). As this case study is of a university campus, we expect the percentage to be even higher thus giving us a greater confidence in terms of representativeness.

Summary results for the data in question allows some basic observations. We observe over the course of the week there were almost 1.7 million connections by 60,500 devices to over 1,048 APs throughout the campus. It is worth noting that connections under 5 min are not recorded, thus erasing some noise due to traveling through intermediate AP within a trip. Maximum connection time recorded was 5 days and 20 minutes, surely a fixed devices such as desktop computer with a very have long connection time.

As mentioned above, there are many buildings on the two campuses of the university. 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 the largest building during the most representative day of the week (04/02/15).

While we were provided with AP log data, we were not provided initially with the locations of APs throughout the campus. At the same time the AP IDs allowed us to determine the building and floor where the APs were located. While accurate location information would have been ideal, we decided to explore what we could do with this non-localized data. As a result, we decided to focus our study on the inference of activities surrounding APs only using device connection information available through the logs without knowledge of the buildings, activities, or AP locations.

3.1.2 Connections

3.1.3 Inter-Ap transitions

Trips in the set of connection data are associated to transitions for a user from one AP to another as shown Table 3.2. We call transition any switch in AP connection : the previous pojnt is the origine while the new one is the destination. We are first interested in the distribution of the trip origins and destinations within the MB building AP subset.

Figure **??** shows the transition distributions observed on Wednesday the 4th of February, while removing trips with the same origin and destination. We grouped the transitions with identics pairs of origin and destination and display the number of tansition found. The AP's IDs allowed us to order them by floor resulting in the pattern observed in Figure 3.2. Here, we observe the presence of square blocks located on the diagonal of the graph corresponding to the inner floor trips. Parallel to those trips, we can observe vertical and horizontal lines corresponding to alternative paths to all floors successively which are probably elevators. The Figure 3.2 is an interesting visualization tool to analyze the main path through the facility if (as in our case) we are forced to work on a symbolic space model (model without geometric localization of the objects).

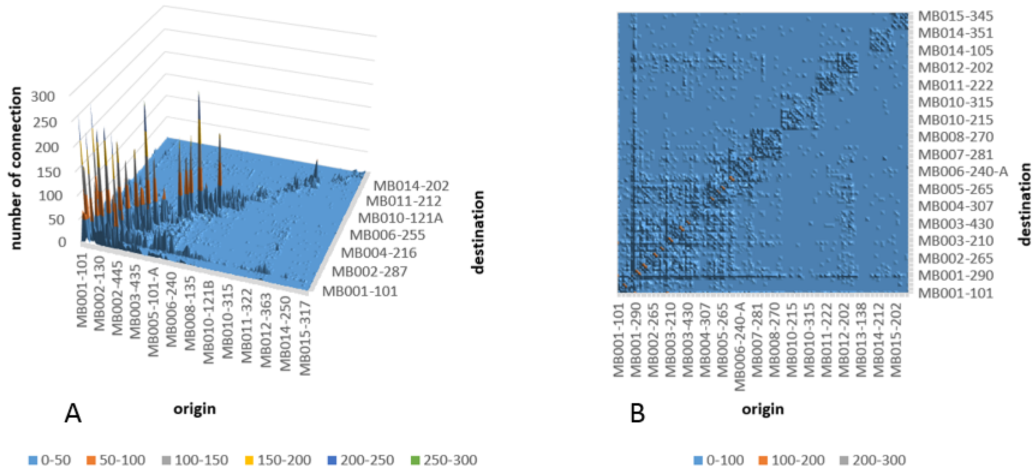

Figure 3.2 Trips OD matrix

3.1.4 Limitation of the trips

In this last section, we discuss the temporal aspect of trips which appears to be one of the main limitations of this connections dataset. Indeed, while the use of signal strength to localize device allows us to follow the user through the network and to localize trips in the time dimension, pure connection data do not seem well suited to do this.

The main reason comes from symbolic space used to locate the user : APs are associated to areas to which the user could be connected. A trip in such a space corresponds to a transition from one area to another. Working with data with dense AP networks, which is our case, becomes a burden because time transitions between APs are mostly instantaneous. In this figure we see that in the vast majority of transitions (or hops) between APs, transition time is 0. As a result, trips between distant APs representing totally distinct portions of space are decomposed into successive elementary trips between close APs, thus making total travel time the sum of a series of instantaneous transitions, and thereby null or very close to it.

By focusing on the user behavior, we reach one of the limits of the network connection analysis. Indeed, if a strength of these data lies in their potential to continuously furnish global information about the whole population of entire facilities, they appear to be weak in characterizing precisely individual mobility behaviors if not enhanced with localization data such as signal strength. In the rest of the thesis, we will focus on the location based analysis, studying the connections perceived from one fix AP. We'll propose a methodology to infer the main type of activity performed at each AP using the connection behavior of the devices around them.

3.2 Pre-processing of the data

A fundamental step in KDD process is the analysis of input data, identification of the potential errors and formatting. This action is essential to ensure the consistency of the data and appropriate character of the algorithm used on them. We first discuss the different biases found in the dataset with the solution proposed to answer them. Then we describe the variables we computed as an input of the clustering algorithm.

3.2.1 Error within the data

First common issue found in databases is the missing values in some fields. The exploration of the database show that this issue only appears for the location field creating some non located connections (no AP attached to them). A generic value filed "unknown" has been created to avoid errors during the data processing. Then, considering the high number of locations and the complexity of destination inference, as shown in the literature, it appears difficult to compute the missing values using the rest of the database. The missing values representing only 0.08% of the database, we chose to take these records away from our analysis, supposing that these modifications will have a really minor impact on the data treatment. Considering that we study a sample located in a specific area, deleting the records would havee the same impact on the study.

Another issue particular to our database is the aggregation of some connection time on our data : the connections with a real duration under five minutes are recorded with a five minute time duration or are not recorded. This phenomenon can be due to the data collection process or a pre-processing realized to alight the size of the data. We observe the absence of any connection with a time duration inferior to five minutes in our database. This phenomenon not representative of the actual behaviors of devices that connect for few seconds when they are transitioning between APs in a corridor, for example. The fit between the disconnection time of the connection and the connection time of the next connection prove that the connections under five minutes are not simply removed from the database. If it was the case, we would observe short time gaps between these successive connections. It appears that the connection under five minutes are aggregated with the five minutes connection as shown in Figure 3.3. The consequence of such an approximation is the loss in the accuracy of the low duration connections. In our methodology, we aggregate our data with a step of 5 minutes to have a comparable level of accuracy between our different variables.

As mentioned in the literature, one of the most common errors encountered in the Wi-Fi log data is the ping pong effect : a user located can alternatively switch between two APs when

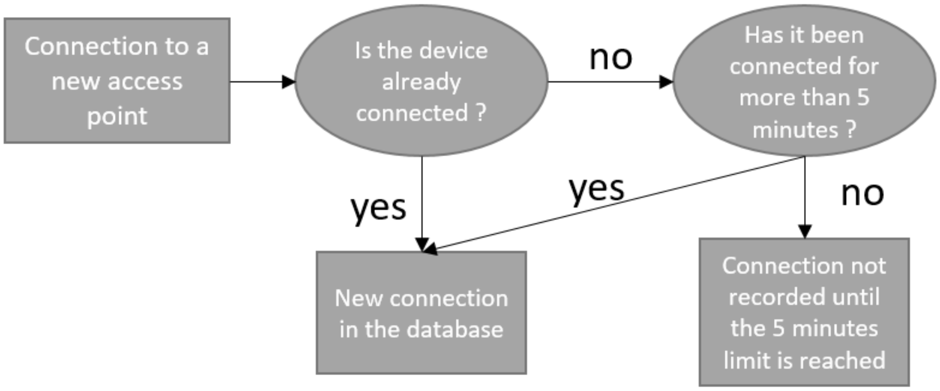
Connection to a new access point
Is the device already connected ?
no
Has it been connected for more than 5 minutes ?
yes
yes
no
New connection in the database
Connection not recorded until the 5 minutes limit is reached

Figure 3.3 Time aggregation process

the strength of the signal received is comparable. A consequence of it is the creation of non representative trips within the database. Different methods have been proposed to identify and limit the records of such false trips in the database, especially a spatial aggregation process within the APs. In our case, the previous aggregation of the connections under five minutes already smooth the high frequency changes of APs. The disaggregate spatial scale of our study does not allow us to spatially merge APs. In consequence, we aggregate temporally the connections showing a periodic and short switch between two points. If a device shows a series of connections of five minutes alternating between two APs, we report all sets of connections to the points representing the higher amount of time.

3.2.2 Indicators

In this document we analyze different indicators generated for the clustering and discuss their physical interpretations. The purpose of these indicators is to provide a way to characterize the user's behaviors around each AP with interpretable variables. Some indicators generated are sensitive to the level of aggregation used. We first aggregated our connection data to five minutes to respect the constraint given by the input data as mentioned previously. Then, the indicators are computed for given duration (a day in the case of our article) creating a second aggregation bias. If the variables used to characterize the variation the number of users might seem simplistic compared to some signal treatment process, they allowed to take into consideration other variables in the clustering as the time spent by users. The limitation of the day aggregation of the indicators is answered in the discussion section of the thesis, using the decomposition of the days into smaller time intervals to improve the accuracy of the characterization of the AP use. In the following section, we use the data of the Friday 06 of

February as an example, considering this day is also used in the following article. Variations of the indicators through the week exist and will be discussed at the end of the section.

a. Connection number

This family of indicators aim to characterize the amount of a given place. To describe this value, we use the number of connections recorded at each AP during the day coupled with the number of unique devices detected. The number of connections recorded during the day is representative of the number of trips or stops performed within the place. This number depends on the location of the AP (transition area, corridor or entrance) and of the activity performed around the location. We can easily imagine that a laboratory or private office will attract fewer users than an auditorium. In the data of Friday Figure 3.4, we observe that more than 27% of the AP records fewer than 50 connections a day and 50% record fewer than 90. There is an important drop in the distribution after 110 connections a day (approximately 60% of the cumulated population of AP), with a little peak around 180 connections a day. The distribution is barely uniform after this point, counting between one and two AP every ten connections. We count only 2 AP overpassing the limit of 1000 connections with one over 2000. These points are irregularities in the data and correspond, as the map data will show later to the entrance of the building. These considerations are particularly important during the clustering process : we must choose between taking these points out of the dataset, or accept the eventuality that these points will create clusters containing just themselves. Then this indicator is a very easy to interpret, and bring pertinent information concerning the pedestrian flow within the place, but is limited when it comes to clustering. We just observe one large cluster under 110 connections a day and a little one around 180, given the 7 cluster parameter we used.

The number of unique devices detected within a place during the day is linked to the number of connections, but depends on the number of connections these devices perform. The study of the unique devices detected allow a more accurate study of the users' behaviors, defining the population of devices actually responsible for the connections observed. This analysis is allowed by the presence of the MAC addresses in our database, allowing to track users through the day. The distribution observed on Friday, shown Figure 3.5, has a relatively similar shape than the connection distribution which is validated by the high correlation value presented in the cross analysis section. However, this distribution shows better defined clusters than the previous one. We observe a high number of APs recording a little amount of devices : 63% of the APs record fewer than 80 devices a day. Then a second group of APs representing almost 20% of the entire population records between 80 and 140 devices. The

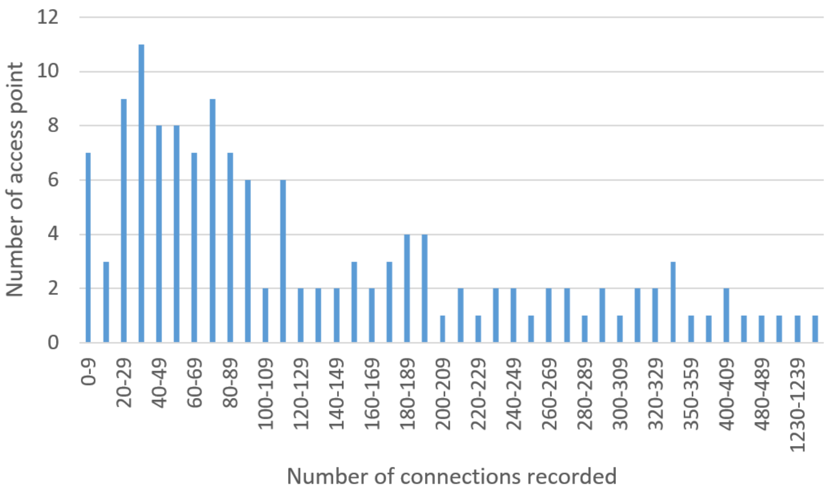

Figure 3.4 Distribution of the access points' indicator "number of connections" on Friday

third group representing 9% of the APs is located between 140 and 210 devices. As for the last variable, we find a little population of the AP showing some extreme behaviors that we'll have to take into consideration before clustering.

Finally, we propose to discuss the number of connections per device which is not used in the clustering process because being the combination of the two previous variables, but that we judge interesting to mention. The value shown Figure 3.6, are the average on the set of devices which were recorded by the AP. It is then not surprising to find only a little portion of the APs recording close to one connection. However 63% of the population of APs show an average number of connections at a given place between 1 and 2 connections per device and 83% under 3 connections per devices. These numbers let us think in the first place that the ping pong effect mentioned in the literature is relatively limited in our case.

b. Derivative of the number of connection through time

The aggregate character of the indicators that we chose to use is helpful to determine the number of users circulating in an area, but don't give information concerning the variations of these numbers through the studied time interval. To answer this question, we study the function corresponding to the number of connections to a given AP through time in the day, and its first and second derivatives function depending on time. The derivatives are calculated discretely using the local variation between the intervals given by the aggregation step. We propose a representation of the connection profile of one of the APs of the building on Friday. To facilitate the references, we will call F the function of the connected number through time, F' and F" respectively its first and second derivatives though time.

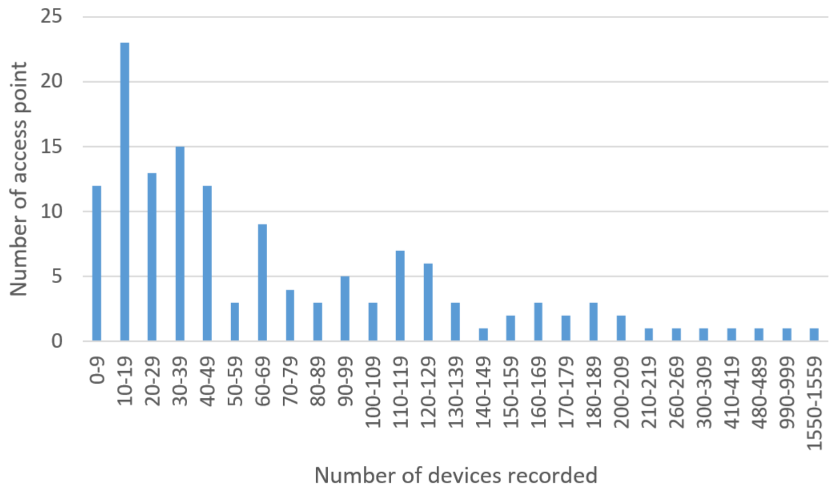

Figure 3.5 Distribution of the acces points' indicator "number of unique devices" on Friday

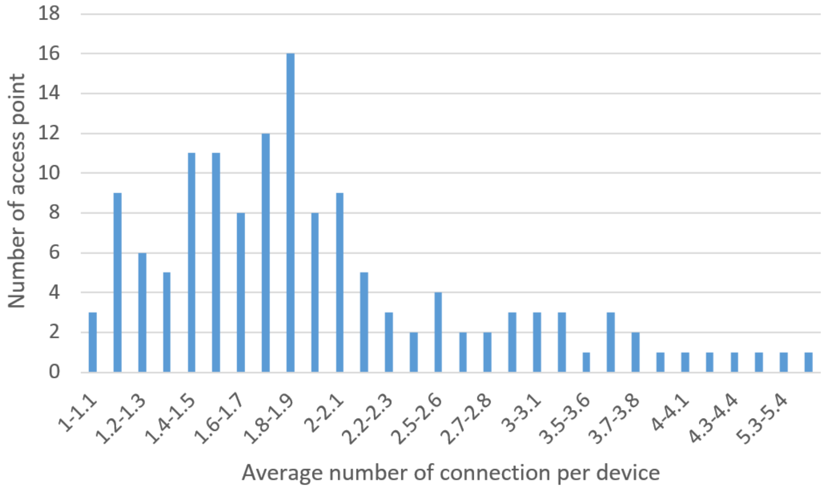

Figure 3.6 Distribution of the access points' indicator "number of connection per device" on Friday

As for the previous set of indicators, we are looking for variables for each AP through one day allowing to characterize the variation of the connected devices. The mean has a limited pertinence in an interval as a day where the number of connected devices start and finish at zero in most of the APs leading to a null derivative for both the derivatives. However, it is an interesting tool to spot the unbalanced arrival and departure on an interval, which is a more common case if the time interval studied is shorter as proposed in the discussion section.

Then, considering a very small mean, the standard deviation :

$s = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}$

with N number of points, x the average value and x_i the value of the point number i

is equivalent to the quadratic mean used to quantify the power in signal analysis.

$lim x \rightarrow 0(\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}$

In our case, with F', we use it to quantify the amount of variations in the connection profile F. A high value indicates important variations through the day while a low value corresponds to a relatively constant amount of connected users. As usual for the standard deviation, the unit is the same than the initial measure, or devices/min in our case. The values are hard to interpret independently and must be analyzed relatively to each another. The Figure 3.7 show four relatively well defined clusters within the distribution, indicating a good potential for clustering. The first and most important group of APs containing the values under 0.2 user/min represent 60% of the AP, showing by that way a smallest amount of variation in their number of connected devices through time. It is interesting to notice the discontinuity in the distribution at the limit value 0.2. The second group representing 25% of the APs correspond to variations twice more important than for the previous group. We found then two smaller groups of 5% of the population with higher variations followed by the extreme APs highlighted previously.

While the quadratic mean allows to quantify the amount of variations in the given interval, the amplitude of the function indicates the boundaries of these variations. In fact, a high variation can be caused by shorts high variations, or smaller continuous variations. The distribution presented Figure 3.8 show less defined groups of APs that the standard deviation. The first group representing 50% of the population show a range inferior to 2 users/min indication a low flow of people passing through this place. The AP with a F' range value comprised between 2 and 4.5 users/minutes don't show special trends and the one above this value are uniformly distributed. This variable is easier to interpret but will probably be less efficient for the clustering.

The second derivative F" correspond to the variations of F' and inform us concerning the

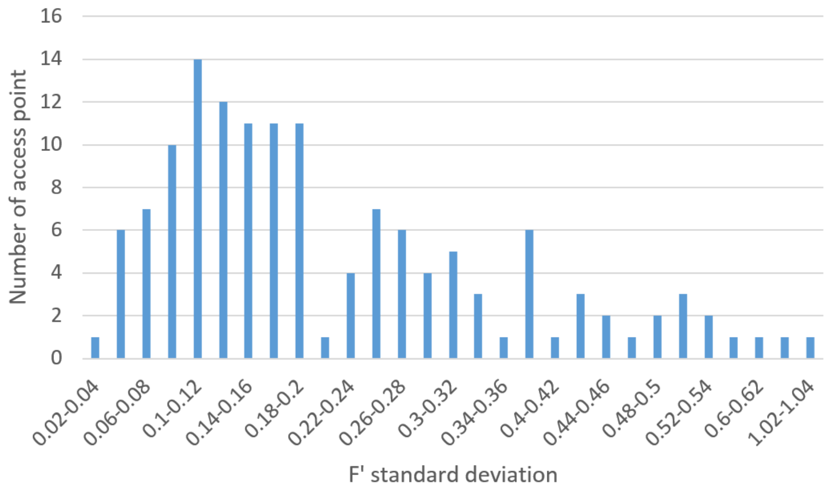

Figure 3.7 Distribution of F’ standard deviation

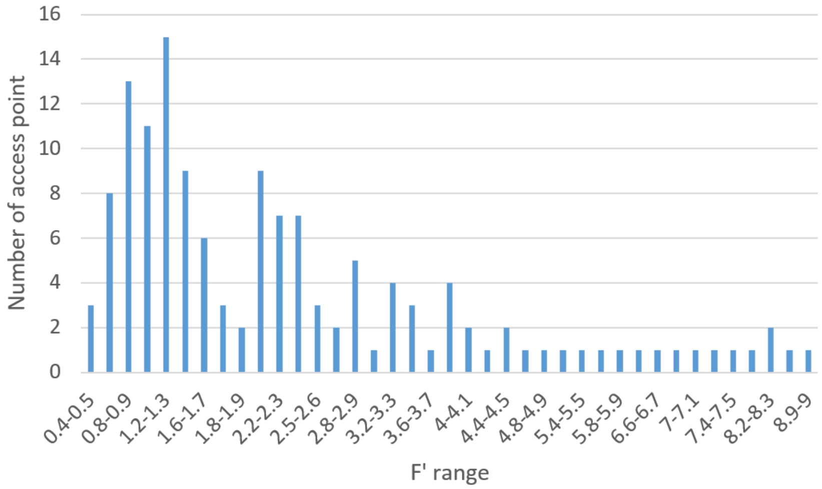

Figure 3.8 Distribution of F’ range

speed in the variations of the number of connected users. The purpose of such a variable is to differentiate AP showing brutal variations in their number of connected devices from those varying continuously. As for the first derivative F, the standard deviation calculated with an approximately null mean can be associated with the quadratic mean of the function or the amount of variation of F' in our case. Then the range of F" correspond to the interval of the F' variations.

c. Connection time

The time is an essential part of mobility. When it comes to activity performing, the time is highly correlated to the activity performed. To take this variable into account, we compute the set of activity taking places around each AP and compute average, standard deviation and maximum time of the connection performed. We remind here the issue concerning the potential missing data of the database and the aggregation of the connection time under five minutes leading to a probably under evaluate amount of APs with short connection time. The distribution of the average connection time Figure 3.9 shows only 6% of the APs with an average under 10 minutes. Then, the distribution increases with the average duration to reach a maximum of 10% of the population having a connection time between 20 and 22 minutes. The distribution then decreases to reach a maximum of 86 minutes average duration. Here, the variable is easy to interpret and bring a valuable information concerning the mobility behaviors of the devices connected. However, the distribution doesn't show the presence of identifiable groups of APs around defined value which would optimize the clustering process. We note than 63% of the APs have an average connection time under 20 minutes, which is relatively short and would be associated with short activities like transitioning from places, or waiting.

In the case of the connection time, the standard deviation, shown Figure 3.10, is associated to the average distance from the recorded points to the average duration. The study of this value allows to analyze the variability of the connection time compared to the mean previously discussed. The standard deviation calculated here is the relative one, corresponding to the usual standard deviation divided by the mean. The distribution of the standard deviation shows an interesting behavior with an important discontinuity around 1.2 minutes. The values located around 4% of the APs pass to 18% to decrease later. Despite this discontinuity, the distribution show one only group of AP centered around 1.3 times the mean which is relatively high considering the duration studied previously. This value shows an important variability in the connection time of devices within a place. Similarly to the average connection time, we have a variable representing an interesting help considering the semantic interpretation

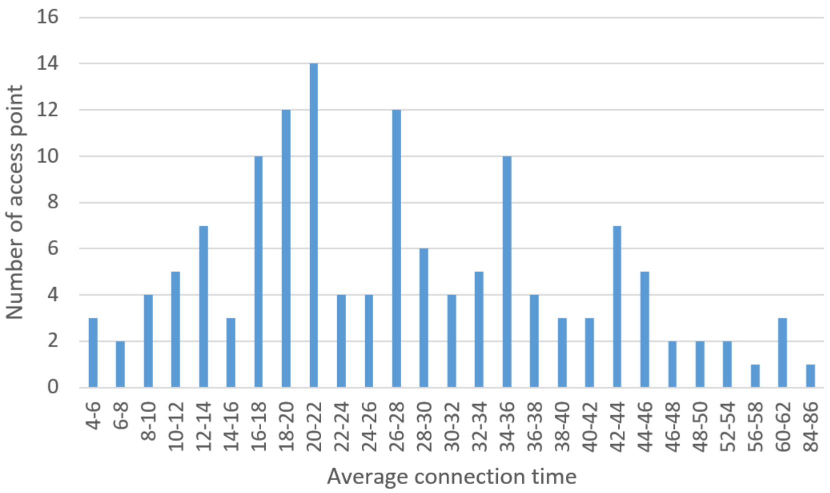

Figure 3.9 Distribution of the average connection time

of the activity performed, but which doesn't show an optimal potential for clustering.

Finally, we decided to take into consideration the maximal connection time that we judge representative of the configuration of the location. Some place like offices or libraries would allow a very long connection of static devices while corridors or transition spaces don't offer the comfort for an extended connection. The distribution of the maximum connection time relatively continuous and centered in 200 minutes connection.

Finally, it appears that the indicators concerning the number of connected users and its variation for a given APs are harder to interpret and associate to activities, they present a better fit for the clustering process. On the other hand, the time indicators are an essential component of the activity bringing a stronger information concerning the type of activity performed, they do not show the presence of obvious groups of APs which is expected to optimize the clustering process. In our methodology, we chose to make the compromise between the performance of the partitioning algorithm, and the possibility to interpret the results furnished by the computation.

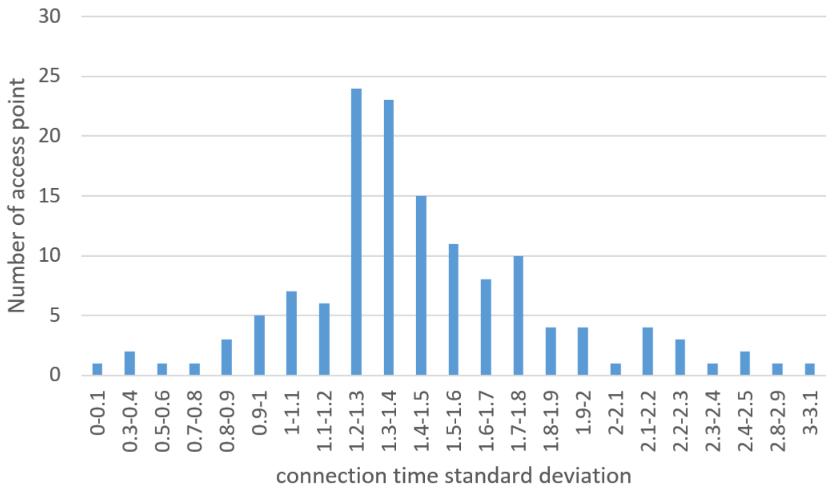

Figure 3.10 Distribution of the standard deviation of the connection time

3.3 PCA and clustering

3.3.1 Correlation analysis

As mentioned previously, the variables (defined as indicators) we choose to use for the clustering come from a compromise made between their pertinence and their interpretability. The choice of variable can lead to a set of correlated variables, decreasing the accuracy of a clustering process. We present some of the global characteristics of the indicators distributions in Table 3.3. As mentioned previously, the means calculated are not pertinent in the case of a daily time interval, the values being null.

Table 3.3 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

In Figure 3.11, we propose the correlation matrix of the different indicators, calculated using the Pearson correlation coefficient between each pair of indicator. We first observe two groups of indicators showing important correlation between them, but almost none between the groups. The indicators concerning the amount of connection in one hand and the ones concerning the connection time in the other.

Such a level of correlation would not allow a direct clustering of them, we could expect that the redundant component highlight by the correlation would be given too much importance compared to the other information in the database. To answer this problem, we propose to use a principal component analysis to select a subset of variables, containing the main part of the variations but showing a lower correlation.

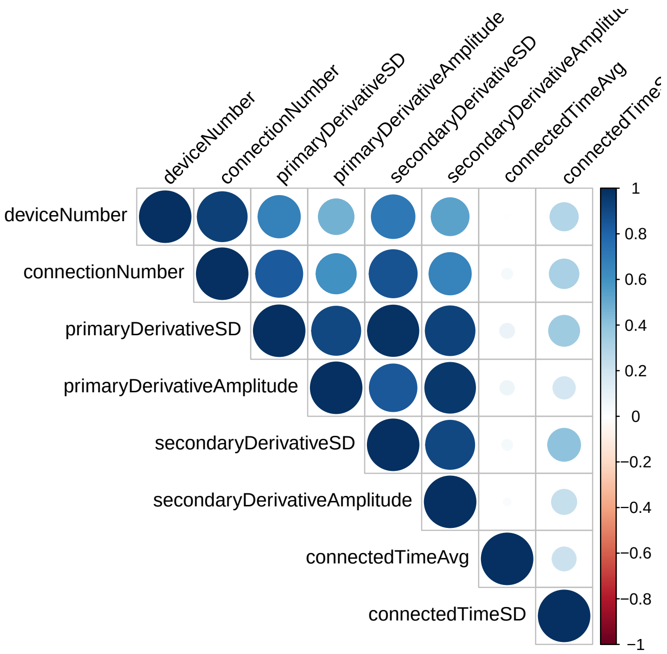

Figure 3.11 Correlation matrix

3.3.2 Principal component analysis

PCA is a statistical process widely used in number of areas when working with potentially correlated data. Generally, this process furnishes a way to extract the relevant information from complex sets of data and the eventual ability to spot hidden structures and dynamic underlying them.

The objective of a principal component analysis is to find the most meaningful basis to express a noisy and unclear set of data. Finding the principal axis allows to figure out which components are essential and which ones are redundant. Defining the number of different types of variables recorded as the dimension of the dataset, each sample of data becomes a vector of the n-dimensional space. With this basis, we can re-express the goal of the principal component analysis as the search of another basis, linear combination of the previous one which allows to express our vector more clearly. This process relies on the main assumption of the linearity of the system studied. While one could say that the activity pattern of connections of wi-Fi APs and human behavior is nonlinear, we suppose a local linearity for our study which is a common approximation, considering higher order terms to vanish when the perturbations are not too important. The results obtained are shows Table 3.4, and show the characteristics of the five first principal components.

In our case we are applying the PCA on vectors of indicators in order to decrease the correlation between the clustering variables. Once the principal components are generated, we express the vectors of indicators in the space based on the PC. We finally select the first components of these new vector based on the variance contained by each PC. We decide to limit ourslef to the first components showing a cumulated variance of 95 %.

To quantify the redundancy in the data, the covariance between each pair of variables is computed and try to be set at zero in the new space. The most basic way to create the new basis is to build successive orthogonal vectors maximizing the variance. A first normalized direction is selected, maximizing the variance along it. Then, a second normalized direction is selected, maximizing the variance and respecting the orthogonality condition with the previous vectors selected. The process is the iterated until the new basis reach the dimension of the first space. In practice, linear algebra furnish more efficient tools to build the basis as eigenvectors decomposition or single value decomposition. We choose to use the single value decomposition in our study because generally showing a better numerical accuracy, the results are shown Figure 3.12 using th Pearson correlation coefficients between each principal component.

3.3.3 K-means algorithm

The idea is to classify a given set of data into k number of disjoint clusters, where the value of k is fixed in advance. The algorithm consists of two separate phases : the first phase is to define k centroids, one for each cluster. The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Euclidean distance is generally considered to determine the distance between data points and the centroids. When all the points are included in some clusters, the first step is completed and an early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to a change in the cluster centroids. Once we find k new centroids, a new binding is to be created between the same data points and the nearest new centroid, generating a loop. As a result of this loop, the k centroids may change their position in a step by step manner.

Table 3.4 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 SD	-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 SD	-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
SD 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

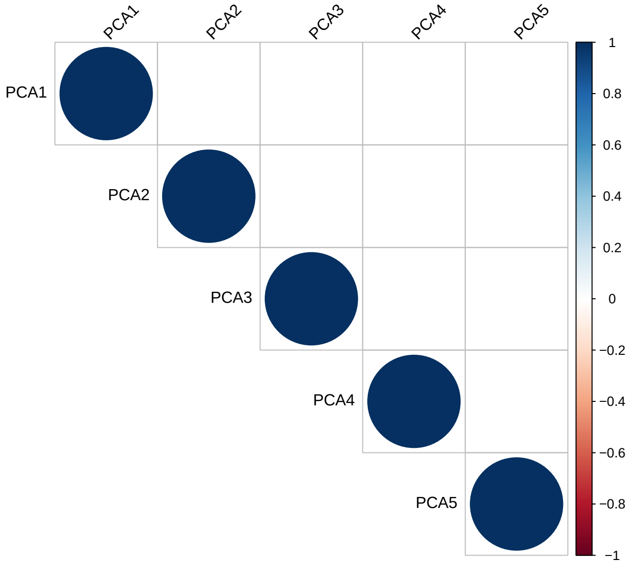

Figure 3.12 Correlation matrix after rotation

Eventually, a situation will be reached where the centroids do not move anymore.

The k-means algorithm is the most extensively studied clustering algorithm and is generally effective in producing good results. The major drawback of this algorithm is that it produces different clusters for different sets of values of the initial centroids. Quality of the final clusters heavily depends on the selection of the initial centroids. The k-means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations.

Once the vecors of indicators have been expressed in the space of the principal components as mentioned in the previous section, the space is reduced to the first components encapsulating 95% of the vaariance computed during the PCA. vector are used as the input for the clustering.

As describe in the article following, we empirically determined the number of clusters used in our case by taking into consideration the evolution of the errors with the number of clusters, and the size and consistency of the clusters during the week. During this process, we perform the clustering algorithm using R with a maximum iteration number of 1000 which we judge large enough to ensure the convergence. We finally arrived at a set of seven clusters with one capturing an extrema point, two difficult to identify and four main clusters representative of a type of activity performed within the building.

The initialization of the centroids can have an influence on the resluts of the algorithm. As we didn't want to use initialization data (which would come from the building infrastructure information), we proceed to 1000 trial with random initialization and keep the best result.

Each day clustering is performed independently of each other. The analysis through the week is based on the hypothesis that the clusters of activity remain the same through the week. Thia hypothesis facilitate the computation of the association algorithm described in Section 4.4.4. From there, the main challenge is to be able to associate the similar clusters through the day using their characteristics. We propose a modified version of the K-mean clustering constraining the algorithm to furnish seven complimentary group of clusters, each containing seven clusters, one from each day.

3.4 Visualization and maps

This last section concerns the geographic data we used to validate and visualize the results of our methodology. The maps of the building are not part of the activity clustering process. While geographical data are useful to get informations concerning the supply furnished by the building, visualizing a set of data can be a pertinent way to spot some tendencies and patterns difficult to identify with a database.

The data concerning the infrastructure of the building being critical and private for security purpose, we cannot show the references of the buildings or display entire map.

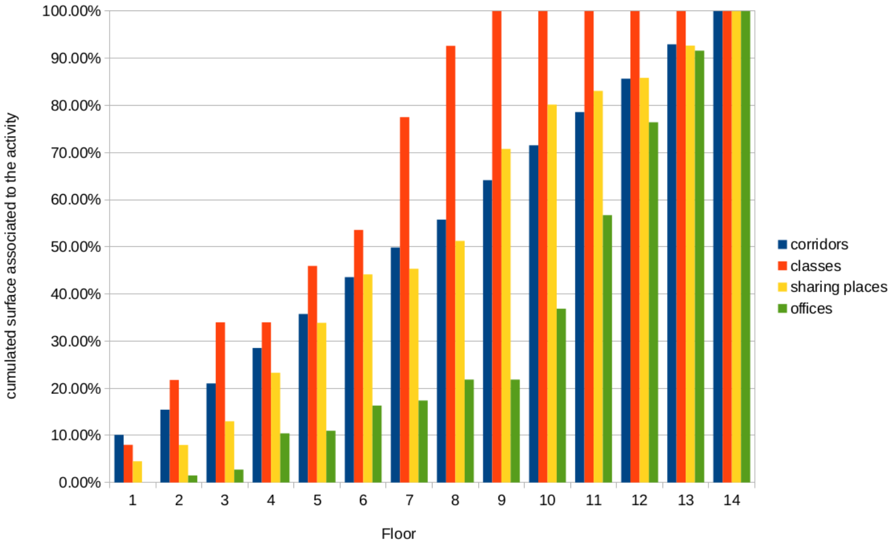

Figure 3.13 Cumulated surface of the building

3.4.1 Codification of the maps

We were given the maps of the buildings of the entire campus of Concordia university. A first necessary step was to digitalize this data converting the images into GIS nodes and areas allowing automated computations and treatment. Considering the need to create a visualization tools in the project, we chose to develop a Java application allowing to create the GIS entities and then display them rather than using an already existing CAD software. The image is displayed in the program and the different pertinent points, lines and areas of the building are manually entered with a point and click system. The points, first created in a local space, are then converted in a global GIS system using the location of the building and the scale of the picture. If this methodology can sound fastidious, automated conversion of the picture was not possible due to the complexity of the map images.

The location of the boundaries of the building and the different rooms being set, we needed to infer the main activity purpose of these locations to use them as validation data. While the geolocation of the different points has a relatively good accuracy (few meters), the label associated to the different location is more subjective. We used four main types of activities found during the clustering process : classrooms, sharing places, offices and corridors. However, the separation between little classrooms, meeting rooms and open workspace is sometimes weak and could be discussed. We chose to consider meeting rooms and open workspaces (except personal offices) as sharing places. The result obtain show the distribution of the space through the building as shown in Figure 4.6. The analysis of these data is done further in the article.

A further step is the analysis of the distribution of the space at a more local scale, analyzing the supply around each AP. We used the locations of the APs furnished through the maps and model the likely range of it to define the area associated to it. To do so we make the approximation that the area of diffusion is circular (not affected by the walls and interferences), just spread at the associated floor, and has a range of 10 meters. Intersecting this disc with the areas previously coded, we obtain the data presented in Figure 3.15.

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 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 classrooms presenting a far higher density of devices than offices.

It appears that with the range we used, which was necessary to be able to cover all the used floor space, the corridors are over represented as shown Figure 3.14. In fact they are located in the periphery of all types of locations.

3.4.2 Visualization of the results

While numeric data allow accurate quantitative analysis, visualization can bring qualitative analysis and description of the phenomenon. We developped a visualization program representing the evolution of the connections over time in the building. The different APs are represented by spheres with a diameter proportional to the number of users connected as shown Figure 3.16. The clusters of activity generated through our methodology are represented by the color code.

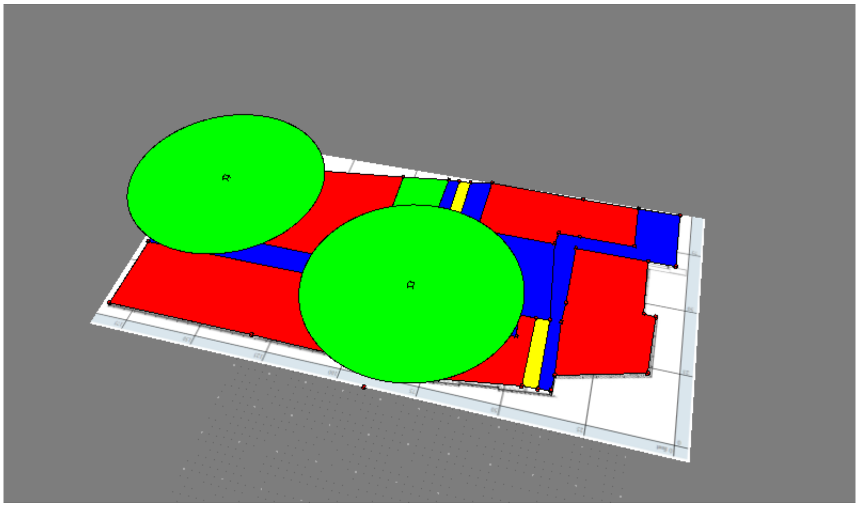

Figure 3.14 access point range

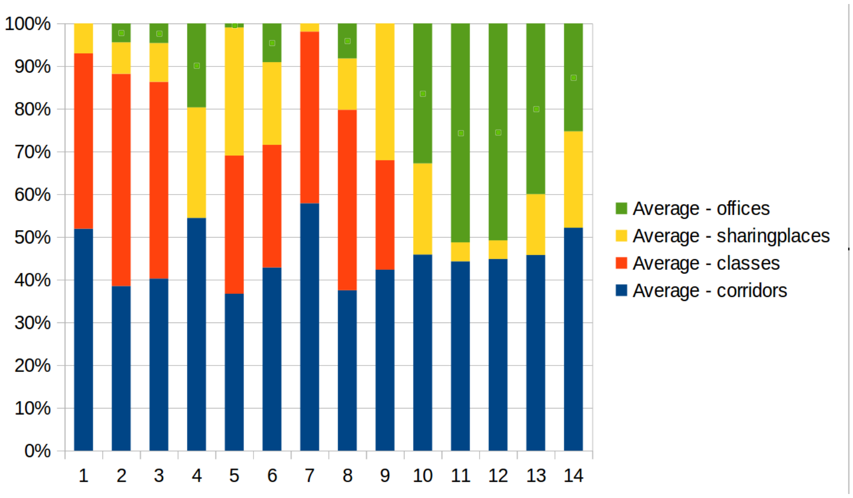

Figure 3.15 Average dedicated space around access points

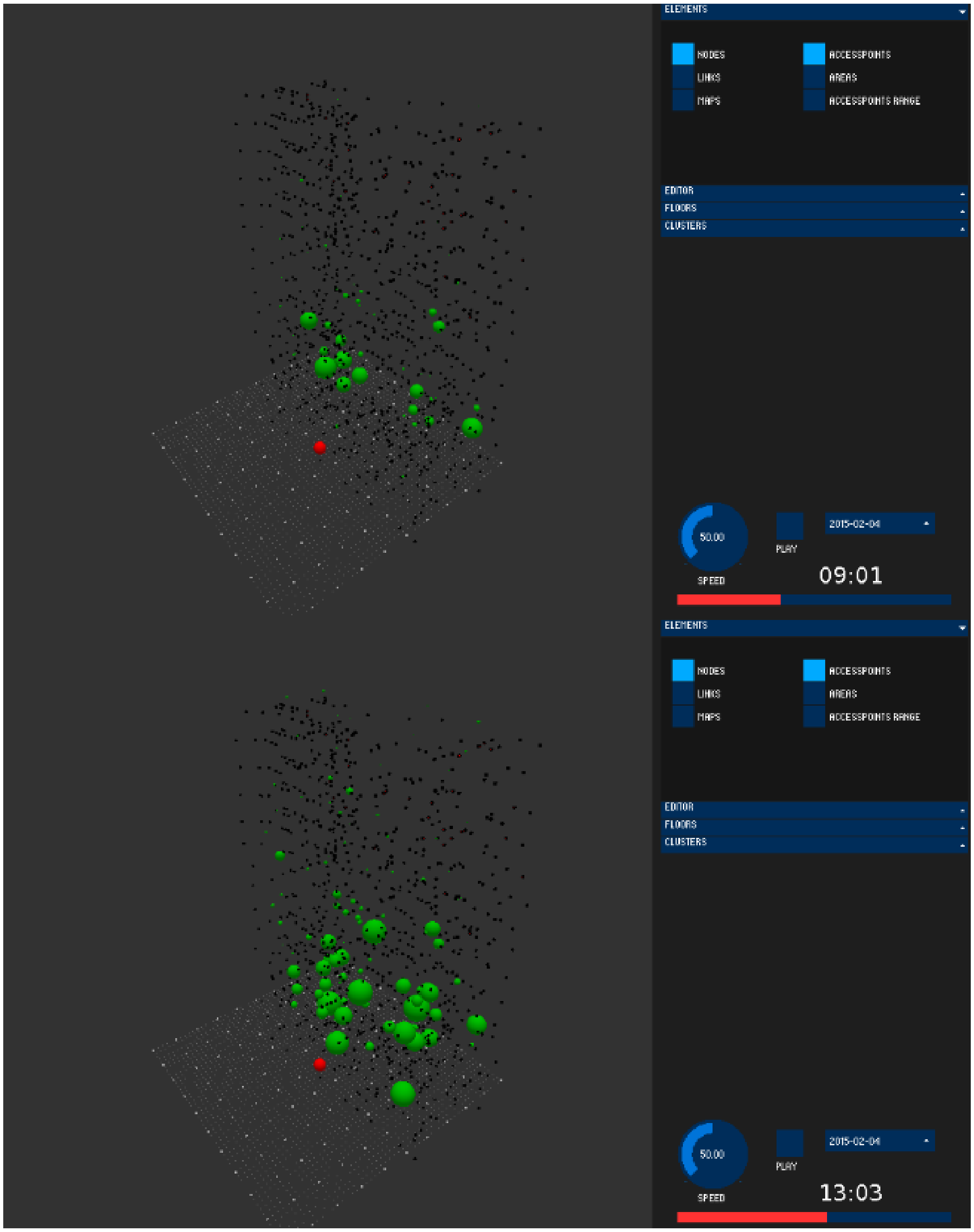

Figure 3.16 Visualization of the connections on Wednesday 04/02/2015

CHAPTER 4 ARTICLE 1 : URBAN ACTIVITY PATTERNS MINING IN WI-FI ACCESS POINT LOGS

This section is the article presented at the TRB 2016, currently under review in the Journal of Computers, Environment, an Urban Systems, and accepted pending minor corrections. The article was written by Guilhem Poucin, Bilal Farooq and Zachary Patterson.

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 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.

4.1 Introduction

Most traditional efforts to collect data on mobility and human behavior involve surveys based on sampling a small proportion of the known population. 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 and the reliability of data reported by respondents may suffer due to difficulty in recalling the activities. 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 et al. (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 (Grapperon et al., 2016; Danalet et al., 2014; Calabrese et al., 2010a). 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 centre, etc.) and collectively comprise a Local Area Network (LAN). Our methodology only uses 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.

4.2 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.

4.2.1 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 et al., 2008) and daily routine (Song et al., 2010). Some work, for instance Wang et al. (2011) studies 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). Even if progress has been made in data treatment to identify individual home and work location, personal characteristics of users are still weak. In parallel, other studies, such as (Jiang et al., 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.

4.2.2 Network traces

As suggested in Aschenbruck et al. (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 et al. (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 et al. (ming) did to infer transit itineraries from smartphone data. However, as mentioned in Su et al. (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 et al., 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 et al., 2013). The improvement of the relatively low accuracy of the locational data obtained is considered in Mao et al. (2007) and Wymeersch et al. (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 the work of 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 Blue-tooth 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 and 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 Su et al. (2004) and Su et al. (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 et al. (2011), where phone Blue-tooth activity at a festival allowed the estimation of the size of the entire population of festival goers using a statistical algorithm derived from biology.

4.2.3 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 and 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 et al. (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 and Moreira, 2012).

Another important issue that network-centred Wi-Fi data processing encounters is the so-called Ping-Pong effect. This arises when the user is connected and disconnected to different Wireless Access Points (WAPs) 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 (Yoon et al., 2006; Aschenbruck et al., 2011).

4.2.4 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 et al. (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 and Savage, 2005), anticipation of resource allocation (Katsaros et al., 2003), improvement of routing protocols (Su et al., 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 and Castro, 2003), hospitals (Prentow et al., 2015), or even cities (Afanasyev et al., 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 WAPs 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 et al., 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 WAPs, Calabrese et al. (2010b) introduce the possibility of identifying the type of activities taking place around WAPs, 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 WAPs. Calabrese et al. (2010b) 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 et al. (2010b) 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 et al., 2015).

4.3 Concordia University Wi-Fi Access Point Log Dataset

This study is based on WAP 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 WAPs 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 minutes were not recorded. Altogether, over the course of the week there were 1.7 million connections by 60,500 devices to 1,048 Access Points. The minimum connection time was 5 minutes, and the maximum 5 days and 20 minutes. 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 et al., 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.

4.3.1 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.

4.4 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. During the preliminary analysis, the descriptive variables that we calculated showed varying degrees of correlation. As a result, a principal component analysis (PCA) was used as a space reduction technique. This process improved the efficiency of the K-means algorithm and decreased 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 week and then observe the variation of clusters found across the days.

4.4.1 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 et al., 2010b) 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 personal user characteristics. Indicators 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 select two main components of usage : density/attendance of the place and time spent by users. 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.

In consequence, the first 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. 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 to respect the constraint described in Section 4.3.

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 and 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 Figure 4.1. 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.2. A possible interpretation of the selected indicators is proposed in Table B.1.

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 Figure B.1 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.

4.4.2 Principal component analysis

Principal component analysis is a multivariate analytic tool applied to data containing intercorrelated 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 and Williams, 2010)

As mentioned in (Xu et al., 2015), the K-means algorithm is prone to falling into local minima. This phenomenon tends to become more likely with the increase in the number of dimensions. However, it has been proven that an optimal solution lies in PCA subspace making it wise to perform the clustering method on the subspace of the main components generated (PCA-guided clustering).

Ding and He (2004) shows that because of the close relationship existing between PCA

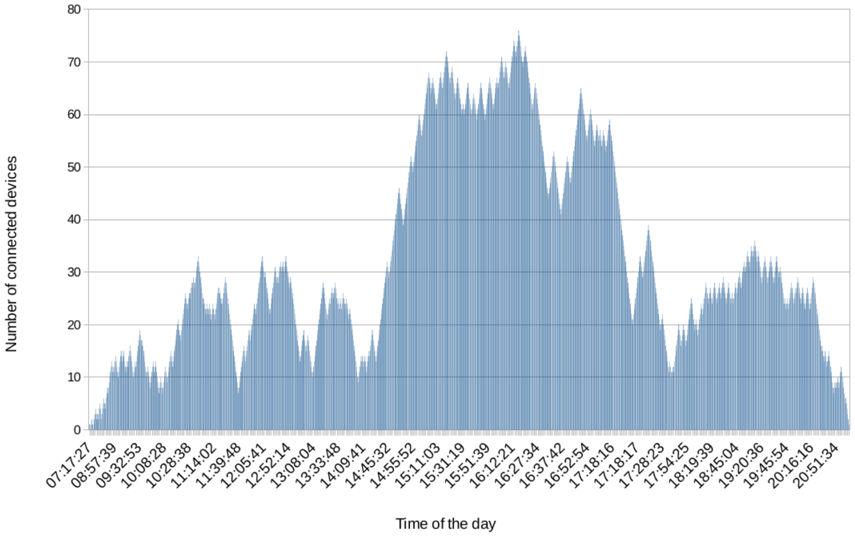

Figure 4.1 Connection profile of a representative AP during one day

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.

4.4.3 Clustering

Izakian et al. (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.

4.4.4 Analysis of clusters over the course of a week

We decided to develop the methodology at the scale of a day rather than a week. This way, the resulting daily clusters were independent of each other–centroid positions, spread, and other properties of each cluster may vary over the week. Applying clustering on a whole week 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, one of the challenges we encountered was the fact that the r daily clusters are not linked across different days of the week. Our hypothesis is that clusters are the same along the week 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, $i \in [1,7]$ days in the week, and $j \in [1,r]$ cluster number

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

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

We define a category of clusters as :

$V_q = \{V^1_q, ..., V^7_q\} \in C_1 \text{x} C_2 \text{x} ... \text{x} C_7$

A solution to our combinatorial problem can be expressed as :

$V = \{V_1, ..., V_7\} \ / \ V^p_n \neq V^p_m \ \forall p \in [1,7], \forall (m,n) \in [1,7] \text{x} [1,r]$

We are looking for the solution :

$min_{V \in V_T} (\sum_{q=1}^{7} \sum_{(i,j) \in [1,7]*[1,r]} d(V^i_q, V^j_q))$

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 week.

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. In the next section, we'll bring another set of data to put the results obtain in perspective.

4.5 Results

4.5.1 Principal component analysis

We summarize the results obtained by applying the algorithm for 1 day of the week in Table 4.1. 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 4.2 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.

4.5.2 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 Figure 4.2. 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 Figure 4.2 where at $r = 7$, the slope begins to level out.

Table 4.1 Importance of the components

Components	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	2.41	1.53	1.12	0.96	0.56	0.46	0.24	0.20	0.13	0.08	0.05
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	0.529	0.742	0.857	0.941	0.969	0.989	0.994	0.997	0.999	0.999	1.00

Table 4.2 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 SD	-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 SD	-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
SD 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

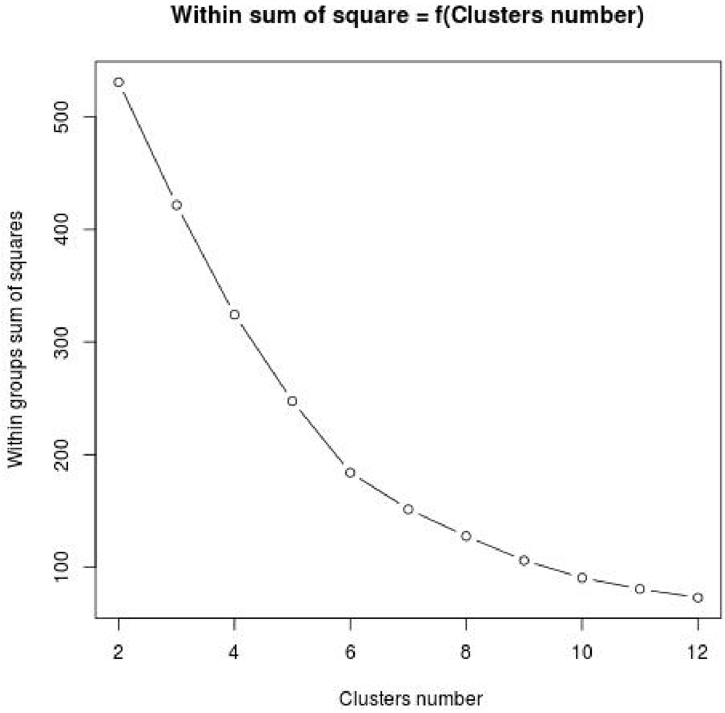

Figure 4.2 Within sum of square of error versus number of clusters

Figure 4.3 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 4.6 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.

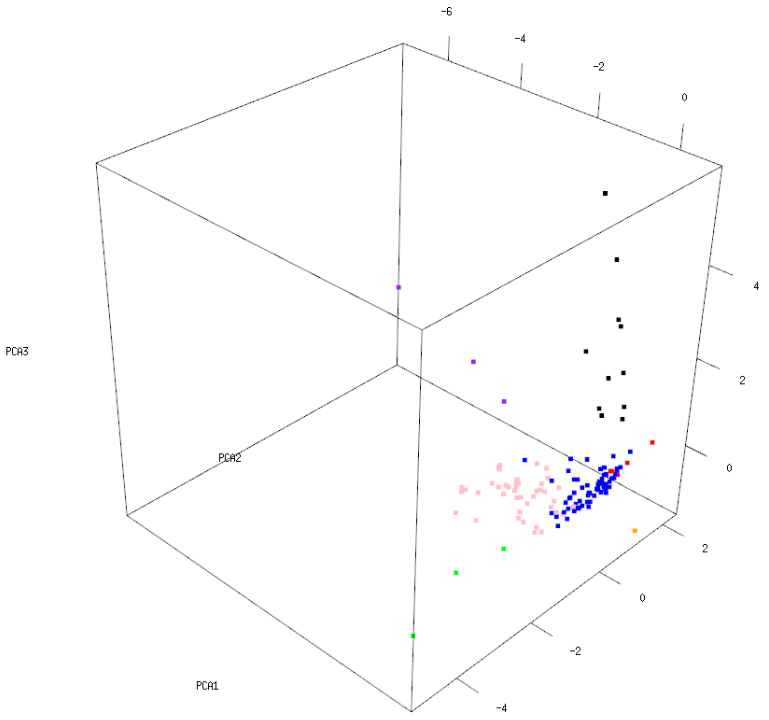

Figure 4.3 Representation of the clusters as a function of first three principal components

Table 4.3 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 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.

4.5.3 Analysis of clusters over the course of a week

We present in Figure 4.4 the evolution of the number of APs in the 4 stable clusters (excluding the *entrance* cluster, as behaviorally there 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 have classes during

Table 4.3 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 SD	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 SD	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 SD	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

weekdays and they 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.

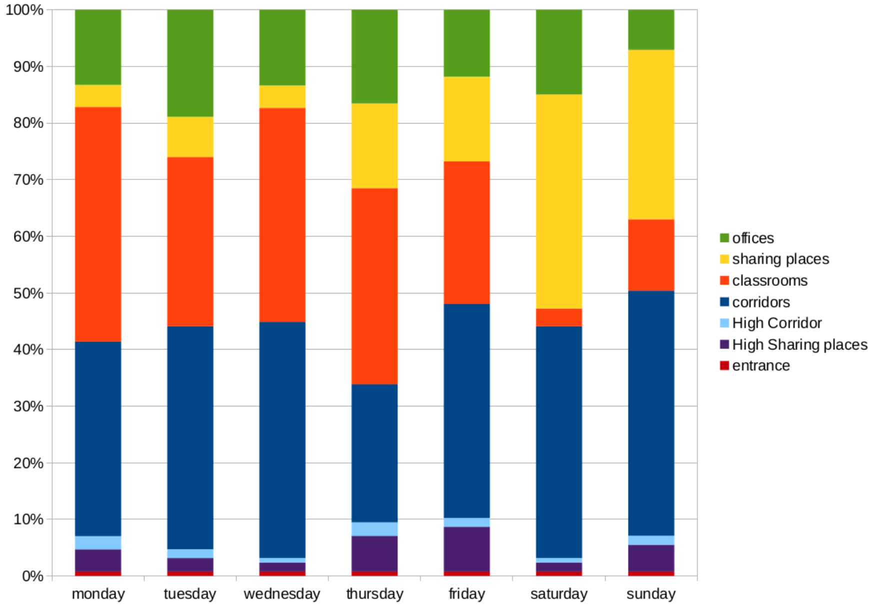

Figure 4.4 Size of the clusters over the week

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.

4.6 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.

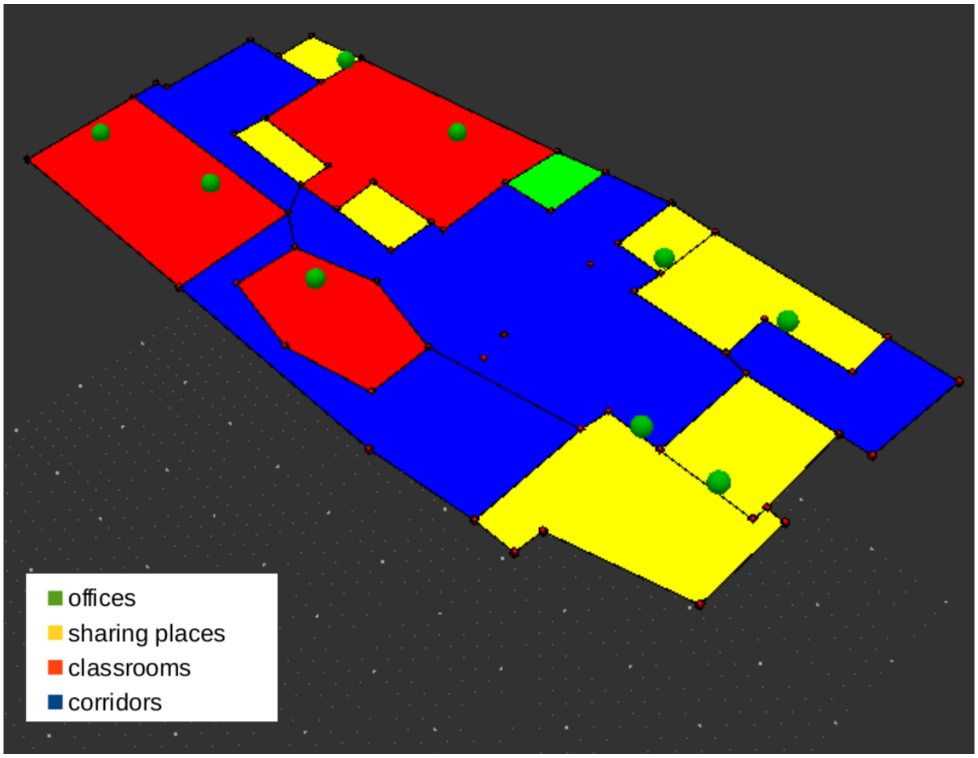

Figure 4.5 Codification of the maps

4.6.1 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 Figure 4.5.

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. Figure 4.6 shows the summary results of this analysis. We can notice that classes are located on the 10^{th} 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 10^{th} floor. It seems that the building has two different parts in term of activities.

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

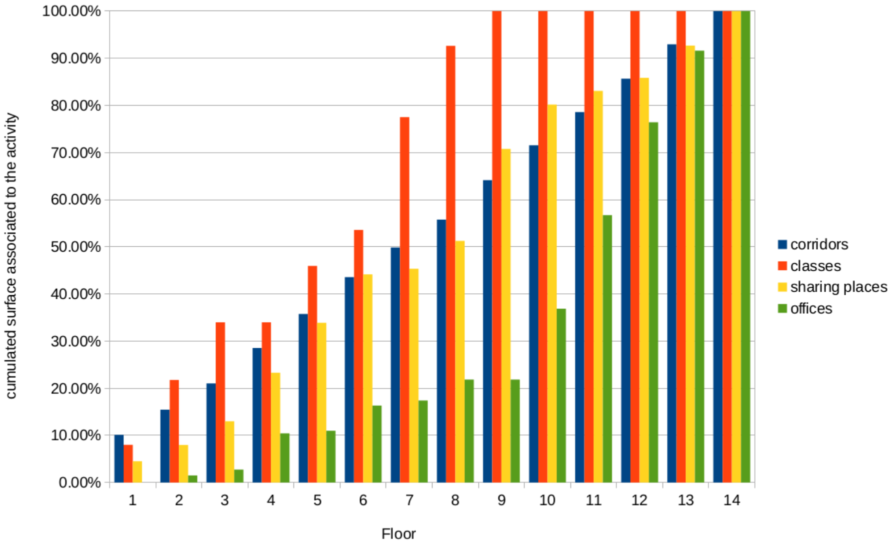

Figure 4.6 Cumulated surface of the building

classroom presenting a far higher density of devices than offices.

4.6.2 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 actual, aggregating both at the level of the floor. A summary of these results appear in Table 4.4. 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.

Table 4.4 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

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 Figure 4.7.

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. The surfaces were computed by intersecting the range of each AP (associated to a disc) with the occupation of the floors available in the maps, the results are presented in Table 4.5. A first observation is the dominance of corridors around the APs which is explained by the global occupation of the floors Figure 4.7. 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. 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.

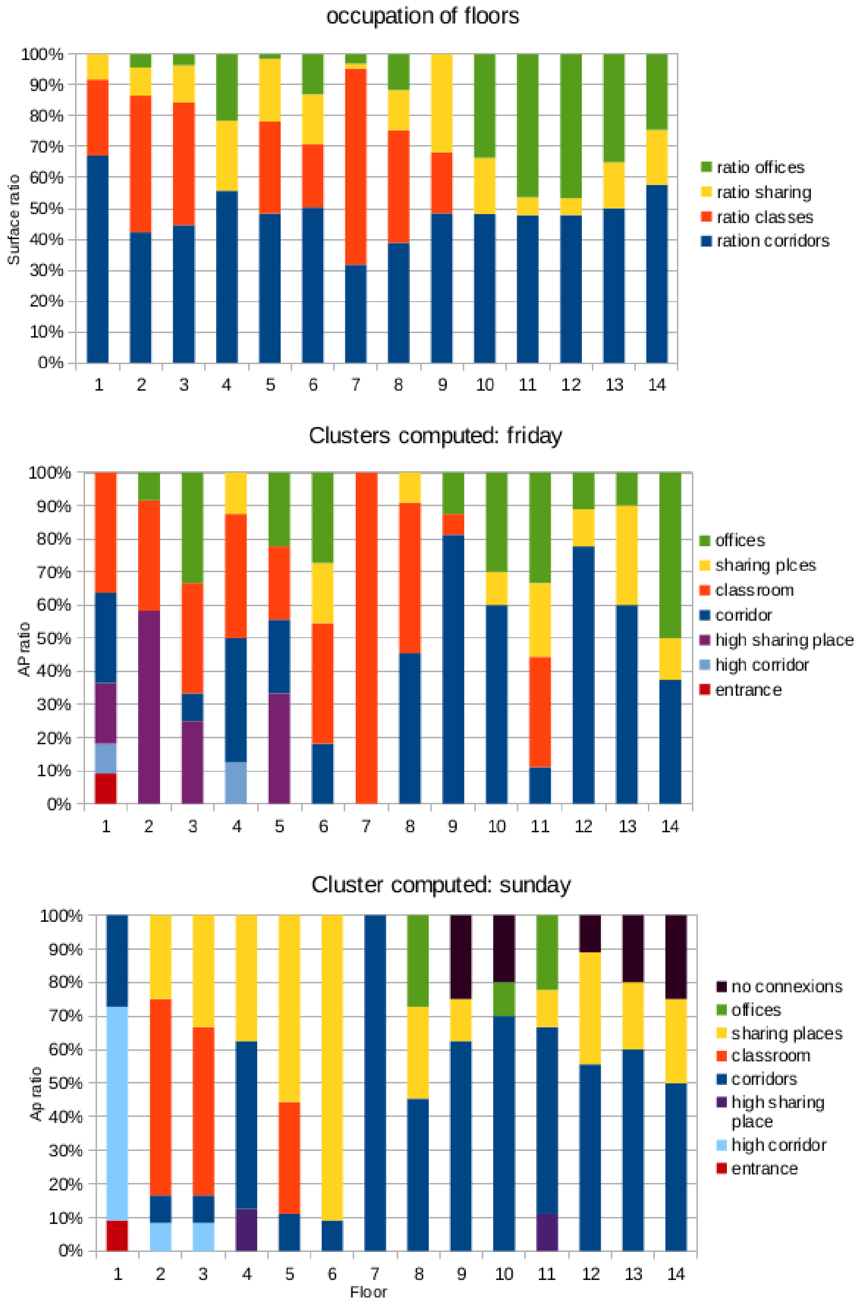

Figure 4.7 Comparison between supply and demand

4.6.3 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 others, as long as we could ensure they were not correlated with the others or adjust for their correlation using techniques like PCA.

Table 4.5 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%

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 activity computed.

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.

4.7 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 very 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. One possibility that has been proposed in some mobility models is to divide the day. 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.

In recent years, there have 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 a great potential for automatic monitoring of urban infrastructure usage and activity pattern analysis. The development of methodology in this paper allows us to take raw, rich, and readily available Big Data and develop in-depth analysis of activity and mobility behavior. This therefore represents a great opportunity to furnish detailed and low-cost analytics. Applied to large scale Wi-Fi networks, it could passively monitor a city and furnish the scale and accuracy of the data that conventional surveys cannot achieve.

CHAPTER 5 GENERAL DISCUSSIONS

5.1 General discussions

The methodology proposes a process to enrich the data passively collected concerning the communication network, adding the activity based mobility dimension to the database. In this study we adopt the network centered view, positioning ourself at a location and describing the characteristics of connections performed there. The connection data we used are not geographically referenced to allow the methodology to increase the objectivity of the activity clustering. Such an approach furnishes a process to monitor the usage of infrastructures independently of its architecture.

Some previous work on the topic used a mathematically more accurate characterization of the number of connections as function of time through eigen decomposition, we choose a method allowing to take time into account. As show the important correlation between the different indicators selected, the different descriptive variables are mainly linked to two main factors : the number of users attending the place, and the time spent at the location. If using a principal component analysis reduce the bias introduced by the correlation, bringing less correlated input data would likely improve the quality of the results. The next section discusses potentially different variables improving the quality of the clustering.

As mentioned, the principal component analysis is used to limit the effect of the correlation of the input variables. If the computation of the clustering with and without the principal component analysis shows a slightly better fit without it, the PCA allow to obtain a better distribution of the clustering variable. Indeed, as show 5.1, four of the clusters generated without PCA have very close centroids making harder the differentiation of activities.

The clustering is applied on data aggregated at the scale of a day. We make the hypothesis that there is only one main activity performed at a give location and that this activity doesn't change through the day. If such an hypothesis is false, the machine learning process furnishes groups showing similar behaviors without any semantic meaning. If we simplify the qualifications of these clusters associating one main activity, these clusters correspond to a same mix of activities. The error brought by the daily aggregation could be answered by dividing the day in smaller independent time intervals and applying the same methodology to the whole set (clustering a number of records equal to the number of APs multiply by the number of time intervals).

As mentioned in the article, the validation process shows some weakness as we do not possess

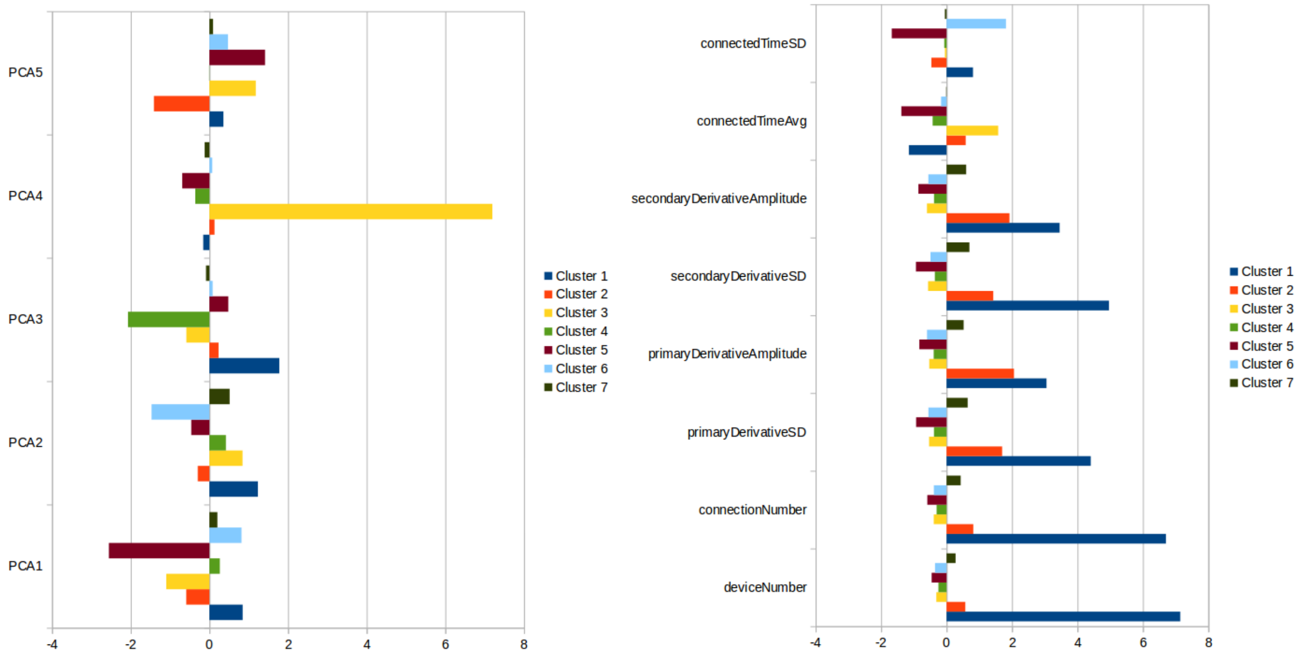

Figure 5.1 distribution of the centroids

information concerning the actual activities performed by the users. Such information could be obtained through surveys, or accurate users-centered monitoring (smartphone application). In our case, we possess the information concerning the type of infrastructures around the APs, and by deduction the main activities performable within these locations. While our methodology produce the demand of the users, we only possesses the supply furnished in the buildings.

5.2 Improvement proposed : set of indicators

We propose in this section a new set of indicators to improve the model previously described. We first replaced the values linked to the derivatives (which were highly correlated to the connection number) by the count of the arrivals. This indicator is more accurate than the variation of the number of connections, the derivative being potentially null if the entering and going out flow were equal. If the time component linked to the time spent within a location is still taken into consideration, we add a new time linked variable i.e. the "location attachment" (or importance of the location for a given user). We define the attachment of a user to a location as the ratio between the time spent by the user at this location during the week, divided by the total time spent by the user in the building. We then compute, for each location, the average attachment of the daily users to this place. This variable, if obviously correlated to the time spent at the location, aim to bring some information concerning the users.

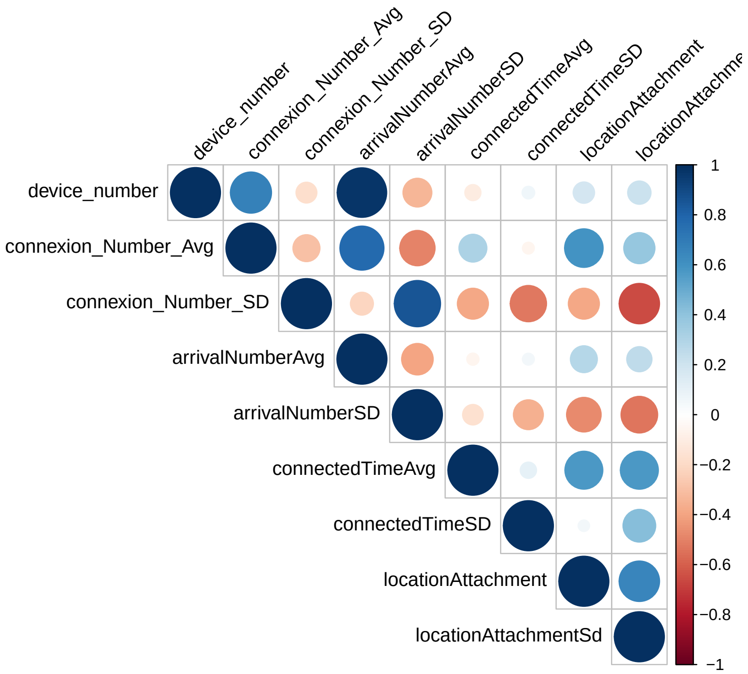

Figure 5.2 Correlation matrix

Figure 5.2 shows the correlation of the new indicators is far lower with the new set. As a consequence, the test with and without principal component analysis show that the PCA is less pertinent in that case. Then the analysis of the results given by the silhouettes are lower than the ones from the previous set of indicator (which could be explain by the lower correlation). However, the dimension brought by the "attachment of people to the places" bring an valuable information to analyse the activity of the generated clusters, as it is easily interpretable. Then, the compareason with the supply infrastructures show a better fit with the activities generated, as shows Table 5.1. For these reasons, we kept this set of indicator for the next step described in the next section.

5.3 Improvement proposed : disaggregation of the day time

As we pointed at it before, one of the main limitaion of our work are the time and spatial aggregation of the activities in our model. So far, we associated one type of activity for a defined space, for a day. This hypothesis allow to facilitate the calibration of models, but

Table 5.1 Average surface ratio around the APs

Cluster of activity	ratio of corridors	ratio of sharing spaces	ratio of classrooms	ratio of offices
Entrance	71%	06%	23%	00%
Corridors	53%	10%	13%	24%
Sharing place	39%	39%	14%	8%
Offices	44%	00%	22%	33%
Classrooms	41%	17%	21%	22%
High attendance/classroom	45%	29%	23%	03%

doesn't take into account the dynamic and heterogeneous character of urban environement.

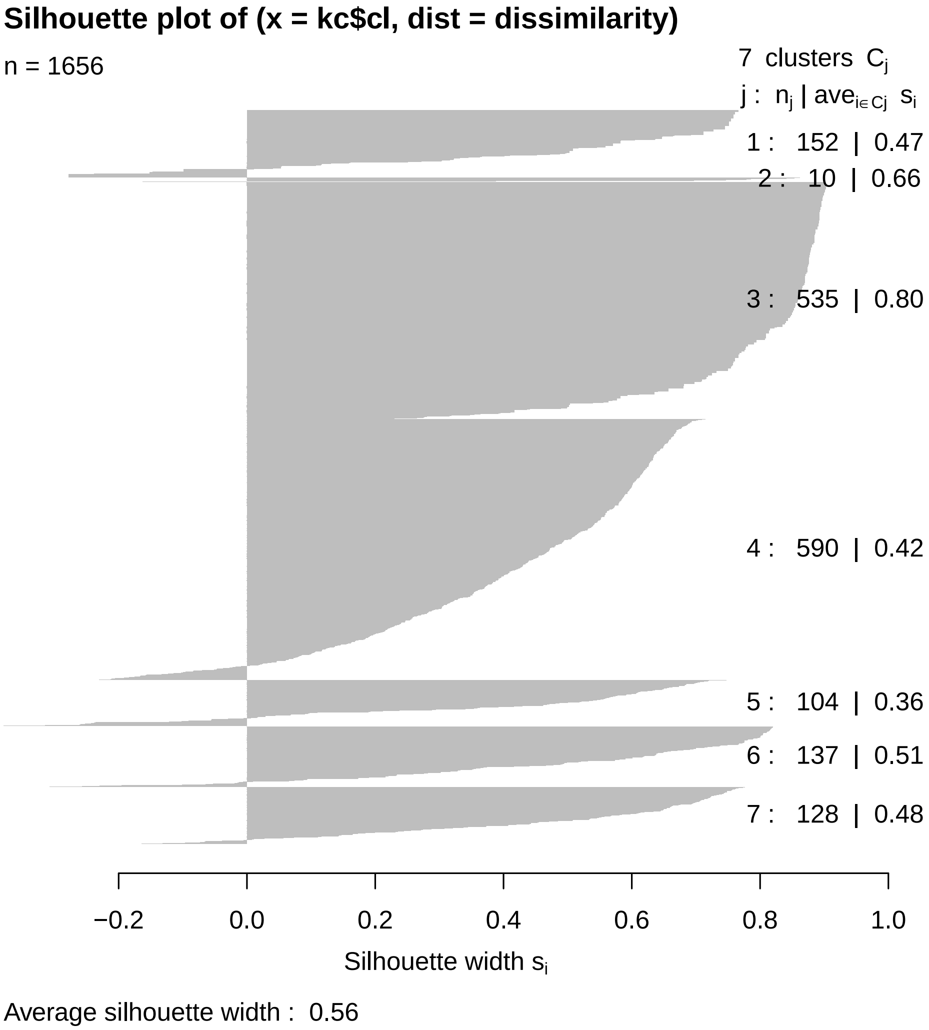

Figure 5.3 Silhouette plots of the clustering result

As a final section of our work, we proposed a step further in our model, decomposing the days into sub-time intervalles, allowing the activity to change through the day. We divide each day into n equals intervalles of time and compute the indicators for each one at each access point. We then apply our methodology, clustering the whole set of values as if they were independant APs. Doing so, we cluster n time more points than previously. The length of the time intervalles influence the impact of the aggregation on the data, long intervalles would smooth the noise while short ones would describe more accurately the behaviour within the building. We define it empirically at 2 hour, using the Silhouette analysis, trying to optimize the fit of the clustering. As shows Figure 5.3, the results obtained show a better quality for the clustering.

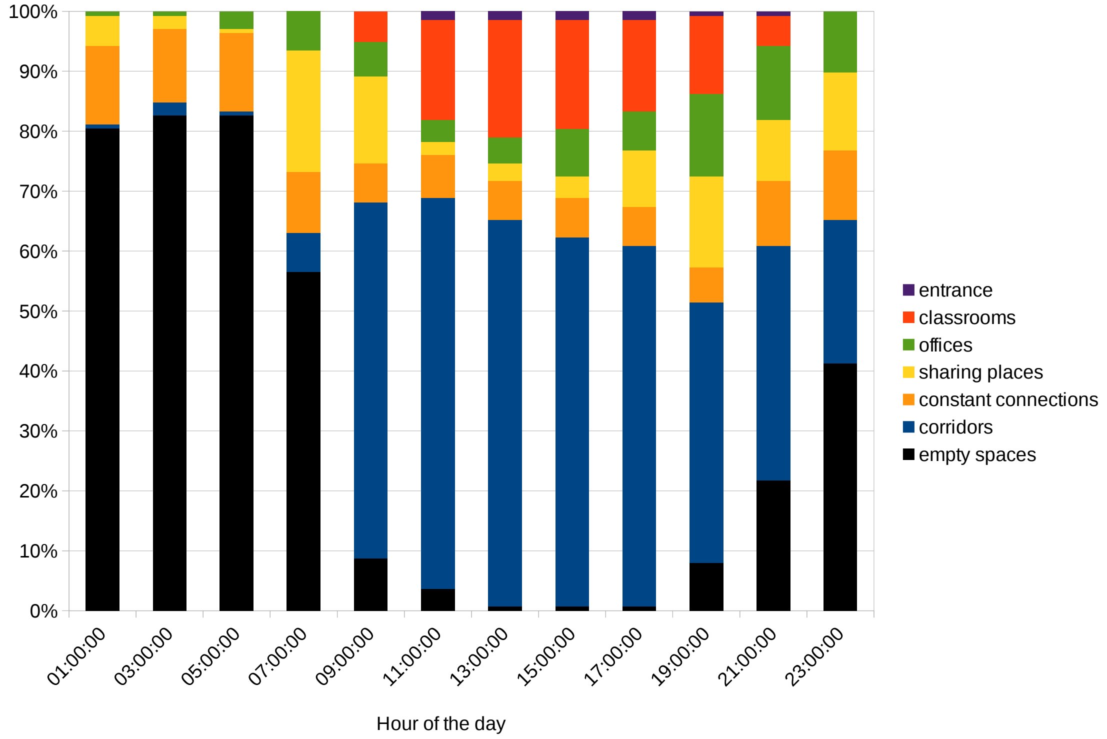

Figure 5.4 Building activities through the day

Once generated, we can observe the evolution of the activity computed through the day as show Figure 5.4. We associated the sementic meaning in the same way that previously explained. A first observation is the important presence of "empty activities" corresponding to the absence of connections. This phenomenon appears mostly during night, but also during day. These phenomenon was shadowed by the aggregation in the previous part of our work. A second observation is the presence of a cluster of permanently connected devices with a very high "location attachmet". These APs have probably fixed devices connected to them.

Such a processs give us an additional tool to understand and analyse the behaviours of users/devices within the building. However the arbitrary choice of the duration of the time intervalle studied, and their start/end time can bias the results by splitting a given activity time like a class, between twod time intervalles with different activities. Then, the compareason with the supplies his made harder highlighting again the of working with a single activity per AP.

CHAPTER 6 CONCLUSION AND RECOMENDATIONS

A methodology has been proposed to enrich the data collected through Wi-Fi network monitoring, bringing the concept of behavior activity to the different locations. This work differentiates itself from the previous ones in the field by applying to smaller scale areas with less contrasted types of activities. To characterize these relatively close types of activities, we bring the notion of time spent in the location and its distribution between the different places for each user.

The results obtained represent the evolution of the main activities performed in each location through the week. The validation process is limited by the lack of data concerning the real activities performed. However, the activities computed are compared to the infrastructure of the building and show satisfying similarities. In consequence, the result obtained give interesting overview of the human activities within infrastructures, but an additional set of validation data should be added to be able to furnish a usable calibrated model.

6.1 Limitation of the proposed solution

The work described previously is based on a certain amount of hypotheses limiting the robustness of the methodology and the accuracy of the results produced. Some of these hypotheses are essential due to the structure and nature of the data studied, some others could be improved by increasing the complexity of the models.

The first limitation is due to the internal errors present in the Wi-Fi log database. As described in Section 1.2, a challenge pointed in various studies is the non representativeness of the some connections records compared to the actual movement of the device. One of the main phenomenon, called the Ping Pong effect, correspond to the record to a succession of alternative connections/disconnection between two APs while the user is static. The impact of such phenomenon can be decreased through time or space aggregation, but bring an error in the data. Then, we observe an other limitation linked to the input data is the 5 minutes aggregation of the connections, which is here particular to our dataset. This phenomenon discussed in the thesis, aggregate the connections with a duration inferior to five minutes, decreasing the accuracy of the description of short connection, and so, of short activities. If we considered such an agregation acceptable and adapted our time analysis to this uncertainty, the access to a more accurate data set would increase the accuracy of the description of the short activities as transitions in corridors.

After having considered the limitations brought by the input data, we made hypotheses concerning the users' behavior allowing to build our methodology. The absence of similar methodologies in this field pushed us to arbitrarily select the clustering variables (so called indicators). We try through this work to validate the hypothesis that these variables can determine the type of activities performed through clustering process. As shown in the previous section, these variables can be modified to improve the methodology. The problematic part is to evaluate the pertinence of these variables considering that data mining algorithm always give result, whatever is the quality of the input data or parameters.

Then our work is based on the concept of activities performed by users at the different locations of the facilities. To simplify the computation, we supposed that each location had one main activity which is fixed for the day. This assertion is obviously false and limiting while the different locations of a facilities often host different simultaneous activities varying through the day. The possible solutions to this problem are discussed in the next section.

While we defined a specific methodology for a day, we simplified it when applying it to the week to facilitate the analysis of the evolution of the activities. While we propose for each day to define the appropriate number of activity clusters, we considered that these clusters were the same for the whole week. In consequence, we took the parameters of the day showing the higher number of clusters, and used these parameters for the other days. Making this hypothesis of the persistence of the type of activities is the basis for the identification algorithm proposed in the article. However, this limiting hypothesis could be dropped by using a different number of cluster for each day, and adapting the algorithm linking the clusters through the week in consequence.

The last limitation we point out, and one of the most important one is the relative weakness of the validation process due to the lack of appropriate validation data. As discussed few times previously, the validation work as been made comparing the supply (activity computed) and the demand (infrastructure architecture) available. If these variables are highly correlated, they often differ and are in fact the results showing a real application, allowing to process to space and schedule optimization. The methodology result would benefit a lot from getting the appropriate validation data set which could be obtained through a survey or phone monitoring.

6.2 Further improvements

A first improvment for the work would be an extensive sensitivity analysis in the different parameters choice and hypothesis made. Such a study would be essential to evaluate the transferability of the methodology to different infrastructure and environments.

Considering the points raised in the last section, few improvements could be brought to improve the methodology, with a reasonable amount of work. These modifications correspond to the decomposition of the daily connections record into smaller time intervals containing different types of activity. The decrease in the time period study could take place transforming the daily connection of one AP (one entity of the clustering) into n independent time complementary period of connections (n entity for the clustering). Doing so, we would cluster a higher amount of "access points" defining independently for each time of the day, a type of activity performed. One of the issues of such a process is the establishment of the time period boundaries : the process of clustering being optimized if these boundaries correspond to the real change of activity. If setting some arbitrary boundaries would appear as the first step, it could be imagined to set these time boundaries as a parameter of a heuristic optimization algorithm. However, such a concept would raise a certain amount of limitations, in particular the definition and implementation of an objective function judging the quality of the day divisions and its impact on the clustering process.

A third point, would be the translation from the computation of one main activity per location to the description of multiple activities. This process could start with the disaggregation of the connections at the APs. Instead of taking all the connections and generating global variables, the clustering would be done one the connections themselves. Such a process would furnish the distribution of the types of activities for each AP, where our process furnishes the global trend. However, a closer analysis of the individual users behavior through their trip chains and habits should be done to bring enough information to the clustering to be applicable to the connections themselves. While our work is mainly network centric, these studies would need a deep user centric analysis to be performed.

REFERENCES

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

Adler, T. and Ben-Akiva, M. (1979). A theoretical and empirical model of trip chaining behavior. *Transportation Research Part B : Methodological*, 13(3) :243–257.

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

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1) :47.

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

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

Bettstetter, C. (2001). Mobility modeling in wireless networks : categorization, smooth movement, and border effects. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(3) :55–66.

Boldrini, C., Conti, M., and Passarella, A. (2008). User-centric mobility models for opportunistic networking. In *Bio-Inspired Computing and Communication*, pages 255–267. Springer.

Cafarelli, D. A. and Yildiz, K. O. (2004). Method and apparatus for filtering that specifies the types of frames to be captured and to be displayed for an ieee802. 11 wireless lan. US Patent 6,693,888.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., and 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., and Ratti, C. (2010a). Human mobility prediction based on individual and collective geographical preferences. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 312–317. IEEE.

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

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

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

Danalet, A., Farooq, B., and Bierlaire, M. (2013). Towards an activity-based model for pedestrian facilities. In *13th Swiss Transport Research Conference*, number EPFL-CONF-186042.

Danalet, A., Farooq, B., and 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. and He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, page 29. ACM.

Eagle, N. and Pentland, A. S. (2009). Eigenbehaviors : Identifying structure in routine. *Behavioral Ecology and Sociobiology*, 63(7) :1057–1066.

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

Farooq, B., Bierlaire, M., Hurtubia, R., and Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B : Methodological*, 58 :243–263.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3) :37.

Femijemilohun, O. and Walker, S. (2013). Empirical performance evaluation of enhanced throughput schemes of ieee802. 11 technology in wireless area networks. *arXiv preprint arXiv :1309.2789*.

Francois, J.-M. (2007). Performing and making use of mobility prediction. *Universität Liège*.

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

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

Han, J., Pei, J., and Kamber, M. (2011). *Data mining : concepts and techniques*. Elsevier.

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

Herrmann, K. (2003). Modeling the sociological aspects of mobility in ad hoc networks. In *Proceedings of the 6th ACM international workshop on Modeling analysis and simulation of wireless and mobile systems*, pages 128–129. ACM.

Iqbal, M. S., Choudhury, C. F., Wang, P., and 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., and Abraham, A. (2016). Automated clustering of trajectory data using a particle swarm optimization. *Computers, Environment and Urban Systems*, 55 :55 – 65.

Jain, A. K. (2010). Data clustering : 50 years beyond k-means. *Pattern recognition letters*, 31(8) :651–666.

Jain, R., Lelescu, D., and Balakrishnan, M. (2005). Model t : an empirical model for user registration patterns in a campus wireless lan. In *Proceedings of the 11th annual international conference on Mobile computing and networking*, pages 170–184. ACM.

Jiang, S., Ferreira, J., and 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., and 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., and Manolopoulos, Y. (2003). Clustering mobile trajectories for resource allocation in mobile environments. In *Advances in Intelligent Data Analysis V*, pages 319–329. Springer.

Kononenko, I. and Kukar, M. (2007). *Machine learning and data mining : introduction to principles and algorithms*. Horwood Publishing.

Kusakabe, T. and Asakura, Y. (2014). Behavioural data mining of transit smart card data : A data fusion approach. *Transportation Research Part C : Emerging Technologies*, 46 :179 – 191.

Liu, L., Andris, C., and Ratti, C. (2010). Uncovering cabdrivers' behavior patterns from their digital traces. *Computers, Environment and Urban Systems*, 34(6) :541 – 548. GeoVisualization and the Digital CitySpecial issue of the International Cartographic Association Commission on GeoVisualization.

Liu, T., Bahl, P., and Chlamtac, I. (1998). Mobility modeling, location tracking, and trajectory prediction in wireless atm networks. *Selected Areas in Communications, IEEE Journal on*, 16(6) :922–936.

Long, Y. and 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.

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

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

Munizaga, M. A. and 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.

Musolesi, M. and Mascolo, C. (2007). Designing mobility models based on social network theory. *ACM SIGMOBILE Mobile Computing and Communications Review*, 11(3) :59–70.

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

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

Patterson, Z. and 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., and Kjærgaard, M. B. (2015). Spatio-temporal facility utilization analysis from exhaustive wifi monitoring. *Pervasive and Mobile Computing*, 16 :305–316.

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

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

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

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

Su, W., Lee, S.-J., and Gerla, M. (2001). Mobility prediction and routing in ad hoc wireless networks. *International Journal of Network Management*, 11(1) :3–30.

Wanalertlak, W., Lee, B., Yu, C., Kim, M., Park, S.-M., and Kim, W.-T. (2011). Behavior-based mobility prediction for seamless handoffs in mobile wireless networks. *Wireless Networks*, 17(3) :645–658.

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

Watts, D. J. (1999). Networks, dynamics, and the small-world phenomenon 1. *American Journal of sociology*, 105(2) :493–527.

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

Xu, Q., Ding, C., Liu, J., and Luo, B. (2015). Pca-guided search for k-means. *Pattern Recognition Letters*, 54 :50 – 55.

Yamamoto, T., Fujii, S., Kitamura, R., and Yoshida, H. (2000). Analysis of time allocation, departure time, and route choice behavior under congestion pricing. *Transportation Research Record : Journal of the Transportation Research Board*, (1725) :95–101.

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

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

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

APPENDIX A Building

Buildings map

We provide the space distribution of the entire building in Figure A.1

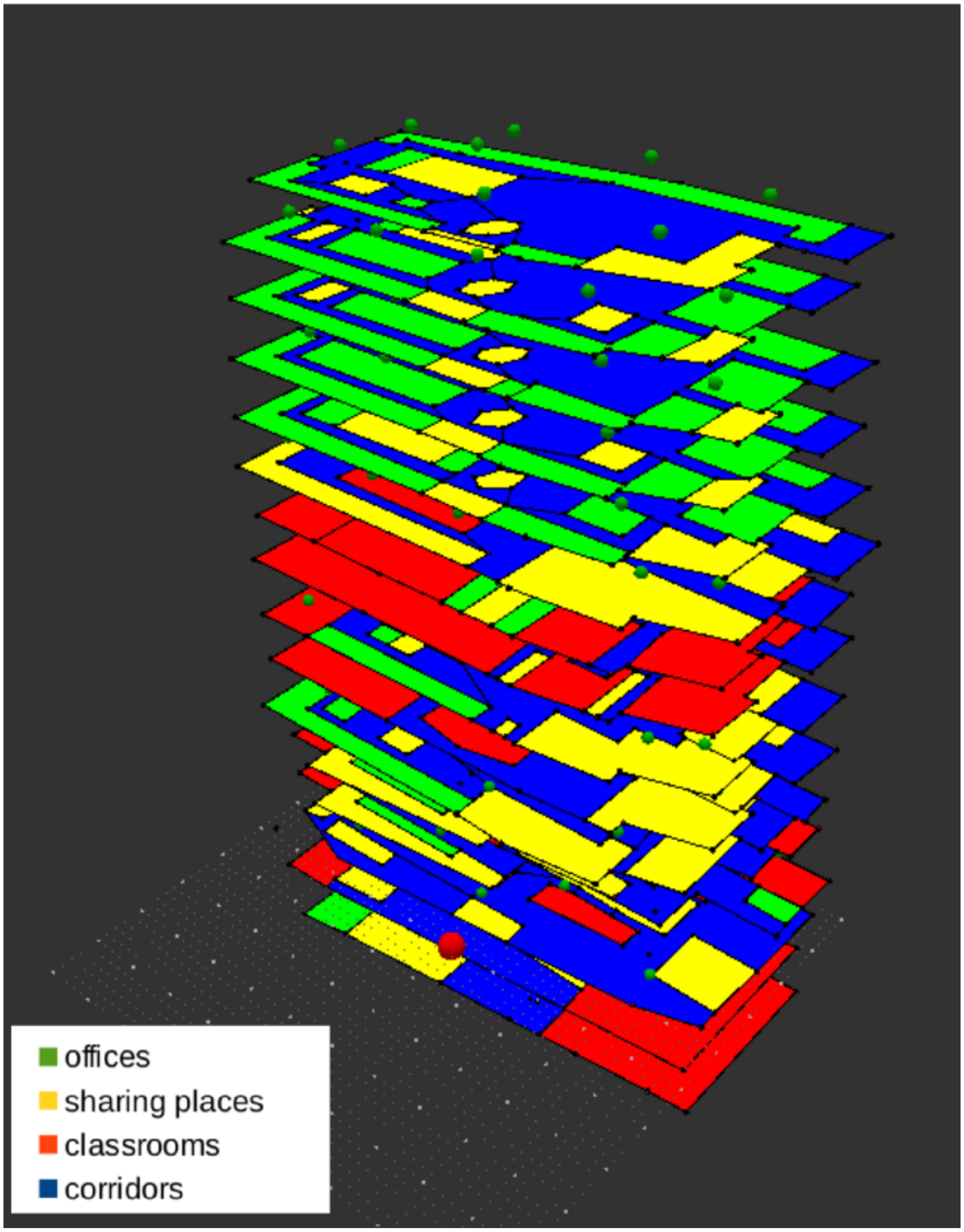

Figure A.1 Building space distribution map

APPENDIX B Statistics

Indicators statistics

In Table B.2 we present the descriptive analysis of the used variables. While in Figure B.1 we demonstrate the existence of high correlation among the variables.

Table B.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

Table B.2 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

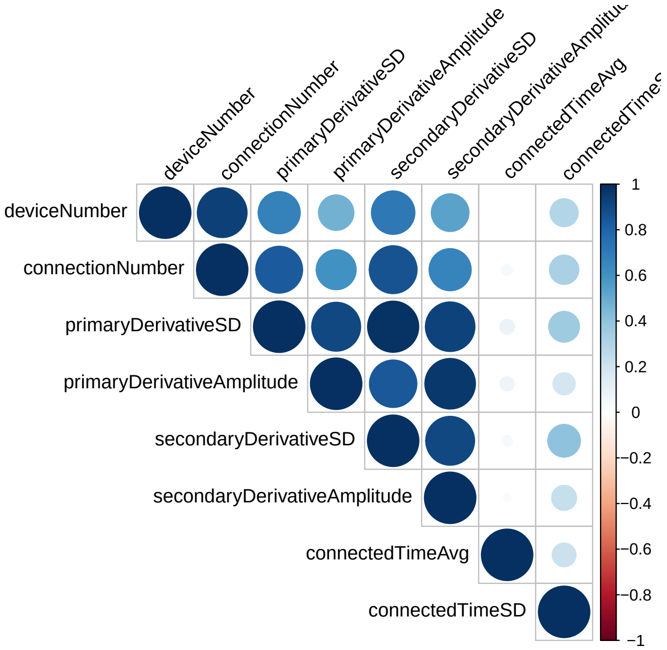

Figure B.1 Correlation matrix