OCCUPANY DETECTION & PREDICTION USING WI-FI DATA

Department of Building, Civil and Environmental Engineering
Winter 2019

Advisor : Dr. Mazdak Nik-Bakht
Co-advisor : Dr. Mohammed Ouf
Research Student : Ali Arbab

ABSTRACT

Buildings are considered as the significant players when it comes to the energy consumption, and occupancy detection is playing a critical role to improve efficiency of building management systems. Conventional occupancy detection techniques have their own share of drawbacks; therefore, inaccurate occupancy information could result in a low comfort level and energy waste. In this study, I tried to investigate the Wi-Fi data and detect the access points which behave similarly in the Engineering and Information Technology Complex (EITC) at the University of Manitoba Fort Garry campus in Winnipeg, Manitoba, Canada. I tried to detect the pattern and identify the impacts of the events on the number of connections throughout the academic year, and determine the major predictors in this domain. Also, I attempted to predict the number of occupants on my study span, but it suffered from the inadequate information in the original dataset, unapproachable building, and discontinuous Wi-Fi communication of devices.

Table of Contents

Abbreviations

- *ANN: Artificial Neural Networks*
- *AP(s): Access Point(s)*
- *ARMA: Auto-Regressive Moving Average*
- *ASCC: Associated Client Count*
- *AUCC: Authenticated Client Count*
- *BLE: Bluetooth Low Energy*
- *BRM: Base Radio Mac Address*
- *CART: Classification and Regression Trees*
- *CNN: Convolutional Neural Network*
- *CSI: Channel State Information*
- *DBI: Davies Bouldin Index*
- *Dif: The difference between the ASCC and AUCC*
- *DNTWI: Dynamic Markov Time-Window Inference*
- *DT: Decision Tree*
- *ELM: Extreme Learning Machine*
- *FS-ELM: Feature Scaled Extreme Learning Machine*
- *GHG: Greenhouse Gas*
- *HMM: Hidden Markov Model*
- *HVAC: Heating, ventilation, air conditioning*
- *KNN: K-Nearest Neighbor clustering*
- *LDA: Linear Discriminant Analysis*
- LHVAC: Lighting, Heating, Ventilation, Air Condition
- *LHVAC: Lighting, Heating, Ventilation, Air Conditioning*
- *LR: Logistic Regression*
- *LRF: Local Receptive Fields*
- *PEM: Percentage of nonzero Elements*
- *PIR: Passive Infrared*
- *RF: Random Forest*
- *RFID: Radio Frequency Identification*
- *RSSIs: Received Signal Strength Indicators*
- *SAD: Sparse Auto-encoder*
- *SVM: Support Vector Machines*
- *SVR: Support Vector Regression*
- *UWB: Ultra-Wideband*

1. Introduction

Background

The building sector is responsible for a significant amount of energy usage worldwide. Buildings in the United States use up to 41% of entire electricity consumption [1], [2], which are responsible for 39% of the whole carbon dioxide emissions [1]. In Canada, buildings are a key area of opportunity. According to the Energy Efficiency in the Buildings Sector, 1/4 of Canada's greenhouse gas (GHG) emissions are from residential, commercial, and institutional buildings. The Commercial building sector is a significant energy user and producer of carbon emissions. It accounts for 14% of end-use energy utilization and 13% of the country's carbon emissions. Canada's commercial building sector is complex and includes a wide range of building types from hospitals. This brings us to the fact that energy efficiency in buildings touches the responsibility of all levels of government in Canada. Moreover, space heating is the major use of energy for the sector [3]. Several attempts have been taken to improve building design and performance for energy preservation purposes in past decades. Based on recent studies, building occupancy detection has a remarkable potential for supporting smart control of building energy systems, which ultimately brings more efficient energy management in both residential and commercial buildings [4]–[6]. The most widely used sign of occupants' behavior is for the scheduling; thus, the occupancy schedule is able to assist the facility managers in optimizing the heating, ventilation and air conditioning (HVAC) system, which ultimately provides satisfactory comfort to the occupants.

Problem Statement

Occupancy schedules are notably based on the approximate presence of people throughout the buildings. In essence, it is excessively complicated to employ a real-time demand rate into buildings a mechanical system. Fundamentally, inaccurate data and errors could result in unacceptable service quality and waste of energy. To put it into perspective, over-cooling or over-heating not only are they able to cause higher electricity consumption, but also, they impact the occupants' comfort level. With that being said, occupancy detection with spatial information can increase the level of accuracy regarding the input information (i.e., the load that each room or spaces are experiencing), which eventually leads to address the problems mentioned earlier. To look on the bright side, this can help the environmental control of the facilities.

There are numerous studies established regarding positioning systems [7]–[10], which they could detect the users' location using Radio Frequency Identification (RFID), ultra-wideband (UWB) and other detection technologies. Also, with using cameras, we are capable to collect the number of occupants in a room. Besides, these studies have their own share of drawbacks and limitations. There are huge barriers associated with these systems such as expensive resources for large scale deployment and anonymity which make the integration process complicated.

Moreover, occupancy detection also identifies the operation of the lighting system [11], [12]. Leephakpreeda specifies that there are potentials in energy saving of lighting systems that need to be considered – around 35% to 75% – based on the proposed methodology [13]. Additionally, people presence and tracking are essential for the security management and emergency evacuations [14].

Our environment presently deals with significant threats, including climate change, fossil fuel depletion, etc. The academic sector as one of the primary drivers should consider this since the educational buildings and campuses are significant players in energy consumption (e.g., library, academic, research space, laboratories, open space, office). Accurate detection of the people presence is crucial and important to improve the performance of the building management system. The occupants' behavior, the prediction of occupants in each space, the spaces load detection, access points load is not determined yet. The primary intent of this research is to focus on occupants' behavior and patterns in an educational building using the Wi-Fi data collected from the access points (AP), which installed throughout the Engineering and Information Technology Complex (EITC) at the University of Manitoba Fort Garry campus in Winnipeg, Manitoba, Canada. However, the results in this research will not be contributed as a linkage between the Wi-Fi data and the HVAC system.

Objectives

1. Inspecting the APs' behavior and what attributes impact the number of connections
2. How the behavior changes for an AP over the weeks

3. Identifying the APs that are acting similarly i.e., similar pattern
4. Predicting the number of occupancies based on the data
5. Using data mining and statistical method to see how many weeks it takes for an AP to converge to the mean value

2. Literature review

Nowadays, occupancy detection, prediction and estimation has emerged into an active field of research. There are many reasons and considerable benefits associated with this topic and here, I am going to contemplate an overview for building occupancy detection and estimation. Several solutions exist for this purpose using different sensors.

2.1. Technologies

2.1.1. Passive Infrared (PIR)

The passive infrared (PIR) sensors are capable of identifying the infrared radiation alterations which resulted by displacement of object. That being said, they are capable of detecting the presence of occupants. Dodier et al. came suggested a network using PIR sensors [15]. In fact, three PIR sensors were responsible for sensing the people's presence separately. Subsequently, Bayesian probability employed to derive the absence of the occupants in an area. Moreover, Duarte et al. inspect a protracted period occupancy variation of several rooms using PIR sensors. After investigating the occupancy patterns, it turns out that actual patterns are substantially different from the standardized occupancy versatilities which are frequently used in energy simulation tools.

A Hidden Markov Model (HMM) employed for occupancy detection using PIR sensors in [16]. Few other studies straightly applied the active state of PIR sensors to develop the statistical model for the building occupancy [17][18]. Besides, when it comes to counting the number of occupants, PIR sensors can be used for this purpose. Wahl et al. proposed a system grounded on PIR sensors that can identify directions that people are moving in an area [19]. An uncomplicated direction-based algorithm along with a probabilistic distance-based algorithm applied to achieve this objective. Employing only one PIR sensor, Raykov et al. abled to analyze and estimate the number of occupants [20]. They began with deriving the motion patterns of people from raw data with adapting an infinite HMM model. Then, the patterns applied for estimating the number of people presence using the regression models. That being said, the PIR sensor are low-cost system to be utilized in the buildings/facilities. Nevertheless, this sensor has its own share of drawback, which can only detect the moving people and it does not consider the static occupants.

2.1.2. Environmental sensors

Environmental sensors such as temperature, CO_2, pressure, and humidity are well-known and widely-used for the varieties of purposes. Occupants directly impact the indoor environment characteristics. There are numerous advanced studies which focused on occupancy detection using the environmental sensor. Among all the sensors, CO_2 frequently used for demand-controlled HVAC systems.

2.1.2.1. CO_2 sensors

The first and foremost reason why these sensors are being employed is the high correlation between the CO_2 levels and occupancy level, which makes them a significant indicator of occupancy detection [21]. Furthermore, some researchers implement only the data from CO_2 sensors for the building or facility occupancy estimation/prediction.

A simple method was proposed by Ansanay-Alex, in which they only considered the alteration in the CO_2 level. This method indicates only the presence or absence of people inside an area. However, Wang and Jin [22] developed a dynamic occupancy recognition, which works by measuring the carbon dioxide concentration of the return air and the outdoor air flow rate, and they have contrasted it with a stable-state. A straightforward method has been provided by Szczurek et al. which can forecast the number of people implementing the statistical pattern matching. The underlying concept is that for a half an hour timeframe, they used the statistical indices of CO_2 concentrations. In particular, autocorrelation function and correlation coefficients. In this respect, they could easily give an approximate estimation of occupancy level. However, little attention has been devoted to the running this in long-term. That means the timeframe is notably long for the real-time control of lighting, heating, ventilation, air condition (LHVAC) systems. Not to mention the considerable intricacy of the CO_2 concentrations, an unequivocal modeling of CO_2 with regard to number of occupants is arduous and imprecise.

In this context, it is worthwhile to consider the data-driven methods, which are capable of modelling the connection between the input and output using machine learning algorithms. Zuraimi et al. investigated the performance of physical and statistical models in forecasting the occupant counts using Support Vector Machines (SVM) and Artificial Neural Networks (ANN) in CO2 concentrations [23]. This research indicates that in a room with up to 200 occupants, the above-mentioned machine learning algorithms performed the best. On the other hand, Jiang et al. [24] developed an indoor occupancy estimator through the use of feature scaled extreme learning machine (FS-ELM) algorithm, and using CO_2 measurements.

Prior researches have suggested that their conclusions can be considered for forecasting the occupancy using ANN with CO_2 concentrations. Having that said, time delay as a consequence of slow-spread of CO_2 is the primary challenge of this feature since the outcomes are always arriving with delay and is not proper for the real-time analysis.

2.1.2.2. Multiple sensors

Coupled with CO_2 sensors, pressure, humidity, light, and temperature are also considered as an environmental sensor which can be beneficial when it comes to occupancy detection. Candanedo and Feldheim [25] developed occupancy detection system using light, temperature, humidity, and CO_2 sensors, along with the implementation of the Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART) and Random Forest (RF) for identification [12]. They achieved an acceptable performance with proper excerpt of features and learning algorithms. Throwing light on temporal interdependencies of environmental time series data, Ertuğrul suggested that higher precision can be achieved using recurrent extreme learning method (ELM) method.

The major drawback of using these sensors as a tool for forecasting the number or range of occupants is that the raw data is extremely noisy. A few studies contemplate employing the feature engineering which is highly used for the machine learning models [26]. That means the researchers applied this method to generate informative representations following the domain knowledge. However, the majority of earlier studies focused on the manual feature extraction, which demands the field knowledge. That being said, Zhu et al. [27] developed a system using local receptive fields (LRF), which is by learning in both time and frequency. Regarding this notion, they have employed the CO_2, temperature, humidity, and pressure sensors. Besides, they discovered that feature learning, along with supplementary frequency domain data, could notably improve the occupancy estimation system. Deep learning likewise is an effective method for automatic feature learning [28]. Sparse auto-encoder (SAD) has been employed by Liu et al. for the occupancy identification system using environmental sensor with the help of automatic feature learning [29]. This technique is capable of learning features from the raw data; then, the learned feature has supplied the classifiers to discover the presence of the occupants in the considered area. Although multiple environmental sensors are broadly available for occupancy detection, estimation, and prediction, the major drawback is the delayed estimation and their limited performance.

2.1.3. Camera

High precision seems to be a common problem in the occupancy detection purpose. That being said, cameras are well-known for this purpose. Fleuret et al. developed a multi-camera system which not only delivers the number of occupancy in an indoor area but also with using individual processing trajectories over long sequences; they avoid confusing the occupants with on another [30]. A vision-based system proposed by Benezeth et al. [31] for the occupancy detection, which was based on the video analysis, using static camera. This system includes three main steps i.e. background deduction, detecting mobile objects, and identification. Other studies focused on building an occupancy model using cameras. Erickson et al. [32] integrated their developed model into building control system in order to reduce the energy consumption. The authors of [33] suggested a framework using a cascade classifier. The pre-classifier employs three-frame difference algorithms to look for a non-head area. Besides, the primary classifier uses the convolutional neural network (CNN) to classify head areas with excellent recall and precision i.e., up to 95.3% can be obtained. A recent study by Peterson et al. suggests that a high accuracy is procurable by counting the entering and leaving people in a room using unsupervised image processing techniques [34].

Although there are many studies that they implemented the camera as a solution for the occupancy detection, there are several types of research available which they employed camera for labeling the ground truth. Nonetheless, this approach suffers from specific weaknesses such as privacy concerns, excessive computational complexity, environment conditions, i.e., lack of proper lighting system.

2.1.4. Bluetooth Low Power (BLE)

Bluetooth Low Energy (BLE) is commonly available in the smartphones. Researchers widely considered using this as a tool for the detecting the occupancy. Conte et al. [35] developed an occupancy detection system using iBeacon[1]. They have employed KNN and decision tree (DT) to classify the people presence in different rooms with reference to their received signal strength indicators (RSSIs) from disparate iBeacons. Moreover, in [36], they endeavored to increase the performance of their system by 10% using iBeacon. That means the modern solution leverages on BLE standards, which then the SVM algorithm employed to classify the occupants into different areas. Likewise, Filippoupolitis et al. implemented the SVM algorithm for tackling the occupancy estimation problem with BLE beacons [37]. As opposed to employing the RSSI values, they studied some statistical features for occupancy classification. Also, they designed an experiment based on using the BLE beacons in a greater area with three learning algorithms i.e., SVM, KNN and logistic regression (LR) for occupancy estimation [38]. The most considerable drawback for BLE based systems are not only occupants might turn off the Bluetooth of their devices, but also the extra cost associated with installing and maintaining the BLE beacons in an indoor environment.

2.1.5. Wi-Fi

The occupancy estimation, detection, and prediction using Wi-Fi have mainly been studied, and many solutions have been found. Lately, it is noticed that people's movements impact Wi-Fi signals. Therefore, researchers intended to tackle this problem using RSSIs [39] and channel state information (CSI) [40].

These approaches have been influential in the field because of the availability of Wi-Fi signals in indoor environments. Besides, since the smartphones are broadly available for the occupants, it can be the proper candidate for detecting the number of occupancies using this

[1] A protocol which is developed by Apple in 2013, and it is based on the BLE proximity sensing by transmitting universally unique identifier.

method. Balaji et al. employed an estimation system with regards to the existing Wi-Fi infrastructure [41]. They presented a system that utilizes the Wi-Fi infrastructure and smartphones to provide fine-grained occupancy-based HVAC actuation. A research has provided a system for the room level occupancy detection [42]. This method captured a snapshot occupancy by pairing the received signal strength (RSS) determinations among the measurements of anchors in different zones. Then the estimation performance was improved when the temporal correlations of continuous snapshot occupancy and historical data came into work. In [33], in place of handling Wi-Fi RSS on smartphones, they employed routers to scan Wi-Fi-enabled smartphones in indoor conditions. A localization algorithm of online sequential ELM was implemented to derive the occupancy data. Likewise, Wang et al. provided a well-defined occupancy inference system using Wi-Fi scanning [43]. The authors desired to have the occupancy dynamics as a Markov process; thus, they have employed a dynamic Markov time-window inference (DMTWI) approach for the occupancy estimation. The underlying reason is due to unstable signal and unpredictable occupant behavior in which the study addressed the time-series and stochastic features of detected signals and presented a modern model for prognosticating a reliable occupancy. The authors then examined and contrasted the conventional Auto-Regressive Moving Average (ARMA) and Support Vector Regression (SVR). Considering the refined information of Wi-Fi or channel state information (CSI), authors in [44] outlined a percentage of nonzero elements (PEM) in the CSI matrix as characteristics for occupancy evaluation. Then the monotonic relationship explicitly formulated by the Grey Verhulst Model, which is used for crowd counting without a labor-intensive site survey.

The foremost problems are the facts that the majority of research works assumed each occupant would have a Wi-Fi enabled smartphone, which is not constantly true in the real world.

2.2. Building occupancy estimation

The majority of research in this field concerned with forecasting the distribution of occupants in indoor environments. The occupancy inference is often leading to generate and design of new algorithms, i.e., Counting algorithms, Tracking algorithms, Detection algorithms. Using k-nearest neighbor clustering or classification using support vector machines (SVM) in detection algorithms.

Or, a wide variety of algorithms spanning both supervised and unsupervised learning have been implemented in counting algorithms [45].

Prediction algorithms employed for forecasting the occupancy pattern inside of a building environment. For instance, the authors of [46] use a recursive algorithm for improving the accuracy of daily occupancy forecasting in a hotel environment. Another system has been presented by Page et al. based on stochastic models [47]. By considering occupant presence as an inhomogeneous Markov chain interrupted by occasional periods of long absence, the model generates a time series of the state of presence (absent or present) of each occupant of a zone, for each zone of any number of buildings. As regards, Markov models provide a proper place to start for these algorithms which are helpful in simulation stochastic interactions of occupants. Algorithms in occupancy detection have been broadly reviewed in [48].

Wi-Fi-based detection approaches have been thoroughly discussed since Wi-Fi access points (APs), and wireless devices are ubiquitously available in indoor environments. But there are several issues and limitations associated with this method. A challenging problem that arises in this domain is the system endeavor to identify the number of smart devices, i.e., smartphones, watches, tablets, and laptops, which could be found in an individual, and ultimately, significant errors in the occupancy estimation. Also, it should be highlighted that these systems can only identify the moving occupants; hence, the immobile ones would be neglected, but this will not be discussed here in this research.

3. Methodology

3.1. Understanding Data

The datasets presented in this research were collected on an hourly basis for 230 days. The time frame started from October 15th, 2014, to June 1st, 2015. This means that there were 230 CSV files, including hourly data points for each access point. All datasets merged, sorted by dates, and merged into a single file to simplify the interpreting stage, producing the training set and the test set, visualizing, modeling, handling the missing features, text and categorical attributes, and experimenting the attribute combinations.

As of the beginning, it is difficult to arrive at any conclusions concerning all of the data points for the three buildings. Therefore, this study aims to investigate a single floor with the assumption that everyone would carry at least one smart device (i.e., smartphone, laptop, tablets, smartwatch, etc.). Therefore, since E2 is the newest building among all of the engineering buildings on campus, I decided to focus on the third floor of this building, which contains a variety of rooms, including classrooms, offices, corridors, and mechanical/washrooms. Also, this floor includes 12 APs which are located all over the area. The initial dataset (which had the data for the three buildings) had 684023 data points and only five features; AP names, event time, base radio mac address, associated client count, and authenticated client count.

The date/time attribute has the date and the hour that the system collected the data related to the specific AP. The base radio mac address (BRM) shows the address associated with the AP and each AP has its own single specific address. The associated client count (ASCC) is the total number of people/devices trying to connect to the wireless network through the APs, and the authenticated client count (AUCC) is the number of successful connections which was recorded by the system.

The evaluation of the data presented in this work is done with a Python development environment, Rapidminer, and Microsoft Excel.

3.2. Pre-Processing

Generating/Excluding features and filtering examples

Due to the complexity of the engineering building and fluctuations in Wi-Fi data, at first point, I decided to focus on single floor for observing the occupancy pattern. The dataset then filtered to only APs located in building E2, 3rd floor, which ultimately, the final dataset has 62,304 data points. Also, a list of data points which have been taken out from the main dataset is provided in the appendix.

Moreover, it is essential to determine the attributes which can be valuable for the process. That being said, some attributes might have a tail-heavy distribution or might not be very useful. Thus, from the attributes mentioned above, BRM excluded since it merely presents an id for each AP. The loaded dataset has an attribute named event time that has day name, day, month, year, and hour altogether. This has been split into day_name, date (dd-mmm-yyyy), and an hour in order to make dataset ready whenever there is need to extract the information out of it

3.2.1. Authenticated (AUCC) and Associated (ASCC) Difference (Dif)

It goes without saying that ASCC always exceeds AUCC. However, a new attribute called “dif has been generated, which stands for the difference between the ACSS and AUCC to observe any significant contrast among these two values during the day. The attribute range varies from 0 to 17. Firstly, it splits into two categories: zeros and ones. Zeros indicate those in which there was no gap between the ASCC and AUCC, and the ones present the difference range from 1 to 17 (Fig 1). It has 75% zeros and 15% ones.

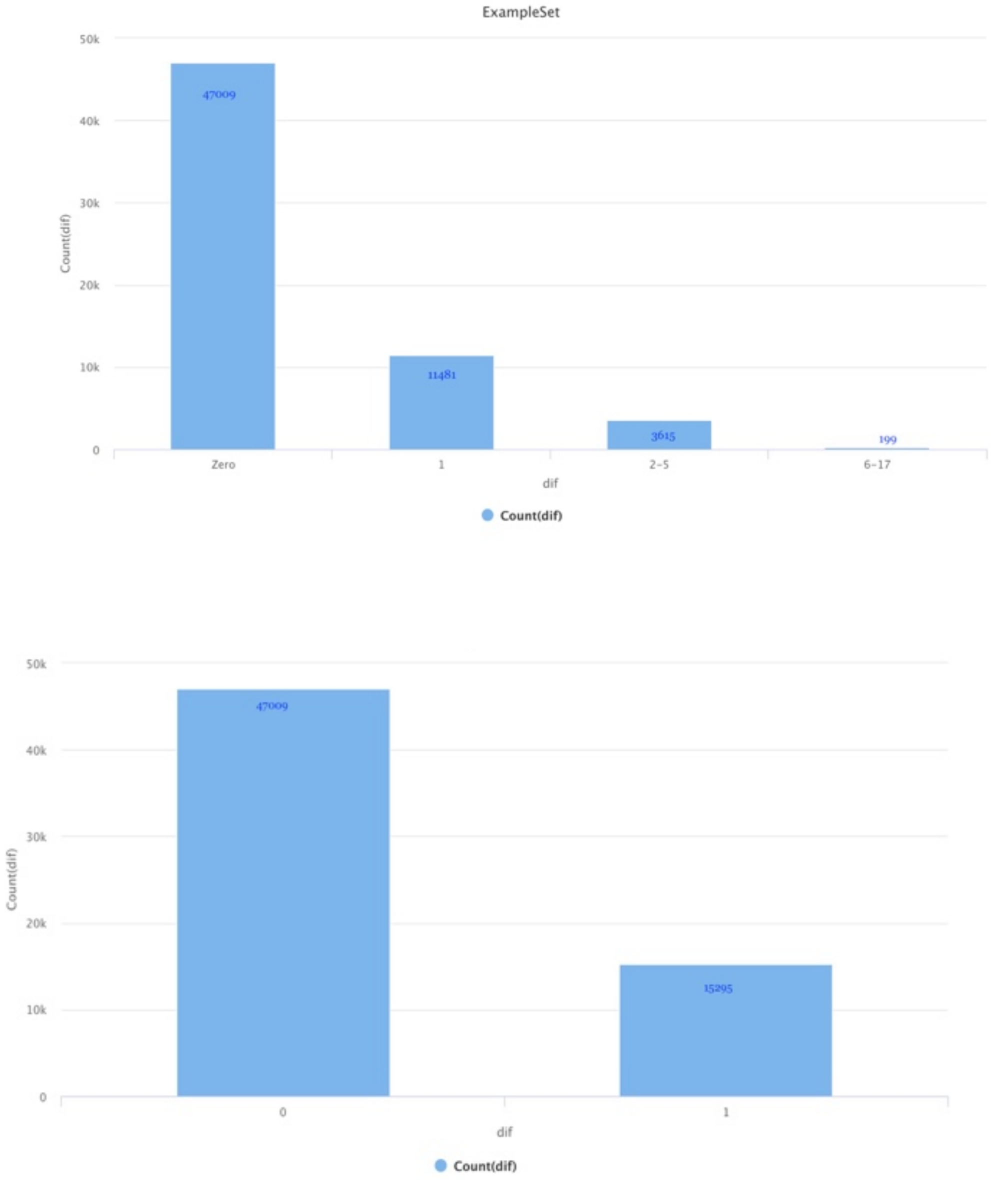


Figure 1 - dif distribution (the difference between ASCC and AUCC)

The scope for "dif" then changed to zero (no gap), 1, 2-5, and 6-17. As demonstrated in Fig. 1, a significant number of devices connected to the APs (i.e., no difference between the ASCC and AUCC). Also, 11481 ones prove that 18.4% of the collected data had only a slight difference from the ASCC number. On the other hand, around 6% and 0.3% calculated for a range of 2-5 and 6-17 respectively.

3.2.2. Weekday

This binary attribute created to merely separate the weekdays from weekends, which facilitates the investigation of the occupancy patterns throughout the weekends or weekdays.

3.2.3. Holiday

Given that here, I am investigating the data related to the educational building, several holidays need to be excluded. That means the university is closed during the holidays, and no classes/examinations will be held, according to the University of Manitoba academic calendar. This binary feature generated to observe the patterns, preparing the data for the modeling stage, and visualizing the APs behavior during the different times of the year.

Furthermore, during the data visualization stage, several time windows have been observed in which no connections were happening (Fig. 2). Data visualization was put into work, and the mean of ASCC throughout the building has been studied within the same periods. The 4th floor had a significant shift within the marked periods (Fig. 3), which means the traffic on 3rd floor has been distributed on to the 4th floor. Consequently, these periods have been excluded from the primary dataset.

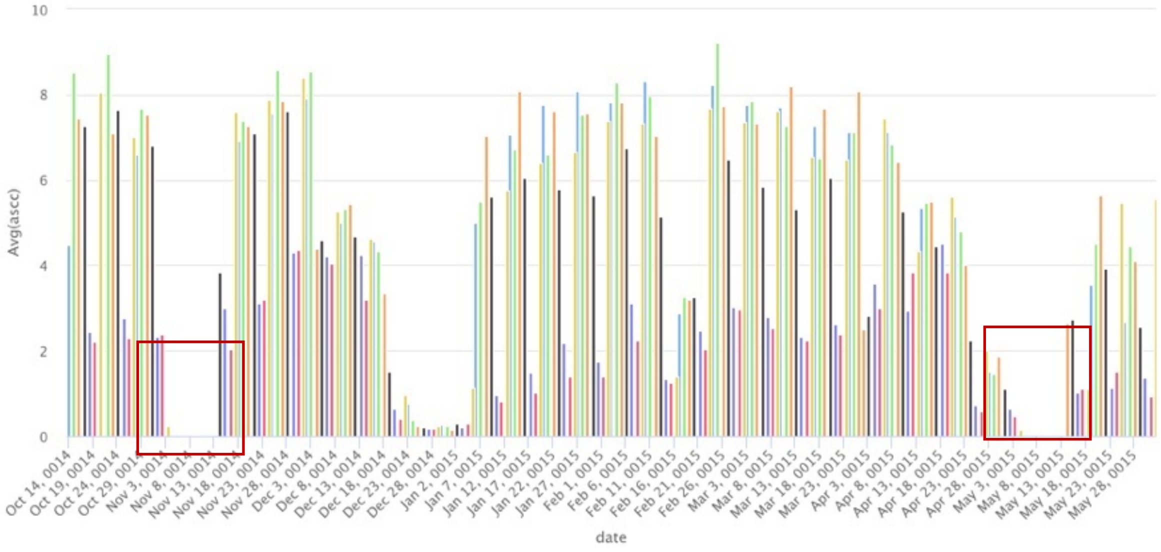


Figure 2 - ASCC was 0 during the time mentioned by red boxes

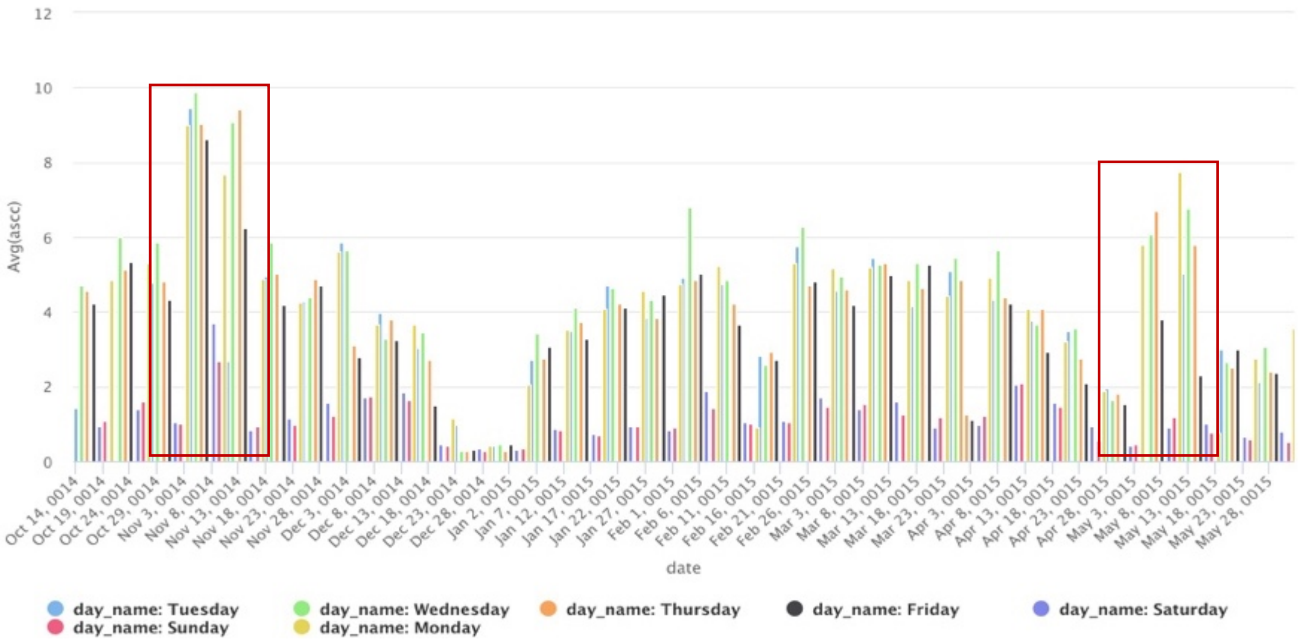


Figure 3 - Increasing the number of connections on the 4th floor during the time that Wi-Fi was shut down on 3rd floor

3.2.4. Shift

Generally, the university is running from 8:30 AM to 4:30 PM. However, this is merely the opening hours and the students might be available across the building for the rest of the day. An attribute called "shift" has been created in order to separate the day-shift (8 AM – 5 PM), night shift (6 PM – 24 PM) and overnight (1 AM – 7 AM). The limitations will be discussed further, but at this point, I settled to focus on day-shift time window, since it is the most crowded period of the educational buildings.

3.2.5. Cumulative average

The primary intent of creating this attribute is to answer this question "how many week/days is needed for ASCC to converge to the mean value?". With that being said, this value has been observed individually for each AP as shown in the (Fig. 4). However, this will not going to be included as an attribute during the t-test, clustering and regression process as a value which represents a detail for a data point.

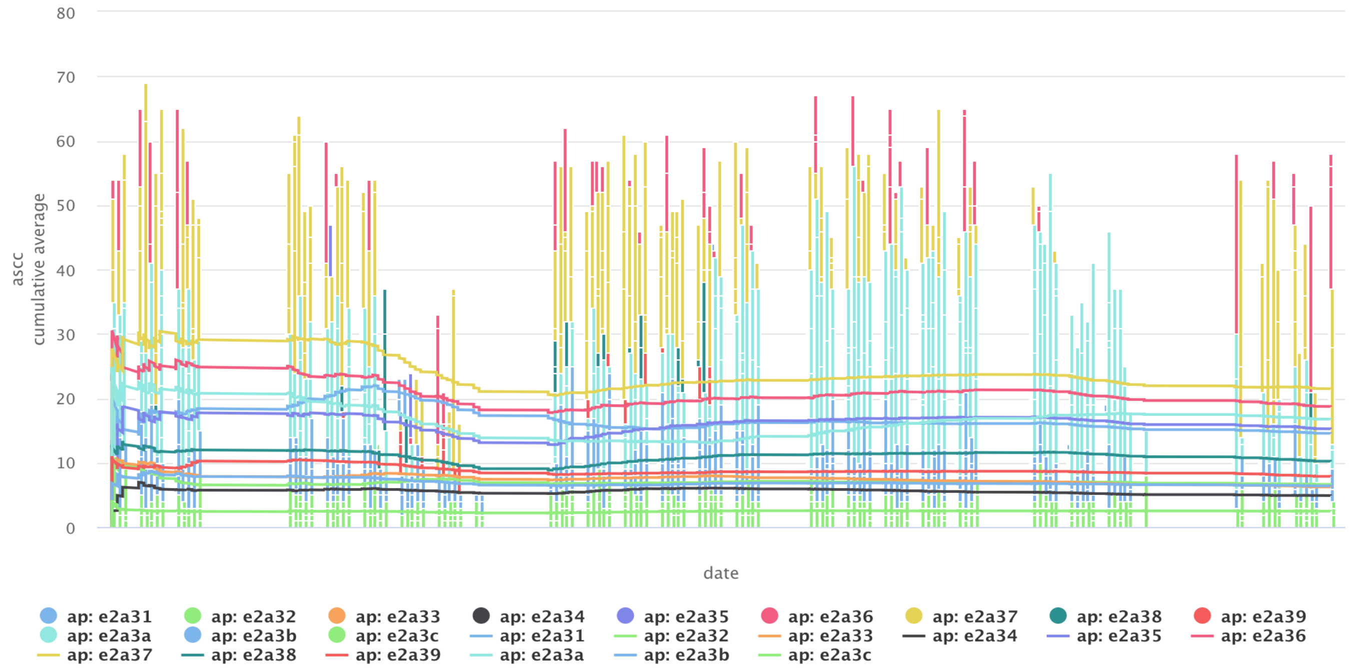


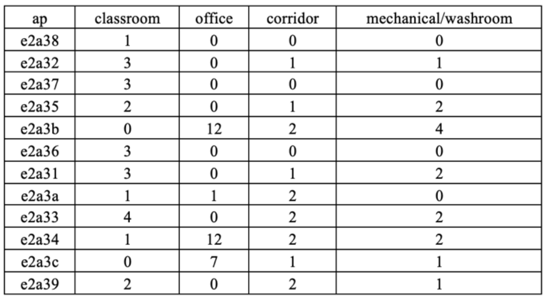
Figure 4 - Cumulative average for each AP during the time that the dataset covers, the y axis demonstrates ASCC, and the horizontal lines shows cumulative averages for each AP throughout the time

The cumulative average has been visualized for non-holidays, day shift, and weekdays. Each AP has its own cumulative average, and they vary from one another. This attribute was also applied on each AP over the Fall and Winter semester, which will be discussed further. (see Appendix A for the cumulative charts for Fall and Winter for each AP).

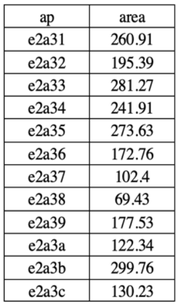
3.2.6. Classroom, office, corridors, mechanical/washrooms and area

APs are located all over the floors, and each AP is covering a particular area. Despite that APs have overlaps with one another, I assigned an exclusive space for each AP (Polygon method). Accordingly, a table developed which shows the number of classroom, office, corridors and mechanical/washrooms (Table #). These values then assigned to the entire dataset as a Boolean attribute. Besides, an attribute which shows the coverage for each AP also has been created in the main dataset (Table 1).

Table 1 – APs' spatial details and area coverage. This shows how many and what types of rooms each AP is covering. Also, the coverage area (m^2) for each AP is stated in the right chart.



ap	classroom	office	corridor	mechanical/washroom
e2a38	1	0	0	0
e2a32	3	0	1	1
e2a37	3	0	0	0
e2a35	2	0	1	2
e2a3b	0	12	2	4
e2a36	3	0	0	0
e2a31	3	0	1	2
e2a3a	1	1	2	0
e2a33	4	0	2	2
e2a34	1	12	2	2
e2a3c	0	7	1	1
e2a39	2	0	2	1



ap	area
e2a31	260.91
e2a32	195.39
e2a33	281.27
e2a34	241.91
e2a35	273.63
e2a36	172.76
e2a37	102.4
e2a38	69.43
e2a39	177.53
e2a3a	122.34
e2a3b	299.76
e2a3c	130.23

3.2.7. Semester and Week Number

As mentioned before, the data gathered on an hourly basis for 230 days. To have better intuition, the data categorized into three semesters. An additional attribute called "week number" generated, which specifies the week number of a particular semester. This would benefit us for comparing the APs behavior in similar weeks of each semester. For instance, by numbering the weeks, it is easier to analyze the AP's behavior in academic weeks between Fall and Winter semester.

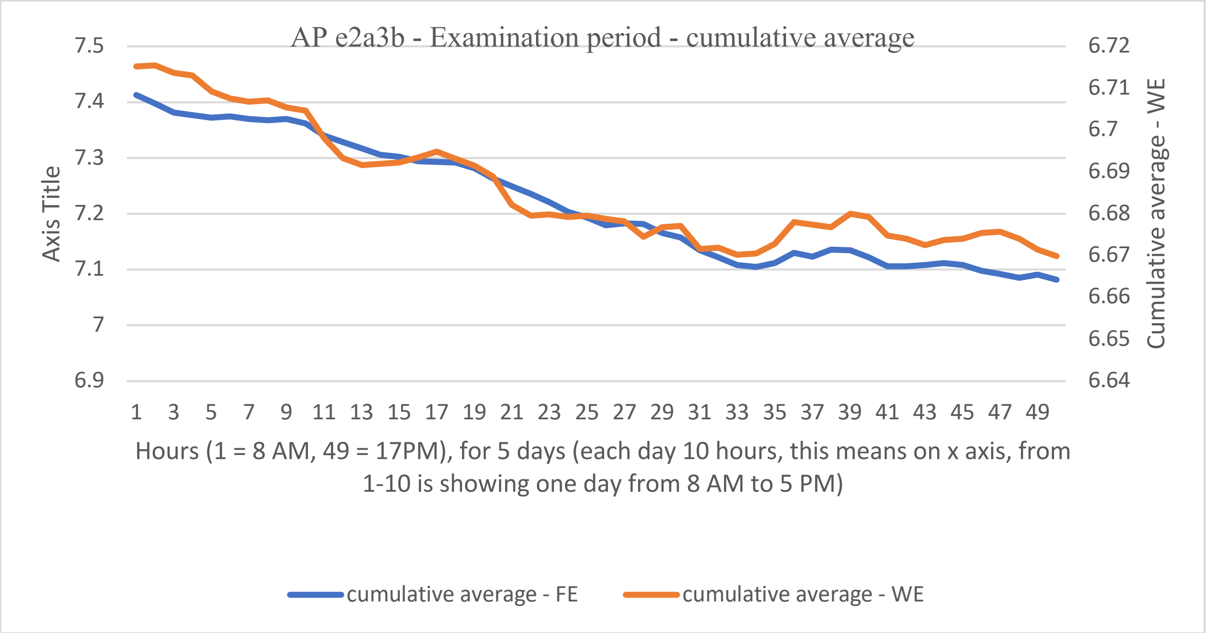


Figure 5 – Cumulative average alteration over time

As shown above, the blue line assigned for the AP's behavior over the Fall examination period and the orange line for a similar purpose in the Winter examination period. The x-axis demonstrates we are moving from hour 8 to 17 each day (10 hours per day, five days per week, therefore 50 hours per week), and the y-axis presents the cumulative average. Although the cumulative average ranges are slightly different, it is clear that they almost have identical behavior. Besides, the detailed plots regarding the selected weeks for each semester will be discussed further in this research.

3.3. Polygon Method and Room Assignment

In the first place, there are two perspectives which the data can be categorized. AP-based and room-based. For the latter category, each AP needs to be assigned to certain rooms. Therefore, Polygon method put into practice. In essence, nearby APs have been connected, then the perpendiculars of each line drawn and extended to the edges of the floor. The intersections of perpendiculars have been considered reasonably for each area to allocate particular space for a single AP (Fig. 6). Cyan lines are responsible for connecting the APs, and red lines are the perpendiculars of the connection lines. Furthermore, the APs cover a horizontal radius of 50ft; hence, this measure also has been checked to see if the assumption for the Polygon method was right or not. The limitations associated with this method is that APs are covering the area in a spherical shape, and they have overlaps with each other. This method cannot consider those overlaps and accurate coverage.

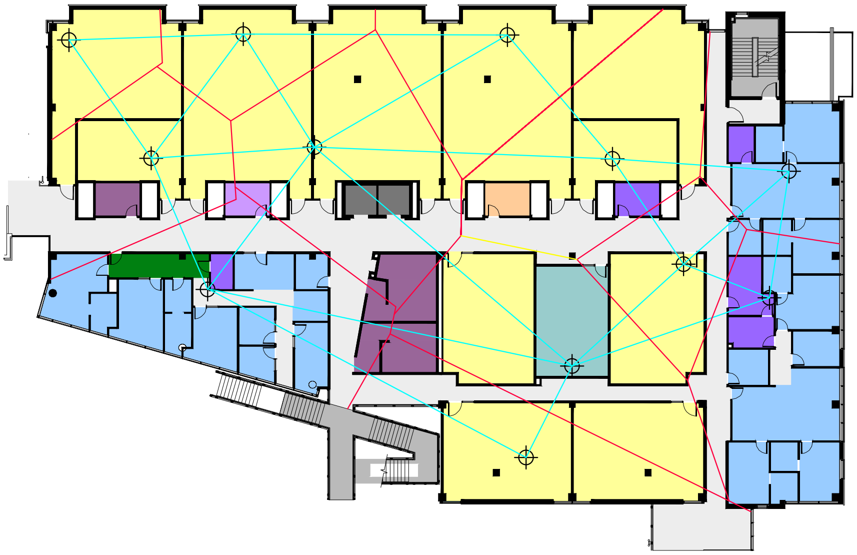


Figure 6 – Developing coverage area for each AP using Polygon method

3.4. Similarity tests

The primary intent of the statistical analysis is to identify the trends and similarities in the dataset. This method could be implemented in order to find the patterns in unstructured or semi-structured data, which ultimately led the research to a better experience and more insightful intuition about the principal concept. Also, statistical methods are required to ensure that data are interpreted correctly, and those apparent relationships are meaningful with no chance occurrence.

After a careful investigation the entire dataset, it has been concluded that the complete data is available for the weeks 7, 8, 13, 14 and 15. This means that the pattern recognition or tests can be carried out on the equivalent week numbers throughout the Fall and Winter semester.

3.4.1. T-test

The purpose of any statistical test is to identify the probability of a value in a sample, given that the null hypothesis is correct. Therefore, a t-test is generally used in case of small samples and when the test statistic of the population follows a normal distribution. There are two types of t-tests, which are one sample and two samples. Here I am going to employ the two samples t-test, which can compare the mean values of both samples (i.e., ASCC in Fall and Winter). The focus was on the number of associated connections in both Fall and Winter, in weeks 7, 8, 13, 14, 15 and from 8 AM to 5 PM which is considered as a day-shift window in the dataset. Regarding the null hypothesis, it has been assumed that the means are going to be the same. Besides, the alpha has been set to 0.05, and if the p-value is less than alpha, we can state that there is a statistically significant difference between the means of the samples, or to be more specific, we can reject the null hypothesis. Also, I then compared the t-critical value with t-value. This means that if the t-value is larger than the t-critical value, again we can reject the null hypothesis. The two-tail value for each AP has been considered since I could not tell if the mean for Winter is higher than Fall or vice versa. The results discussed further in the results and discussion part.

3.4.2. Clustering

K-means clustering is one of the well-known unsupervised machine learning algorithms. Generally, unsupervised algorithms make inferences from datasets using merely input vectors without referring to known or labelled outcomes. The primary purpose of K-means is to group similar data points together and detect the underlying patterns. That being said, K-means need a fixed number (k) of clusters in the dataset. I decided to apply the K-means clustering using Rapidminer for the data preparation and having the visual workflow while modeling.

Here, it is necessary to define a target number *k,* which refers to the number of centroids I need in the dataset. Although identifying this parameter is one of the challenges of this method, but in this case, since the focus is on an educational building, it is assumed that there might be two types of behavior. The first one is the people who are working at the university, and the second ones are students who have class during the day on that specific floor.

In essence, a centroid is an unreal location representing the center of the cluster. To process the learning data, the K-means algorithm in the data starts with first randomly selected centroids. At the beginning of each cluster, this iterative process occurs, and then the calculations for optimizing the position of the centroids take place after. For this purpose, the data set prepared with a different structure. This means that for each AP, there are 10 data points, and each data point represents the associated client counts on each week in a Fall or Winter semester from 8 AM to 5 PM. Therefore, with 12 AP, there are 120 data points. On the other hand, the attributes are the day names starting from Monday to Friday which is shown with the hours of each record, i.e., Monday-8 means Monday at 8 AM, and Monday-17 means Monday at 5 PM. The following features have been utilized in stage of modelling.

- APID (dtype: Polynominal)
- Week_number (dtype: Polynominal)
- Semester (dtype: Polynominal)
- Days-hours (from Monday to Friday, 8 AM to 5 PM, dtype: integer, presents ASCC)

There are 24 missing values in the dataset, which belong to the Wednesdays at 9 AM in week 15th both in Fall and Winter. Missing values have been replaced with the mean value of the same day at 9 AM for the week 14th since these 2 weeks was the examination period.

It is necessary to turn categorical attributes (i.e. week numbers, APID, semester) into numeric values since the K-means algorithm does not accept the nominal values as an input. In this regard, I employed an operator called “nominal to numerical”, which can change the type of selected non-numeric attributes to a numeric type. There is also a parameter inside of this operator called “coding type”. This parameter indicates the coding which is used for transforming nominal attributes to numerical attributes. Here I am going to use dummy coding. This means that for all values of the nominal attribute, a new attribute is created. In every example, the new attribute which corresponds to the actual nominal value of that example gets value 1, and all other new attributes get value 0.

The next step is normalizing the values in the dataset. There are several methods exist for normalizing data. Firstly, z-transformation. This is also called statistical normalization. This normalization subtracts the mean of the data from all values and then divides them by the standard deviation. That being said, the distribution of the data has a mean of zero and variance of one. This is a useful normalization technique, which preserves the original distribution of the data and is less influenced by outliers. Nonetheless, there is another method called "range transformation", i.e., mapping all values to a range between min and max. This method can be influenced by outliers. The reason for this is bounds move towards these values. However, this method keeps the original distribution of the data points. To put it simply, I employed both transformation methods, and since the range transformation proved higher accuracy, I decided to stay with this method.

The next phase is to employ the multiply operator, to merely multiply the results of the data preparation to use them as an input for different *k* for the clusters. Afterward, several K-means clustering operator have been employed in the design canvas to examine different values for *k.* As for the measure types, the Bregman Divergences has been utilized, and also the divergence has been set to Squared Euclidean Distance. The question here is, what is a good value of *k* that has a better performance? I then utilized the Performance operator to check the Davies Bouldin Index (DBI) and the Average within centroid distance to measure the goodness of the clusters.

3.5. Regression

The interest lies under finding out which characteristics have the most substantial impact on the number of connections on the floor. The study aims to suggest ways of predicting this number, which approximately indicates the number occupants and optimizes the HVAC system to operate based on the space and occupants' demand. The cluster analysis performed to cluster AP readings according to essential characteristics such as week number, semester, and the connections on each hour. I could then do a regression analysis separately for each of the 2 or 3 clusters identified to determine which of the remaining characteristics are most influential for each cluster. The reasoning behind this is that certain features definitely impact the number of connections.

For this section, second data set has been, and the attributes are listed below:

- *Cluster*
- *Day_name*
- *Hour*
- *ASCC*
- *Classroom*
- *Office*
- *Corridor*
- *Mechanical/washroom*
- *Semester*
- *Week number*
- *Area*

In addition, I want to also include the categorical attribute such as day name and week number. Additional steps are needed to ensure that the outcomes are interpretable, and these steps include recoding the categorical variable into several separate variables (i.e., dummy coding).

Now, it is required to split the data for both cluster_0 and cluster_1 to the training set and test set. A 10% of each clusters' dataset has been set aside for the test set. This means that there are 3138 examples for training set in cluster_0 and 348 examples for the test set, and 2241 examples for the training set and 249 examples for test set in cluster_1. Moreover, split validation employed, which randomly splits up the example set and test set and evaluates the model. In essence, this technique merely measures the performance of a learning operator, and it is mainly used to determine how accurately a model will perform in practice.

Afterward, the linear regression operator is used inside of the training section of the split validation nested operator, and apply model and performance operator in the testing part. Besides, I also left the 'eliminate collinear features' check-mark on since it automates the process of finding c-dependent attributes and setting them aside from the regression procedure. Also, the linear model does not fundamentally need to pass through the origin, therefore, I left the "use bias" parameter on as well. Feature selection method also needs to be put into consideration; thus, different methods tested here, which will be further discussed.

Although performance operator provides us a quite large number of metrics for linear regression, I am going to focus on the Root Mean Squared Error (RMSE) and Squared Correlation which is equivalent to R^2 in Rapidminer as for the performance metrics in this step. Last but foremost, I am going to deploy the trained model against the 'unseen' data. The results and details in this step will be discussed further in the next section.

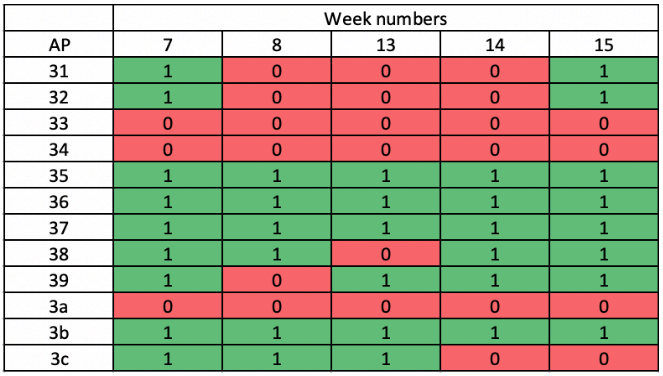
4. Results and Discussion

4.1. Similarity tests

4.1.1. T-test

As mentioned earlier, I performed a t-test to observe the statistical difference between the values of an AP in each semester. A hypothesis was: “There is no difference in the number of connected devices in Fall and Winter.” As shown in the following chart, the APs which had a considerable difference in their behavior in the Fall and Winter colored with red. This means the data support the hypothesis that the populations' means are different. However, greens demonstrated the ones that the alternative hypothesis has accepted, which means there is no significant difference between their values. This should be noted that, here, this test assumes that the data is normally distributed in both groups (ASCC in Fall and Winter), which is one of the limitations of employing this test in this project.

Table 2



	Week numbers				
AP	7	8	13	14	15
31	1	0	0	0	1
32	1	0	0	0	1
33	0	0	0	0	0
34	0	0	0	0	0
35	1	1	1	1	1
36	1	1	1	1	1
37	1	1	1	1	1
38	1	1	0	1	1
39	1	0	1	1	1
3a	0	0	0	0	0
3b	1	1	1	1	1
3c	1	1	1	0	0

4.1.2. Clustering

As mentioned earlier, the different values for *k* have been tested in order to come up with the most desirable number of clusters. In the first place, it is assumed that two behaviors exist for the APs. This merely implies that there is a a group of low and high traffic AP. However, before going deep into the details, let’s bring the light to the spatial characteristics of APs such as location, area coverage, and the type of the room which they are covering.

Spatial characteristics:

In 3.2.6, the values which are with each AP as their spatial features associated presented the values which are associated with each AP as their spatial features. Nevertheless, the location of each AP also needs to be put into consideration.

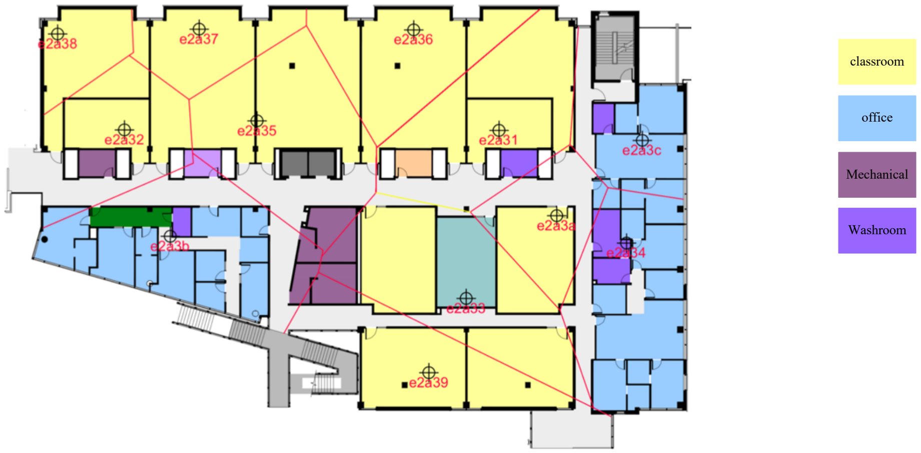


Figure 7 – Map details and info

As demonstrated above, the twelve APs' are arranged on the floor so that they can cover all the areas. Some APs are located in the office spaces, while some are placed in classrooms to provide the Wi-Fi to all the occupants. In 3.2.6, APs' coverages are mentioned. In essence, APs e2a3c, e2a34 and e2a3b are providing service to the office areas, and APs e2a39, e2a31, e2a3a, e2a38, e2a37, e2a36, e2a35, e2a32, e2a33 are providing the internet to classrooms. This is just an overview of the APs' location and coverage before the clustering analysis starts.

K = 2 (k: number of clusters that K-means looks for)

As noted earlier, the main parameter to be decided on is *k*. *K* = 2 might be sound the most basic case, but this varies in each study. Here, it is assumed that there are two different behavior exists on this floor. One for offices (low traffic) areas, and one for classrooms (high traffic). I ran the model and here is the result:

```
Cluster 0: 91 items
Cluster 1: 29 items
Total number of items: 120
```

Two clusters generated, and each cluster has its own centroid table, which provides the information of centroids of each cluster. After investigating each centroid's values, it is concluded that cluster_0 is a representative for the APs located in the office areas, and cluster_1 is a representative for APs located in classrooms or high traffics. The clusters do not present that each AP belongs entirely to either cluster_0 or cluster_1. But if I study the centroid table and the clusters' plot, I can determine that the low traffic APs are located in the right, bottom, and top left area of the map. As indicated below, the APs in the green line area belong to cluster_0 and other APs belong to cluster_1.

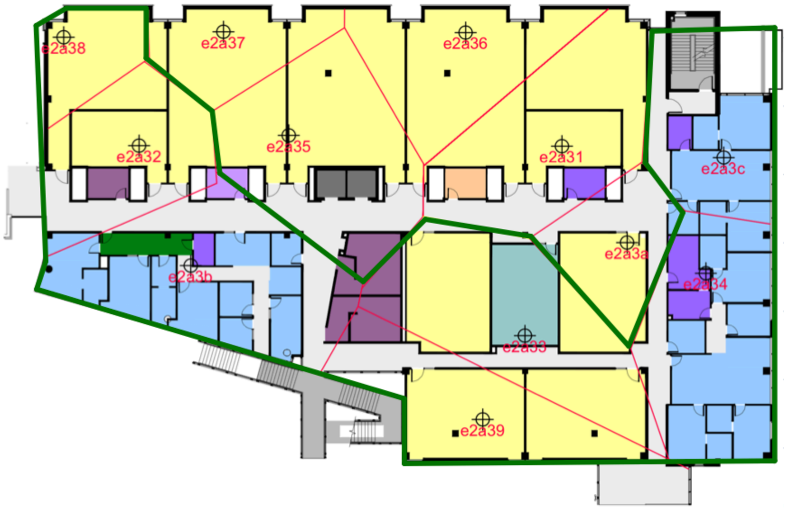


Figure 8 – Green line indicates the area covered by cluster_0

Moreover, the clusters' plot specifies that the number of connections for the e2a37, e2a36, e2a35, e2a31 is higher than connections for the rest of the APs. There is no significant difference between these clusters when it comes to the semester attribute. Indeed, the APs which belong to cluster_0 are located in either the office areas or at the corners of the map. For instance, e2a38 has low traffic since it is located at the edge of the classroom, or e2a32 has low traffic since it is also covering the mechanical/washroom spaces, and the connections might have been distributed between these two APs. The cluster distance performance operator took this centroid cluster model and clustered set as input and evaluates the performance of the model based on the cluster centroids. Here, the two most common performance measures are supported: Average within-cluster distance and Davies-Bouldin index. The clustering algorithm that processes a collection of clusters with the smallest Davies-Bouldin index is considered the best algorithm based on this measure. The performance results are written below.

```
Performance Vector (k=2, range transformation: min = 0, max = 1):
Avg. within centroid distance: 3.257
Avg. within centroid distance_cluster_0: 2.777
Avg. within centroid distance_cluster_1: 4.766
```

Davies Bouldin: 1.705

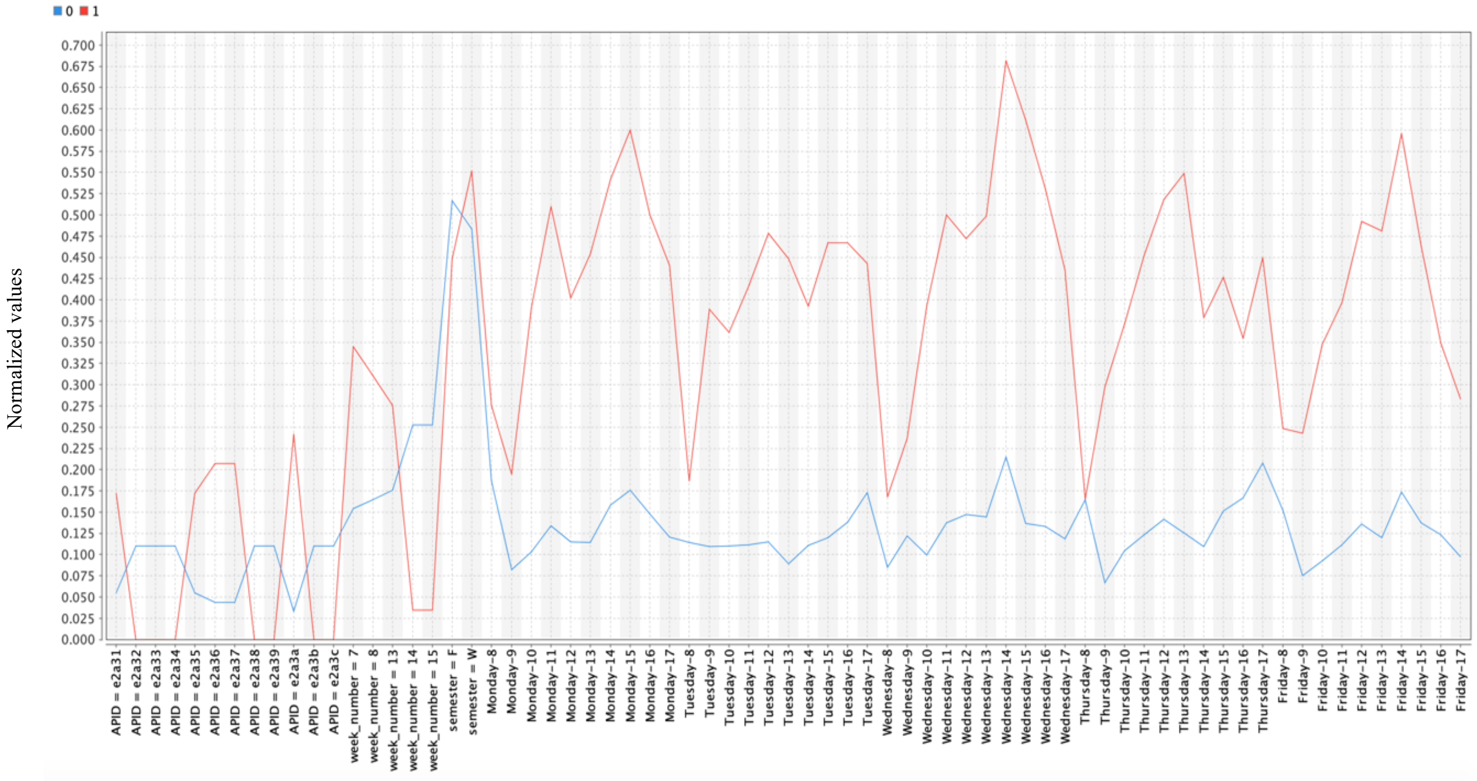


Figure 9 – Clusters' plot, k=2

K = 3, range transformation as for normalization method

When I tried *k=3*, I achieved the third cluster, which presents the APs located in the office and low traffic spaces for, i.e., the same as cluster_0 in *k=2.* Cluster_1 and cluster_2 differ in the semester attribute. This means that one shows the APs in Fall, and the other one shows the same APs for Winter, and given the different values for the centroids of these two clusters, they are approximately identical clusters. That being said, here are the performance results and the plot for the second step.

Cluster 0: 29 items
Cluster 1: 44 items
Cluster 2: 47 items
Total number of items: 120

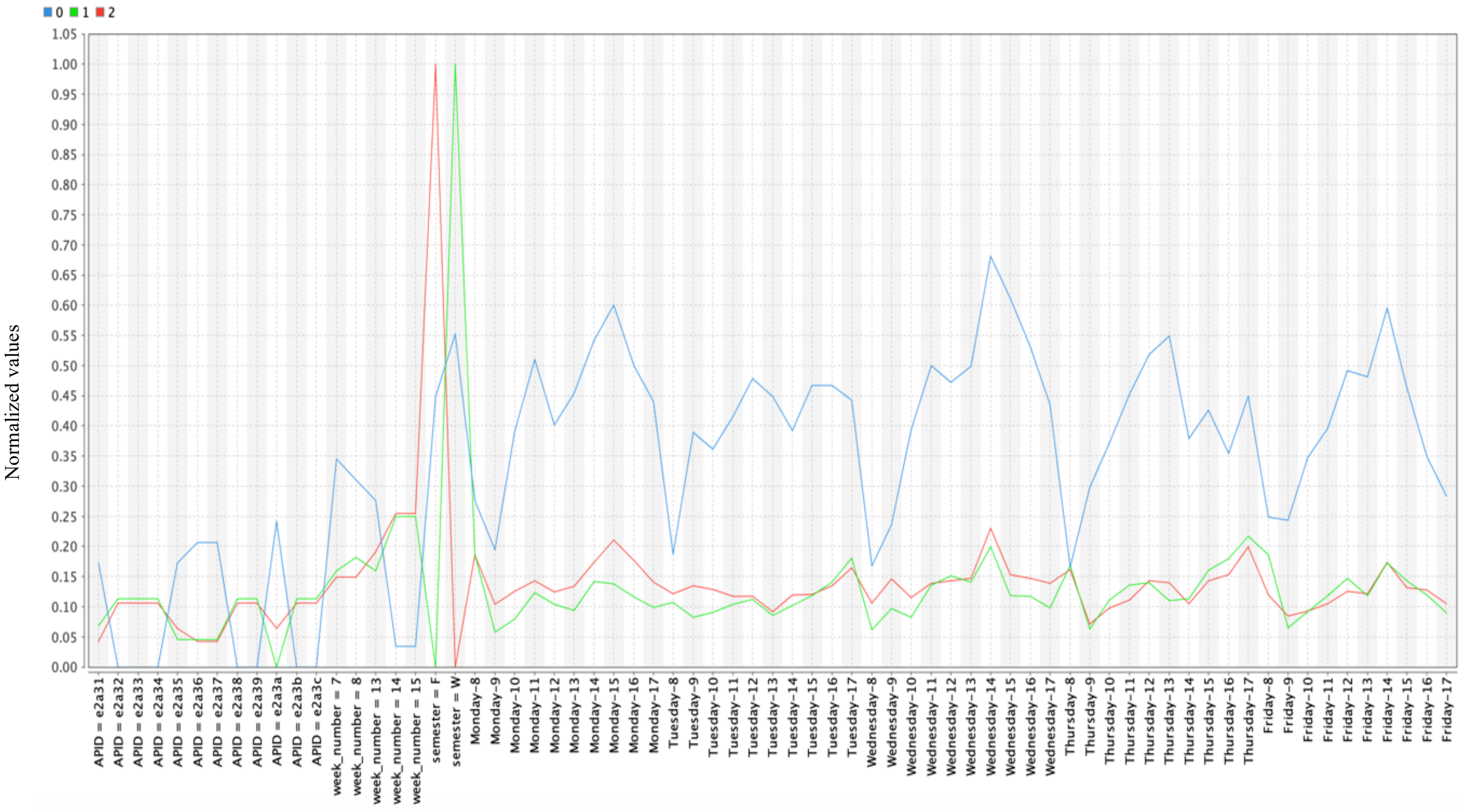


Figure 10 – Clusters' plot, k=3

Performance Vector (k=3, range transformation: min = 0, max = 1):
Avg. within centroid distance: 2.869
Avg. within centroid distance_cluster_0: 4.766
Avg. within centroid distance_cluster_1: 2.240
Avg. within centroid distance_cluster_2: 2.287
Davies Bouldin: 1.919

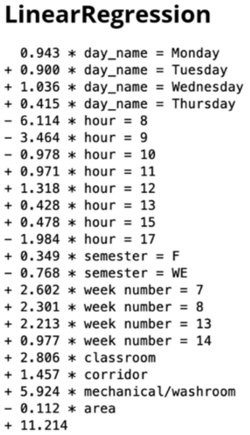
Compare to the previous section, a higher Davies-Bouldin index is achieved, and lower Avg. within centroid distance. For cluster_1 in *k=2,* which is equivalent to cluster_0 when *k=3*, the distance value is the same (4.766). On the other hand, a lower Avg. within centroid distance obtained for the second group, but still, *k=2* is the preferred value for the clustering analysis. However, let's not forget that it is not possible to blindly follow these numbers and claim reaching the best clusters. Issues such as overfitting are always a challenge for tasks such as clustering.

4.2. Regression

The results for regression can be examined in multiple ways. The results for cluster_0 (low traffic – office spaces) and cluster_1 (high traffic – classrooms) are provided in a table which has all the attributes, coefficient, Std. Error, Std. Coefficient, Tolerance, t-Stat, p-Value and Code. I begin with sorting this table based on the "Code" attribute, which ranks the attributes as four stars (****), three stars (***), etc. After trying different methods for the feature selection technique, I decided to employ "greedy" method for this parameter due to the higher performance in R^2 and RMSE. Here are the results for the cluster_0:

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
day_name = Monday	0.943	0.248	0.078	0.997	3.808	1.43E-04	****
day_name = Tuesday	0.900	0.250	0.074	0.999	3.603	3.21E-04	****
day_name = Wednesday	1.036	0.250	0.085	1.000	4.149	3.45E-05	****
hour = 8	-6.114	0.309	-0.371	0.976	-19.782	0	****
hour = 9	-3.464	0.316	-0.205	1.000	-10.969	0	****
hour = 12	1.318	0.304	0.082	0.979	4.337	1.50E-05	****
hour = 17	-1.984	0.313	-0.118	0.998	-6.333	2.85E-10	****
week number = 7	2.602	0.300	0.214	0.998	8.666	0	****
week number = 8	2.301	0.301	0.189	1.000	7.656	2.74E-14	****
week number = 13	2.213	0.300	0.183	1.000	7.385	2.08E-13	****
week number = 14	0.977	0.251	0.082	0.938	3.892	1.02E-04	****
classroom	2.806	0.174	0.796	0.992	16.159	0	****
corridor	1.457	0.252	0.219	0.938	5.780	8.45E-09	****
mechanical/washroom	5.924	0.430	1.439	0.923	13.777	0	****
area	-0.112	0.008	-1.748	0.983	-13.947	0	****
(Intercept)	11.214	0.554	NaN	NaN	20.232	0	****
hour = 10	-0.978	0.306	-0.060	0.997	-3.192	0.00143077	***
hour = 11	0.971	0.299	0.061	0.983	3.242	0.00120174	***
semester = WE	-0.768	0.251	-0.063	0.945	-3.061	0.002232803	***
day_name = Thursday	0.415	0.250	0.034	0.993	1.659	0.0972008	*
semester = F	0.349	0.205	0.033	0.957	1.707	0.087939767	*
hour = 13	0.428	0.301	0.027	0.986	1.420	0.155772726	
hour = 15	0.478	0.303	0.030	0.986	1.576	0.115113073	

```
Performance Vector:
root_mean_squared_error: 4.045 +/- 0.000
squared_correlation: 0.350
```



```
LinearRegression

  0.943 * day_name = Monday
+ 0.900 * day_name = Tuesday
+ 1.036 * day_name = Wednesday
+ 0.415 * day_name = Thursday
- 6.114 * hour = 8
- 3.464 * hour = 9
- 0.978 * hour = 10
+ 0.971 * hour = 11
+ 1.318 * hour = 12
+ 0.428 * hour = 13
+ 0.478 * hour = 15
- 1.984 * hour = 17
+ 0.349 * semester = F
- 0.768 * semester = WE
+ 2.602 * week number = 7
+ 2.301 * week number = 8
+ 2.213 * week number = 13
+ 0.977 * week number = 14
+ 2.806 * classroom
+ 1.457 * corridor
+ 5.924 * mechanical/washroom
- 0.112 * area
+ 11.214
```

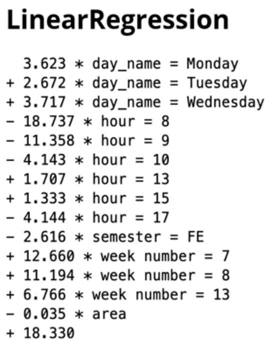
Figure 11 - Standard format of linear regression model,

The standard format of the linear regression model as demonstrated in (Fig. 11). The RMSE value is a great metric for comparison, but this may not add much value. Thus, I consider R^2, which can provide some sense for the goodness of fit. Generally, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased. However, unfortunately 35% for R^2 is not impressive. It is necessary to check the residual plots, which can reveal unwanted residual patterns that indicate biased results more effectively than numbers. But for now, let's also investigate the results for the cluster_1.

Table 4 - Significance levels and metrics for cluster_1

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
day_name = Monday	3.623	0.675	0.106	0.998	5.368	9.06378E-08	****
day_name = Tuesday	2.672	0.676	0.078	0.994	3.951	8.10345E-05	****
day_name = Wednesday	3.717	0.688	0.106	0.999	5.401	7.53394E-08	****
hour = 8	-18.737	0.868	-0.411	1.000	-21.584	0	****
hour = 9	-11.358	0.904	-0.238	0.997	-12.564	0	****
hour = 10	-4.143	0.882	-0.089	0.990	-4.697	2.85267E-06	****
hour = 17	-4.144	0.868	-0.091	0.992	-4.771	1.98685E-06	****
semester = FE	-2.616	0.791	-0.076	0.857	-3.308	0.00095778	****
week number = 7	12.660	0.783	0.374	0.997	16.160	0	****
week number = 8	11.194	0.797	0.322	1.000	14.037	0	****
week number = 13	6.766	0.795	0.195	0.986	8.508	0	****
area	-0.035	0.004	-0.177	1.000	-9.859	0	****
(Intercept)	18.330	0.972	NaN	NaN	18.858	0	****
hour = 13	1.707	0.887	0.037	0.985	1.924	0.054562637	*
hour = 15	1.333	0.868	0.029	0.977	1.535	0.124844351	

Performance Vector:
root_mean_squared_error: 10.417 +/- 0.000
squared_correlation: 0.447



LinearRegression

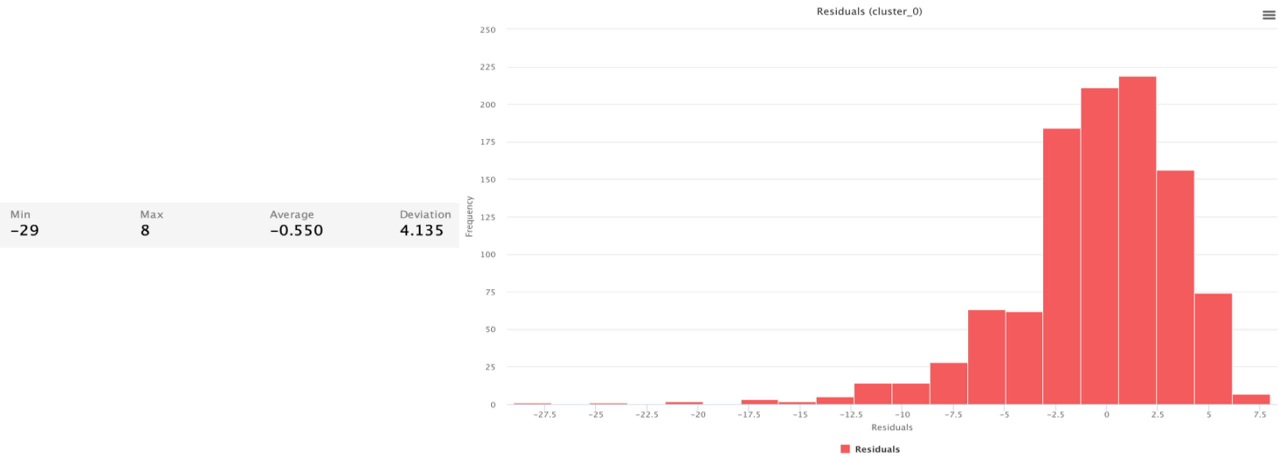
```
  3.623 * day_name = Monday
+ 2.672 * day_name = Tuesday
+ 3.717 * day_name = Wednesday
- 18.737 * hour = 8
- 11.358 * hour = 9
- 4.143 * hour = 10
+ 1.707 * hour = 13
+ 1.333 * hour = 15
- 4.144 * hour = 17
- 2.616 * semester = FE
+ 12.660 * week number = 7
+ 11.194 * week number = 8
+ 6.766 * week number = 13
- 0.035 * area
+ 18.330
```

Figure 12 - Standard format of linear regression

Although 44.7% achieved for R^2 in cluster_1, but still it is allowed to determine whether the coefficient estimates and predictions are biased. In addition, this performance metric does not indicate if a regression model provides an adequate fit to the data.

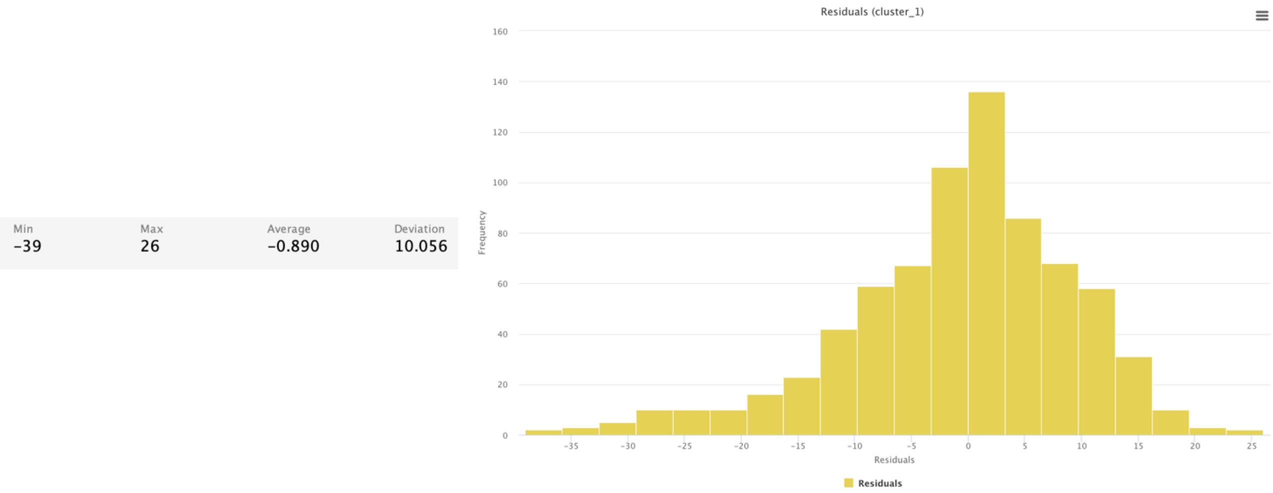
The trained model should be deployed against the ‘unseen’ data which previously has been set aside. Therefore, I am going to apply the trained model of each clusters to their unobserved data, and then calculating the difference between prediction (ASCC) and observed (ASCC), which is called “residual”.

Using residual plots, we can assess whether the observed error is consistent with the stochastic error. The residuals should not be either systematically high or low. Hence, the residuals should be centered on zero throughout the range of fitted values. In other words, the model is correct on average for all fitted values. Not only can we study the histogram of the residuals attribute, but also average and standard deviation can be retrieved and considered as indicators of accuracy and stability.



Min	Max	Average	Deviation
−29	8	−0.550	4.135

Figure 13 – Histogram of residuals, cluster_0



Min	Max	Average	Deviation
-39	26	-0.890	10.056

Figure 14 - Histogram of residuals, cluster_1

As mentioned earlier, it is required to ensure all error terms in the model are normally distributed. The mean needs to be close to 0, and a low standard deviation is expected. Nevertheless, unfortunately, the values I achieved for the deviation and the mean is not entirely satisfactory. In cluster_0, it seems that some outliers exist (for which residual ≈ -27, -25 and -17.5), and also the same thing happens in cluster_1 (for which residual ≈ -35). However, setting these aside, the distribution looks more or less close to normal. The high value for SD and unremarkable R^2 can, however, be right motivations to continue advancing the model. (See appendix B for the regression modeling process)

5. Conclusion

Occupancy counting and detecting are becoming more common in today's buildings as an indicator of the emerging concept of smart buildings. Also, this concept has captured growing attention in terms of improving building energy efficiency. This, along with the data collected over Wi-Fi and other opportunistic context sources, contributes to rapidly increasing amounts of collected data. Big Data analytics can provide insight into such data, along with providing efficient methods of storage and retrieval. In this research, the Wi-Fi data, which was collected by the APs located in an educational building has been investigated. The objectives have been specified in the first place; however, the accuracies suffer due to several limitations. In 4.1 and 4.2, I referred to the attributes which can impact the number of connections. Moreover, by numbering the weeks and investigating the data, I specified how access points react to the events held during the academic year. By employing unsupervised learning algorithm, I then recognized the APs which are acting similarly. However, due to the constraints and limitations mentioned below, it was not possible to predict the number of occupancies desirably and accurately identify the number of weeks that it takes for AP to converge to the mean value.

There are practical challenges that exist when applying the Wi-Fi-based occupancy detection in real-life. In this case, the precise data regarding the spatial characteristics such as the exact APs' location, APs' coverage, and unclarity regarding some types of the rooms, inaccessibility to MAC addresses due to privacy issues, lack of information regarding the precise academic calendar, lack of data on the entire academic year, inadequate features in the original dataset, unapproachable building due to the great distance between Montréal and Winnipeg, determining and excluding the noises on the dataset were the major limitations. The discontinuous Wi-Fi communication of smartphones suggests a practical challenge in occupancy detection, and non-human Wi-Fi communications also bring in the irrelevant or the unwanted amount of connections in the data. This should be noted that Wi-Fi only my not handle every complicated scenario.

On the other hand, the spatial characteristics can be modified in the future research. This means that the memberships or shares of each AP from any spaces can be generated as a percentage value. For instance, an AP can have a 30% of a classroom while another AP can have a 70% coverage of the same classroom. This would be an improvement when it comes to considering the spatial characteristics.

Moving average can be employed to analyze data points by creating a series of averages of different subsets of the full dataset. This can also be applied on this dataset to smooth out short-term fluctuations and highlight longer-term trends or cycles, and it needs to be used with the greatest value for the moving window.

In this project, the focus and analysis applied on the morning shift, which is based on the working hours (8 AM to 5 PM). Due to the scope of this project, this was one of the main limitations during this research since the analysis could be applied to different time intervals. Furthermore, although the t-test results have been discussed, but I believe that there is a place for improvement regarding the APs that rejected the null hypothesis. This means that those can be excluded, and treated differently.

Generally, I believe that not only the development of approaches but also collecting a higher resolution of Wi-Fi data can significantly improve the information available from the environment. The scalability of this research is depending on how one can overcome the problems and limitations mentioned earlier. This means that by increasing the accuracy and your insight regarding one floor, you can extend your focus to the entire educational building. Besides, by using multiple sensing data sources, accuracy and reliability can remarkably get increased due to the literature. Also, the power of deep learning has started to emerge recently. That being said, the current applications of deep neural networks in the field of building occupancy detection is scarce; thus, there lies a promising future in this field.

6. References

[1] U. S. E. I. Administration, “Monthly Energy Review,” Washington,DC, 2019.

[2] L. D&R International, *Building Energy Data Book*. Washington, DC: U.S. Department of Energy, 2011.

[3] National Round Table on the Environment and the Economy and Sustainable Development Technology Canada “*Geared for Change: Energy Efficiency in Canada's Commercial Building Sector,”* Ottawa, Ontario, 2009.

[4] E. Azar and C. C. Menassa, “Evaluating the impact of extreme energy use behavior on occupancy interventions in commercial buildings,” *Energy Build.*, vol. 97, pp. 205–218, Jun. 2015.

[5] T. V. Vytautas Martinaitis, Edmundas Kazimieras Zavadskas, Violeta Motuzienė, “Importance of occupancy information when simulating energy demand of energy efficient house: A case study,” *Energy Build.*, pp. 64–75, 2015.

[6] S. S. Wilhelm Kleiminger, Friedemann Mattern, “Predicting household occupancy for smart heating control: A comparative performance analysis of state-of-the-art approaches,” *Energy Build.*, pp. 493–505, 2014.

[7] C. Zhang, M. Kuhn, A. E. Fathy, and M. Mahfouz, “Real-time noncoherent UWB positioning radar with millimeter range accuracy in a 3D indoor environment,” *IEEE MTT-S Int. Microw. Symp. Dig.*, no. 10.1109/MWSYM.2009.5165971, pp. 1413–1416, 2009.

[8] T. Cheng, M. Venugopal, J. Teizer, and P. A. Vela, “Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments,” *Autom. Constr.*, vol. 20, no. 8, pp. 1173–1184, Dec. 2011.

[9] A. W. Jochen Teizer, Manu Venugopal and A. Walia, “Ultrawideband for Automated Real-Time Three-Dimensional Location Sensing for Workforce, Equipment, and Material Positioning and Tracking,” *Transp. Res. Rec.*, pp. 56–64, 2008.

[10] N. Masoudifar, A. Hammad, and M. Rezaee, “Monitoring occupancy and office equipment energy consumption using real-time location system and wireless energy meters,” in *Proceedings - Winter Simulation Conference*, 2015, vol. 2015-Janua, pp. 1108–1119.

[11] T. A. Nguyen and M. Aiello, “Energy intelligent buildings based on user activity: A survey,” *Energy Build.*, vol. 56, pp. 244–257, Jan. 2013.

[12] L. M. Candanedo and V. Feldheim, “Accurate occupancy detection of an office room from light, temperature, humidity and CO2 measurements using statistical learning models,” *Energy Build.*, vol. 112, pp. 28–39, Jan. 2016.

[13] T. Leephakpreeda, “Adaptive Occupancy-based Lighting Control via Grey Prediction,” *Build. Environ.*, vol. 40, no. 7, pp. 881–886, Jul. 2005.

[14] Z. Chen, C. Jiang, and L. Xie, “Building occupancy estimation and detection: A review,” *Energy Build.*, vol. 169, pp. 260–270, Jun. 2018.

[15] R. H. Dodier, G. P. Henze, D. K. Tiller, and X. Guo, “Building occupancy detection through sensor belief networks,” *Energy Build.*, vol. 38, no. 9, pp. 1033–1043, Sep. 2006.

[16] P. Liu, S. K. Nguang, and A. Partridge, “Occupancy Inference Using Pyroelectric Infrared Sensors Through Hidden Markov Models,” *IEEE Sens. J.*, vol. 16, no. 4, pp. 1062–1068, 2016.

[17] F. Castanedo, H. Aghajan, and R. Kleihorst, “Modeling and discovering occupancy patterns in sensor networks using latent dirichlet allocation,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6686 LNCS, no. PART 1, pp. 481–490, 2011.

[18] F. Castanedo, D. Lopez-de-Ipina, H. Aghajan, and R. Kleihorst, “Building an occupancy model from sensor networks in office environments,” pp. 1–6, 2011.

[19] F. Wahl, M. Milenkovic, and O. Amft, “A distributed PIR-based approach for estimating people count in office environments,” *Proc. - 15th IEEE Int. Conf. Comput. Sci. Eng. CSE*

2012 10th IEEE/IFIP Int. Conf. Embed. Ubiquitous Comput. EUC 2012, pp. 640–647, 2012.

[20] Y. P. Raykov, E. Ozer, G. Dasika, A. Boukouvalas, and M. A. Little, “Predicting room occupancy with a single Passive infrared (PIR) sensor through behavior extraction,” *UbiComp 2016 - Proc. 2016 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.*, pp. 1016–1027, 2016.

[21] W. J. Fisk and A. T. De Almeida, “Sensor-based demand-controlled ventilation: A review,” *Energy Build.*, vol. 29, no. 1, pp. 35–45, 1998.

[22] S. Wang and X. Jin, “CO2-based occupancy detection for on-line outdoor air flow control,” *Indoor Built Environ.*, vol. 7, no. 3, pp. 165–181, 1998.

[23] M. S. Zuraimi, A. Pantazaras, K. A. Chaturvedi, J. J. Yang, K. W. Tham, and S. E. Lee, “Predicting occupancy counts using physical and statistical Co2-based modeling methodologies,” *Build. Environ.*, vol. 123, pp. 517–528, 2017.

[24] C. Jiang, M. K. Masood, Y. C. Soh, and H. Li, “Indoor occupancy estimation from carbon dioxide concentration,” *Energy Build.*, vol. 131, pp. 132–141, Nov. 2016.

[25] Ömer Faruk Ertuğrul, Yılmaz Kaya, and Mehmet EminTağluk, “Detecting Occupancy of an Office Room by Recurrent Extreme Learning Machines,” *Int. Artif. Intell. Data Process. Symp.* , 2016.

[26] H. Liu and H. Motoda, “Feature extraction, construction and selection: A data mining perspective,” *Springer Sci. Bus. Media*, 2001.

[27] Q. Zhu, Z. Chen, M. K. Masood, and Y. C. Soh, “Occupancy estimation with environmental sensing via non-iterative LRF feature learning in time and frequency domains,” *Energy Build.*, vol. 141, pp. 125–133, Apr. 2017.

[28] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, p. 436, May 2015.

[29] Z. Liu, J. Zhang, and L. Geng, “An Intelligent Building Occupancy Detection System Based on Sparse Auto-Encoder,” pp. 17–22, 2017.

[30] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, “Multicamera people tracking with a probabilistic occupancy map,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, 2008.

[31] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger, “Towards a sensor for detecting human presence and characterizing activity,” *Energy Build.*, vol. 43, no. 2–3, pp. 305–314, 2011.

[32] V. L. Erickson, M. Á. Carreira-Perpiñán, and A. E. Cerpa, “PAWS: Passive Human Activity Recognition Based on Wi-Fi Ambient Signals,” in *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2011, pp. 258–269.

[33] J. Zou, Q. Zhao, W. Yang, and F. Wang, “Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation,” *Energy Build.*, vol. 152, pp. 385–398, Oct. 2017.

[34] S. Petersen, T. H. Pedersen, K. U. Nielsen, and M. D. Knudsen, “Establishing an image-based ground truth for validation of sensor data-based room occupancy detection,” *Energy Build.*, vol. 130, pp. 787–793, 2016.

[35] G. Conte, M. De Marchi, A. A. Nacci, V. Rana, and D. Sciuto, “BlueSentinel: A first approach using iBeacon for an energy efficient occupancy detection system,” *BuildSys 2014 - Proc. 1st ACM Conf. Embed. Syst. Energy-Efficient Build.*, pp. 11–19, 2014.

[36] A. Corna, L. Fontana, A. A. Nacci, and D. Sciuto, “Occupancy detection via iBeacon on Android devices for smart building management,” *Proc. -Design, Autom. Test Eur. DATE*, vol. 2015-April, pp. 629–632, 2015.

[37] A. Filippoupolitis, W. Oliff, and G. Loukas, “Occupancy detection for building emergency management using BLE beacons,” *Commun. Comput. Inf. Sci.*, vol. 659, pp. 233–240, 2016.

[38] A. Filippoupolitis, W. Oliff, and G. Loukas, “Bluetooth Low Energy Based Occupancy Detection for Emergency Management,” *Proc. - 2016 15th Int. Conf. Ubiquitous Comput. Commun. 2016 8th Int. Symp. Cybersp. Secur. IUCC-CSS 2016*, pp. 31–38, 2017.

[39] Y. Gu, F. Ren, and J. Li, “PAWS: Passive Human Activity Recognition Based on Wi-Fi Ambient Signals,” *IEEE Internet Things J.*, vol. 3, no. 5, pp. 796–805, 2016.

[40] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, “Tool release: Gathering 802.11n traces with channel state information,” *Comput. Commun. Rev.*, vol. 41, no. 1, p. 53, 2011.

[41] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, and Y. Agarwal, “Sentinel: Occupancy based HVAC actuation using existing Wi-Fi infrastructure within commercial buildings,” *SenSys 2013 - Proc. 11th ACM Conf. Embed. Networked Sens. Syst.*, 2013.

[42] X. Lu, H. Wen, H. Zou, H. Jiang, L. Xie, and N. Trigoni, “Robust occupancy inference with commodity Wi-Fi,” *Int. Conf. Wirel. Mob. Comput. Netw. Commun.*, 2016.

[43] W. Wang, J. Chen, and X. Song, “Modeling and predicting occupancy profile in office space with a Wi-Fi probe-based Dynamic Markov Time-Window Inference approach,” *Build. Environ.*, vol. 124, pp. 130–142, Nov. 2017.

[44] W. Xi *et al.*, “Electronic frog eye: Counting crowd using Wi-Fi,” *Proc. - IEEE INFOCOM*, pp. 361–369, 2014.

[45] D. Chen, S. Barker, A. Subbaswamy, D. Irwin, and P. Shenoy, “Non-Intrusive Occupancy Monitoring using Smart Meters,” pp. 1–8, 2013.

[46] Z. Schwartz, M. Uysal, T. Webb, and M. Altin, “Hotel daily occupancy forecasting with competitive sets: a recursive algorithm,” *Int. J. Contemp. Hosp. Manag.*, vol. 28, no. 2, pp. 267–285, 2016.

[47] J. Page, D. Robinson, N. Morel, and J. L. Scartezzini, “A generalised stochastic model for the simulation of occupant presence,” *Energy Build.*, vol. 40, no. 2, pp. 83–98, 2008.

[48] H. Saha, A. R. Florita, G. P. Henze, and S. Sarkar, “Occupancy sensing in buildings: A

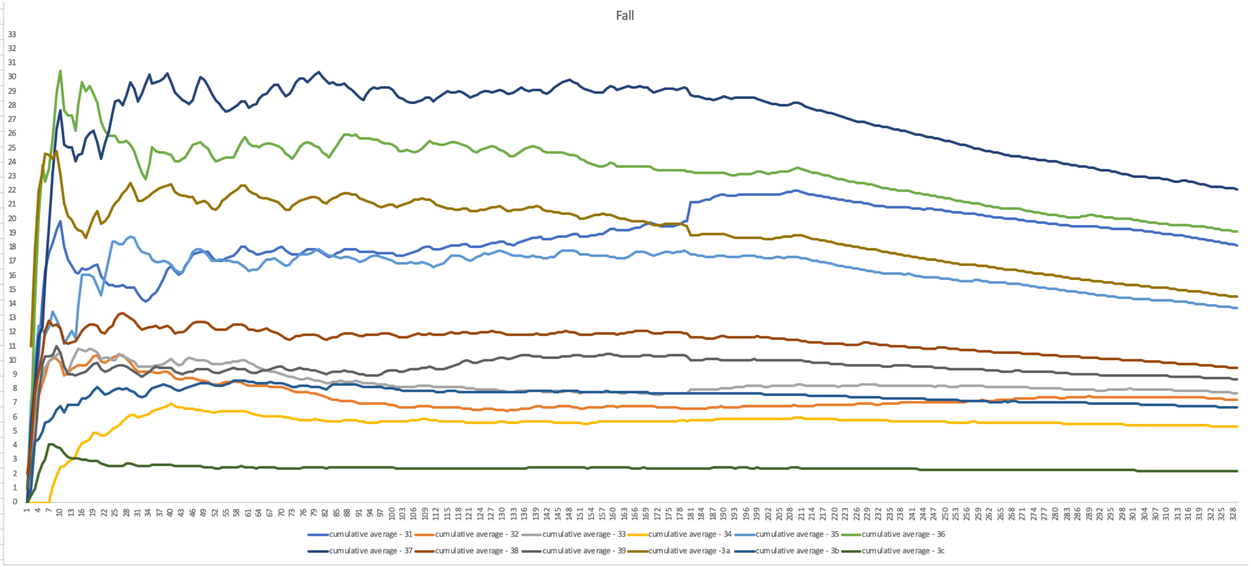
review of data analytics approaches," *Energy Build.*, vol. 188–189, pp. 278–285, Apr. 2019.

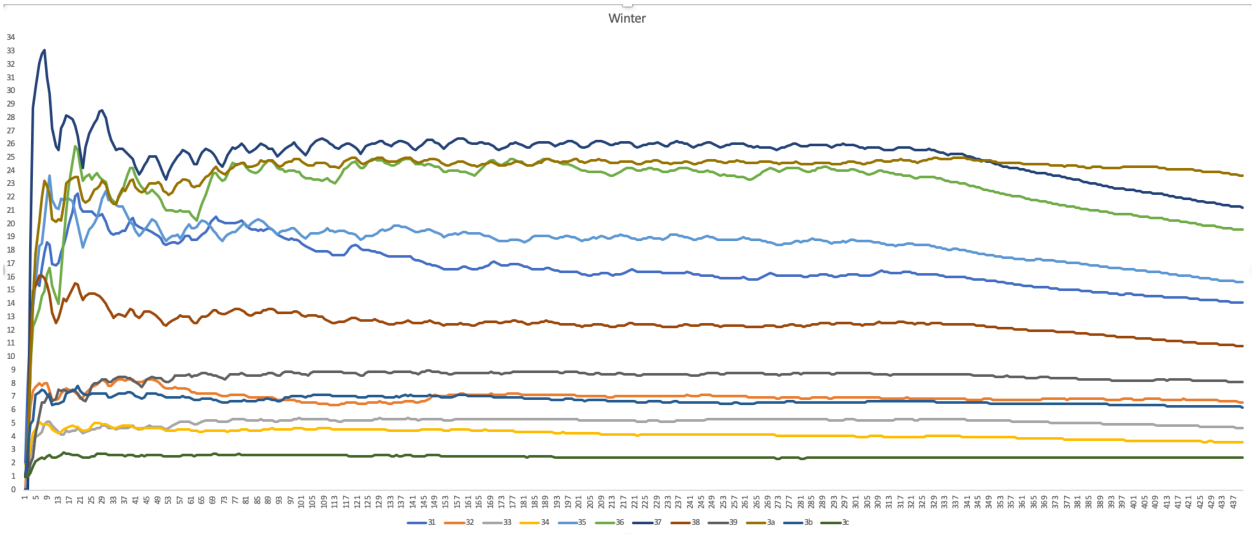
Appendices

Appendix (A)

Cumulative average plots in Fall and Winter for each AP

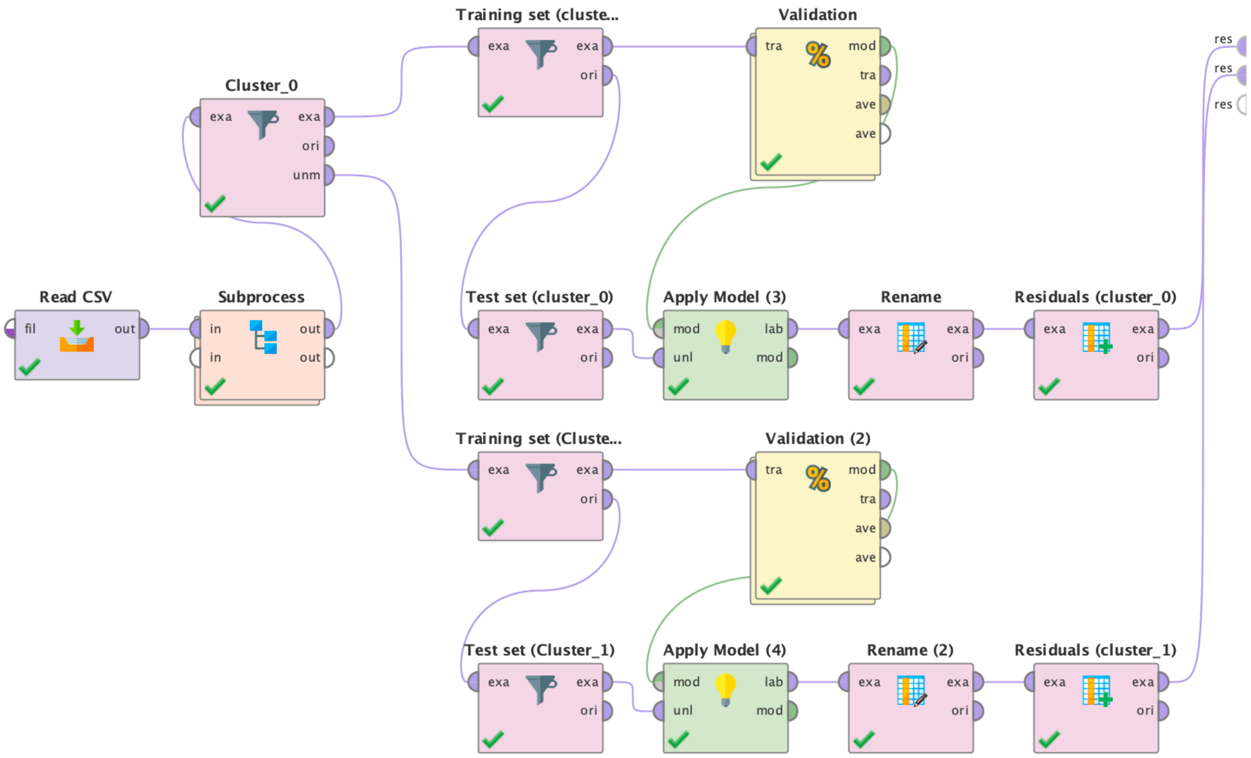
Cumulative average calculated on the associated connections demonstrated in the following chart. These charts provide an overview of the overall performance of APs during each semester. There several APs existed in the third floor which has the low traffic and some have high traffic. However, a general intuition can be extracted from these charts such as drops in the means when reaching the end of the semester.





Appendix (B)

Regression modeling process in Rapidminer


Training set (cluste...
exa
exa
ori
Validation
tra
mod
tra
ave
ave
res
res
res
Cluster_0
exa
exa
ori
unm
Read CSV
fil
out
Subprocess
in
in
out
out
Test set (cluster_0)
exa
exa
ori
Apply Model (3)
mod
unl
lab
mod
Rename
exa
exa
ori
Residuals (cluster_0)
exa
exa
ori
Training set (Cluste...
exa
exa
ori
Validation (2)
tra
mod
tra
ave
ave
Test set (Cluster_1)
exa
exa
ori
Apply Model (4)
mod
unl
lab
mod
Rename (2)
exa
exa
ori
Residuals (cluster_1)
exa
exa
ori