



# Understanding Customers and Their Grouping via WiFi Sensing for Business Revenue Forecasting

Vahid Golderzahi<sup>(✉)</sup> and Hsing-Kuo Pao

Department of Computer Science and Information Engineering,  
National Taiwan University of Science and Technology, Taipei 10607, Taiwan  
golderzahi@gmail.com, pao@mail.ntust.edu.tw

**Abstract.** Emerging technologies provide a variety of sensors in smartphones for state monitoring. Among all the sensors, the ubiquitous WiFi sensing is one of the most important components for the use of Internet access and other applications. In this work, we propose a WiFi-based sensing for store revenue forecasting by analyzing the customers' behavior, especially the grouped customers' behavior. Understanding customers' behavior through WiFi-based sensing should be beneficial for selling increment and revenue improvement. In particular, we are interested in analyzing the customers' behavior for customers who may visit stores together with their partners or they visit stores with similarly patterns, called group behavior or group information for store revenue forecasting. The proposed method is realized through a WiFi signal collecting AP which is deployed in a coffee shop continuously for a period of time. Following a procedure of data collection, preprocessing, and feature engineering, we adopt Support Vector Regression to predict the coffee shop's revenue, as well as other useful information such as the number of WiFi-using devices, the number of sold products. Overall, we achieve as good as 7.63%, 11.32% and 14.43% in the prediction on the number of WiFi-using devices, the number of sold products and the total revenue respectively if measured in Mean Absolute Percentage Error (MAPE) from the proposed method in its peak performance. Moreover, we have observed an improvement in MAPE when either the group information or weather information is included.

**Keywords:** Customer behavior · Group behavior  
Received Signal Strength Indicator (RSSI) · Revenue forecasting  
WiFi sensing

## 1 Introduction

Nowadays most stores provide WiFi services for customers who are equipped with WiFi functioning smartphones and interested in accessing Internet. It is well known that WiFi signals, along with other video or non-video-based technology

may be helpful in understanding people's behavior for people located in a smart space, or in particular the customers' behavior in stores [3,12]. In this work, we propose a method based on WiFi sensing given customers' behavioral inputs for store revenue forecasting. In particular, we are interested in the customers' group information where we can observe friends who find each other to go for a drink together or different individuals may share similar visiting behavior even they do not know each other.

Compared to online shopping where all the surfing and purchase behaviors from customers are automatically logged, the marketing in brick-and-mortar business usually face the challenges as they need to deploy the customer analytics framework to the physical realm. In one way or the other, the traditional stores must find solutions to keep track of customers such as when customers may visit the stores, what they prefer to own and what they really purchase in the end based on their judgment between the product quality and price. To understand targeted customers as much as they can to boost the stores' revenue, two major technologies offer the answers: the video and non-video-based approaches. To avoid the privacy leaking issues, the non-video-based approaches are generally favored from the customers' side because they keep the customers' information to its minimum for business analytics. Among the various non-video-based approaches, WiFi-sensing is a major choice due to its popularity. Existing WiFi sensing, which is a cost-effective as well as privacy-preserving technology, can be appropriate for customer behavior analysis [1,3].

Signal-based indoor sensing for human tracking and business analytics are generally categorized into several categories [14]. A rich set of IoT (Internet of Things) technology with sensors such as passive infrared sensor (PIR), ultrasound, temperature sensors, as well as various vision-based devices can be deployed in the indoor environment for human counting, tracking and activity recognition to name a few. In general, we need to spend efforts on the device deployment physically and the device calibration and threshold setting may not be straightforward for this kind of technology. On the other hand, there are also some devices that we need the humans located in the indoor environment to carry to make the sensing possible. Some wearable devices and smartphones fall into this category. Apparently, we prefer a scenario that is: (1) easy to deploy in the indoor environment, (2) providing high sensing accuracy, and (3) with enough covering rate among people. In another word, we look for a sensing technology where we can: (1) easily implement both in its hardware and software, (2) find convincing tools for analytics and (3) detect as high percentage of people as possible in a given environment where each of the targeted people carries a device that is necessary for sensing. We propose a WiFi-based sensing method [13] where we only assume smartphone carrying from the customers for the indoor customer detection and tracking for business revenue forecasting. By having the technology, we keep the deployment efforts to the minimum and at the same time, we enjoy a decent sensing performance.

The proposed method is realized in a coffee shop where we track and analyze customers' behavior and the associated group information with a WiFi AP. For customers who visit the coffee shop with functioning WiFi, we can collect the

WiFi related information and use it to summarize customers' behavior. One of the reasons why we choose the coffee shop for our study is because drinking coffee and visiting the coffee shop is considered not a mandatory but an optional activity for people where we may choose to have with our friends and when we have certain mood for relaxation or doing business in the environment. This coffee shop is located in Da'an district in Taipei City and close to a university area. Most of its customers are students who may spend their time to have fun with their friends, or work on their homework/projects individually or with a group. From time to time, the coffee shop owner may provide some special discounts to students to encourage them coming to the shop, which could lead to revenue increment.

We use the WiFi related information to track the coffee shop's customers and analyze their behavior using RSSI signals captured via the WiFi AP. We monitor the coming and leaving time for each customer as well as their duration of stay given the RSSI signals. Occasionally, the AP may grab some data from people who pass by the coffee shop or stay in a store nearby. We address these noise data by applying some filters on RSSI signals and the duration of customers' stay. Furthermore, we will extract frequent customers and analyze their behavior to detect the groups of frequent customers. Customers may form a group if they come to the shop together. On the other hand, we also consider a group if customers from the group often come to the coffee shop at some similar time or stay for similar duration. For instance, some people may come the shop before going to work or stay in the shop for almost the whole day long. We believe that they could have similar working patterns or share similar income levels and should behave similarly in their visiting and purchase behavior. In the end, we discuss both of the cases where we may not include or may include the group information as described above in the feature set for prediction. We take turn to predict the total number of customers' devices, the total number of sold products and the total revenue. The prediction model is Support Vector Regression (SVR).

We should emphasize the main contributions and what differentiate the proposed method from the previous solutions for indoor human sensing and business revenue estimation as follows:

- The proposed method is based on an easy-to-deployed scenario where we only assume smartphone carrying from the customers. Moreover, the proposed approach is a *passive* approach where we need customers to open no special software to activate the sensing. On the indoor environment, we need only a tuned AP for WiFi signal collection. By having this property, we can easily convince business stores for its realization.
- We focus on using the customers' group information for revenue forecasting. The group information separates customers from different groups, such as loyal customers, customers with different vocations, customers with different product preferences and customers with different daily or weekly schedule. Knowing the above information may improve the business revenue as the business should have more understanding about its customers.

- The proposed method respects the privacy issue. Unlike many indoor tracking strategies, we collect the information only the part for *signal broadcasting* from customers. Usually, we can assume customers have no objection on releasing the information. It could be hard to hide the broadcasting information in general when a handshaking communication is needed.

The remaining of the paper is organized as follows. An overview of WiFi-based and non-WiFi-based sensing approaches is provided in Sect. 2. Afterwards, we discuss the proposed method along with all the necessary procedures in Sect. 3, which is followed by the experiment results and evaluation in Sect. 4. Finally, we conclude our work in Sect. 5.

## 2 The Past Work

The goal is to adopt indoor sensing on customers for business revenue forecasting. There are a variety of technology that has been developed for this purpose. As we briefly described, the major strategies can be separated into several groups based on whether we need to deploy certain devices or system on the indoor side and whether we assume any devices from the customers to carry to make the sensing possible. In this section, other than the research that we have discussed in Sect. 1, we mainly discuss the approaches that are directly related to this work. We emphasize that what we plan to detect and track is more than a handful customers where we may not assume any limit for the number of customers. Moreover, identifying the tracked customers is valuable to have in this application. Therefore, the IoT solutions such as PIR, ultrasound and temperature sensing are not precise enough to solve the problem. On the other hand, the vision-based methods may not be the best choice due to the privacy concerns from the general public. We turn our attention to the approaches where we assume customers carrying devices and the devices provide enough information for detection, tracking and analytics.

The user-carrying device approach can be divided into smartphone and non-smartphone categories. The former represents a scenario in which users carry their own smartphones, thereby they are trackable and their identifiable information would be extractable through the smartphones [1]. The latter relies on additional wearable devices that should be carried by users such as bracelets, smart glasses, RFID, etc. For instance, Han et al. [3] implemented a Customer Behavior IDentification (CBID) system based on passive RFID tags. Their system includes three main parts; discovering popular items, revealing explicit correlations, and disclosing implicit correlations to understand customers' purchase behavior. The technology is mainly focused on a small set of people and may have difficulty when we have a large number of unknown people to track and therefore hard to implement in the crowded situation [4, 7].

On the category of smartphones, we have all-in-one devices which have the identifiable information as well as a various set of equipments, sensors and apps for information collection and environmental monitoring. People may prefer to

carry smartphones simply because the smartphones play such a role of combining many functionalities in a single device [6, 8]. That implies using smartphones as the assumed carrying device for customer sensing should provide enough covering rate when we use smartphone-related signals to estimate the existence of customers. Among all possibilities, WiFi-equipped smartphones can be considered one of the best solutions to be carried by unknown people or a large number of customers who intend to communicate with public devices due to the built-in identifiable characteristics in the smartphones. By having that, we aim to detect, track and analyze people with their existence and group behavior [6, 11].

Zeng et al. [12] proposed WiWho, which is a method to identify a person using walking gait analysis through the WiFi signals. WiWho consists of two endpoints, a WiFi AP and any WiFi-equipped device for communicating and collecting Channel State Information (CSI). It has some limitations such as assuming the straight walking paths from customers and should have the performance limit while the tracked person turns. Vanderhulst et al. [6, 11] discussed a framework to detect human spontaneous encounters in which spontaneous and short-lived social interactions between a small set of individuals have been detected. It leverages existing WiFi infrastructure and the WiFi signals, so-called “probe” can periodically be radiated by a device to search for available networks. The probes are used to capture radio signals transmitted from users’ devices to detect human copresence. The limitations of the proposed method include device variety, a limited number of participants to be allowed for high accuracy detection, and the required application to be installed on users’ smartphones.

An extended Gradient RSSI predictor and filter was proposed by Subhan et al. [10]. It is a predictive approach to estimate RSSI values in presence of frequent disconnections. The approach predicts users’ positions and movements in terms of their current situations and movements. The distance changes between users’ devices and the AP lead to the increase and decrease of the RSSI values and therefore the targeted users as well as their movements can be detected.

As other similar research, Maduskar and Tapaswi [6] proposed an approach to trace people’s positions and movements using an RSSI measurement of WiFi signals from several APs in predetermined locations. The RSSI-based approaches have the minimum complexity compared to other signal-based indoor localization techniques. In their approach, the larger size of the APs results in more accurate location estimation. The weakness of the approach is that a careful initialization is necessary given a new environment, e.g., customer sensing given an indoor store. Du et al. [2] proposed algorithms for fine-grained mobility classification and structure recognition of social groups using smartphones through their embedded sensors. They have utilized embedded accelerometer to detect group mobility behavior. Afterward, a supervised learning algorithm is applied to recognize different levels of group mobility, such as stationary, walking, strolling, and running. The method can also be used to recognize the relations and structures of a group by monitoring the leader-follower, the left-right relations and distances using smartphones’ sensor data. To compare the above two research work, the localization technique is basically not a must to have in our scenario

because the main purpose of the proposed method is on understanding when customers visit a store instead of what customers prefer to own. Therefore, it is the visiting behavior not the purchase behavior that interest us. In the next section, we discuss the proposed method in details.

### 3 Proposed Method

The goal is to predict the number of customers and revenue on each day given the past selling and customers' behavior history. What is different from previous approaches is that we rely on the WiFi signal collection to help us know further about the visiting customers where the WiFi information may tell us the information from the macro scope such as the total number of customers to the micro scope such as the customers' identifiable information. We demonstrate the whole prediction scenario starting from data collection, data preprocessing to prediction model itself in the next few subsections.

#### 3.1 Data Collection

The data for the proposed method includes two parts: the WiFi-based data and others. The WiFi-based data has been collected using a WiFi collecting device, TP-Link TL-WR703N WiFi router in the coffee shop. It is an access point (AP) which operates in IEEE 802.11n mode to collect the data from customers' devices such as smartphone, laptop, tablet, etc. The extracted data from the received WiFi signals per customer includes:

- the physical address (MAC address),
- service set identifier (SSID), and
- received signal strength indicator (RSSI).

Given the MAC address information, we can calculate the number of devices for each day. Usually the number of devices may be close to the number of customers per day if each customer carries only one smart device (further discussed below in the assumption part). The WiFi data has been extracted using the Wireshark packet analyzer<sup>1</sup>. Based on the collected WiFi information, we also derive some information which may be important for the prediction:

- the come-in time of a customer,
- the leaving time of a customer, and
- the way a customer was served, such as “staying in” or “prepared to go”.

The come-in time and leaving time are recored based on the first and last signals that we can collect for each specific MAC address (identity). How a customer was served is estimated based on the duration of the WiFi signals that we received per MAC address, such as above or below a predefined threshold (further discussed below). Furthermore, there are customer considered as frequent

<sup>1</sup> <https://www.wireshark.org>.

customers. We add a set of group information features, which are extracted via *frequent customers' behavior analysis*. Note that the above three are individual based, collected for each MAC address. On the other hand, the SSID and RSSI are collected following a predefined sampling rate. In addition to the WiFi related information, we also collect some other information which may have influence on the coffee shop revenue. The information includes:

- temperature, and
- rain probability.

On the side, the analyzed dataset also consists of the number of various sold products and the total revenue for each day. The owner of the analyzed coffee shop is kind to provide the valuable information for us to confirm the performance of the proposed method. Some correlation between the number of sold products, the revenue and the number of devices (MAC address) is further discussed below. The dataset was collected from 2016/09/02 to 2016/12/04 in which the training data is set from 2016/09/02 to 2016/11/20 including 78 days, and the test data is from 2016/11/21 to 2016/12/04 including 14 days. Due to some technical difficulty, we have a few days of missing data. The longest of data missing is a gap of eight days from the 6th to the 7th weeks, shown in Fig. 1. We have consulted the average ratio between the number of devices and the number of sold products to fill the missing values for the study.

*Privacy Issues.* We need to emphasize that we try our best to respect the customers' privacy. The data collection procedure is focused only on the part that the customers broadcast to the environment. We do not attempt to construct a data collection procedure where the customers' browsing history, browsing URLs, etc. may be collected through our AP. That is, we do not trick customers by creating an AP where we may have the above information or even the username or password information from customers.

### 3.2 Data Preprocessing

The first issue of customer behavior analysis is to identify real customers. In this study, we filter the devices (identified via the MAC address) detected by the WiFi AP by setting thresholds of duration from five minutes to three hours. That is, we assume that each customer stays in the coffee shop no shorter than five minutes and no longer than three hours ( $5 \text{ min} \leq \text{staying time} \leq 3 \text{ h}$ ). The detected devices with duration shorter than five minutes are assumed to be passing by devices and the devices with duration longer than three hours are likely to be the staff of the coffee shop or neighboring shops. The thresholds are decided based on our visual estimation when we visited the coffee shop. Also, we set a threshold applied to RSSI where we include only the RSSI greater than  $-70 \text{ dBm}$  in the data collection ( $-70 \text{ dBm} < \text{RSSI}$ ).

As customers reach the entrance door of the coffee shop, they are in the range of our data collection AP and the RSSI keeps increasing as customers moving

into the coffee shop. We record the devices and identify the come-in time when the devices show the above pattern. Recording the customers' leaving time is the opposite. The stay duration for customers can be used for customers' behavior analysis such as the reasons they visit the coffee shop (for study, meeting friends or web surfing, etc.) and whom they go with.

### 3.3 The Proposed Prediction Method

The complete step-by-step procedure of the proposed method includes: (1) data collection, (2) feature extraction, (3) clustering for group information extraction, (4) model building and (5) prediction. We start by collecting the WiFi data through our modified AP. The WiFi data are compiled into several features and they are combined with the non-WiFi features such as weather information to form a complete feature set for model training. On the side, we have additional information provided by the coffee shop owner such as revenue related information to confirm our evaluation.

Analyzing customers' behavior or more specifically finding customers' group information is a key contribution of the proposed method. We assume customers who are classmates, partners, or colleagues may go to the coffee shop together frequently as a group. On the other hand, some customers, even they may not know each other can behave similarly such as they may visit the shop at similar time or on similar days (all coming in the morning, after lunch, after work or coming during weekdays or weekend), or with similar frequency (once per day or once per week). Given the above group behavioral inputs, we would like to extract a set of features called *group information* to describe different customers. By including those features, we may have a better chance to understand different customers and thus a better chance to predict the revenue of a business.

Given a set of customers' features, we adopt a hierarchical clustering method called Unweighted Pair Group Method with Arithmetic mean (UPGMA) to find customers' group information. Specifically, we have a set of features to describe customer  $i$  as:

$$\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iK}) \quad (1)$$

where we have  $K$  days to consider in our customer analysis and we should use  $K$  binary attributes to indicate the presence of customer  $i$  in the coffee shop on different days. That is,

$$z_{ik} = \begin{cases} 1 & \text{if customer } i \text{ visits the coffee shop on day } k, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Given the above inputs, UPGMA builds a rooted tree (dendrogram) that reflects the structure of pairwise similarities between different customers [5]. To describe the similarity between two clusters  $C_i$  and  $C_j$ , we utilize a proportional averaging formulation written as:

$$\sigma_{ij} = \frac{1}{|C_i| \cdot |C_j|} \sum_{p \in C_i} \sum_{q \in C_j} \sigma_{pq}, \quad (3)$$



where  $|C_i|$  and  $|C_j|$  represent the cardinality of the set (i.e., the size) for  $C_i$  and  $C_j$  respectively; also,  $\sigma_{pq}$  measures the similarity between two entities  $p$  and  $q$  from  $C_i$  and  $C_j$  respectively. We measure the similarity between two customers  $p$  and  $q$  as:

$$\sigma_{pq} = \sum_k \delta(z_{pk}, z_{qk}) \cdot \delta(z_{pk}, 1), \quad (4)$$

where the function  $\delta(x, y)$  outputs 1 if  $x = y$  and outputs 0 if  $x \neq y$ . That is, we count 1 when two show up in the coffee shop on the same day and count 0 otherwise. All pairs of customers are compared through the pairwise computation to form a similarity matrix in the end. Then, a pair of elements with the maximum similarity are recognized and clustered together as a single grouped pair first. Afterwards, the similarity between this pair and all other elements are recalculated to form a new matrix. We go on to find the pair with the maximum similarity for grouping step by step until all are combined into one in the end [4, 5]. The output of UPGMA is a dendrogram and we can find the final grouping result by setting an appropriate number of clusters. In the end, the group information shall be used in building the Support Vector Regression (SVR) model [9] for the prediction on the number of customers' devices, the number of sold products and the total revenue.

### 3.4 Assumptions and Limitations

The goal is to analyze customers' behaviors that are related to coffee consumption. Due to the WiFi-based data collection nature, we first assume that all customers carry WiFi-based devices and their WiFi signals can be detected easily by the deployed AP. That is, the WiFi function must be on at all times when the customers visit the coffee shop, starting from entering to leaving the coffee shop, for all customers. Based on the assumption, we could capture customers' existence, in particular, we know when customers come to the coffee shop and leave the coffee shop. That is, as soon as we detect RSSI signals for each customer's device, it will be assumed that this is the exact entrance time for the customer. The leaving time of a customer is also assumed to be the time of losing or dropping off of the RSSI signal received from the customer's device. There are also some limitations in this research, such as using just one WiFi AP leads to weak distinguishment of the exact coming and leaving time for each customer. Moreover, we cannot detect the exact location and position of each customer. In the data cleaning phase, removing noisy or irrelevant data is hard especially in a crowded area<sup>2</sup>.

## 4 Experiment Results

We would like to predict the number of customers' devices, the number of sold products and the total revenue given a set of WiFi-based and non-WiFi-based features. There are two scenarios that we discuss:

<sup>2</sup> There is a convenient store right next to this coffee shop.

**Table 1.** The statistics of frequent and non-frequent customers.

	The frequent customer (%)	Non-frequent customer (%)
The number of customers	11%	89%
The number of visits	27%	73%

1. In the first scenario, we take turn to work on three prediction tasks given a sliding window of size  $L$  as well as other features such as the day in a week, the weather information, which consists of the temperature and rain forecasting to build the learning model.
2. In the second scenario, we consider additional features, the group information with the same sliding window as described in the first scenario to build the learning model.

We attempt to analyze how the past presence or purchase records can be used to predict the future presence or purchase. In particular, between the first and the second scenarios, we discuss how the group information can help us for better prediction. We utilized Support Vector Regression (SVR) [9] as the predictive model. The size of sliding window  $L$  is set as  $L = 14$  for this work. The detail result shall be shown below.

#### 4.1 Statistics of Frequent and Non-frequent Customers

Before going on to demonstrate the effectiveness of the proposed method, we first study some basic statistics of the data set. In many retail stores, the transactions from the frequent customers may usually dominate the store revenue. In this case, we also would like to understand the contribution from the frequent and non-frequent customers separately. In Table 1, we show the numbers of frequent and non-frequent customers, which are 11% and 89% out of the whole group of customers who visited the shop during the data collection period. Interestingly, we also observe that this 11% frequent customers contribute 27% of the visiting times in the coffee shop, compared to 73% of the visits from non-frequent customers. It implies a relatively large consumption from the frequent customers compared to the non-frequent ones. When we aim to find a predictive model with good performance, we better to focus more on the prediction of the frequent customers rather than the non-frequent ones. Fortunately, the frequent customers are likely to come to the coffee shop in a regular manner and could be predicted easily if compared to the non-frequent group. Moreover, the prediction on the frequent rather than the non-frequent customers may be easy simply because we usually have a relatively large frequent customers' data in the training set. We discuss more along these two aspects below.

We also analyze the daily visits (in %) from the frequent and non-frequent customers, shown in Fig. 1. From the beginning to the end of data collection period, we observe that the percentage of the frequent customers increases

slightly as time moving forward. It may be due to that the major group of customers includes a significant percentage of students from a nearby university. The students may know each other better and better starting from September (the beginning of the semester) through November and they may have more chances to go for a coffee when they know each other better. Note that we have a few days of missing values due to data collection difficulty in the period.

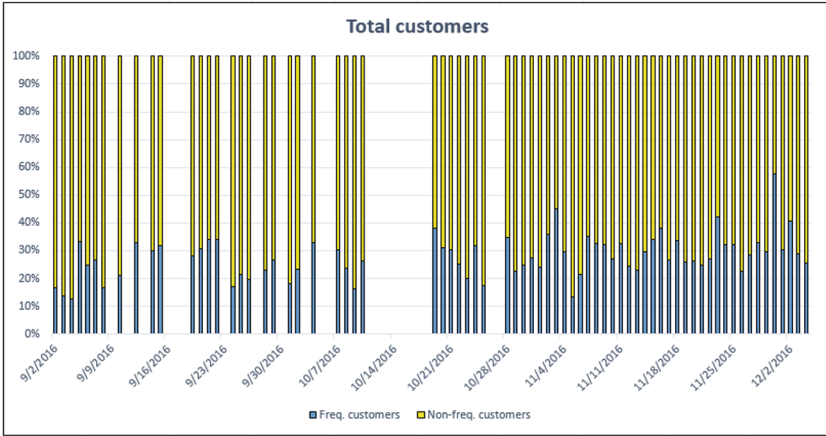


Fig. 1. The percentage of frequent and non-frequent customers per day.

### 4.2 Features and Results

**The Feature Set.** We first discuss the features that are used in this study. Following the description in the beginning of this section, let us use  $\tau_t$ ,  $\rho_t$  and  $day_t$  to describe the information such as the temperature forecasting, the probability of raining and the day in a week on the  $t$ -th day respectively. The sliding window of size  $L$  for the above information (except the day in a week) can be written as:

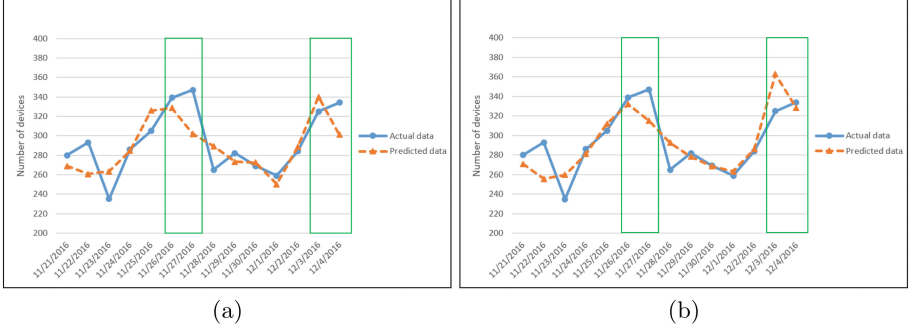
$$\begin{aligned} \mathbf{temp}_{t,L} &= (\tau_{t-L}, \dots, \tau_{t-1}), \\ \mathbf{rain}_{t,L} &= (\rho_{t-L}, \dots, \rho_{t-1}). \end{aligned} \tag{5}$$

To speak of the group information, we set the number of groups for group information extraction as  $K = 4$ . Given the assignment, we have the group features written as:

$$\mathbf{g}_t = (g_{t,1}, g_{t,2}, \dots, g_{t,k}, \dots, g_{t,K}), \tag{6}$$

where  $g_{t,k}$  records the number of customers from group  $k$  who visit the coffee shop on the  $t$ -th day. We can describe the group information with the sliding window of size  $L$  as:

$$\mathbf{g}_{t,L,k} = (g_{t-L,k}, \dots, g_{t-1,k}), \quad \forall k \in \{1, \dots, K\}. \tag{7}$$



**Fig. 2.** The prediction on the total number of customers' devices: (a) without and (b) with the group information (weather information not included). The X-axis is the UTC (Epoch time) format and the Y-axis represents the number of customers' devices.

After all, we also write down the sliding window of size  $L$  for the target value that we want to predict:

$$\mathbf{y}_{t,L} = (y_{t-L}, \dots, y_{t-2}, y_{t-1}). \quad (8)$$

In the end, the overall feature set in this study can be written as:

$$\begin{aligned} \mathbf{D}_{t,L}^g &= (\text{day}_t, \mathbf{temp}_{t,L}, \mathbf{rain}_{t,L}, \mathbf{y}_{t,L}; y_t), \\ \mathbf{D}_{t,L}^{\text{gp}} &= (\text{day}_t, \mathbf{temp}_{t,L}, \mathbf{rain}_{t,L}, \mathbf{y}_{t,L}, \mathbf{g}_{t,L,1}, \dots, \mathbf{g}_{t,L,K}; y_t), \end{aligned} \quad (9)$$

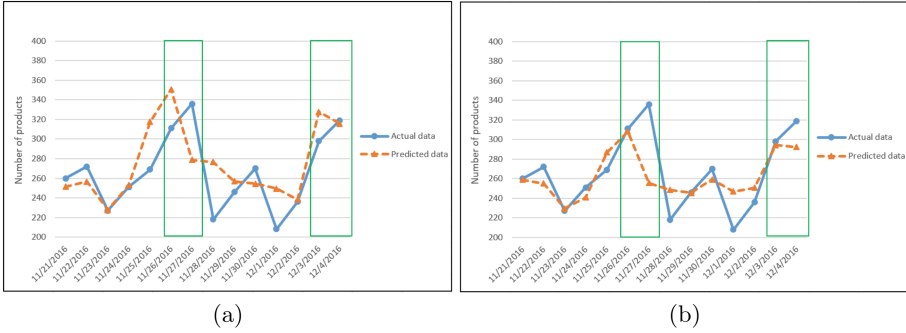
for the case without or with the group information included, respectively. The target value  $y_t$  that we want to predict could be the number of customers' devices, the number of sold products or revenue as described before. From time to time, we may have the feature set described in Eq. 9 too large to create the risk of overfitting. To avoid the situation, we reduce the dimensionality by shrinking the feature size of sliding window as follows. For each sliding window, e.g., the sliding window for temperature, we may choose a pre-defined function such as the mean function to compress a long sliding window to a scalar such as:

$$\text{temp}_t = (\tau_{t-L} + \dots + \tau_{t-1})/L. \quad (10)$$

Some other possible functions for shrinkage include minimization, maximization. Now the complete feature set is shrunk to:

$$\begin{aligned} \mathbf{d}_{t,L}^g &= (\text{day}_t, \text{temp}_{t,L}, \text{rain}_{t,L}, \mathbf{y}_{t,L}; y_t), \\ \mathbf{d}_{t,L}^{\text{gp}} &= (\text{day}_t, \text{temp}_{t,L}, \text{rain}_{t,L}, \mathbf{y}_{t,L}, \mathbf{g}_{t,L,1}, \dots, \mathbf{g}_{t,L,K}; y_t), \end{aligned} \quad (11)$$

for the cases of not including the group information or including the group information respectively. Note that we choose the mean function for the shrinkage on all the sliding windows except for the sliding window for the target value.



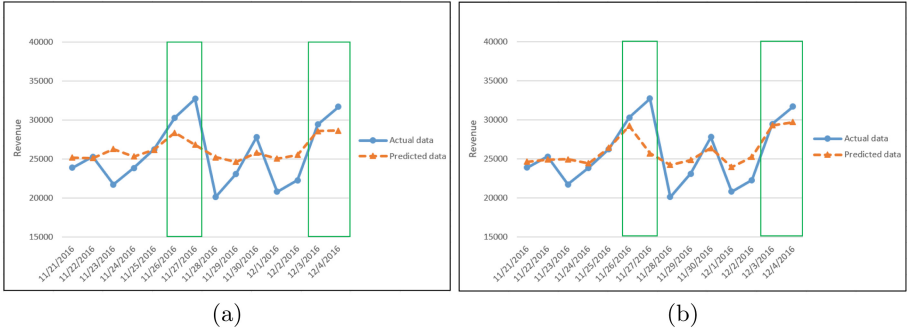
**Fig. 3.** The prediction on the total number of sold products: (a) without and (b) with the group information (no weather information included). The X-axis is the time and the Y-axis represents the number of products.

**Result.** The first experiment is to predict the number of customers’ devices given the features described in Eq. 11. In Fig. 2, we demonstrate the effectiveness of the proposed method by showing the difference between the actual and the predicted result on the total number of customers’ devices spanning two weeks<sup>3</sup>. First, we notice the ups and downs on the selling between different days where we usually have high selling and revenue during the weekends (Nov. 26, 27, and Dec. 3, 4). In fact, those are the days that we have significant gaps between the actual and predicted result. In (a), we have the prediction given the features without the group information and we include the group information for prediction in (b). Overall, we obtain an improvement from 9.11% to 7.63% in MAPE (Mean Absolute Percentage Error) from (a) to (b) if the group information is included and the weather information is not included (also in Table 2).

In the second experiment, we aim to predict the number of sold products. Other than the number of devices which may not be 100% identical to the number of customers, the amount of sold products could be a better quantity to reflect the business profit. In Fig. 3, we can compare between the actual number of sold products and the prediction. Again, we observe more selling during the weekends rather than during the weekdays. The weekend period is also the time that we have worse prediction if compared to the prediction on the weekdays.

To compare between the scenario when we include no group information and the scenario when we do include group information, we found out that including the group information can improve the prediction result from 15.06% to 11.32% in MAPE (without the weather information). It implies that including group information can help us understand more about customers’ behavior on visiting the coffee shop. Intuitively speaking, people often visit coffee shops with their partners. The decision about whether people visit a coffee shop or not may be highly influenced by their partners. On the other hand, the group information

<sup>3</sup> We add random numbers in the Y-axis for Figs. 2, 3 and 4 due to a concern from the coffee shop.



**Fig. 4.** The prediction on the total revenue: (a) w/o and (b) w. the group information (no weather information). The X-axis is the time and the Y-axis shows the revenue.

**Table 2.** The summary of all predictions in MAPE. The boldface numbers show the best performance in each group and the underline, boldface numbers show the best performance across all settings.

	w/o weather		w. weather	
	w/o group info.	w. group info.	w/o group info.	w. group info.
	MAPE (%)	MAPE (%)	MAPE (%)	MAPE (%)
# of devices	9.11	<b><u>7.63</u></b>	<b>7.95</b>	8.43
# of sold products	15.06	<b><u>11.32</u></b>	<b>11.60</b>	12.25
Revenue	18.10	<b><u>14.43</u></b>	14.58	<b>14.51</b>

may also imply a similar behavior on visiting the coffee shop such as the people in the same group may choose to visit the coffee shop on similar days or at similar moments. This kind of group information could reflect the vocations that the customers have or the living style they share. Knowing such information may give us more hints on predicting whether or not certain people visit the coffee shop on a particular day or at a particular moment.

In the end, we discuss the revenue prediction as described in Fig. 4. Again, we have similar result like the prediction on the number of devices and the prediction on the number of sold products. We have the prediction errors improved from 18.10% to 14.43% in MAPE when the weather is not included and from 14.58% to 14.51% in MAPE when the weather is included. Overall, we have the improvement when the group information is included in four out of six different settings, as shown in Table 2 given the settings such as without or with the weather information and for different prediction tasks. In the table, we also noticed the improvement from including the weather information in many of the occasions. Note that including both the weather information and group information may not produce the best result. We believe that too many features may harm the performance due to overfitting and the problem could be eased when more data are collected in the near future.

## 5 Conclusion

We proposed an easy-to-deployed, low cost and privacy-preserving method for business revenue forecasting based on WiFi sensing. A WiFi collection AP was installed in an indoor environment to collect related WiFi signals for us to understand more about customers who visit the business. The case study was done in a coffee shop where we analyzed the WiFi-based and non-WiFi-based data for 12 weeks for the evaluation. We worked on three prediction tasks such as the prediction on the number of devices, the number of sold products and the total revenue. In the experiment study, we found out the improvement when the weather information is included; more importantly, when the group information is included in most of the prediction tasks even with a limited data collection period. The prediction on the number of devices, the number of sold products and the revenue can reach 7.63%, 11.32%, and 14.43% in MAPE in their peak performance. A large scale data collection and study is on the way for more extensive study in the near future.

## References

1. Draghici, A., Steen, M.V.: A survey of techniques for automatically sensing the behavior of a crowd. *ACM Comput. Surv.* **51**(1), 21:1–21:40 (2018)
2. Du, H., Yu, Z., Yi, F., Wang, Z., Han, Q., Guo, B.: Recognition of group mobility level and group structure with mobile devices. *IEEE Trans. Mob. Comput.* **17**(4), 884–897 (2018)
3. Han, J., Ding, H., Qian, C., Xi, W., Wang, Z., Jiang, Z., Shangguan, L., Zhao, J.: CBID: a customer behavior identification system using passive tags. *IEEE/ACM Trans. Netw.* **24**(5), 2885–2898 (2016)
4. Lau, E.E.L., Lee, B.G., Lee, S.C., Chung, W.Y.: Enhanced rssi-based high accuracy real-time user location tracking system for indoor and outdoor environments. *Int. J. Smart Sens. Intell. Syst.* **1**(2), 534–548 (2008)
5. Loewenstein, Y., Portugaly, E., Fromer, M., Linial, M.: Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space. *Bioinformatics* **24**(13), i41–i49 (2008)
6. Maduskar, D., Tapaswi, S.: RSSI based adaptive indoor location tracker. *Sci. Phone Apps Mob. Devices* **3**(1), 3 (2017)
7. Nguyen, K.A.: A performance guaranteed indoor positioning system using conformal prediction and the WiFi signal strength. *J. Inf. Telecommun.* **1**(1), 41–65 (2017)
8. del Rosario, M.B., Redmond, S.J., Lovell, N.H.: Tracking the evolution of smart-phone sensing for monitoring human movement. *Sensors* **15**(8), 18901–18933 (2015)
9. Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222 (2004)
10. Subhan, F., Ahmed, S., Ashraf, K., Imran, M.: Extended gradient RSSI predictor and filter for signal prediction and filtering in communication holes. *Wirel. Pers. Commun.* **83**(1), 297–314 (2015)
11. Vanderhulst, G., Mashhadi, A.J., Dashti, M., Kawsar, F.: Detecting human encounters from WiFi radio signals. In: *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, Linz, Austria, 30 November–2 December 2015, pp. 97–108 (2015)

12. Zeng, Y., Pathak, P.H., Mohapatra, P.: WiWho: WiFi-based person identification in smart spaces. In: 2016 15th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN), pp. 1–12, April 2016
13. Zeng, Y., Pathak, P.H., Mohapatra, P.: Analyzing shopper’s behavior through WiFi signals. In: Proceedings of the 2nd Workshop on Workshop on Physical Analytics, WPA 2015, pp. 13–18. ACM, New York (2015)
14. Zhang, D., Xia, F., Yang, Z., Yao, L., Zhao, W.: Localization technologies for indoor human tracking. In: 2010 5th International Conference on Future Information Technology, pp. 1–6, May 2010