Estimating Room Occupancy in a Smart Campus using WiFi Soft Sensors

Iresha Pasquel Mohottige
School of Electrical Engineering and Telecommunications
UNSW, Sydney, Australia
i.pasquelmohottige@unsw.edu.au

Tim Moors
Oiklo Pty Ltd.
Australia
tim@oiklo.com

***Abstract*—Universities worldwide are experiencing a surge in enrolments, therefore campus estate managers are seeking continuous data on attendance patterns so as to optimize the usage of classroom space. While prior works have measured room occupancy via hardware sensor instrumentation, in this paper we explore the use of pervasive WiFi infrastructure for estimating attendance. In a dense campus environment, WiFi connectivity counts are poor estimators of room occupancy since they are polluted by adjoining rooms, outdoor walkways, and network load balancing. The main contribution of this work is to develop new ways to distinguish and filter out WiFi-connected users outside of the lecture room of interest, and feed such data to a regression analyser to estimate room occupancy. We evaluate our technique across lecture theatres of varying size in our campus, and show that their accuracy approaches that of hardware sensors without incurring cost and effort of installing and maintaining them.**

***Index Terms*—WiFi, people counting, rooms, hardware sensors, machine learning**

I. Introduction

In a modern day university, class attendance can vary widely depending on factors like time-of-day, lecturer engagement, and availability of online content. However, classrooms to which courses are allocated in a university is based on the enrolment while there is ample anecdotal evidence that enrolment may differ from the number of students attending a class. It is, therefore, useful for building management and control systems in a university campus to be able to determine how the classroom spaces are occupied and occupancy sensing plays an important role in allocating classrooms to courses based on attendance levels rather than enrolment.

There are predictions of the growth of global occupancy sensors [1], but special-purpose hardware sensors have a high upfront cost and require efforts in deployment and maintenance whereby limiting their adoption only to large commercial buildings. To this end, there is an emerging need for affordable, reliable, low-cost occupancy sensors. As the wireless infrastructure pervades modern campuses and usage of mobile devices is growing rapidly, WiFi can act as an explicit occupancy sensor in many university campuses. Although WiFi infrastructure has been used to determine occupancy for coarse spatial resolutions (e.g., floors of buildings) and in smaller office spaces with few occupants, it has not yet been successfully used to count large numbers (e.g., hundreds) of occupants in rooms such as lecture theatres in a university. WiFi signals crossing rooms is the major challenge in using WiFi data in estimating occupancy in a heavily populated modern campus. As WiFi signals cross rooms, devices connecting to APs in a room, but carried by WiFi users in nearby rooms or outside walkways corrupt the occupancy estimations. On the other hand, students connect to university network with multiple devices lead to overestimating occupancy while actual room occupants who connect to APs outside the room of interest and occupants who do not have any wireless device connected lead to underestimating the number of occupants estimated based on the WiFi connectivity data.

This paper[1] presents our study to estimate room occupancy using existing wireless infrastructure in a university campus, as might be useful to facilities managers at a university to allocate teaching spaces to courses based on attendance rather than enrolment, thereby leading to optimum utilization of limited available space. We assume that the exam conditions which occur in rooms are considered by schedulers when deciding which room to use for a class since they would be aware of that variation. Throughout the paper WiFi connected user from outside the room of interest is termed as a 'bystander' and the WiFi connected user inside the room is termed as an 'occupant'. We begin by collecting WiFi connectivity data for 790 different classes held across nearly 70 courses from 7 rooms of varying size on UNSW campus over a period of 12 weeks. Our first contribution is the proposal of a rich set of features from WiFi connectivity data that distinguishes occupants from bystanders. The second contribution is the development of classification and regression analysis based supervised learning approach to estimate room occupancy using the proposed features.

The paper begins by reviewing relevant related work in Section II and then describes in Section III, the collection of WiFi connectivity data and the effect of bystanders when estimating room occupancy with WiFi connectivity data. Section IV presents the extraction of features that distinguish the WiFi users as occupants and bystanders and then in Section V we present our approach based on classification and regression to estimate room occupancy. Section VI evaluates the performance of our approach and we present our conclusions

[1] Funding for this project was provided by the Australian Research Council (ARC) Linkage Grant LP150100666.

in Section VII.

II. Related Work

There are studies that use specialized occupancy detection hardware while there are other approaches that require new hardware or modification to existing hardware (e.g. new software installed). Also, there are approaches that uses existing building infrastructure (some rooms may already have CO2 sensors, cameras for security, or WiFi coverage). This section covers the categories aforementioned.

Many approaches to occupancy estimation have been based on data collected from explicitly installed hardware sensors such as infrared (IR), RFID, and camera sensors. In [2] researchers used machine learning techniques such as Support Vector Machine (SVM), Neural Networks (NN) and Hidden Markov Models (HMM) to process the data collected from a network of sensors consisting CO2 monitors and ambient sensors. HMM gave the most realistic results in predicting the number of occupants in offices with 73% accuracy, however it was only tested in small rooms with less than 10 occupants. In their approach to determine occupancy using single passive infrared sensor combined with machine learning techniques Raykov et. al. [3] proposed a low-cost occupancy estimation solution that produced a mean absolute error (MAE) of 1, but was tested only in rooms with 14 or less occupants.

Sgouropoulos et. al. in [4] achieved a MAE of 1.15 by employing camera image processing techniques. However, the complex image processing algorithms require heavy computational resources. Paci et al. [5] utilized camera sensors and thermal comfort sensors combined with Support Vector Regression (SVR) to successfully count number of people inside large lecture rooms. Their approach produced a MAE of 7 people in rooms with 0-150 occupants, but worked well only when there is less movement. Nevertheless, if explicit consent is not obtained, privacy remains an issue for camera image processing based approaches. All the approaches based on special-purpose hardware sensors have the disadvantage of associated costs in deployment and maintenance.

Among the approaches that require both hardware and software, [6] installed mobile phone application to collect Received Signal Strength Indication (RSSI) data from beacons transmitted from Apples iBeacons. The approach in [7] proposes to estimate room occupancy by modifying the iBeacon protocol. Both [6] and [7] displayed near 100% accuracy but they require the cooperation of the occupants. Yoshida et. al. [8] employed a number of WiFi devices (e.g., Raspberry Pi) in a room to collect RSSI from WiFi networks. They related occupancy of a room with changes in signal propagation between APs and devices using linear regression (LR) and SVR to achieve a MAE of 0.471 in estimating occupancy in indoor environments with maximum 8 people.

Most light weight approaches for occupancy estimation use existing infrastructure as soft occupancy sensors. In [9] and [10] authors localize a set of people with high reliability and accuracy using WiFi fingerprints, however required the cooperation of occupants. Melfi et. al. [11] employed occupancy sensing methods such as monitoring of MAC and IP addresses in routers and WiFi APs. Although accuracy was within a 10% confidence interval around the ground truth occupancy for whole buildings, it was inaccurate at floor or room granularity due to the overlap of AP coverage and inconsistent wireless connectivity of devices. Balaji et al. [12] attempted to improve the accuracy issues identified in [11] by using occupant identity. They used WiFi MAC address and AP location from WiFi logs and achieved 86% accuracy in determining occupancy in office spaces in a commercial building.

The closest match to our work is by Redondi et al. [13] where they analysed WiFi activity data from APs inside a room to predict on the presence of people in a room. The study extracted sets of WiFi attributes (e.g., average signal quality, number of connections, etc.) during different time slots of the day to determine the presence using classification-based techniques (e.g., logistic regression, and Linear Discriminant Analysis (LDA). LDA produced best performance with 92% accuracy in predicting the presence of people in a room (empty or non-empty). They encountered the problem of bystanders and made an attempt to filter them out by using a threshold RSSI, without measuring the effectiveness of RSSI or determining if better approaches existed.

In [14], Akkaya et. al highlighted the growing trend to employ implicit sensing infrastructure (e.g., WiFi) to estimate occupancy due to the associated high costs in deployment and maintenance of special purpose hardware sensors. They also emphasized the challenges in estimating room occupancy with WiFi soft sensors, especially in areas such as lecture theatres in a university. To the best of our knowledge, our work is the first to use metadata in WiFi activity combined with intelligent machine learning techniques to address the challenges in using WiFi connectivity data to estimate occupancy in classrooms with large number of occupants in a university campus.

III. Data collection and Need to filter - Campus Testbed

A. Collection of WiFi data and ground-truth

We received daily dumps of WiFi sessions from UNSW IT department for 41 APs located in 7 lecture rooms on UNSW campus. A sample format of the WiFi activity data we received, is shown in Table I. The logs provide a unique identifier for each user, the address or addresses of device of devices that they used, the time when the device or devices associated or disassociated, which AP they associated with, and performance information such as the signal strength, number of retries, and quantity of information sent.

The unique user identifier shown in the first column of Table I is the UNSW student or employee identifier with which the WiFi user logged in to the university WiFi network. We have anonymised the user identifiers as we do not need to identify the user but only needs to find the users with multiple devices which otherwise over count the number of people in the room of our interest. As shown in the Table I,

TABLE I
EXAMPLE OF WIFI DATA LOG

Student ID	MAC address	Association Time	Disassociation Time	Session Duration	AP Name	Bytes Tx	Bytes Rcv	Pkts Tx	Pkts Rcv	SNR	RSSI	Status	Retries
1234567	00:08:22:60:fb:fe	27/06/2017 10:40	27/06/2017 11:15	35 min	mattap1	2717397	1717397	16536	24665	31	-63	Disass	1267
1234567	00:1e:64:d5:43:e6	27/06/2017 10:55	27/06/2017 11:20	25 min	mattap14	473749	2456743	2987	16041	27	-68	Disass	574
1235678	00:34:5c:fb:8d:2b	27/06/2017 11:15	27/06/2017 11:20	5 min	mattap13	1465373	6293826	4692	7832	35	-61	Disass	237
1235290	00:3b:21:5d:fb:80	28/06/2017 06:40	-	20 min	clb1ap17	156318	3462431	3689	2860	49	-45	Ass	453

the data in the first two rows belong to the same user with student identifier 1234567. The specific WiFi user had been connected to WiFi with two different devices during 27 th June 2017 as seen in the second column which specifies the device MAC address. For each session, WiFi logs consist of the associated time and the disassociated time and additionally the session duration 'Session Duration' which could also be computed using the association and disassociation times related to each session recorded in the WiFi logs. The UNSW IT department uses a naming convention for the APs where the APs are identified with relation to their location rather than using the MAC address of the APs and it is shown in 'AP Name' column. Furthermore, the WiFi report includes bytes and packets exchanged between the WiFi user and AP during each session. The 'RSSI' column shows the average of the received signal strength values during the session while average signal to noise ratio is recorded in 'SNR' column. In 'Status' column, $Disass$ indicates a disassociation if the session has been disconnected at the time of generating the report while Ass is possible for any ongoing sessions at the time of report generation. Since the WiFi reports are generated at a fixed time everyday (at 7am) majority of the sessions are recorded as $Disass$. An example with Ass status is shown in the fifth row of the Table I. For such records 'Disassociation Time' is not applicable and session duration is computed as the duration of time between the association time and the report generation time. The column 'Retries' indicates the number of times data frames has been resent to the receiver till the AP received ACK (acknowledgement) during the session. The large values seen in the WiFi logs for retries are desirable due to the interference and multipath fading which are common to WiFi.

As we noted that the WiFi logs provide us the personal information such as unique identifier for each user and the address of their devices, we therefore obtained ethical clearance (UNSW Human Research Ethics Advisory Panel approval number HC17140) to use that information in our research. Our application to the Ethics panel mentioned that one of the terms of use of our WiFi network is that "All activity on the wireless network is monitored" and by agreeing to that term the users grant explicit permission to monitor their activity in university network. The availability of unique user identifiers gave us the opportunity to remove multiple devices of a single user that causes overestimation of occupancy, when using WiFi as an occupancy sensor [15]. The list of enrolled students (class list) for a set of classes are collected as the ground truth for classification while we obtain ground-truth for actual occupancy by employing volunteers to collect head-counts.

B. Motivation of the study

We collected attendance for 40 classes on our campus and graphed class attendance against enrolment as shown in Figure 1. For most of the classes the actual attendance was well below the enrolment. For example, class shown by A in Figure 1 has an enrolment of 247 while the attendance was only 81 students. This presents the opportunity to optimize the usage of classroom spaces in a university campus by understanding attendance patterns.

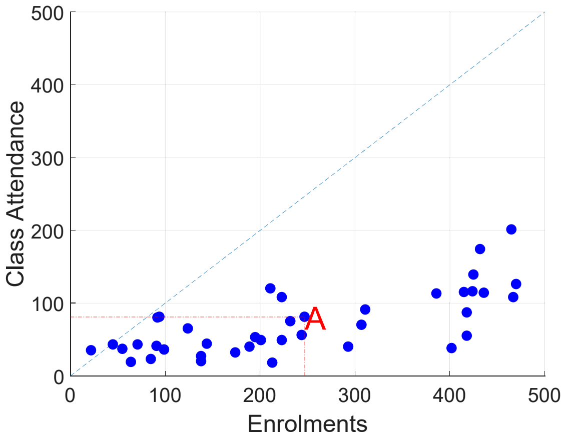

Fig. 1. Observed class attendance is often smaller than enrolments.

C. Necessity of filtering WiFi data

We analysed how effective 'WiFi connectivity counts' (the number of unique identifiers in WiFi data) as estimators of room occupancy in a dense campus environment, so as to determine the necessity of filtering the WiFi data. For the analysis we collected WiFi connectivity data, actual occupancy (ground-truth) and class lists across 40 classes in UNSW campus whereby the samples came from different courses, locations and during different days and times of the day. The WiFi connectivity data for a particular class consists of the WiFi users connected to APs in the room during the class. With the help of class lists for the classes of interest we filter bystanders by assuming, WiFi users who appear in both WiFi connectivity data and in class list are room occupants. We define 'WiFi Occupancy' ($Occupancy_{WiFi}$) as the number of people who get connected to WiFi APs of interest during the class while 'Enrolled WiFi Occupancy' ($Occupancy_{EnrolledWiFi}$) as the number of enrolled students among the $Occupancy_{WiFi}$. In Figure 2, we have illustrated the students enrolled in a class (in class list) as A,

$Occupancy_{WiFi}$ as B and $Occupancy_{EnrolledWiFi}$ as the intersection of A and B.

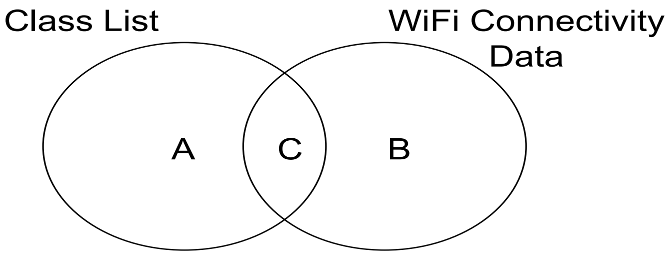

Fig. 2. Definition of Occupancy types for a class

As shown in Figure 3, all the data points lie below the 45 degree trendline showing that $Occupancy_{WiFi}$ was always higher than the $Occupancy_{EnrolledWiFi}$. This indicates that the $Occupancy_{WiFi}$ include both the students enrolled in the class and students in adjoining rooms or outside walkways.

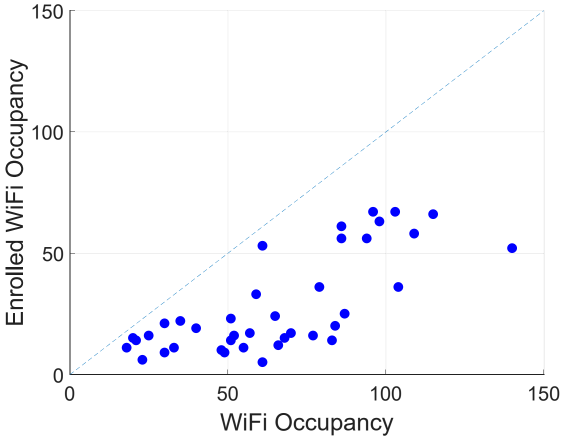

Fig. 3. $Occupancy_{WiFi}$ is higher than the $Occupancy_{EnrolledWiFi}$

Furthermore, we graphed $Occupancy_{EnrolledWiFi}$ with actual occupancy to find out that $Occupancy_{EnrolledWiFi}$ was lower than the observed actual occupancy for many classes (Figure 4). This is due to the actual room occupants who do not have any device connected to university network (e.g.,connected with private high-speed internet connections or turned off devices during lectures). We also identified few odd data points (A,B in Figure 4), where the actual occupancy is less than the $Occupancy_{EnrolledWiFi}$. The reason would be the rare cases of WiFi connections of enrolled students who do not actually attending the class but staying around the particular room.

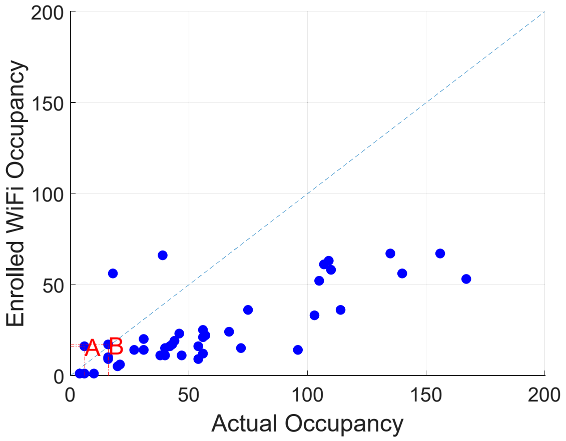

Fig. 4. Actual Occupancy is not an exact match to $Occupancy_{EnrolledWiFi}$

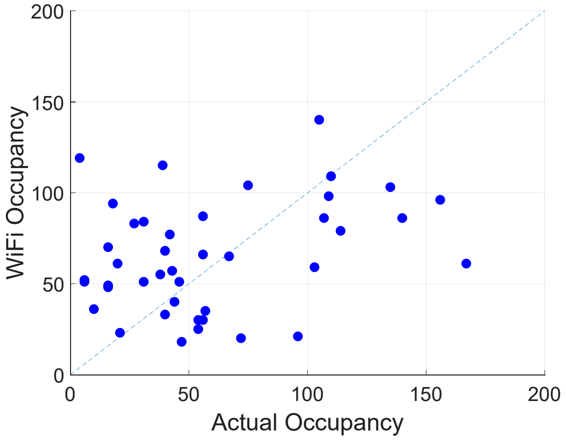

Fig. 5. Actual Occupancy correlates better to $Occupancy_{EnrolledWiFi}$ than to $Occupancy_{WiFi}$

A visual comparison between Figure 4 and Figure 5 shows that the actual occupancy shows a lower variability with Enrolled WiFi Occupancy than with WiFi Occupancy. The Pearsons correlation coefficient between the actual occupancy and the $Occupancy_{EnrolledWiFi}$ was 0.77 and actual occupancy and the $Occupancy_{WiFi}$ was 0.35, showing that a higher correlation is achieved when the bystanders are filtered out with class lists. Our analysis, therefore provides insights that filtering WiFi connectivity data would better relate the WiFi estimated room occupancy with the actual room occupancy. Even though using class lists to remove bystanders from WiFi data seems a good solution, it faces the difficulty in manually combining class list information every semester and also such approach is not generalizable to scheduled events such as meetings and seminars where class lists do not exist. In our approach we propose a generalizable approach to filter bystanders in estimating room occupancy in a campus environment. We use timetabling information which is available to access to general public but get rid of the need to access class lists.

IV. Feature Extraction

In general, the bystanders would often differ from occupants in the way they use WiFi. This section covers the set of features (extract from WiFi connectivity data) that is helpful to distinguish occupants and bystanders.

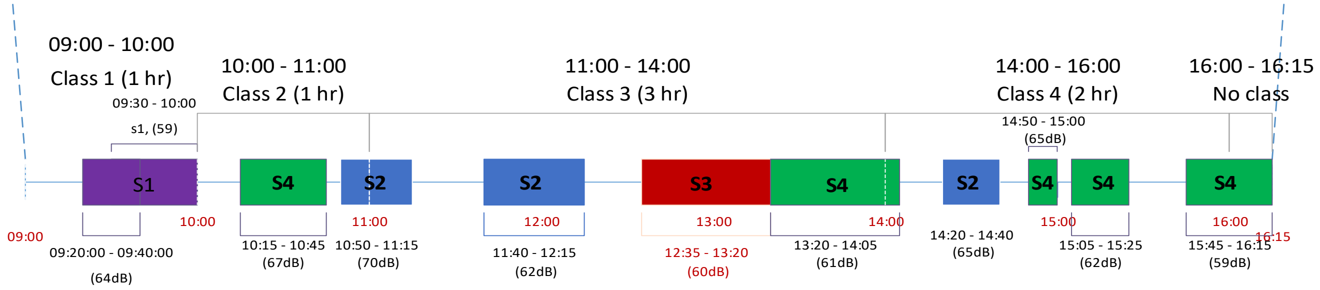

Fig. 6. WiFi activity of students

1) RSSI - Average of RSSI values across number of sessions associated with a user during the class of interest. Bystanders in general are expected to receive less signal strength than occupants.
2) Arrival delay - Time difference between the class start time and the WiFi user's first appearance in WiFi during the class of interest. A student who is attending the lecture is more likely to come to the classroom around the start time of the class, hence expected to have low arrival delays.
3) Number of sessions - Number of associations during the class of interest. There is a high chance for a student attending a lecture to have multiple associations during the class due to inconsistent WiFi connectivity of mobile devices as highlighted in [11].
4) Number of devices - Number of devices used to connect to WiFi during the class of interest. A bystander, walking past by is highly likely to get connected only with the mobile phone while student attending a class has a high probability of using multiple devices (mobile phone, tablet and laptop) to connect to WiFi.

Also, we derive two other time related features from WiFi activity data as explained below.

5) Percentage of 'in time' (t_{in}) - Percentage of a user's WiFi access that occurred inside the class time during the class of interest. Bystander who is walking past by the room may have less connected time to WiFi.
6) Percentage of 'out time' (t_{out}) - Percentage of user's WiFi access that occurred outside the class of interest. This is normalised by subtracting the class duration from the time in which the lectures are usually scheduled during the day (9am - 9pm) on our campus. Bystanders who connect to APs in a room who are working in nearby offices or study spaces typically have high t_{out} values.

To better understand these features, consider the activity of students S1 to S4 shown in Figure 6 which gives an example of a class timetable for one room with the timing of some user associations (shown in coloured boxes). The corresponding features are computed and summarized in Table II. S1 is a user who connects to WiFi with multiple devices, having overlapping sessions and maintains the connection for a short time even after the class. S2 probably has two classes in the same room on that day. S3 has only connected with one device during a class while S4 is seen throughout the day, hence likely to be someone who is working in the area, but may not be inside the room. During class 1 which lasted one hour, user S1 has two connections: One from 9.20am 9.40am and the other from 9.30am 10.00pm. We compute the non-overlapping connected time during class to be 40 minutes. S1 has spent 10 minutes out of the class during the day. Similarly, during class 3 which lasted for three hours, user S2 has two sessions having spent 50 minutes in class and he has a out time of 30 minutes. Another user S3 has spent 45 minutes in class 3, however has a (t_{out}) of 0 since he does not have any connection out of the class time of class 3 . However, S4 during class 3, has spent 40 minutes with a out time of 85 minutes displaying higher outside class presence.

TABLE II
FEATURES CALCULATED FOR EXAMPLE USERS

User	Class Duration	t_{in}	t_{out}	RSSI (dB)
S1	1-hour	40/60 = 67%	10/660 = 1.4%	61.5
S2	3-hour	50/180 = 27.8%	30/540 = 5.6%	66
S3	3-hour	45/180 = 25%	0	60
S4	3-hour	40/180 = 22.2%	85/540 = 15.8%	62

For the WiFi user groups namely occupants and bystanders, we plot the distributions of proposed features as shown in Figure 7 and compare the mean values of the two distributions of each feature; t_{in}, t_{out}, arrival delay, number of devices, number of sessions and average RSSI to understand how well the features would distinguish the different behaviours of the two user groups.

The analysis showed that the occupants in a room can be characterized by higher t_{in}, (67.9% vs. 27.3%), lower t_{out}, (3% Vs 25.1%), and lower 'arrival delay' (13.1 vs. 29.1 minutes). Furthermore, occupants used multiple devices to connect to WiFi (1.47 vs.1.08) and connected multiple times (more number of sessions) during a class (2.19 vs. 1.34). Although we expect RSSI to be one of the key features to identify user's presence in a room, we discovered that average RSSI of a user across the number of sessions he or she connected during the considered class does not exhibit much differentiation between the distribution of occupants and bystanders (59.4 vs. 66.4). This is due to the fact that devices

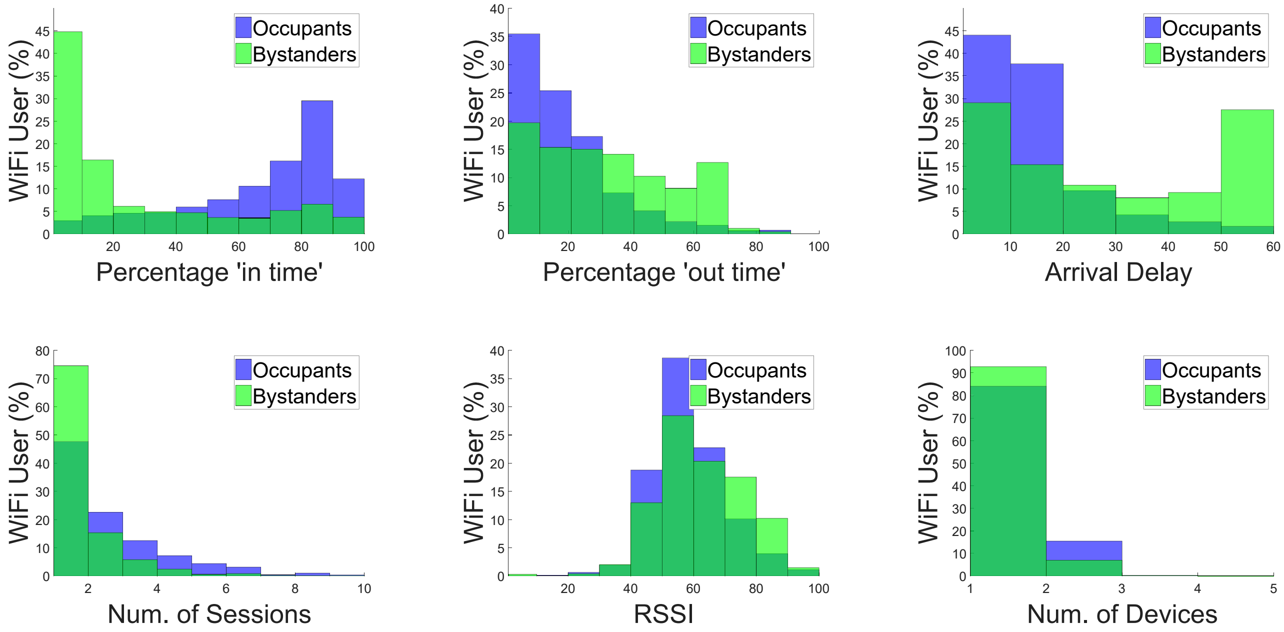

Fig. 7. Histogram of features for Occupants and Bystanders

in general get connected to the strongest AP regardless of its location and the received signal strength vary based on number of factors such as the type (e.g., laptop, mobile phone) and the vendor of the device. In addition, the session RSSI recorded in WiFi report is an average across the whole session which limits the insights to variations of RSSI.

V. Approach - Modelling of room occupancy

Our approach is two-fold. First, we detail how the features defined in Section IV, are used for classification to filter out bystanders. We then explain the use of regression algorithms to predict the room occupancy which compensate for the room occupants who are not captured by WiFi soft sensors. Figure 8 illustrates an overview of the proposed approach.

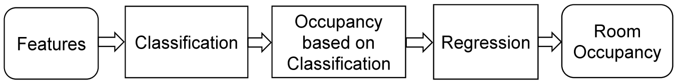

Fig. 8. Overview of the approach

A. *Classification of WiFi users*

We collected a data set of 10000 WiFi users across number of classes and compare the performance of logistic regression, SVM and LDA, which are widely used methods in binary data classification and compare their performance in the Section VI. We have shown in Figure 4 and Figure 5 that WiFi occupancy would better relate to the actual occupancy when bystanders are removed.

For each WiFi user (unique identifier appears in WiFi data) we extracted the features mentioned in Section IV;

1) Percentage of ‘in time’ (t_{in})
2) Percentage of ‘out time’ (t_{out})
3) Arrival delay
4) Number of sessions
5) Number of devices
6) RSSI

The above features are fed as inputs to the model that classifies a WiFi user as an occupant or a bystander. Based on the assumption that students who appear in both WiFi connectivity data and class list are in fact inside the room, we labelled such WiFi users as occupants and others as bystanders.

B. *Regression Analysis*

There are room occupants who have no connectivity to university network (students who do not have a connected device or students who have their own high-speed internet connections) as shown in the analysis in Section III, hence the occupancy obtained using the classification alone is not sufficient to estimate the room occupancy. Therefore, we employ a regression stage to better relate the output of the classifier to the actual room occupancy.

During 12 weeks of our study, we extracted 790 classes spanning different courses and 7 classrooms on the UNSW campus. In the sample, 46% of classes lasted one hour, 43% lasted two hours, 8% lasted three hours, 2% lasted one and a half hours, and 1% lasted four hours. The rooms are scheduled for lectures most of the time while paper-based exams are also occasionally possible, therefore we expected anomalous periods with little WiFi use. However, we omitted the data from weeks when classes were not held (e.g., mid-semester break). For each class, we predicted individual WiFi user’s presence in the room through classification and computed the number of occupants to be fed to regression analyser as the input variable. The training set was labelled using the actual occupancy of the room.

Linear Regression establishes a relationship between single or multiple independent variables with a dependent variable by fitting a best line, while SVR is another widely used regression method. We identified that both LR and SVR are commonly used in literature ([5], [8]), hence evaluate their performance in predicting room occupancy in Section VI.

VI. Evaluation of proposed approach

In this section, we first present the performance of classification and regression algorithms and then compare our approach with related work. Lastly, we compare the accuracy of different occupancy estimation methods and show how the accuracy vary across rooms with different capacities and classes with different enrolments, and classes held during different times of the day.

A. Performance of classification models

To evaluate the performance of the different classification algorithms, we used a test set that included 1500 WiFi users and determined the accuracy of them being predicted bystanders or occupants. Among the classifiers, LDA exhibits the best performance. For the WiFi users who were actual occupants, the model correctly classified them 84% of the time while this accuracy dropped to 81% in bystander prediction. Table III provides a comparison of different classifiers and the confusion matrix in Table IV shows the performance of LDA.

TABLE III
Comparison of classification methods

	Predicted Occupant	Predicted Bystander
Logistic Regression	79%	76%
SVM	83%	70%
LDA	84%	81%

TABLE IV
Confusion matrix for LDA classification

	Predicted Occupant	Predicted Bystander
Actual Occupant	84%	16%
Actual Bystander	19%	81%

B. Performance of Regression models

We used the samples of actual occupancy collected for the analysis in Section III, to evaluate performance of regression methods. Root Mean Squared Error (RMSE) (1) and Mean Absolute Error (MAE) (2) were used as relative measures of accuracy. According to RMSE and MAE criteria, the smaller the error, the better the forecasting ability of the method.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(F_i - A_i)^2}{n}} \quad (1)$$

$$MAE = \frac{\sum_{i=1}^{n} \mid F_i - A_i \mid}{n} \quad (2)$$

where A_i is the actual observed value and F_i is forecasted value for ith regression input of n inputs.

Table V shows a comparison of LR and SVR where MAE and RMSE values computed for the two models indicate that both of them would display nearly the same forecasting performance. We only show the results from LDA classification based LR in the evaluations in subsections C, D and E, however emphasize that SVR is another potential regression method for our approach.

TABLE V
Comparison of regression methods

	RMSE	MAE
Support Vector Regression (SVR)	25.5	17.8
Linear Regression (LR)	25.4	17.5

The Linear Regression equation obtained for our training dataset is shown in Equation 3.

$$Y = 10.3 + 1.25 \times X \quad (3)$$

where Y is Room Occupancy and X is occupancy computed by LDA classification. According to the obtained equation, occupancy predictions by LR model inflate the occupancy computed based on LDA classification so as to predict an occupancy value closer to the actual occupancy value. This can also be explained by the observations in Section III (Figure 4) where for most cases actual occupancy was higher than the WiFi occupancy filtered using class lists.

C. Performance Comparison with related work

We compare the performance of our method with the existing studies in which errors were computed and presented in terms of mean absolute error (MAE) by computing the normalized mean absolute error as shown in Table VI. In the table, we have also shown the level of cost requirement in terms of deployment, maintenance and computational complexity.

TABLE VI
Comparison of errors with related work

Study	Occupants	Normalized MAE	Sensor	Cost
[4]	0 - 8	0.29	Camera	High
[8]	0 - 8	0.12	Raspberry Pi + WiFi	Low
[3]	0 - 14	0.14	PIR	Medium
[5]	0 - 150	0.09	Camera + Ambient	High
Our Method	0 - 250	0.14	WiFi	Zero

Many of the related work were tested only for fewer number of occupants in rooms and also require additional efforts and costs of installation and maintenance. The performance of our method which can be deployed at zero cost and has been tested in rooms with occupants range from 0 to 250 is therefore appealing as it displays a similar normalized MAE with existing studies.

D. Performance Comparison of different methods

In a parallel research to our work [16], the same rooms considered in our study were instrumented with EvolvePlus wireless beam counters to count the room occupancy. We compare the accuracy of the occupancy output by LDA classification, LDA classification based linear regression, raw WiFi connectivity counts (unique student identifiers appeared in WiFi data during class of interest) and the beam counters. We compute the occupancy output of LDA classification by summing up the number of WiFi users predicted as occupants while occupancy output by regression is computed by using the output of the LDA classification as the input to linear regression. Since it is intuitive to interpret results in percentage terms we used the symmetric mean absolute percentage error (sMAPE)(4) in our comparison.

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{| F_i - A_i |}{| F_i | + | A_i |} \tag{4}$$

where A_i is the actual value, F_i is the forecasted value for ith regression input of n inputs.

The sMAPE computed for different approaches are shown in the Table VII.

TABLE VII
ERROR RATES (SMAPE) OF DIFFERENT OCCUPANCY ESTIMATION METHODS ACROSS ALL ROOMS

	WiFi Counts	LDA	Our method	Beam Counters
sMAPE	26.3%	20.15%	13.3%	13%

The objective of the classification is to remove the bystanders which corrupts the WiFi connectivity counts in estimating room occupancy in a dense campus environment. To compensate for room occupants who are not captured by WiFi we proposed employing a regression step. Regression after classification yielded better accuracy displaying the importance of having a two-stage approach so as to remove bystanders and also to capture the actual room occupants who are not captured by WiFi soft sensors. A closer look at the predictions of regression showed that it inflates the result of classification such that it gets closer to the actual occupancy. The lowest percentage error was obtained with beam counters which was approximately half that of the error produced by raw WiFi counts and followed by our WiFi estimators, while raw WiFi counts showed the highest error. However, our approach reduced the error rates of raw WiFi counts approaching to that of the accuracy achieved by hardware beam counter sensors. Figure 9 illustrates the estimated number of people using our approach with the actual number of people for the test set.

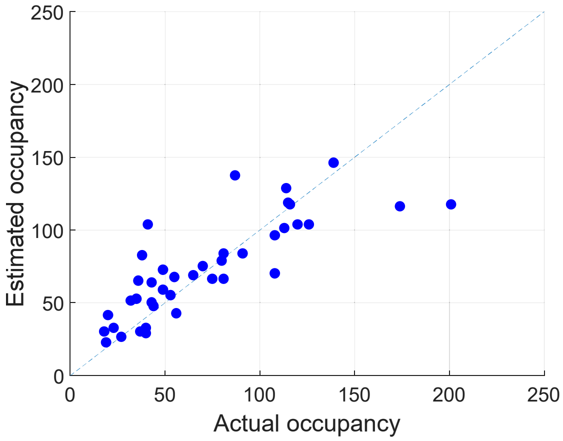

Fig. 9. Actual and Estimated occupancy by LDA based LR

E. Generalization ability of the approach

In Table VIII, we have shown how percentage error (sMAPE) varied across rooms in our test set. Similarly, Figure 10 shows the percentage error (sMAPE) for the estimations of our test set across different enrolment values range from 0 to 500. The slight variations of error rates across rooms of varying size and classes with varying enrolment indicates that our approach is generalizable to the whole campus.

TABLE VIII
PERCENTAGE ERROR (SMAPE) VARIED SLIGHTLY ACROSS ROOMS

	Capacity	LDA based LR
Room 1 (MAT 227)	42	9.1%
Room 2 (MAT 228)	42	8.6%
Room 3 (MAT C)	110	9.7%
Room 4 (CLB 8)	231	15.4%
Room 5 (MAT B)	246	16.1%
Room 6 (MAT A)	472	9%
Room 7 (CLB 7)	497	12.5%

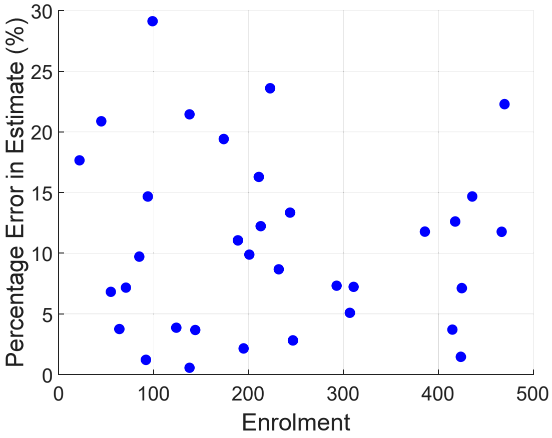

Fig. 10. Distribution of Percentage Error (sMAPE) across Enrolment

Furthermore, it is possible that afternoon and evening classes may have more students whose battery is low, hence do not connect to WiFi. We computed the sMAPE for classes held in the morning, afternoon and evening classes. In Table IX, we have shown how percentage error (sMAPE) varied across classes held in the morning (9am - 12pm), afternoon (12pm - 4pm) and evening (4pm - 9pm). There is only a small variation of sMAPE for classes held in morning and afternoon while we see the error rate has reduced to that of half during the evening classes. The reduced error for evening classes could be explained as majority of the classes held in the evening are postgraduate classes where occupants may have not been in the university from the morning but arrived for the evening classes only. The spatial distribution of WiFi access points is another factor that affects accuracy at scenarios where there is no access points in the room. While this could be an issue for other venues, it did not arise on our campus since we have a high density of APs and classrooms are usually equipped with multiple APs. However, with the growing popularity of wifi networks in the world, our approach is applicable to majority of the venues.

TABLE IX
EVENING CLASSES SHOW LOWER PERCENTAGE ERROR (sMAPE) THAN MORNING AND AFTERNOON CLASSES

	Duration	sMAPE
Morning	9am - 12am	13.6%
Afternoon	12pm - 4pm	14.1%
Evening	4pm - 9pm	7.1%

VII. CONCLUSION

In this paper, we have proposed and evaluated a novel approach to estimate room occupancy using pervasive wireless infrastructure in a university campus. We propose a set of features extracted from WiFi connectivity data to distinguish bystanders and filter them out using classification-based algorithms. LDA showed the best classification performance having 84% accuracy in predicting actual room occupants and 81% accuracy in predicting bystanders. The percentage error in estimating room occupancy based on standalone classification was 20.15%. Then we feed the LDA classification output to a regression model to compensate for the occupants who are not captured by WiFi. Out of the regression methods we employed, both LR and SVR displayed similar forecasting performance having a percentage error of 13.3%. The comparison of forecasting performance between raw WiFi connectivity counts, our LDA based LR estimator and hardware beam counter sensors indicated that our approach estimates occupancy in rooms with hundreds of occupants with an accuracy approaching the accuracy obtained by special-purpose sensors such as beam counter sensors. There was only a slight variation in sMAPE observed across classes with different enrolment and rooms with different capacities, therefore displaying the possibility of generalizing our approach. The study demonstrated our hypothesis that WiFi activity could be a non-intrusive, reliable soft sensor to estimate occupancy in rooms with large number of occupants in a university campus, hence proving the rooms can be dynamically allocated to courses based on the attendance rather than enrolment. The obvious variation in attendance which occur during exams can be considered by schedulers when deciding which room to use for a class since they would be aware of that variation. We intend to address the variation, more generally, by determining how wifi use and occupancy may vary by the type of class (lecture, tutorial and lab) in future work. Although our approach is intuitive in institutional buildings where there is timetable information, there is good scope to apply it to other buildings with office, meeting and open spaces by adjusting the identified features further in future work.

REFERENCES

[1] "Occupancy sensors market - industry trends, opportunities and forecasts to 2023," https://www.researchandmarkets.com/research/w4kck9/global_occupancy?w=5, Dec. 2017, [Online; accessed 27-03-2018].

[2] B. Dong, "Occupancy detection through an extensive environmental sensor network in an open-plan office building," in *Eleventh Int. IBPSA Conference*, no. January 2009, 2014.

[3] 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," *Proc. of the 2016 ACM Int. Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, pp. 1016–1027, 2016.

[4] D. Sgouropoulos, E. Spyrou, G. Siantikos, and T. Giannakopoulos, "Counting and Tracking People in a Smart Room : an IoT Approach," in *2015 10th Int. Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, 2016.

[5] F. Paci, D. Brunelli, and L. Benini, "0, 1, 2, many - A classroom occupancy monitoring system for smart public buildings," *Conference on Design and Architectures for Signal and Image Processing, DASIP*, vol. 2015-May, 2015.

[6] Y. Yang, Z. Li, and K. Pahlavan, "Using iBeacon for Intelligent In-Room Presence Detection," in *2016 IEEE Int. Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2016.

[7] G. Conte, A. A. Nacci, V. Rana, and P. Milano, "BlueSentinel : a first approach using iBeacon for an energy efficient occupancy detection system," in *1st ACM Int. Conference on Embedded Systems For Energy-Efficient Buildings (BuildSys) 2014*, no. Nov, 2015.

[8] T. Yoshida, "Estimating the number of people using existing WiFi access point in indoor environment," *6th European Conference of Computer Science (ECCS '15)*, pp. 46–53, 2015.

[9] Y. Jiang, X. Pan, K. Li, Q. Lv, R. P. Dick, M. Hannigan, and L. Shang, "ARIEL: Automatic Wi-Fi based Room Fingerprinting for Indoor Localization," *Proc. of the 2012 ACM Conference on Ubiquitous Computing*, pp. 441–450, 2012.

[10] N. T. Nguyen, R. Zheng, and Z. Han, "UMLI : An Unsupervised Mobile Locations Extraction Approach with Incomplete Data," in *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, 2013.

[11] R. Melfi, B. Rosenblum, B. Nordman, and K. Christensen, "Measuring building occupancy using existing network infrastructure," *2011 Int. Green Computing Conference and Workshops, IGCC 2011*, 2011.

[12] B. Balaji, J. Xu, A. Nwokafor, R. Gupta, and Y. Agarwal, "Sentinel: occupancy based HVAC actuation using existing WiFi infrastructure within commercial buildings," *Proc. of the 11th ACM Conference on Embedded Networked Sensor Systems*, p. 17, 2013.

[13] A. E. Redondi, M. Cesana, D. M. Weibel, and E. Fitzgerald, "Understanding the WiFi usage of university students," *2016 Int. Wireless Comm. and Mobile Computing Conference, IWCMC 2016*, pp. 44–49, 2016.

[14] K. Akkaya, I. Guvenc, R. Aygun, N. Pala, and A. Kadri, "IoT-based occupancy monitoring techniques for energy-efficient smart buildings," *2015 IEEE Wireless Communications and Networking Conference Workshops, WCNCW 2015*, no. March, pp. 58–63, 2015.

[15] Y. Wang and L. Shao, "Understanding occupancy and user behaviour through wi-fi-based indoor positioning," *Building Research & Information*, vol. 0, no. 0, pp. 1–13, 2017.

[16] T. Sutjarittham, H. Habibi Gharakheili, S. Kanhere, and V. Sivaraman, "Data-Driven Monitoring and Optimization of Classroom Usage in a Smart Campus," *Proc. of the 17th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN 2018)*, 2018.