SD-HOC: Seasonal Decomposition Algorithm for Mining Lagged Time Series

Irvan B. Arief-Ang(✉), Flora D. Salim, and Margaret Hamilton

Computer Science and Information Technology, School of Science,
Royal Melbourne Institute of Technology (RMIT), Melbourne, VIC 3000, Australia
{irvan.ariefang,flora.salim,margaret.hamilton}@rmit.edu.au

Abstract. Mining time series data is a difficult process due to the lag factor and different time of data arrival. In this paper, we present Seasonal Decomposition for Human Occupancy Counting (SD-HOC), a customised feature transformation decomposition, novel way to estimate the number of people within a closed space using only a single carbon dioxide sensor. SD-HOC integrates time lag and line of best fit model in the preprocessing algorithms. SD-HOC utilises seasonal-trend decomposition with moving average to transform the preprocessed data and for each trend, seasonal and irregular component, different regression algorithms are modelled to predict each respective human occupancy component value. Utilising M5 method linear regression for trend and irregular component and dynamic time warping for seasonal component, a set of the prediction value for each component was obtained. Zero pattern adjustment model is infused to increase the accuracy and finally, additive decomposition is used to reconstruct the prediction value. The accuracy results are compared with other data mining algorithms such as decision tree, multi-layer perceptron, Gaussian processes - radial basis function, support vector machine, random forest, naïve Bayes and support vector regression in two different locations that have different contexts.

Keywords: Ambient sensing · Building occupancy
Presence detection · Number estimation · Cross-space modeling
Contextual information · Human occupancy detection
Carbon dioxide · Machine learning

1 Introduction

Data mining technology is assimilated in human life and it helps solve many problems that could not be solved before. The problem we will consider in this paper is to do with building operational costs. From the U.S. Department of Energy, 35%–45% of the total operational costs within a building are spent on heating, ventilation, and air conditioning (HVAC) [1]. Due to this, substantial investment in the energy usage research area is needed to reduce HVAC costs in buildings based on their occupancy patterns. Reducing HVAC usage is

Y. L. Boo et al. (Eds.): AusDM 2017, CCIS 845, pp. 125–143, 2018.
https://doi.org/10.1007/978-981-13-0292-3_8

equivalent to reducing the overall energy consumption. Furthermore, a Building Management System (BMS) can then intelligently adjust the HVAC based on the occupancy pattern so the comfort of the dwellers is not sacrificed.

Using sensor data to detect human's precence is the current trend in ambient sensing research area [2–6]. Yan higlighted the importance of occupant related research [7]. In [8], it was highlighted that carbon dioxide (CO_2) is the best ambient sensor predictor for detecting human presence. By using only CO_2, 91% accuracy was achieved for binary prediction, knowing the room is occupied or vacant [9] and have 15% accuracy for recognising the number of occupants. A hidden Markov model (HMM) was implemented for CO_2 dataset to predict human occupancy and 65%–80% range of accuracy was achieved for predicting up to 4 occupants [10].

In this paper, we propose a new algorithm for decomposing large datasets to extract the relevant features to be used for prediction and identification of seasonal trends. We can then apply the computations of these trends to various incomplete sets matching the time series to predict the relevant future features in the new dataset. We identify relevant seasonal trends in the data over time and apply these to the new dataset and use them to predict future trends in the data.

We have found this to be particularly useful for sensor data where we can extrapolate the CO_2 data to indoor human occupancy prediction with promising accuracy. We can match the sensor measurements for zero occupancy, at various times, possibly overnight and tune our predictions to optimise individual comfort and the overall carbon footprint of the building.

We apply our new feature transformation algorithm to the prediction of the number of people in a room at a particular time through the measurement of the carbon dioxide. Human occupancy prediction is an significant problem for the building industry because it enables the automation of heating, cooling and lighting systems. If it is known that certain rooms are empty or underutilised during certain times, operational costs and carbon footprint can be reduced with better planning and scheduling. When the rooms are not occupied, the building system can also adjust these facilities to keep the inhabitants comfortable. This framework is called seasonal decomposition for human occupancy counting (SD-HOC).

SD-HOC pre-processes the data and integrates various machine learning algorithms. The experiment is conducted on two different locations. There are two stages we have defined in our experiment. Firstly, SD-HOC result is compared with a variety of other data mining prediction algorithms such as decision tree, multi-layer perceptron, Gaussian processes - radial basis function, support vector machine and random forest. The second stage of our experiment compares SD-HOC with one of the best data mining prediction accuracy to predict the human occupancy number on different number of prediction days. There are three advantages of utilising SD-HOC:

1. It employs low equipment cost due to pre-installation;
2. SD-HOC ensures that users' privacy is protected;
3. It only uses CO_2 data, reducing the chance of errors caused by data integration.

The remainder of the paper is organised as follows. Section 2 presents the related work on current state-of-the-art indoor human occupancy methods. Section 3 covers the problem definition. Section 4 covers the features and Sect. 5 introduces SD-HOC framework. Section 6 describes experiments we conducted concerning set-up machine and multiple datasets. It also contains the results and comparisons with other data mining algorithms. Section 7 discusses the results and Sect. 8 concludes the paper with directions to the future work.

2 Background and Related Work

When using image processing techniques [11,12], the levels of accuracy for human occupancy detection can reach up to 80%. Unfortunately, these image processing methods raise privacy concerns. Research communities have been doing their best to propose various methods to detect human occupancy without using cameras or image processing.

We focus on utilising only CO_2 sensors alone and data to estimate the indoor human occupancy number. The main reason is because CO_2 sensors are already integrated with the BMS and ventilation infrastructure and are commonly installed in buildings.

Machine learning algorithms including a hidden markov model (HMM), neural networks (NN) and support vector machine latent (SVM latent) were implemented in [10] by using CO_2 data with the sensors deployed both inside and outside room. By feature engineering CO_2 data with first order and second order difference of CO_2, the accuracy achieved is between 65%–80%.

A mass balance approach was implemented to predict both human occupancy and occupant activity using CO_2 and door sensors in [13]. The authors mentioned that sources of error and uncertainty in this method are part of the limitation of this approach. CO_2 based occupancy detection in office and residential buildings was implemented in [14]. Binary occupancy accuracy prediction is 95.8% and the people counting accuracy is 80.6% for 2–3 person in each room.

PerCCS is a model with a non-negative matrix factorization method to count people [9] using only one predictor in CO_2. In predicting vacant occupancy, they achieved up to 91% but only 15% accuracy in predicting the number of occupants.

Overall, sensor-based detections have higher accuracy compared to radio-based detections. For example, Wi-Fi and RSSI signals achieved 63% accuracy for indoor detection with 9 occupants [15]. For occupancy counting, CO_2 sensors only have been experimented with the maximum of 42 occupants and accuracy limit of 15% [9]. A domain adaptation technique has been implemented for human occupancy counting with CO_2 and the prediction accuracy increases up to 12.29% compared to the baseline [16].

3 Problem Definition

Given significant motivations in our research, this paper presents the problem on how can we use data mining techniques and feature selection to predict the number of people by using a single CO_2 sensor. We would take the results to have similar accuracy to the state-of-the-art techniques in the occupancy detection field. In Fig. 1, the data shows there is a dependency between CO_2 and occupancy data.

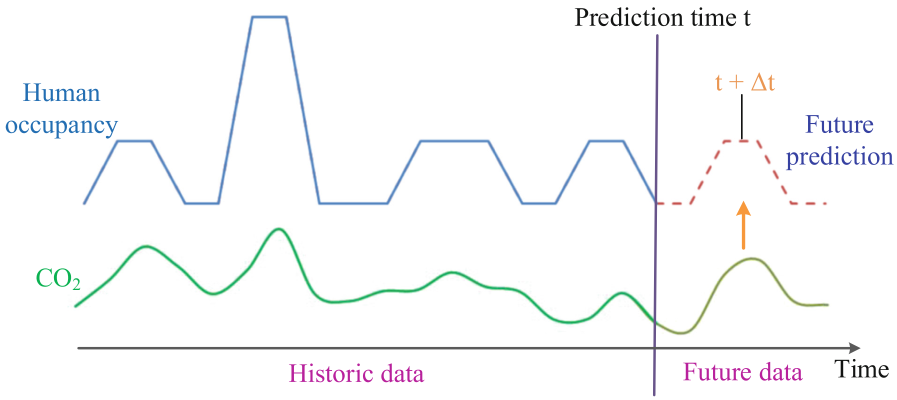

Fig. 1. Real-time prediction scenario for continuous t showing the amount of CO_2 fluctuations. The fundamental task is to predict the number of occupants at time $t + \Delta t$.

3.1 Scenario Assumption

Assume $|TS|$ represents the length of a time series, $TS = \{ts_1, ts_2, \ldots, ts_{\mathrm{q}}\}$, where q means the number of sample points. In our time series datasets, we have two aspects:

- Carbon dioxide (CO_2) concentration C, defined as $C = \{C_1, C_2, \ldots, C_q\}$
- Indoor human occupancy O, defined as $O = \{O_1, O_2, \ldots, O_q\}$

Our framework only depends on CO_2 data set to calculate the prediction. This is where the challenge lies as the model needs to extract more features from a time series, which may seem simple, but it contains hidden trends. We introduce a term ‘lagged time series’ as a set of data in regression time series where each value relate to a situation in a surrounding context but does belong to different time frame.

3.2 Time Series

In time series prediction, analysing one-step-ahead prediction is different from analysing multi-step-ahead prediction. Predicting multi-step-ahead needs a more complex method due to the accumulation of errors and the number of uncertainties increasing with time. We focus on multi-step-ahead prediction with the support of one dependent variable to reduce uncertainties.

We have two different types of datasets: CO_2 concentration C and indoor human occupancy O. In order to explore the relationship between both factors above, we need to identify the relevant features by exploring the correlation between CO_2 concentration and indoor human occupancy and all of their decomposed components to find what correlations exist between each component.

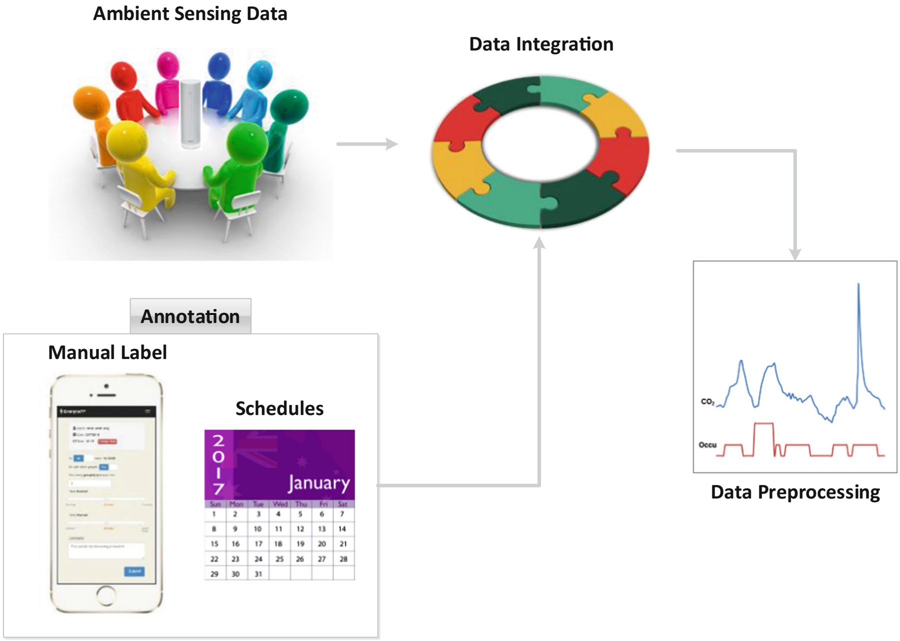

Fig. 2. Data collection and analysis framework.

4 The Features

This section explains our data pre-processing time series components, cross-correlation and line of best fit. Data pre-processing is crucial for our model as this step will further increase prediction analysis with various machine learning algorithms that we implemented in the experiment section. We collected data about both the CO_2 concentration from the sensor data and the number of humans in the room as shown in Fig. 2. Both data are pre-processed and integrated using our novel pre-processing method described below in the Subsect. 4.1. We transformed each data set using feature engineering into more features and applied our prediction model, SD-HOC, described in Sect. 5 to predict the indoor human occupancy.

4.1 Time Delay Components

Time delay issue is a problem because it takes time for the concentration of CO_2 to build up enough to measure a person. To model a real time delay we need the value of a time series regression function obtained after specific time lag. This issue normally happens in the majority of sensor data analyses as data obtained from sensor readers needs to travel to a sensor reader before it can be captured in storage. In our study, when one person enters a room, it will take some time before the CO_2 level in the air increases proportionally. Due to this reason, we must pre-process the data to fix a time delay between CO_2 data and the indoor human occupancy number.

4.2 Cross-Correlation and the Line of Best Fit

Before analysing the data between CO_2 and the number of occupants, the data lagging issue needs to be considered. Data lagging means that it will take a certain time for CO_2 to populate the room as there is delay between the time of people exiting (or entering) the room and the decrement (or increment) of the CO_2 value on the air. To find out how much data lagging need to be implemented, first we need to find upper bound value (UB). UB is a maximum value calculated based on the room volume. UB will be used to calculate the time lag value and is defined by the formula in Eq. 1.

$$UB = |(RL * RW * RH)/C| \tag{1}$$

UB	upper bound value
RL	room length
RW	room width
RH	room height
C	constant value (100)

For each dataset from 0 min time lag to UB minutes time lag, the correlation of CO_2 data with the number of occupancies is measured. If the room size is small, the UB value will be 1. The larger is room is, the bigger the UB value is. In our case study, for the small room A, the UB value will be 1 and for big room B, the upper bound of N is 60. This value is aligned with the explanation above due to the large size of the big room B.

To calculate a line of best fit, we need to calculate the slope value between CO_2 and occupancy data, defined by Eq. 2.

$$SL = \frac{\sum(O_t - \bar{O_t})(C_t - \bar{C_t})}{\sum(O_t - \bar{O_t})^2} \tag{2}$$

SL	slope of the linear regression line
O_t	occupancy value
$\bar{O_t}$	sample means of the known occupancy value
C_t	CO_2 value
$\bar{C_t}$	sample means of the known CO_2 value

Next, the intercept value between both data sets needs to be calculated using the formula in Eq. 3.

$$IC = \bar{C}_t - SL * \bar{O}_t \tag{3}$$

IC	intercept of the linear regression line
$\bar{C}_t$	sample means of the known CO_2 value
$\bar{O}_t$	sample means of the known occupancy value

The main formula for the line of best fit (LBF) is shown in Eq. 4.

$$LBF = (O_t - (SL * C_t + IC))^2 \tag{4}$$

O_t	occupancy value
SL	slope of the linear regression line
C_t	CO_2 value
IC	intercept of the linear regression line

4.3 Time Lag

For each line of best fit from Subsect. 4.2, we calculate the mean squared error (MSE), root-mean-square deviation (RMSD) and the normalised root mean squared error (NRMSE). The formula for calculating NRMSE is shown in Eq. 5.

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(C_t - \bar{C}_i)^2}}{O_{max} - O_{min}} \tag{5}$$

$NRMSE$	normalized root mean square error
t	total number of data set
C_t	CO_2 value
$\bar{C}_t$	sample means of the known CO_2 value
O_{max}	maximum occupancy value
O_{min}	minimum occupancy value

This step is repeated UB times for each time lag. For time lag analysis, we use least square regression to compare each NRMSE from time lag 0 until time lag UB. We pick the lowest number of NRMSE value as our time lag value (TL). The TL value formula is shown in Eq. 6 and it performs as our baseline time lag for the data analysis.

$$TL = min(NRMSE) \tag{6}$$

For the academic staff room, the TL value is 0. This value represents no time lag is needed for this analysis. For the cinema theatre, the lowest number of error value happens at time lag TL = 32 as shown in Fig. 3. This TL value is our base for the cinema theatre data analysis. So for the entire cinema data analysis process, we use time lag 32.

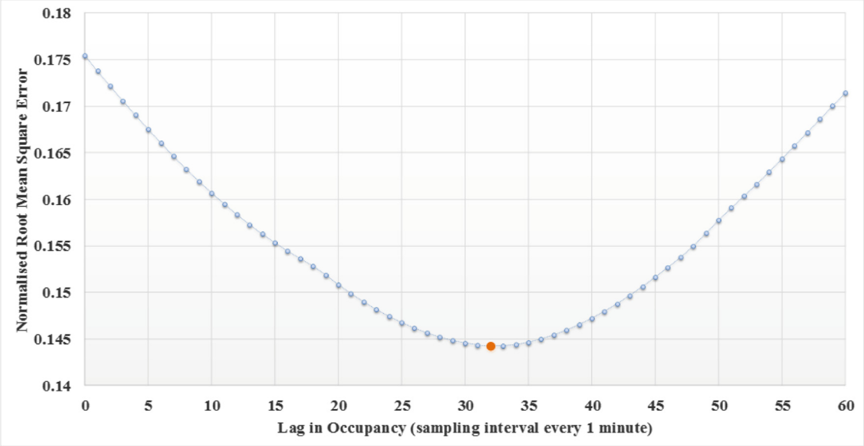

Fig. 3. Ordinary Least Square Regression Normalised Root Mean Squared Error (NRMSE) between CO_2 data and actual occupancy for 60 min time lag.

5 The Framework

There is no linear relationship between CO_2 and indoor human occupancy. For this reason, we introduce a new SD-HOC analysis framework in Fig. 2 to address this non-linear correlation issue by decomposing both CO_2 and occupancy data shown in Fig. 4. In this paper, for the main decomposition method we are using is known as seasonal trend decomposition (STD).

The core feature transformation prediction model will be explained in the next following subsections. The first subsection discusses STD in detail. The next subsection explains the correlation model for trend, seasonal and irregular features. The last subsection presents zero pattern adjustment (ZPA), a new method for analysing conditions when the room is vacant. ZPA method can increase the overall accuracy. This model needs to be re-trained for different locations to obtain the most optimal accuracy results.

5.1 Seasonal-Trend Decomposition

STD is a decomposition technique in time series analysis. X-11 method with moving average is one of the most famous variants [17] and X12-ARIMA is the most recent variant [18]. STD is an integral part of our framework.

To understand each time series data, we utilise STD to decompose the data into four main features: trend, cyclical, seasonal and irregular. The trend feature (T_t) represents the long-term progression of the time series during its secular variation. The cyclical feature (C_t) reflects a repeated but non-periodic fluctuation during a long period of time. The seasonal feature (S_t) is a systematic and regularly repeated event during short period of time. And the irregular feature (e_t also known as error or residual) is a short term fluctuation from the time series and is the reminder after the trend, cyclical and season features have been removed.

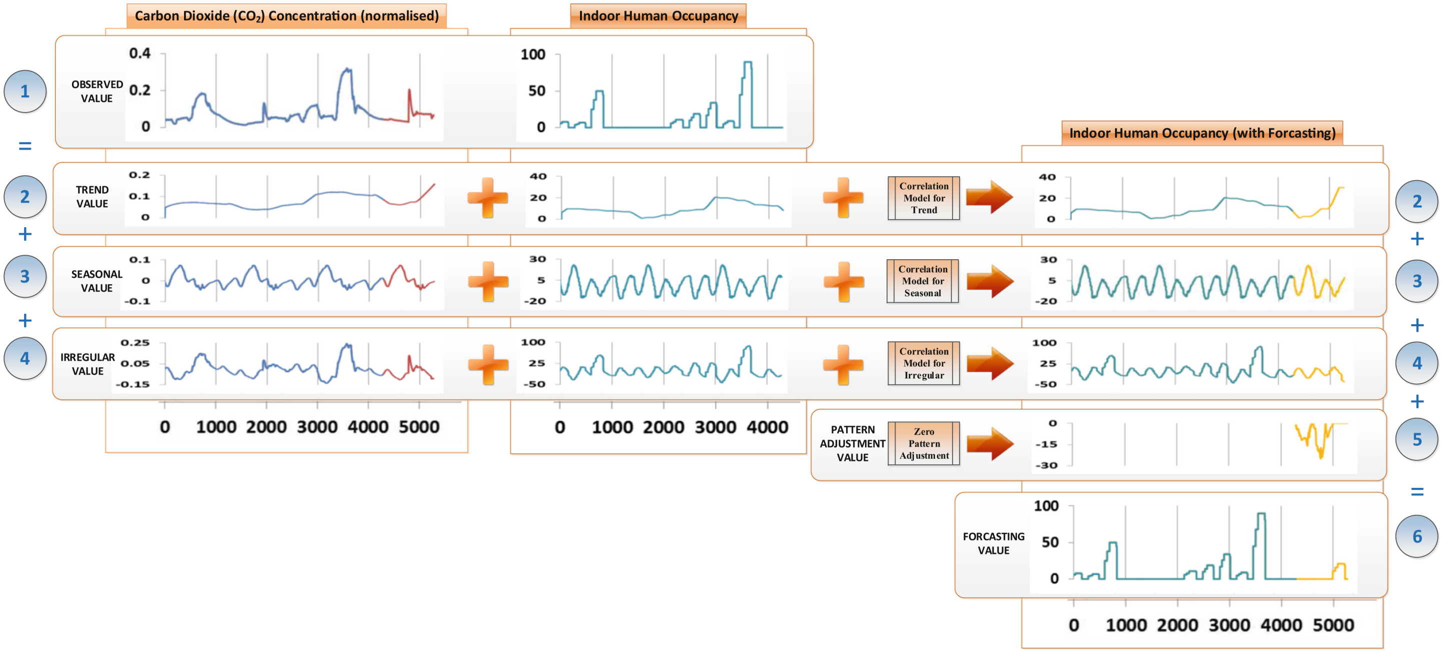

Fig. 4. Seasonal decomposition for human occupancy counting (SD-HOC) analysis framework.

In this paper, we decide to combine the cyclical feature into trend feature due to its similarity to make the model simpler without sacrificing the accuracy.

Below is the core logic for seasonal trend decomposition:

1. Calculate 2×12 moving average in the raw data (both CO_2 and occupancy datasets) to obtain a rough trend feature data T_t for all period (12 is the default due to there are 12 months in a year).
2. Calculate ratios of the data to trend, named "centred ratios" (y_t/T_t).
3. To form a rough seasonal feature (S_t) data estimation, apply separate 2×2 moving average to each month of the centred ratios.
4. To obtain the irregular feature (e_t), divide the centred ratios by S_t.
5. Multiply modified e_t by S_t to get modified centred ratios.
6. Repeat step 3 to obtain revised S_t.
7. Divide the raw data by the new estimate of S_t to give the preliminary seasonal adjusted series, y_t/S_t.
8. The trend feature (T_t) is estimated by applying a weighted Henderson moving average [19] to the preliminary seasonally adjusted values.
9. Repeat step 2 to get new ratios by dividing the raw data by the new estimate of T_t.
10. Repeat Steps 3 to 5 using the new ratios and applying a 3×5 moving average instead of a 3×3 moving average.
11. Repeat step 6 but using 3×5 moving average instead of a 3×3 moving average.
12. Repeat step 7.
13. Finally the reminder feature is obtained by dividing the seasonally adjusted data from step 12 by the trend feature obtained in step 8.

Our customised STD formulation is:

$$STD_t = f(T_t, S_t, e_t) \tag{7}$$

t	time
STD_t	actual value of a time series at time t
T_t	trend feature at t
S_t	seasonal feature at t
e_t	irregular feature at t

In this paper, we decided to use additive decomposition. Additive decomposition is chosen because is the simplest to give the first approximation. Our overall STD formula becomes:

$$STD_t = T_t + S_t + e_t \tag{8}$$

This general STD formula will be applied to both CO_2 time series dataset and human occupancy time series datasets:

$$C_t = T_t^C + S_t^C + e_t^C \tag{9}$$

$$O_t = T_t^O + S_t^O + e_t^O \tag{10}$$

To predict O_{t+1} up to O_{t+n}, we need to create a model to systematically predict each of T_{t+1}^O, S_{t+1}^O and e_{t+1}^O up to T_{t+n}^O, S_{t+n}^O and e_{t+n}^O and then reconstruct the new prediction dataset using additive method.

5.2 Correlation Models

There are three correlation models for features of trend, seasonal and irregular in the following subsections.

Correlation Model for Trend Feature ($\boldsymbol{T_t}$). The definition for the trend feature (T_t) is the long-term non-periodic progression of the time series during its secular variation. Due to this, we assume that the trend feature for the CO_2 dataset (T_t^C) will be similar to the trend feature for indoor human occupancy (T_t^O) because there is dependency between both dataset.

Correlation model for trend feature start with checking the similarity between both trend features. We use Pearson product-moment Correlation Coefficient (PCC) to validate it as shown below:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{11}$$

$$\begin{array}{l} r \text{ correlation coefficient} \\ x \text{ dataset x} \\ y \text{ dataset y} \\ n \text{ number of sample points} \end{array} \tag{12}$$

Correlation coefficient Pearson's r value ranges from -1 to $+1$. If the value is >0.7, the correlation between both datasets is strongly positive. If the correlation is less than 0.7, data pre-processing needs to be redone to find the new TL value (Eq. 6).

Once it passes the validation step, polynomial M5 linear regression is implemented. We chose the M5 method because it will build trees whose leaves are associated with multivariate linear models and the nodes of the tree are chosen over attributes that maximise the expected error reduction, given by the Akaike Information Criterion (AIC). AIC is a measure to check the relative goodness of fit of a statistical model [20]. The purpose of using AIC is to evaluate the model. The value for each of trend feature needs to be a positive value so we put the absolute value on both the CO_2 ($|T_t^C|$) and human occupancy trend features ($|T_t^O|$). The main formula for trend feature correlation is shown below:

$$|T_t^O| = |\alpha_0 + \alpha_1(T_t^C) + \alpha_2(T_t^C)^2 + \ldots + \alpha_n(T_t^C)^n + \epsilon| \tag{13}$$

Linear regression with M5 will output each α_n and ϵ value. With these parameters, the future trend for T_{t+n}^O can be obtained.

Correlation Model for Seasonal Feature ($\boldsymbol{S_t}$). The seasonal feature (S_t) is a systematic and regularly repeated event during short period of time. Due to this characteristic, every seasonal feature can be fitted by a finite Fourier series. To correlate S_t^C and S_t^O, we use Dynamic Time Warping (DTW), a pattern matching technique to score the similarity between the shape of specific signal within certain duration [21]. The full correlation algorithm is implemented in Algorithm 1 to find regularly repeated events within each S_t.

Algorithm 1. Finding a repeated event inside seasonal feature

```
procedure REPEATED_EVENT($S_t$)
    $s_t^{temp}, s_t^{fin} \subset S_t$
    $len \leftarrow 0$                                          ▷ $len$: Length for $s_t^{temp}$
    $a \leftarrow S_t[len]$                                     ▷ $a$: Start Point
    for each node $i \in S_t$ do
        $len$++
        $s_t^{temp} \leftarrow s_t^{temp} + S_t[i]$
        if $a = S_t[i]$ then
            if DTW($s_t^{temp}$,$S_t[i+1..i+len]$) $> 95$ then
                $s_t^{fin} \leftarrow s_t^{temp}$
                break
            end if
        end if
    end for
    return $s_t^{fin}$
end procedure
```

Once we find repeated event in s_t^{fin} for both the CO_2 and occupancy seasonal features, we compare the length of $s_t^{fin(O)}$ and $s_t^{fin(C)}$. If the length of $s_t^{fin(O)} < s_t^{fin(C)}$, we apply an interpolation method inside $s_t^{fin(O)}$ so both have the same length. If the length of $s_t^{fin(O)} > s_t^{fin(C)}$, we apply data reduction method so finally both have the same length. The final regression equation for seasonal feature correlation is shown below:

Correlation Model for Irregular Feature (e_t). Due to similar characteristics between trend and irregular features, we apply the same correlation method from the trend feature:

$$|e_t^O| = |\beta_0 + \beta_1(e_t^C) + \beta_2(e_t^C)^2 + \ldots + \beta_n(e_t^C)^n + \gamma| \tag{14}$$

The only difference from the trend feature is that we do not need to validate it using PCC as the shape of the irregular feature will depend more on its trend and seasonal features.

5.3 Zero Pattern Adjustment

In human occupancy prediction research, inferring knowledge when a room is vacant is paramount. By minimising false positives, the accuracy prediction can be improved. The Zero pattern adjustment (ZPA) method learns the behaviour from previous historical data and makes some smart adjustments for a vacant room when the normal algorithm returns incorrect prediction. The ZPA technique overlays all previous dataset and puts them on a single 24-h x-axis chart to determine the earliest start and end points when the room is vacant each day during the night to dawn period. We symbolise ZPA as zpa_t^O.

For our main occupancy model, we integrate each feature to get the occupancy prediction value.

6 Experiments and Results

In this section, our model is assessed for two different locations with distinct contexts to ensure the model's adaptability to various conditions. The first location is a small room A, belonging to one staff member at RMIT University, Australia. This room is chosen for human occupancy prediction since a controlled experiment can be conducted for an extended period of data collection.

The second dataset was collected inside a cinema theatre in Mainz, Germany [22]. Cinema theatre is chosen as another setting due its nature of having fluctuating numbers of people throughout the day. The numbers of people in the audiences can reach hundreds and can decrease to zero within a few hours. We will address this room as big room B.

6.1 Experiment Setting

Small Room A. We use a commercial off-the-shelf Netatmo urban weather station(Range: 0–5000 ppm, accuracy: ±50 ppm) to read and collect ambient CO_2 data. The experiment took place between May and June 2015. The dataset is uploaded to a cloud service for integration purposes. We selected two weeks data from the whole dataset and used them in the further analysis. The room size is 3×4 m.

Big Room B. The cinema dataset were collected between December 2013 and January 2014 [22]. The dataset was collected using mass spectrometry machinery installed on the air ventilation system. The air flows from the screening room via the ventilation system to the mass spectrometer for data analysis.

Experiment Tool. We utilised WEKA, MATLAB and R to help us perform this experiment. WEKA is used for polynomial linear regression with M5 method for both correlation models for trend and irregular features (Subsect. 5.2). We also used WEKA for majority of data mining algorithms such as multi-layer perceptron, Gaussian processes (with kernel RBF), support vector machine, random forest, naïve Bayes, decision tree (with random tree) and decision tree (with M5P). MATLAB code is run for the baseline method, SVR and its prediction result. We used R to integrate all the data, including decomposition of STD and the majority of data pre-processing.

6.2 Experiment Parameters

SD-HOC model predicts each future value for the whole period of time based on specific time window. To understand this model better and how well it performs compared with the baseline, we define x, accuracy error tolerance parameter. Zero units error tolerance means only the exact number recognised is considered as true positive. For example with ten units error tolerance, if the real indoor human occupancy is 150 people, the prediction shows as low as 140 or as high as

160 is considered correct as it is within ±10 units error tolerance. The parameter x value will be different based on the size of the room.

Each machine learning algorithms data has been preprocessed using the same method as in Sect. 4 to ensure that the comparison is fair.

Experiment for Small Room a Dataset. For an academic staff room dataset, we used 5-min time window. Total data that we gathered from this room are 4,019 data spread in 14 days. Due to the small room size, we decided not to use time lag for data analysis as there is a negligible period between exhaling process and sensor reading. For this room, we have seven pairs of the training-test dataset. It starts with seven days of training dataset and seven days of test dataset. It ends with 13 days of training dataset to predict one-day test dataset.

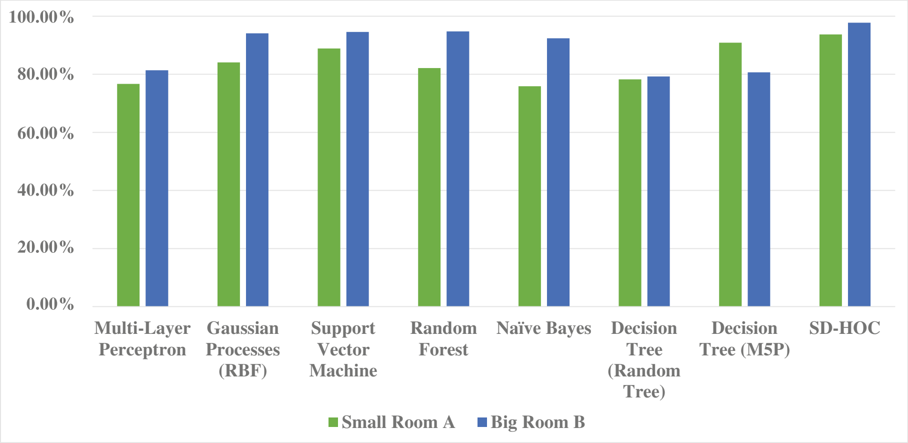

Fig. 5. Accuracy result of various machine learning algorithms.

Experiment for Big Room B Dataset. For cinema theatre dataset, we use 3-min time window for data analysis. Data that we gathered from this cinema theatre consisting of 68,640 instances spread over 23 days. The cinema theatre capacity is up to 300 people and for this experiment, we run the line of best fit for time lag 0 to time lag 60. The lowest NRMSE is at time lag 32 and we use time lag 32 as time lag baseline. This time lag is appropriate as bigger room need larger time lag for the model to have a better accuracy. For this room, we decided to use December 2013 data for training and January 2014 data for testing. Then we replicated it in the similar method by giving one day from testing dataset to training dataset and ran the model again. This method is repeated until the test dataset only consisted of one day of data.

6.3 Experiment Results with Other Data Mining Algorithms

From Table 1 and Fig. 5, we run each data mining algorithms and compare the result with our novel model, SD-HOC. SD-HOC have the highest accuracy prediction with 93.71% accuracy for staff room and 97.73% for cinema theatre.

Table 1. Accuracy result of various machine learning algorithms.

Machine learning	Small Room A	Big Room B
Multi-layer perceptron	76.69%	81.39%
Gaussian processes (RBF)	84.09%	94.09%
Support vector machine	88.86%	94.55%
Random forest	82.16%	94.75%
Naïve Bayes	75.89%	92.40%
Decision tree (Random tree)	78.23%	79.26%
Decision tree (M5P)	90.87%	80.68%
SD-HOC	**93.71%**	**97.73%**

6.4 Experiment Result with SVR on Different Number of Prediction Days

Support vector regression (SVR) have the highest prediction accuracy compared to the other data mining algorithms. Due to this reason, we run the experiment with the different training and testing to compare SD-HOC and the state-of-the-art machine learning algorithm baseline, SVR.

Evaluation and Baseline. To evaluate the result, we divide the data into 2 equal parts. The first part is the training dataset and the second one is the test dataset. To be able to understand how well the model fits for a longer duration, we repeat the division of training and test dataset by adding one day data from the test dataset to the training dataset. This replication is repeated again until the test dataset has only one single day and the rest belong to training dataset. This incremental days of training and reduction in testing evaluation method ensure the robustness of model.

Experiment Result for Small Room A Dataset. From Fig. 6, SD-HOC performed better than the baseline on average by 4.33%. As the last two days are Saturday and Sunday, both SD-HOC and the baseline models correctly predict zero occupancy for each day. In fact, they are vacant for the whole day.

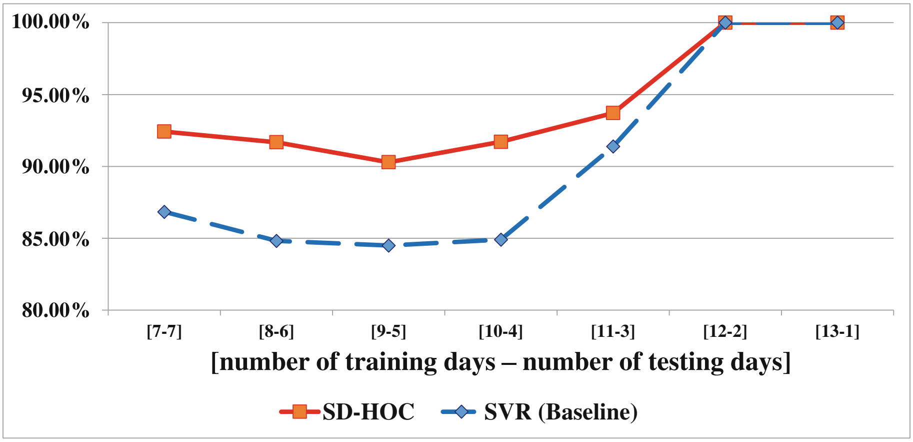

Fig. 6. Small room A dataset - The comparison for indoor human occupancy.

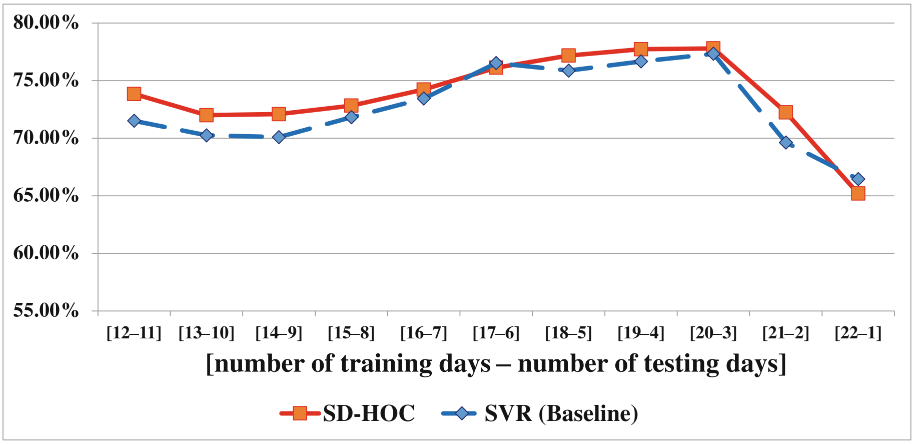

Fig. 7. Big room B dataset - The comparison for indoor human occupancy.

Experiment Result for Big Room B Dataset. For the cinema dataset, the comparison accuracy result is shown in Fig. 7. SD-HOC method performed better than the baseline method and on average SD-HOC method has 8.5% higher accuracy in predicting indoor human occupancy. The highest prediction accuracy was found when we used 22 days data for training to predict the number of human occupants the next day.

The results from Fig. 7 show that SD-HOC method is more accurate in predicting indoor human occupancy. This result is encouraging. Furthermore, we can observe that the accuracy for less number of days prediction is higher than for more days prediction, which is aligned with the results from academic staff room experiment.

7 Discussion

Our experiment shows that our new framework has the highest accuracy than most of data mining algorithms for both small and large rooms as shown in Fig. 5. Furthermore, this SD-HOC model is robust enough to handle different scales of data, proven by performing evaluation of our proposed model in two environments with different contexts such as room size and maximum number of occupants.

Compared with the baseline, SVR, our framework shows a better prediction accuracy over differing numbers of days for both the training and testing periods. This result demonstrates that seasonal decomposition can be utilised for predicting indoor human occupancy. The SD-HOC model can be used in many applications and is not limited to human occupancy prediction as it is based on seasonal decomposition methods.

SD-HOC performs well in comparison to other machine learning algorithms due to the feature transformation step, where, for each transformed feature, a set of relevant algorithms is run. For the small room A, SVM and decision tree method are the next best in prediction accuracy after SD-HOC. It is due to the fact that the number of people in this room fluctuated less from hour to an hour. There was also a stable CO_2 concentration for an extended period.

For the big room B, SVR and random forest are the next best in prediction accuracy after SD-HOC. Random forest behaves well when irrelevant features are present or these features have skewed distributions. The number of people in the big room could fluctuate from zero, or a vacant room to hundreds of people within 10–15 min. The SVM technique enables accurate discrete categorical labels to be predicted. This is why SVR is chosen as a baseline in Sect. 6.4.

8 Conclusion, Limitation and Future Work

Data mining algorithm roles in human life are becoming more important and its technology can be assimilated in human daily life. SD-HOC utilises several data mining algorithms and contributes to building and room occupancy counting. By understanding and knowing the numbers of people within a building, the heating, cooling, lighting control, building energy consumption, emergency evacuation, security monitoring and room utilisation can be made more efficient.

Although research in the human occupancy area has been studied with various methods including the use of ambient sensors, occupancy models that have been studied in previous work require the use of many sensors. In this experiment, we use a single sensor that is commonly available in the BMS to reduce the cost and complexity as more sensors can mean less reliability.

There are many possibilities that can be explored by using this technique. SD-HOC can be used for any time series dataset to predict another time series dataset as long as there is some dependency between those two data sets. SD-HOC is more than a simple correlation model and can solve many problems that a simple correlation model will not be able to solve.

CO_2 that is generated by human beings is affected by levels of physical activity. These different levels of activity such as walking, standing, or sitting could produce distinct CO_2 concentrations for the same individual. Also, the CO_2 rate in nature fluctuates around the day, reaching a higher value during the noon and dipping to a lower value at midnight. These CO_2 related facts could be integrated into the future works.

As our research was focussed on two locations and datasets, we plan to extend this research to other places that have different environmental dynamics and characteristics. For future work, other decomposition models and real-time online learning can be pursued to enhance the performance. Furthermore, from our research, indoor human occupancy could be related to certain events like public holidays and hence this feature could be included in future studies.

Acknowledgements. The authors would like to thank Joerg Wicker from University of Mainz for providing the cinema dataset used in this paper. This research is supported by the Australian Government Research Training Program Scholarship and two RMIT and Siemens Sustainable Urban Precinct Project (SUPP) grants: “iCo2mmunity: Personal and Community Monitoring for University-wide Engagement towards Greener, Healthier, and more Productive Living” and “The Greener Office and Classroom”.

References

1. U.S. Department of Energy (DOE): Building Energy Databook. Technical report (2010)
2. Candanedo, L.M., Feldheim, V.: Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. Energy Build. **112**, 28–39 (2016)
3. Ekwevugbe, T., Brown, N., Pakka, V.: Realt-time building occupancy sensing for supporting demand driven hvac operations. Energy Systems Laboratory (2013)
4. Hailemariam, E., Goldstein, R., Attar, R., Khan, A.: Real-time occupancy detection using decision trees with multiple sensor types. In: Proceedings of the 2011 Symposium on Simulation for Architecture and Urban Design, pp. 141–148. Society for Computer Simulation International (2011)
5. Khan, A., Nicholson, J., Mellor, S., Jackson, D., Ladha, K., Ladha, C., Hand, J., Clarke, J., Olivier, P., Plötz, T.: Occupancy monitoring using environmental & context sensors and a hierarchical analysis framework. In: BuildSys@ SenSys, pp. 90–99 (2014)
6. Leephakpreeda, T.: Adaptive occupancy-based lighting control via grey prediction. Build. Environ. **40**(7), 881–886 (2005)
7. Yan, D., OBrien, W., Hong, T., Feng, X., Gunay, H.B., Tahmasebi, F., Mahdavi, A.: Occupant behavior modeling for building performance simulation: current state and future challenges. Energy Buildings **107**, 264–278 (2015)
8. Ang, I.B.A., Salim, F.D., Hamilton, M.: Human occupancy recognition with multivariate ambient sensors. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 1–6. IEEE (2016)
9. Basu, C., Koehler, C., Das, K., Dey, A.K.: PerCCS: person-count from carbon dioxide using sparse non-negative matrix factorization. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 987–998. ACM (2015)

10. Lam, K.P., Höynck, M., Dong, B., Andrews, B., Chiou, Y.S., Zhang, R., Benitez, D., Choi, J., et al.: Occupancy detection through an extensive environmental sensor network in an open-plan office building. IBPSA Building Simul. **145**, 1452–1459 (2009)
11. Erickson, V.L., Lin, Y., Kamthe, A., Brahme, R., Surana, A., Cerpa, A.E., Sohn, M.D., Narayanan, S.: Energy efficient building environment control strategies using real-time occupancy measurements. In: Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings, pp. 19–24. ACM (2009)
12. Lee, H., Wu, C., Aghajan, H.: Vision-based user-centric light control for smart environments. Pervasive Mob. Comput. **7**(2), 223–240 (2011)
13. Dedesko, S., Stephens, B., Gilbert, J.A., Siegel, J.A.: Methods to assess human occupancy and occupant activity in hospital patient rooms. Build. Environ. **90**, 136–145 (2015)
14. Cali, D., Matthes, P., Huchtemann, K., Streblow, R., Müller, D.: CO_2 based occupancy detection algorithm: experimental analysis and validation for office and residential buildings. Build. Environ. **86**, 39–49 (2015)
15. Depatla, S., Muralidharan, A., Mostofi, Y.: Occupancy estimation using only WIFI power measurements. IEEE J. Sel. Areas Commun. **33**(7), 1381–1393 (2015)
16. Arief-Ang, I.B., Salim, F.D., Hamilton, M.: DA-HOC: semi-supervised domain adaptation for room occupancy prediction using CO_2 sensor data. In: Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments (BuildSys 2017), pp. 1–10. ACM (2017)
17. Shiskin, J., Young, A.H., Musgrave, J.C.: The X-11 variant of the census method II seasonal adjustment program. Number 15. US Department of Commerce, Bureau of the Census (1965)
18. Findley, D.F., Monsell, B.C., Bell, W.R., Otto, M.C., Chen, B.C.: New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. J. Bus. Econ. Stat. **16**(2), 127–152 (1998)
19. Hyndman, R.J.: Moving Averages, pp. 866–869. Springer, Heidelberg (2011)
20. Akaike, H.: A new look at the statistical model identification. IEEE Trans. Autom. Control **19**(6), 716–723 (1974)
21. Petitjean, F., Ketterlin, A., Gançarski, P.: A global averaging method for dynamic time warping, with applications to clustering. Pattern Recogn. **44**(3), 678–693 (2011)
22. Wicker, J., Krauter, N., Derstorff, B., Stönner, C., Bourtsoukidis, E., Klüpfel, T., Williams, J., Kramer, S.: Cinema data mining: the smell of fear. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1295–1304. ACM (2015)