UiT THE ARCTIC UNIVERSITY OF NORWAY

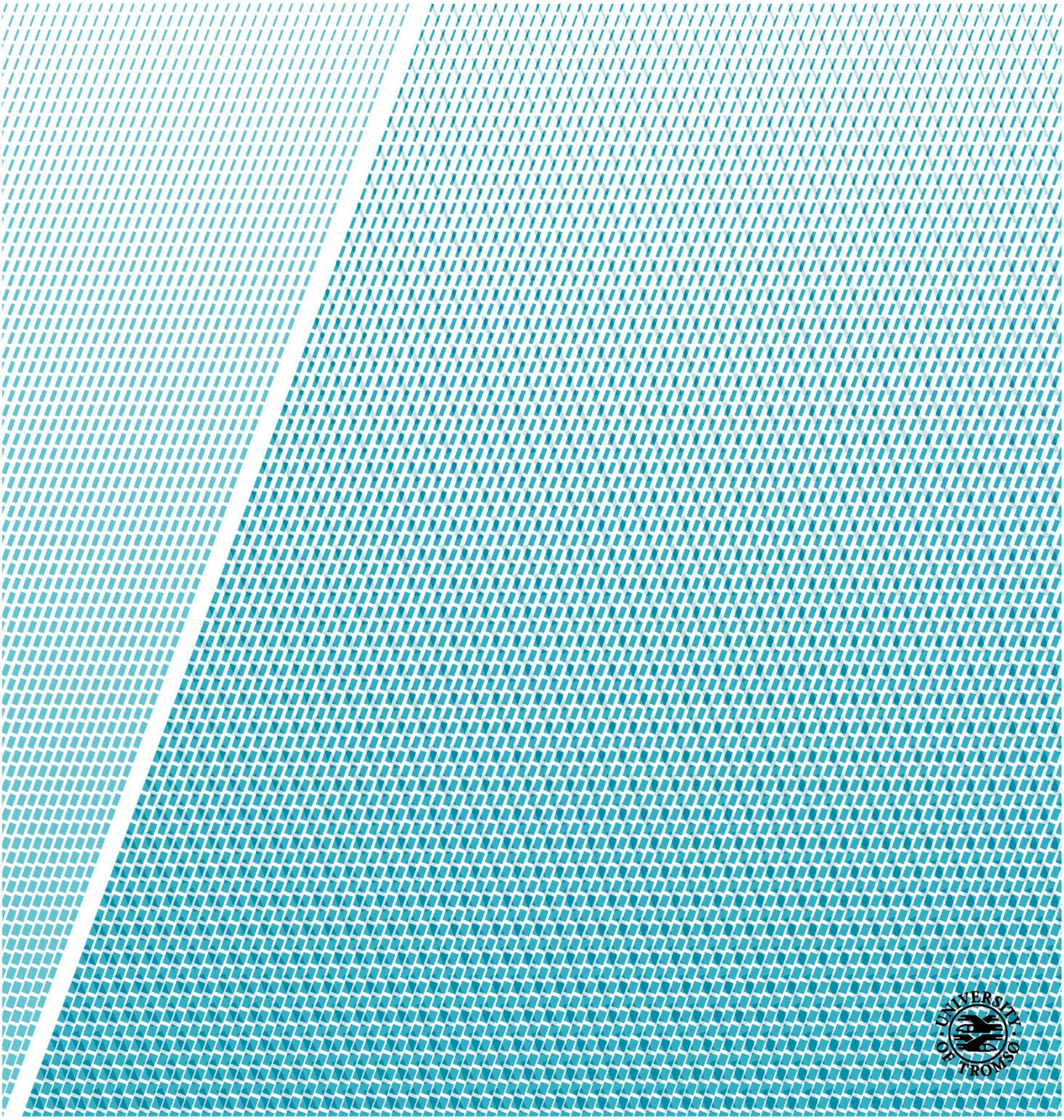
Department of Computer Science and Computational Engineering

Study of area use and floor space occupation in office buildings using ML approach

-----

Fatemeh Heidari

Master’s thesis in computer science, June 2018



Contents

List of Tables

List of Figures

Abstract

This thesis presents a general model to estimate the number of people at office spaces in the given time step. This project represents the description of the several approaches for similar problems, general description of statistical and machine learning models and applying those models for specific building. This work is also cover some suggested dashboards to keep track of space usage. The flowing models are applied to indoor air parameters: multi linear regression, support vector regression, and neural networks such as multi-layer perceptron and long-short term memory. The best performance was achieved by LSTM with 4 hidden layer and 16 number of neurons.

Keywords: Occupancy prediction, Machin learning, principle component analysis, correlation coefficient, Artificial intelligence, LSTM, SVR, MLP

Acknowledgments

I would like to express my sincere gratitude to my supervisor Prof. Bernt A. Bremdal who provided the baseline, and guided me all the time during doing my research. His advices were crucial during development and writing thesis.

Also I would like to thank to Rolv-Møll Nilsen, Serinus technology owner for delivering data and his advises and recommendation during data analysis.

I thank my friend and my family for their encouragement and supports.

1 Introduction

1.1 Background

Nowadays finding a relevant workplace space is important for both tenant and landlord because having insight of space usage can help to reduce cost and energy consumption, but this is not easy to evaluate using traditional ways. One of the important metrics which can lead to energy and cost reduction is to know how many people exist at each part of the building within current time step. To calculate this parameter, ML[1] has effective algorithms. ML methods which are rapidly developing can help to estimate or even forecast number of people within offices at each time step. Using that information leads to developing documentation which represents best matches based on company's number of employees. The main objective of current study is to estimate the number of occupants in each office. For this purpose, 31 sensors are mounted in three floors. The sensors captures IAQ[2] values such as: temperature, CO_2, humidity, noise, motion and light. The sensors used in this study are produced by "Serinus technology AS". Since we are dealing with predicting the number of people, it is possible to refer to human body metabolic parameters like CO_2. Carbon dioxide is produced in the body as a result of cellular respiration. Basically by each inhale we take oxygen then exhale out CO_2, but amount of direct CO_2 production by human is related to wide variety of other parameters such as: activity, food, space, and so on. Thus the amount of CO_2 concentration can be utilized as an indicator for the number of people. A recent study that has used CO_2 as a main indicator of human occupancy counter, has suggested SVR as a ML model to predict number of people (Irvan B. Arief-Ang, 2017).

There are two main motivations for this research. The first motivation is to improve space and room utilisation. By knowing the number of people in each room at a given time, we can monitor which rooms are under-utilised and which rooms are over-utilised. With this knowledge, they can adjust meeting room booking strategy. The second motivation is to support BMS[3] so that it can reduce the energy consumption when the room is not being used. Again knowing the number of people at a given time for each room becomes crucial to achieve this motivation.

This work presents different ML techniques, mainly LSTM[4] to estimate the number of occupants at workplace. Target building consists of an open office and several meeting rooms located on different floors. The building is a modern structure with eco-light certificate which introduces some new parameters to data analysis phase.

1.2 Research question and method

1.2.1 Research questions

Investigation of previous studies to find out which ML method is more relevant for this study and has a better performance, leads to the fact that the performance of ML algorithms is highly related to the complexity of data, selected features, and the used data set. According to (Arief-Ang, 2017), the most

[1] Machine Learning

[2] Indoor Air Quality

[3] Building Management System

[4] Long Short Term Memory

popular algorithms to estimate the number of people based on CO_2 concentration are SVR[5], KNN[6], MLP[7], and NN[8].

The main goal is estimating number of people to find the optimal model which can be used for different building types.

RQ1: Parameters that make an impact on the perceived indoor climate and how they are correlated.

RQ2: Determine the number of people in a room or different office spaces based on these parameters.

RQ3: Determine movements of people within an office area (transient use).

RQ4: Creating a generic empirical model that allows no or less supervised training.

1.2.2 Method

To answer the research questions, we start to analyse raw data. This phase belongs to pre-process part and is supporting phase to identify important events, data set behaviour as well. The next step is calculating correlation coefficients and PCA to find out how data set parameters that recorded by sensor are related to each other.

Correlation and PCA would assist to select feature vector, it is input vector that use on model development. The main model in this study is LSTM, but other models such as, Multilinear regression, Multi-layer Perceptron and Support Vector Regression developed to find the optimal algorithm.

To estimate the model performance, RMSE[9] is used. Root mean squared error is the standard deviation of residuals (prediction error). Residuals are the difference between the actual and estimated values. This value can be positive or negative as the predicted value under or above estimates the actual value. The formula to calculate RMSE is:

$$\text{RMSE} = \sqrt[2]{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

1.3 Project Business aspect

This study is supported by Serinus technology which is a sensor manufacturer, that company main objective is to estimate building energy consumption consequently reduce energy usage level. The parameter that are measured by sensor are temperature, CO_2, relative humidity, light, and movement. Data from sensor is presented live in an online dashboard and may also be reported historically. The sensor data can be exported in .csv format which then it is easy to interpret them. These data are relevant to manage buildings' properties and make estimation on number of people. There are also corporation between Serinus technology and View to ease building management facility system.

[5] Support Vector Regression
[6] K Nearest Neighbor
[7] Multilayer Perceptron
[8] Neural Network
[9] Root Mean Squared

The principle business aspect of project is:

1. Improving BMS (Building management system)
2. Find the quality of building performance to use as service level agreement type
3. Reducing Maintenance cost
4. Identify building condition then reducing buildings energy consumption
5. Find the actual building usage in compare with planed usage
6. Short term aspect: developing "meeting" booking system strategy plan

Therefore, to achieve these objectives again it is required to estimate the number of people at each given time step. Afterward we use this estimation to create simple dashboard that is visualized duration of each rooms usage.

2 Theoretical framework

This chapter describes the theoretical background of current work by describing the fundamental theory of machine learning models which were applied to sensor data, also brief review of related research.

2.1 IEQ

Indoor environmental quality (IEQ) encompasses the conditions inside a building. It has strong relation with resident comfort. In office places always the personal cost of salaries and benefits surpass operation costs of building, strategies to improve tenants' health and productivity over long term could have an acceptable return on investment. IEQ often focus on providing comfortable environment and minimizing the risk of building-related health issues.

The indoor environmental assessment consists of four different parameters, namely thermal comfort (TC), indoor air quality (IAQ), visual comfort (VC) and aural comfort (AC) (A.C.K. Lai a, 2009).

In this research we aim to use IAQ parameter to estimate the number of people consequently make assessment on building usage. Building usage can be metrics to select the most suitable office place respect to occupants. Another important aspect of this project is to help building management system to identify buildings characteristics and issues, therefor prevent or resolve problems in time.

2.2 Machine learning

The act of exploration and development of mathematical models over data to get knowledge about them is called Machine Learning. It covers clustering, regression and classification problems with aim to find optimal mapping between the data domain and the knowledge set and developing the learning algorithms (Suthaharan, 2016). In other word, Machine learning is that domain of computational intelligence which is concerned with the question of how to construct computer programs that automatically improve with experience (Mitchell, 1997). ML model can be supervised or unsupervised. Supervised model refer to the dataset which we have prior knowledge about output value then model can learn through train dataset, while unsupervised models refers to problems that we have no prior data about output. In this study, we are working with supervised regression models.

2.2.1 Regression models

2.2.1.1 Linear regression

The output of linear regression algorithm is the summation of attributes, where weights are applied to each coefficient before adding them. This algorithm is work better when attributes are separable by linear line. Linear regression is one of the simplest machine learning algorithm, since our data somehow express linear relation it is worth to evaluate this model as well. In this study the linear regression from Sklearn library are used, besides backward elimination was performed to improve the model accuracy.

2.2.1.2 Support Vector Regression

Support Vector Regression is a way of interpolating datasets. It learns fast and is systematically improvable. SVR is kind of the support vector machine to the regression problem. It can be categorized in supervised learning algorithm. Same as other algorithm it requires training set that covers domain of interest. The SVM [10]is to approximate the function to generate the training set then we can look at it as interpolation scheme.

2.2.2 Artificial Neural network

The concept of neural network patterned after the behaviour of neurons in the human brain. By this simple concept, ANN[11] tries to simulate interconnected brain cells inside a computer so it can learn things, recognize patterns, and make decisions. ANN can be introduced by collection of connected neurons, that learns incrementally from the dependency between input variables in which simulation process is just collections of algebraic variables and mathematical equations.

Without taking brain analogies in account, the simplest way of thinking about ANN is a mathematical function that maps a given input to a desired output. Therefore, ANN consist of following components:

- An input layer
- An arbitrary amount of hidden layers
- An output layer
- A set of weights and biases between each layer
- A choice of activation function for each hidden layer

2.2.2.1 Recurrent Neural Network

RNN is a type of Artificial neural network which consist of sequential information. This model has memory to remember previous information to produce accurate output, that memorizing occurs by looping process which allows information to be passed from one step to the next step. The standard RNN model has two issues: Long-term dependencies and Vanishing/exploding gradients that doesn't allow to use this model to solve all kind of problems. According to (Graves, Supervised Sequence Labelling with Recurrent Neural Networks, 2012) Long-term dependency means to produce better output, model just need to remember important information for long term not the whole history.

Vanishing/Exploding Gradients occurs when weight or activation functions are:

- $W_{recurrent} < 1$ Vanishing Gradients
- $W_{recurrent} > 1$ Exploding Gradients

Gradient based techniques take in a parameter's an incentive by seeing how a little change in the parameter's value will influence the system's output. In the event that an adjustment in the parameter's value causes little change in the system's output - the system can't take in the parameter successfully, which is an issue. This is what happening when we talk about vanishing gradients, the gradients of system output related to the parameters in the early layers will become much smaller. LSTM model which proposed in 1997 and improved later in 2000 has the ability to overcome that problems.

LSTM made of memory block that consists of three gates, namely input gate, forget gate and output gate, while in some articles called multiplicative units, these units allow memory cell to store and access

[10] Support Vector Machine
[11] Artificial Neural Network

information over long periods of time, as a result of that can extremely reduce the influence of vanishing gradient.

2.3 Related works

Occupancy detection has been explored in the last decade and one of the biggest challenges is to do this without using image processing. In (V. L. Erickson, 2009) they show accuracy can reach up to 80% by using image processing method. Although using image processing can improve the model accuracy but peoples' privacy is more important. Therefore, using camera ML methods help to calculate human calculation without privacy concern.

A report survey by the national human activity indicates that individuals spent an average of 87% of their time indoors, so understanding IAQ and its impacts are of critical importance (Klepeis NE, 2001). Series of recent studies has indicated that the quality of work environment has a direct impact on productivity, because people are spending approximately 80% of their life at workplace so the air condition has an impact on employees' health and their well-being.

Most of studies using CO_2 to evaluate air quality because it is a very good indicator of indoor air quality (IAQ), also to estimate number of people at office places. This has also been explored in prior studies by that even a slightly elevated CO_2 level at workplace can have an impact on how well people work. At that study different level of CO_2 was examined to show those who are working under the heaviest concentration of CO_2 performed 50% worse on cognitive tasks than they did in the low amount of CO_2.

But Most early studies as well as current work focus on Carbon Dioxide (CO_2) as indicator for occupancy. One Study done by (Irvan B. Arief-Ang, 2017) introduces CO_2 as a novel way to estimate number of people. It shows SVR, KNN, MLP and NN are the most popular algorithms which use to estimate occupancies. They have also suggested that SVR registers better accuracy than other model for their case study.

Prior research about "Improving building energy efficiency with a network of sensing learning and prediction agents", it is also well acknowledged that reduce energy usage of HVAC systems by estimating number of people. On that literal they used motion, CO_2, sound level, ambient light, and door state to predict number of occupant in the room. They applied machine learning techniques to find approximation of occupant, therefore they would be able to control HVAC system. In that article, they used" Multilayer Perceptron" and" Logistic Regression" to predict the number of occupant. Also it shows highly correlation between CO_2 level and the number of people in the room (Mamidi, 2012).

3 Research tools and approach

In this chapter we will go through the research tools and approach. The main technical tool is combination of python programming language, Sklearn library and statistical libraries for machine learning. The research approach is organized according to typical supervised learning workflow inspired by article from (John Joseph Valletta, 2017) figure (3-1).

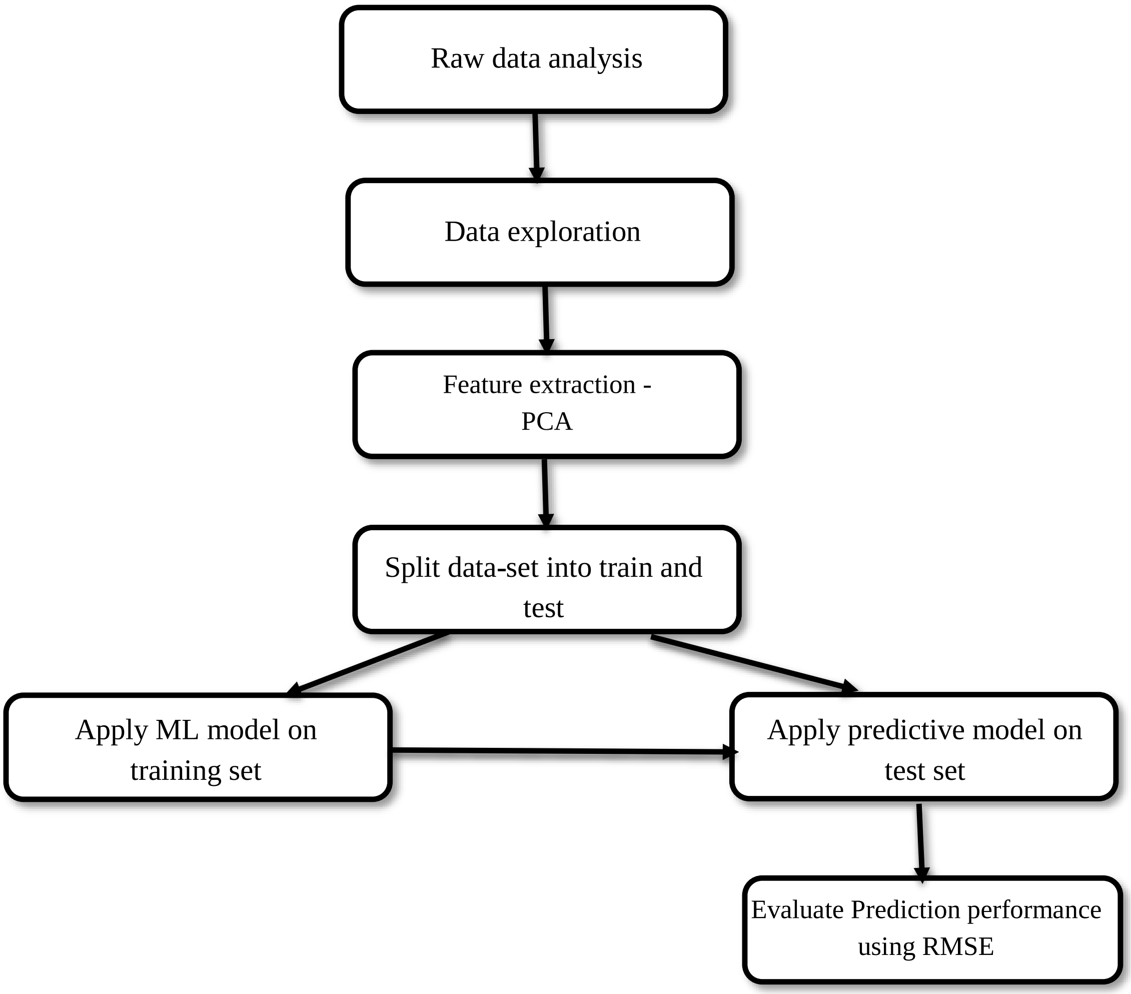

Raw data analysis
Data exploration
Feature extraction - PCA
Split data-set into train and test
Apply ML model on training set
Apply predictive model on test set
Evaluate Prediction performance using RMSE

Figure 3-1 Typical Supervised learning model

3.1 Tools and Instrument

The Scikit-Learn library and Python are used as main tools. The scientific Python libraries that are applied in this paper are:

1. Numpy, provides the ndarray data type to python, an efficient n-dimensional data representation for array-based numerical computation (Luis Pedro Coelho, 2015).
2. Matplotlib: A plotting library tightly integrated into the scientific Python stack. It offers publication-quality data visualization in different formats and is used to generate the plots in this paper.

3. Pandas, is an open source library which provides statistical analysis tools and data structure for python language. Pandas provide two dimensional, mutable size data structure called DataFrame with possibility to perform arithmetic operation along both axis (row and column). In this study we use DataFrame as main data structure (Luis Pedro Coelho, 2015).

Scikit-learn, has wide variety of algorithm for data mining and data analysis. In this research normalization process, splinting dataset to train and test and examination of model skill are developed by this library. ((BSD), 2007 - 2017)

Keras, is high level neural network API with capability to use TensorFlow, CNTK, or Theano as backend. The reason of using this API is that supports recurrent neural network which can be run on both CPU and GPU (Chollet, 2014-2017). In this project multilayer perceptron and main model called LSTM was developed based on this API.

3.2 Research Approach

3.2.1 Data description and experiments

The primary dataset gathered from 31 mounted sensors on three different floors at Gjensidige headquarter in Oslo. First floor is meeting rooms, second floor is library and third floor is open office. This dataset contains temperature, CO_2, humidity, light, motion and noise parameters that gathered from those 31 sensors at every 10 minutes. To train and validate model the researcher spent 7 days at target building to observe the number of people, door situation, rooms booking duration, and main characteristic of office space.

There are also two other sensors which mounted in different buildings to observe some functionality and limitation of sensor. Those building are not modern like the target building.

Final dataset contains (1006*10) *11 means, 1006 observations (rows) for one week with 10 columns. In the original data some of observation was missing. Mostly there was up to ten missing values, in some case the whole CO_2 level was missing because the sensor was low in battery. To deal with missing data the mean strategy was selected. Statistical data description from one of sensors generated, table (3-1) shows the behaviour of dataset, the data are from meeting room at first floor with planned capacity of 12 people. Before developing a model, it is needed to normalize data such that they have the properties of a standard normal distribution.

Table 3-1 Dataset statistic description

Statistic Description						
	Temperature	CO_2 (ppm)	Noise	Light (lux)	Humidity	Motion
count	1004	1004	1004	1004	1004	1004
mean	20.53	487.47	40.36	17.9	20.8	0.23
std	0.72	105.60	6.98	20.96	3.43	0.42
min	19.2	405	29.06	-39.24	13.7	0
25%	20.08	424.75	37.44	3.6	18.15	0
50%	20.5	446	37.94	4.14	19.98	0
75%	20.87	490	37.94	30.06	23.12	0
max	22.31	963	77.14	100.98	28.77	1

Managing various amount of data which recorded by different sensors is not easy. Thus the dataset selection strategy is needed. To select best matches, we decided to find similarity between rooms, one of trends that we take it was CO_2 level, then we divided datasets based on rooms planned capacity. Based on those process three different sun sets was established. The purpose of each floor was differ from another one, for example at first floor space usage and its compatibility with room booking system is an essential question.

3.2.2 Raw data analysis

Sensor data can be analysed then labelled for different events and validate based on researcher observation, what does that mean at this context? Serinus technology which is sensor producers' company provided online monitoring dashboard. That dashboard shows sensor data at real-time, by looking at graphs plotted in their dashboard some relation between events is observable. The moment when the ventilation system is turned off and the CO_2 level sharply goes up, figure (3-2) shows CO_2 level on 11th March. The CO_2 level raises up to 2500 ppm which can be indicator of more than eight people but that day was weekend and the time of event was midnight so by questioning from person in charge, we found out that ventilation was out of service and that's why we see a big spike on CO_2 level. This types of events are beneficial on BMS in a way to make pre alarm system.

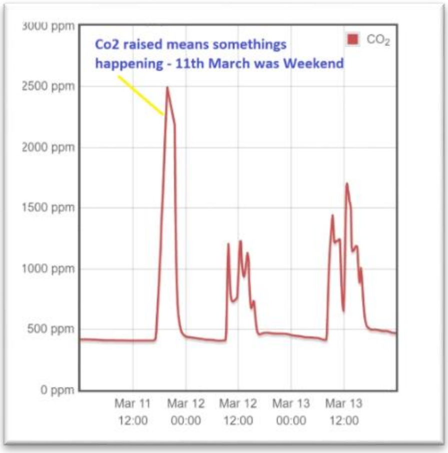


Figure 3-2 level shows peak during midnight which is good indicator for ventilation maintenance

The target building ventilation system and light is sensor based. In that case whenever sensor observe a person the light and ventilation are turning on. However, at first floor tenant can adjust temperature and lights that can generate some challenging events. Figure (3-3) shows CO_2 level and lights, we can see the lights are off but CO_2 level is higher than minimum level and that shows people are present. For that time researcher was observed three persons had meeting but they turned off the lights. This building aspect can cause some outlier on data and final model.

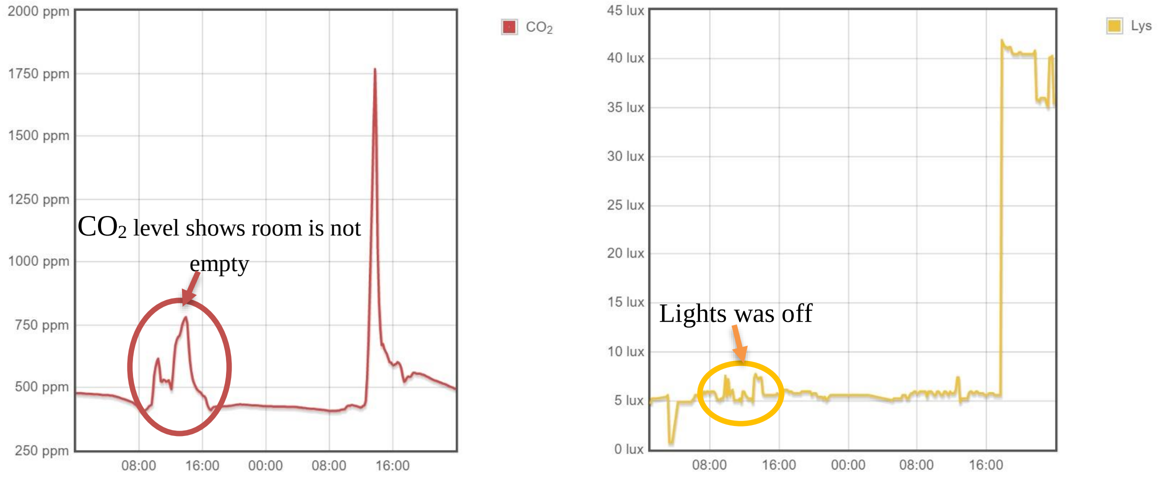


Figure 3-3 lights was off but CO_2 level shows room is in-use

3.2.3 Data exploration – Correlation analysis

Once the data cleaning process done, it is time to understand the information contained within data. Calculating correlation at this phase helps to understand which variable has the highest impact and which one has the lowest impact on dependent and independent variable. It has been wised that spending time at this step will help to boost accuracy of final model. Different strategy was taken by participant such as: Pearson's correlation coefficient using Pandas library, Recursive feature elimination, plotting scaled CO_2 values to compare with number of people from real-observations because CO_2 is one of important indicator for human presence, and calculating covariance of different sensors to answer the fourth thesis question. The covariance between open office sensors has been calculated to address movement transition, but to gain knowledge about movement we need another sensor like camera which is not allowed because of privacy issues.

The Pearson's Correlation coefficient measures the strength and direction of linear relationship between variables, the value of "r" is always between -1 and +1 while negative sign indicates downhill linear relationship and positive sign induces uphill linear relationship. Data closer to -1 shows data are lies on negative slope while closer to +1 shows data are lies on positive slope. To answer the first and second questions, first step would be calculating correlation coefficients to find what values has the most impact on indoor air climate and how they are correlated with number of people. Using Pandas library DataFrame we can perform corr() function then plot that values to see the strength of each variable. Figure (3-4) shows correlation coefficient between every two variables.

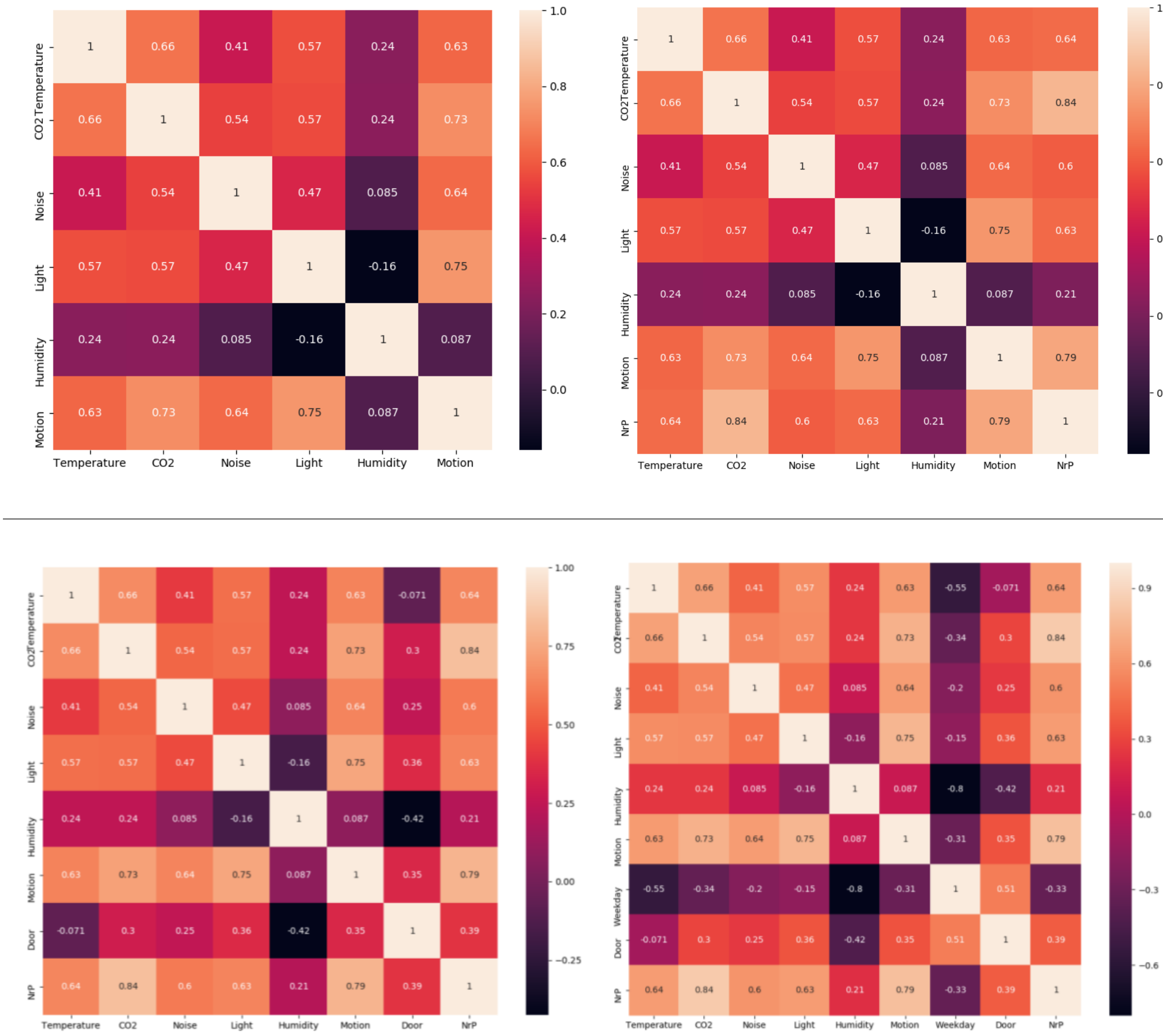


Figure 3-4 Correlation Coefficient between every two variables, NrP is indicator for number of People

Obviously, Temperature, CO_2, Motion and light are all closely correlated, but what if we take number of people in consideration? Well, there is also strong correlation between CO_2, temperature, noise, motion and light with the number of people, but this is just when we take values more than or equal to 0.6 which mean 70% to 36% of Number of people can be describes by these variables.

Since there are 31 different sensors then we should find common strong variables between them, by merging data from more than one sensors and performing correlation function we will see the results in figure (3-5).

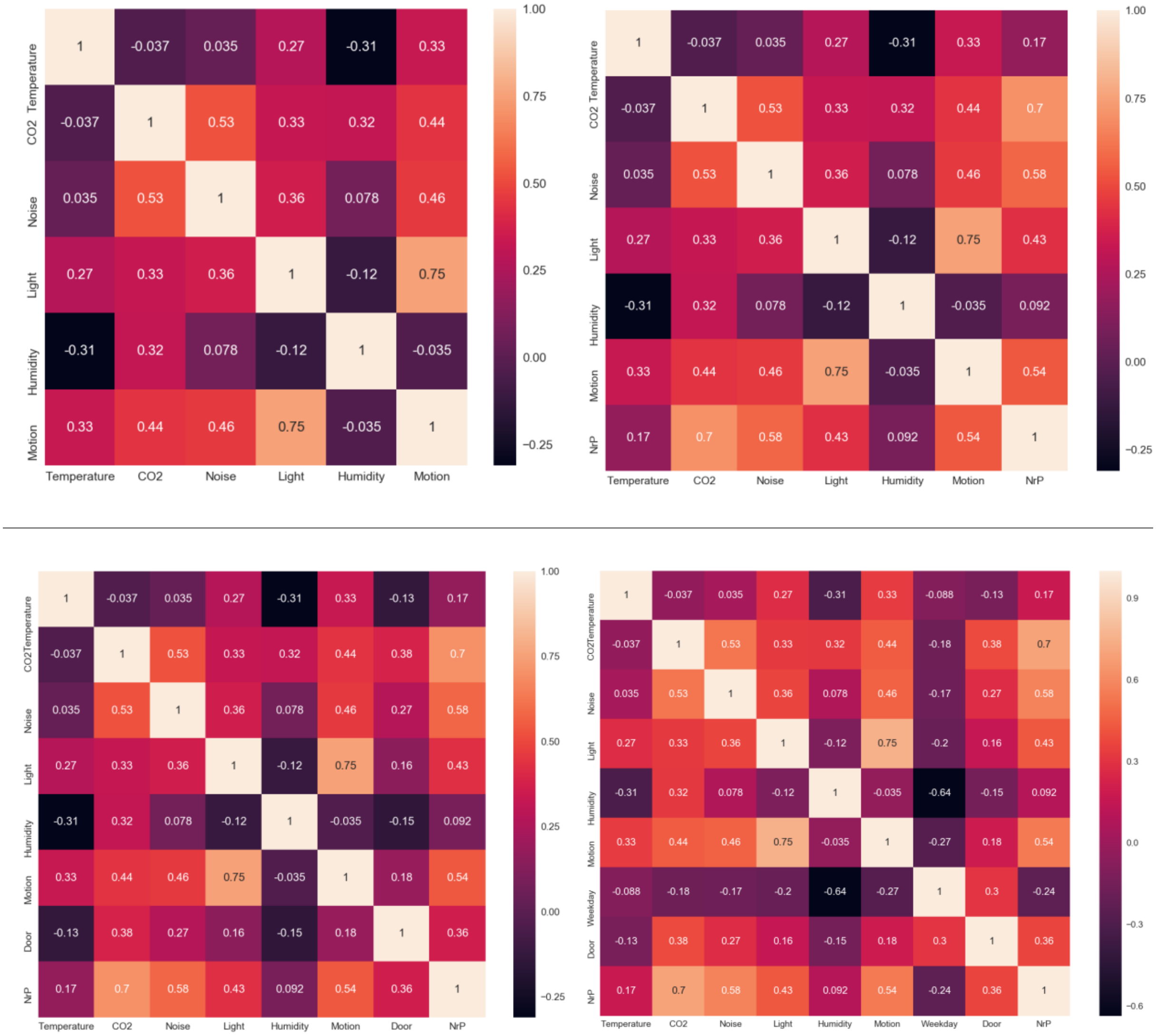


Figure 3-5 Correlation Coefficient between every two variables, multiple datasets, NrP is indicator for number of People

These figures show the strong correlation between CO_2 and number of people, about 49% of Number of people can be describes by this variable. There is high negative relation between humidity and weekdays. In case the dataset is divided to two parts working days and weekends, it is observable that this negative correlation during working days is even higher than weekends figure (3-6). A consequence of this negative correlation probably is indicator to existence of persons. It has shown in figure (3-6) during weekends strong negative correlation between CO_2 and weekend which we do not see that correlation during weekdays, that might be related to CO_2 level on prior day which means during weekdays CO_2 level cannot meet the minimum level before next working day.

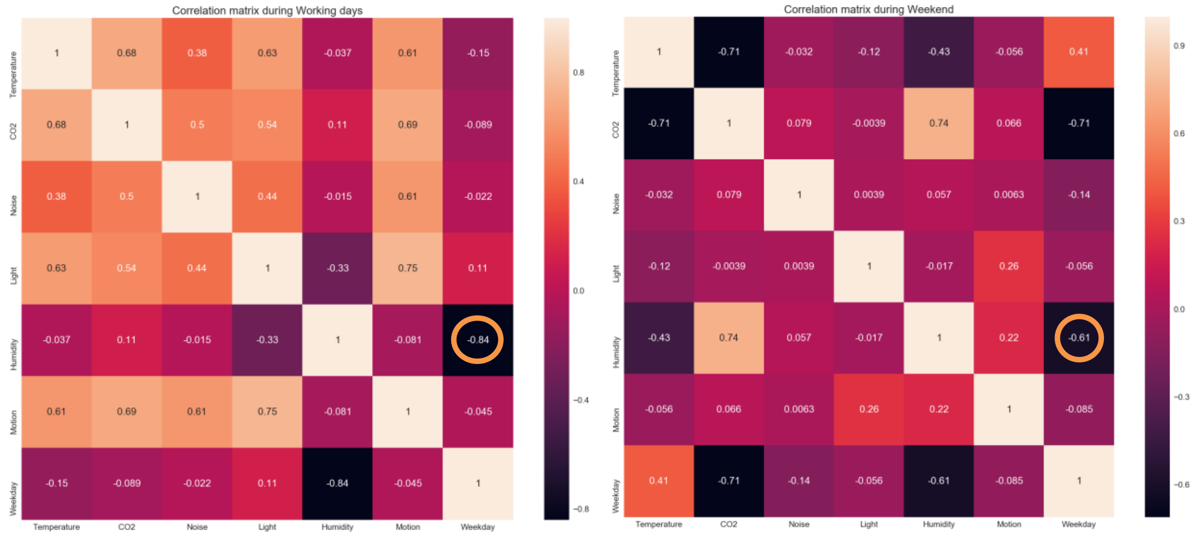


Figure 3-6 Negative Correlation Coefficient between humidity and Weekday type

One of current study objective is to make generalization, thus it is needed to find common feature vector to use in model which can cover all rooms with different characteristics, and this is one of challenges. Despite the other rooms which mostly shows strong correlation between CO_2 level and number of people there is two rooms that behave in different way, these rooms show poor correlation between CO_2 and number of people. Figures (3-7) and (3-8) shows the correlation coefficient for those rooms, it has observable that both shows poor relation between CO_2 and number of people. Sensor number 15 shows strong relation between noise and number of people, also it is relation between temperature and CO_2 which is strong enough to use in input vector.

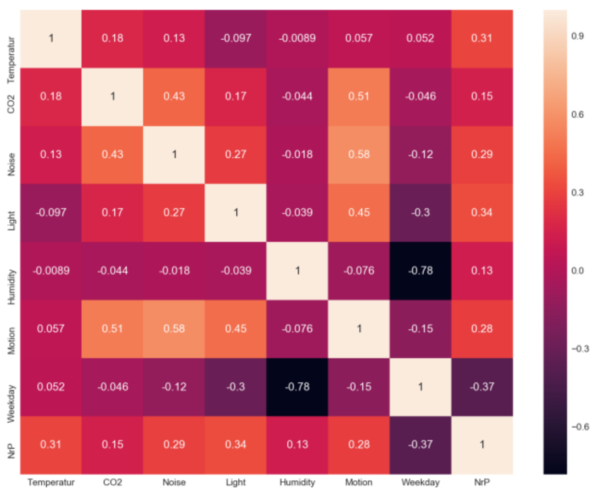


Figure 3-7 Correlation coefficient sensor number 12

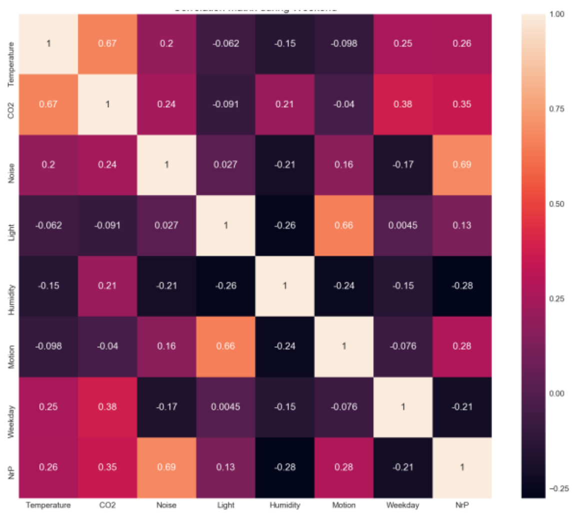


Figure 3-8 Correlation coefficient sensor number 15

By performing RFE[12] which is a feature selection method and it is working by recursively removing attributes. This method shows which variables or combination of them can contribute more to estimate the dependent variable. The result from RFE are: [1 4 3 2 1 1] = [temperature, CO_2, Noise, Light, Humidity, Motion]

To be more accurate about that strong correlation between CO_2 and number of persons, we plot the CO_2 and counted number of people during data gathering week. Figure (3-9) displays when we scale the CO_2 level using StandardScaler function from Sklearn library. That function tries to define a standard normal distribution over data then they have mean of zero and standard deviation of one. By scaling CO_2 level there is similarity between CO_2 level trend and number of people which can be proof for high r value in correlation matrix. This step is just to have evidence for choosing CO_2 level as one of primary input value.

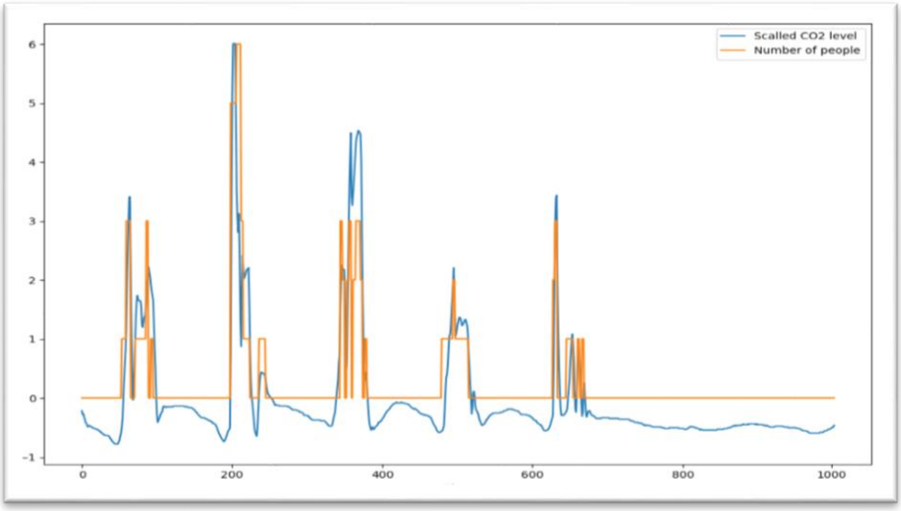


Figure 3-9 Comparison between CO_2 level and number of people

By these two steps, it is obvious that number of people and CO_2 are tightly correlated and CO_2 level can be used as one of input vector values. Also CO_2 is one of human metabolic parameter which is good indicator for people existence.

Figure (3-10) shows cluster map, the number of people and CO_2 has the strong correlation and the next place is Motion, and so on.

Since we are dealing with people estimation temperature and humidity can be added to input feature vector as another body metabolic parameters, but that needs to be evaluated using principle component analysis which is the next section.

[12] Recursive Feature Elimination

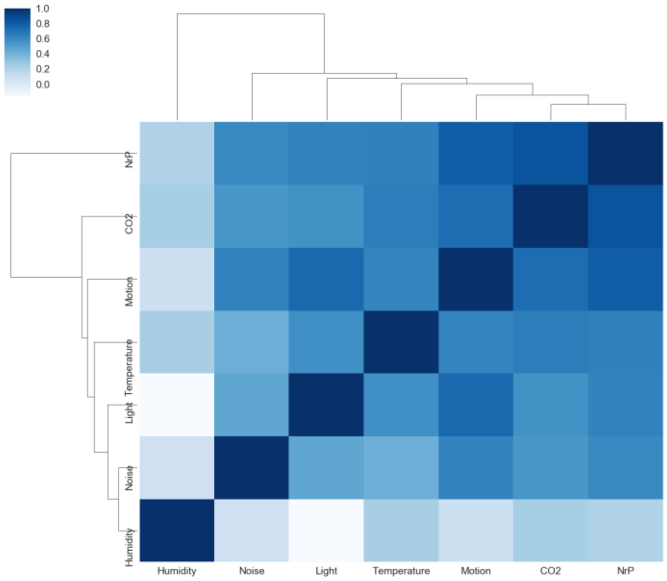


Figure 3-10 Cluster map to show correlation between variables on meeting rooms

Is it possible to use this features at open office as well? To answer this question correlation coefficient and cluster map produced using open office sensors. There poor and challenging relationship is observable. Figure (3-11) shows moderate correlation between CO_2, motion in one way and strong correlation between light and motion in other way.

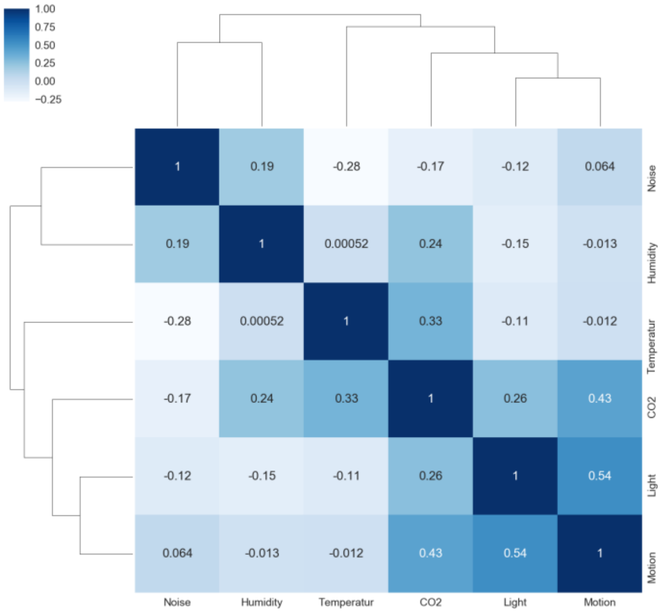
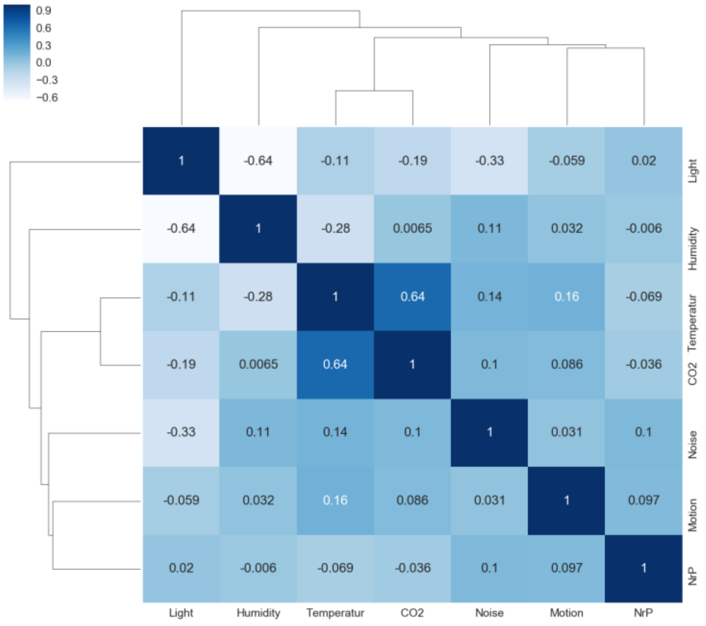


Figure 3-11 Cluster map to show correlation between variables at open office

But it is required to evaluate correlation between these variables and number of people, this is challenging because of intersection between close sensors.



	Light	Humidity	Temperatur	CO2	Noise	Motion	NrP
Light	1	-0.64	-0.11	-0.19	-0.33	-0.059	0.02
Humidity	-0.64	1	-0.28	0.0065	0.11	0.032	-0.006
Temperatur	-0.11	-0.28	1	0.64	0.14	0.16	-0.069
CO2	-0.19	0.0065	0.64	1	0.1	0.086	-0.036
Noise	-0.33	0.11	0.14	0.1	1	0.031	0.1
Motion	-0.059	0.032	0.16	0.086	0.031	1	0.097
NrP	0.02	-0.006	-0.069	-0.036	0.1	0.097	1

Figure 3-12 Cluster map to show correlation between variables at open office and number of people

Figure (3-12) contains correlation coefficient and cluster map which shows which parameter has the most relations, and which one are connected to each other. First positive strongest relation in that figure is between CO_2 and temperature, next is negative strong relation between light and humidity, then it is again negative relation between noise and lights. The most relation respect to number of people is occurred with noise. But we can use inner-relation between variables as well. Figure (3-11) presents proper relation between CO_2 and motion, also there is strong relation between light and motion. There is moderate correlation between temperature and CO_2.

3.2.4 Feature extraction – PCA[13]

PCA is one of known method for dimensionality reduction specially when we are dealing with continuous data sets. PCA is a mathematical procedure that transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components. The first principle component contains the most possibility of variables. According to Sklearn documentation (developers, Principal component analysis (PCA), 2007-2017) PCA is “Linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space.”

Since at this process we are interested on component that gives higher variance. If one variable varies less than another because of their respective metrics PCA might determine that the direction of maximum variance (developers, Importance of Feature Scaling, 2007 - 2017). In our dataset CO_2 level is in ppm[14] and gives quit higher value than temperature which is in Celsius, in other word our dataset is heterogeneous in scale. Therefore, PCA might determine that the direction of maximum variance is more closely corresponds with CO_2 axis. To dominate the effect of CO_2, the data is scaled beforehand.

[13] Principal Component Analysis (PCA)

[14] Parts Per Million

Figures (3-13) and (3-14) shows the performance of PCA before and after scaling. The alter of vectors direction after scaling is visible.

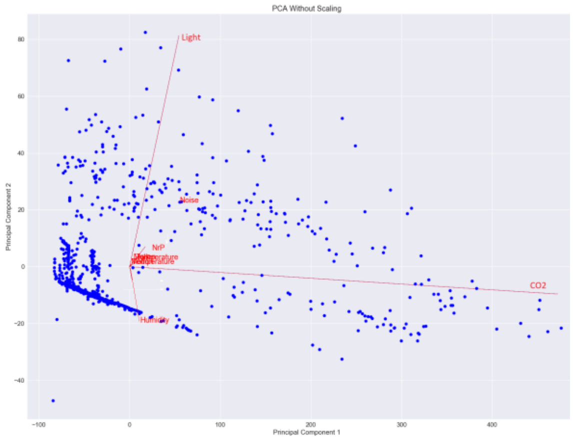


Figure 3-13 PCA before Scaling process

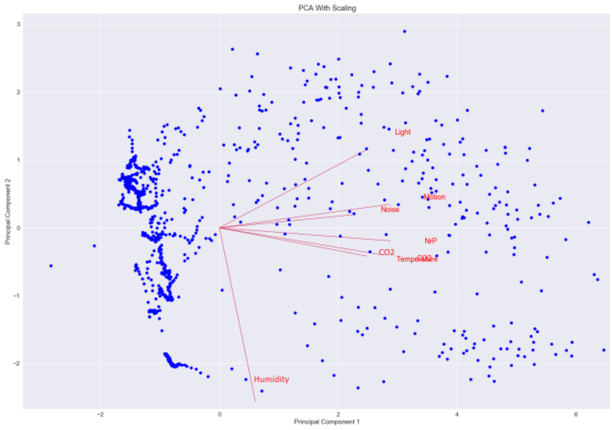


Figure 3-14 PCA after Scaling process

PCA Explained Variance after scaling is: (4.218 and 1.148) with explained ratio of (0.602 and 0.164) which means around 60% of variance lies on first principle component. This results are belonging to sensor number 3 which is meeting room with capacity of 12 people. This process repeated for another sensor to find the best feature vector. That vector is reason for answer of the first question, also it can help to boost performance of suggested model. Selected feature vector based on human metabolic parameters, correlation coefficient and PCA is temperature, CO_2 and humidity, but this can change during model implementation to improve performance and model accuracy.

To check explained variance 3D plot was generated. As it shows the most explained variance is laid on first principal component figure (3-15).

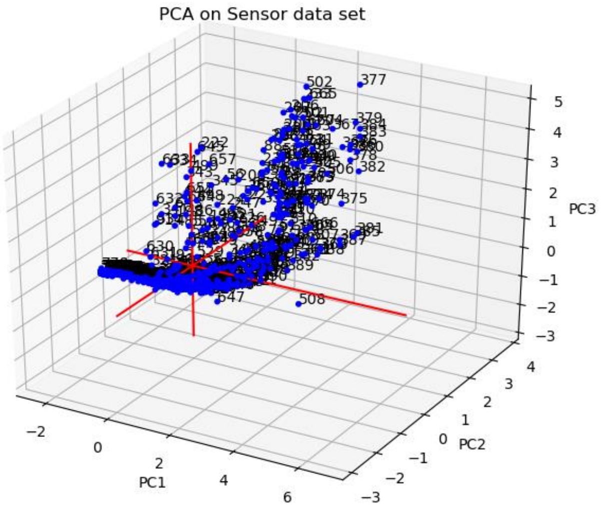


Figure 3-15 PCA 3D plot to show the strength of PC1

Since in this study we have several rooms it is important to identify common attributes between them. To address this issue, one solution is to combine data from different rooms, but the dataset for each room should be expanded to contain information about rooms. The information is rooms' capacity, volume, size, door situation and weekdays. After that process which was done in excel, PCA is performed on new dataset, the result again shows strong relation between CO_2 and the number of people figure (3-16) is evidence for this conclusion.

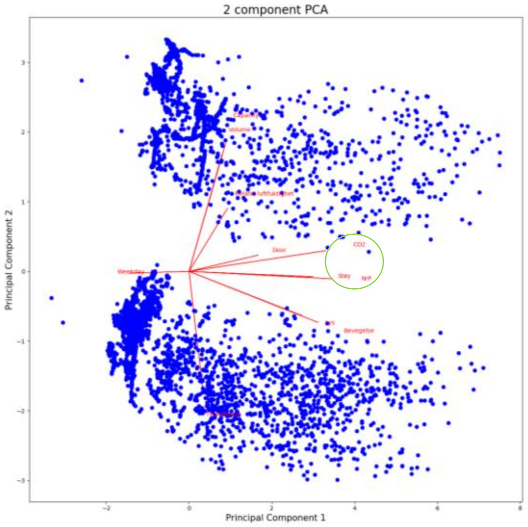


Figure 3-16 PCA performance for all sensors

In addition to meeting rooms which are regular rooms, the behavior of open office spaces has been evaluated. Open office space located at sixth floor in target building, eleven sensors was mounted to collect data every 10 minutes same as meeting rooms. Dataset contains observation of number of people for one day. At first attempt PCA performed on different sensors shows rotation along vectors, this rotation is more likely to happen on sensors that are close to each other figure (3-17).

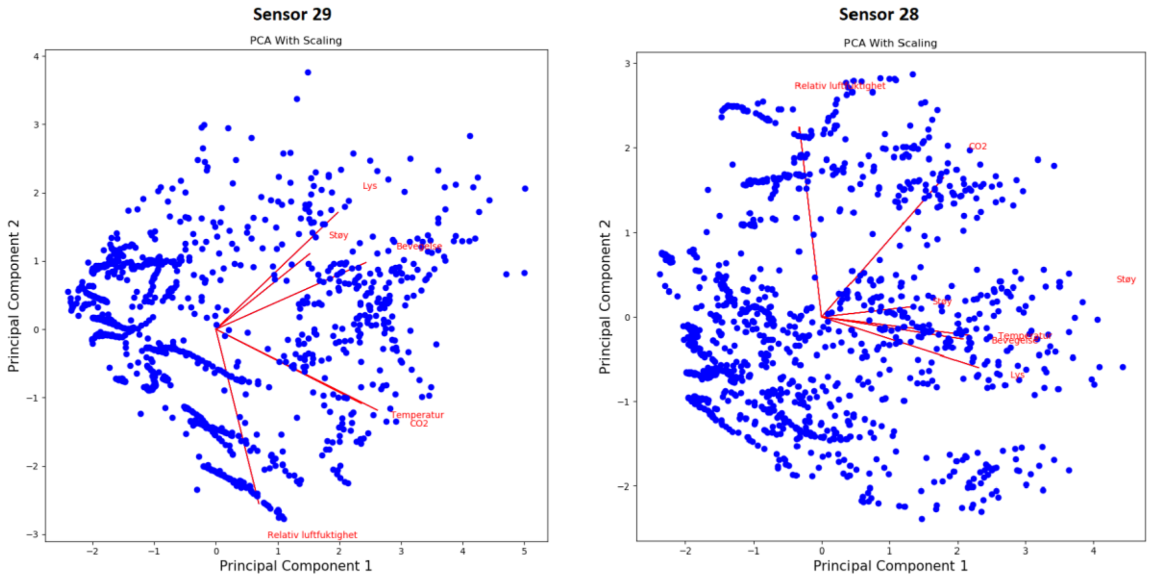


Figure 3-17 PCA shows rotational behavior along different parameters

Next step is to combine observations gathered from different sensors and use PCA to find most variances. Eight of these sensors are located in a big open area, so we will use data from those sensors to get better results.

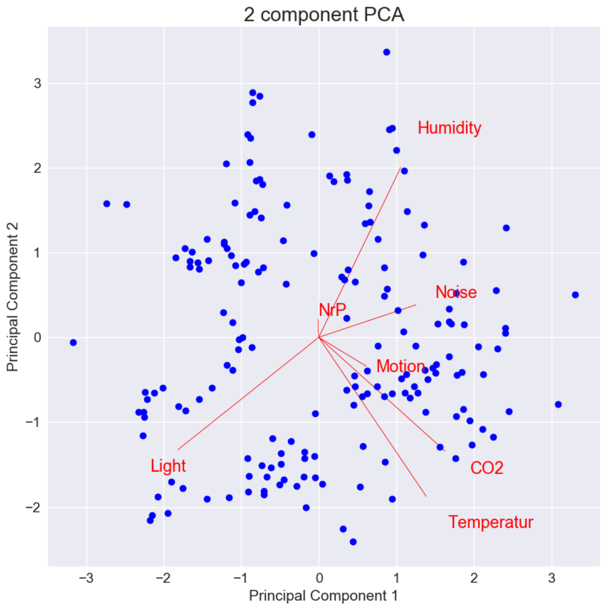


Figure 3-18 Open office PCA

Figure (3-18) displays that the most covariance is laid in first component, which is contain CO_2, temperature, light and motion. If we remove sensor number 31 from that dataset the number of people shows negative relation with CO_2 and motion but it is still on first component figure (3-19).

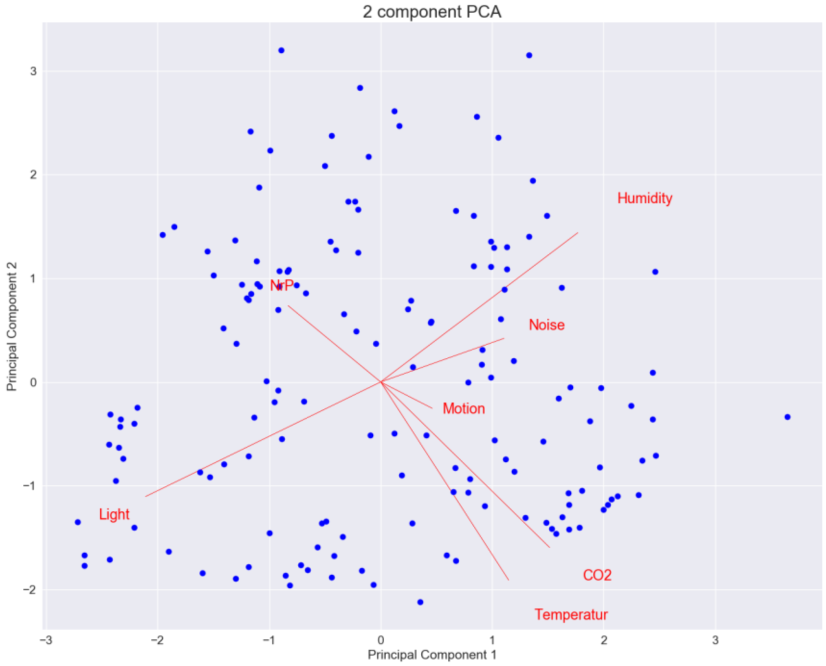


Figure 3-19 PCA with multiple dataset and removed sensor number 31

Accordingly, it is obvious that we can put temperature and CO_2 in input vector for open office to estimate the number of people.

4 Models and results

At this chapter we discuss four models which developed to predict the number of people using selected feature vector. To reviwe the properties of dataset pair plot in figure (4-1) shows the correlation between each pair of variables along with histograms for univariate distributions.A few things about the shape of the data is abvious:

1. The attributes have a range of differing distributions.
2. The Motion attribute is a binary distribution (two values).
3. The CO_2 and NrP[15] attribute looks like it has a same distribution.

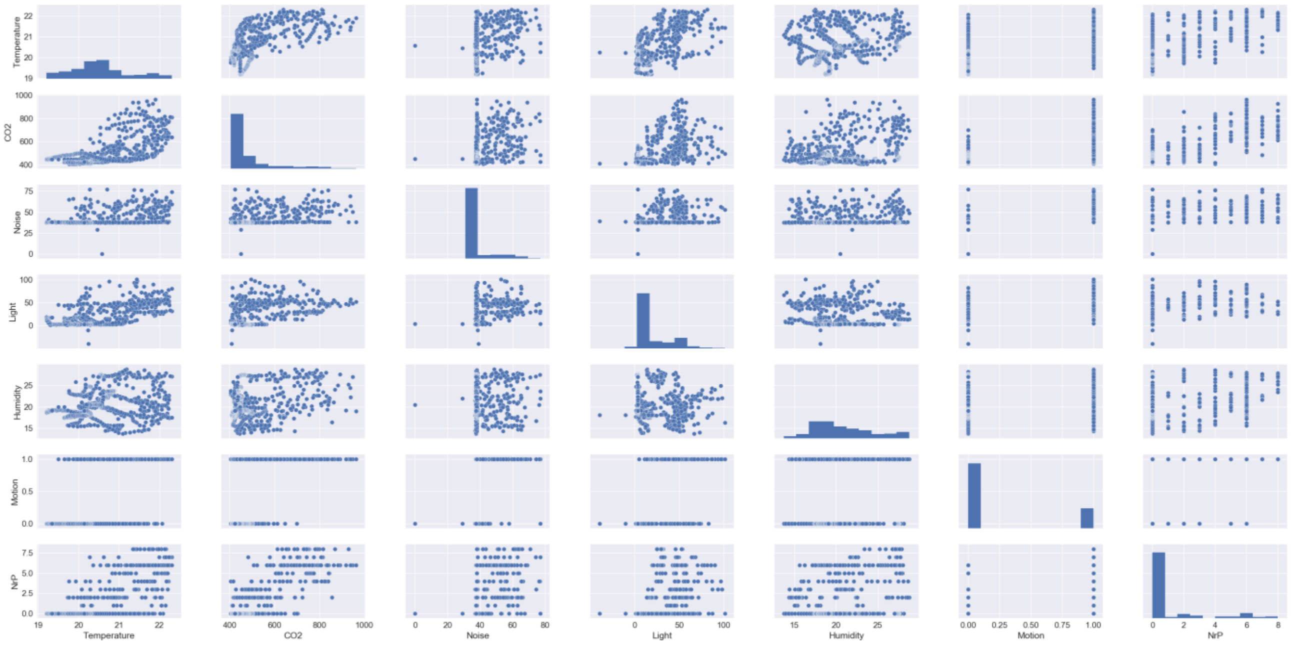


Figure 4-1 scatterplots for joint relationships and histograms for univariate distributions

Figure (4-2) shows the fited regresion line over data set attributes.

[15] Number of People

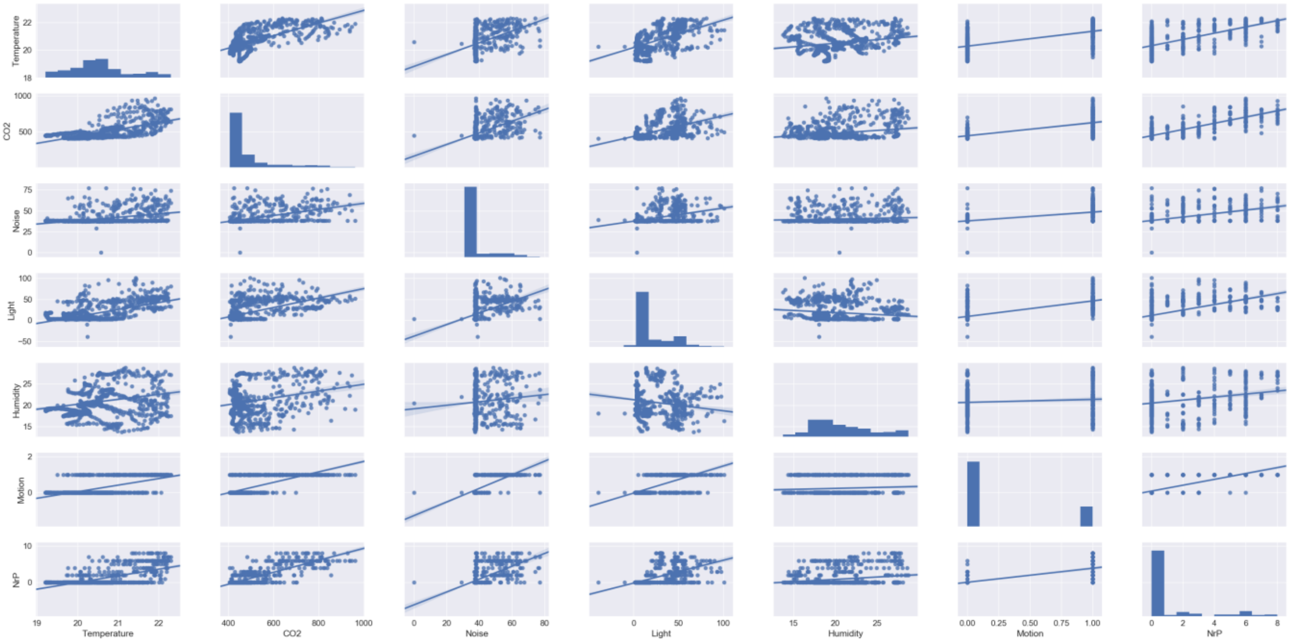


Figure 4-2 Fit linear regression models to the scatter plots

4.1 Linear regression

Linear regression model is one of the simplest machine learning algorithm which is based on linear equation as follow:

$$y = b_0 + b_{1*}X$$

where b_0 is the bias coefficient and b_1 is the coefficient for independent variable which in our case is CO_2. The coefficient y is dependent variable and X is independent variable. The model developed using Sklearn linear model library. The first attempt with just CO_2 as an independent variable gave RMSE ~ 0.969 on test set and RMSE ~ 0.974 on validation set for rooms with capacity of 12 people.

Next model was developed with all sensor data, consist of temperature, CO_2, noise, light, motion and humidity thus this is multiple linear regression, therefore the equation is:

$$y = b_0 + b_{1*}X_1 + b_2 * X_2 + b_3 * X_3 + b_{4*}X_4 + b_5 * X_5 + b_6 * X_6$$

Multilinear regression model shows RMSE ~ 0.915 on test set while introduces RMSE ~ 0.913 on validation set for rooms with capacity of twelves people and RMSE ~ 0.464 on test set and RMSE ~ 0.585 on validation set for rooms with capacity of four people.

To improve skill of model backward elimination method was performed. The method is using ordinary least square techniques to estimate the values of coefficients. Then by looking at summary result which contains p-value, at each step we remove variable with p-value lower than significant level, which in our case SL[16] threshold is 0.05. Afterwards the OLS process are performed again to find parameters with best p-values. This steps should be repeated until finding features with higher p-value. Feature

[16] Significant Level

input vector from this process contains CO_2, noise and motion. It is important that this model works better for rooms with lower planned capacity. Dataset extended to include weekday, rooms planned capacity, room volume and door situation. It is noticeable when linear model trained by combination of several sensor data, also adding capacity and weekday can improve model accuracy. However, the model work better when room has lower planned capacity, around 4 to 6 number of people. The test set RMSE is approximately 0.887 which shows improvement on error. After several test it was observed that RMSE error is between 0.6 to 1 and that is depend on room capacity. Rooms with capacity of four to six again shows better performance with RMSE ~ 0.6.

4.2 SVR

Support vector machine is mostly used in classification problems, since the main problem is classified as regression we need to use Support Vector Regression which is generalized form of SVM. In classification version of this model, the input vector used to define a hyperplane in which separate the two different classes in solution vector. Although, in SVR these feature vector along with solution for training part used to perform linear regression. The vector closest to your test point, or decision boundary are the ones that refer to a support vector (J.Smola, 2004). In SVR, loss and kernel are different from SVM because of the nature of task. Sklearn library is used to develop model over training set and make prediction over test set, to evaluate model accuracy RMSE is performed.

SVR is a model that uses linear and non-linear kernels. We are implemented SVR model and experiment through different kernels to identify which one perform better. In case, input vector is selected in such to contain CO_2 level and Weekday with ‘rbf’ kernel the test RMSE ~ 0.946 and this value on validation set is approximately 0.971 for room with high capacity about 12 people. If we test the model over smaller rooms, capacity vary from 4 to 6, test set RMSE approximately is in range [0.564, 0.82] and on validation set is [0.664, 0.814]. During experimenting on different datasets it has shown that room which are located in same direction have approximately the same RMSE error, rooms with sensor number 4, 8, 18 ,11 and 16 shows RMSE ~ 0.6 while sensors 12 and 7 which located on the other side show RMSE ~ 0.82. When the model trained by data from multiple rooms the RMSE on validation set will be slightly better than previous.

4.3 ANN

The basic idea of ANN is a network of small processing units called node, joint together with weights. Artificial neural networks (ANNs) were originally developed as mathematical models of the information processing capabilities of biological brains (Graves, Neural Networks, 2012). In term of biological model nodes are equivalent to neurons and weights are similar to the strengths of the synapses which make connection between neurons. The way of working is based on input values which spreads through hidden layer. In this research two different kind of ANNs called MLP and LSTM are developed.

4.3.1 Multilayer perceptron

An MLP[17] is neural network with multiple hidden layers with mainly known as feed forward model and can be represented graphically as follows:

[17] Multi-Layer Perceptron

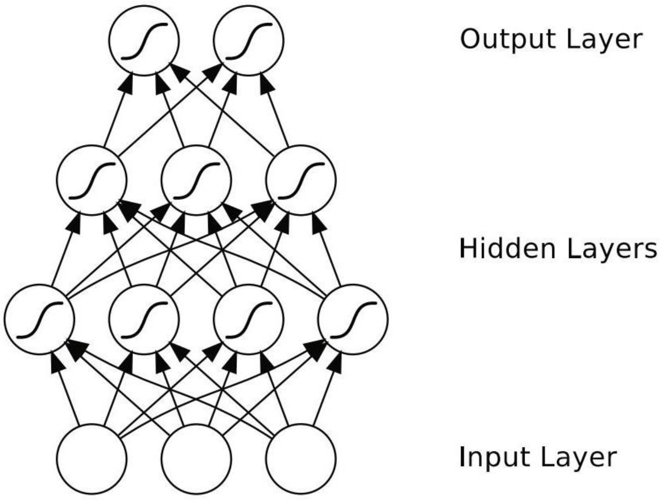


Figure 4-3 multilayer perceptron, the S-shape curve represent activation function on hidden and output layer (Graves, Supervised Sequence Labelling with Recurrent Neural Networks, 2012)

As shown in figure (4-3) this model is a feedforward process because input presented to input layer, then propagated through the hidden layers to the output layer. In developed model, since during feature selection we decide to proceed on temperature, CO_2 and humidity. The final model has three input dimension along with 13 cells at first layer and 9 cells at each other hidden layers. The activation function on hidden layers selected to be rectifier function called relu[18]. The function returns 0 when it receives negative input, however for positive values it returns that value back. Since the aim is to predict number of people, we will have ended up with one output. In output layer sigmoid function used as activation with below equation:

$$F(x) = \frac{1}{1+e^{-x}}$$

The sigmoid function forces values to the range of 0 and 1. To evaluate model before implementation phase, it is popular to use cross-validation technique. The purpose of that phase is to make estimation of mean error over K number of test folds, which mostly K-fold is selected to be ten. As we are using Keras functionality, models can be evaluated by Keras wrapper object, called KerasRegressor for regression problems, provided to use over Sklearn. That wrapper object takes function as argument, thus different MLPs model created to send as input parameter for that object. The MLPs models differ from each other by their number of hidden layers and number of cells at each layer.

A strategy to improve MLP performance is to explicitly change number of hidden layers and cells. To compile model, Stochastic gradient-based optimization, Adam, is used as optimizer while mean squared error defined as loss function. During estimation of first and second moments of the gradients different parameter will be produced such that model can compute adaptive learning rate for each parameter. The important aspect of Adam optimizer is that only requires first-order gradients with little memory requirement (Ba, 2014).

[18] Rectified Linear Unit

Table (4-1) is demonstrated the performance of several architectures of MLP on train, test and validation.

Number of hidden layer	Number of Neurons	Room capacity	Train RMSE	Test RMSE	Validation RMSE
4	6	4-6	0.502	0.502	0.670
		10-12	1.148	1.021	1.024
6	6	4-6	0.488	0.483	0.678
		10-12	1.164	1.084	1.088
12	6	4-6	1.007	1.182	1.179
		10-12	2.190	2.309	1.624
4	13	4-6	0.489	0.489	0.682
		10-12	1.147	1.016	1.024
6	13	4-6	0.489	0.487	0.687
		10-12	1.149	1.013	1.035
12	13	4-6	1.007	1.180	1.179
		10-12	1.129	1.035	1.025
4	18	4-6	0.487	0.483	0.683
		10-12	1.14	0.996	0.999
6	18	4-6	0.491	0.492	0.715
		10-12	1.13	0.981	0.977

12	18	4-6	1.007	1.181	1.179
		10-12	2.189	2.309	1.624

Table 4-1 The performance of different architecture of MLP on train, test and validation sets

Validation it has done using data of different sensor but the rooms capacity is same as the test data. As it has shown on table (4-1) MLP is performing better on rooms with less planned capacity. It is observable that expanding layers and number of neurons it cannot improve models' accuracy. MLP can predict up to five persons but not more than five. It is possible to boost model accuracy by introducing more training data because in this study we just have observation from one week which contains weekend.

To improve model accuracy on bigger rooms we developed these steps, result shown on table (4-2):

1. train model with data collection of rooms with different capacities
2. introducing capacity and weekday to model and expanding input feature vector
3. expanding model by composed layers with different number of neurons

Number of Neurons 1st layer	**Number of Neurons 2nd ,3rd , 4th layer**	**Room capacity**	**Train RMSE**	**Test RMSE**	**Validation RMSE**
6	6	10-12	0.995	0.997	0.982
13	13	10-12	0.875	0.923	0.944
13	9	4-6	0.947	0.954	0.654
		10-12	0.871	0.928	0.939
13	6	4-6	0.907	0.968	0.712
		10-12	0.909	0.897	0.885

Table 4-2 The performance of composed MLP architecture on train, test and validation sets

One more test was performed based on all sensor parameters along with rooms' capacity and weekdays, model architecture was consisting of four hidden layer, 13 neurons on first layer and 9 neurons on other layers. The result shows slightly improvement on biggest rooms accuracy, table (4-3) shows results from expanded feature vector.

Number of Neurons 1st layer	**Number of Neurons 2nd ,3rd , 4th layer**	**Room capacity**	**Train RMSE**	**Test RMSE**	**Validation RMSE**
13	9	4-6	0.941	0.974	0.625
		10-12	0.917	0.973	0.851

Table 4-3 The performance of MLP with expanded feature vector

To sum up, MLP performed better when it is trained with wider dataset, which can be a proof for claim about neural networks, they perform better when they trained on wider dataset. Composed hidden layers are represented better on rooms on rooms with higher capacity.

4.3.2 Long Short Term Memory

In previous section we considered feedforward neural network without cycle between connections so the model cannot remember long term data. To capture pervious history, one solution is adding cycle to network thus it can refer to prior data. RNN is an ANN model which has cycle through and within layers. The simplest RNN model has displayed in figure (4-4).

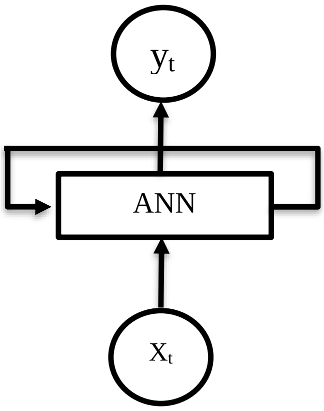


Figure 4-4 Simple RNN which contain loop

By that cycle network can use its history when mapping between input and output is occurred. Since RNN is gradient based algorithm with back propagating, one of issues is access to the limit range of context. The problem is that the influence of a given input on the hidden layer, and therefore on output, either decreases or exceeded exponentially as it cycles around the network's recurrent connections. This affect called vanishing gradient problem. Therefore, it makes model really hard to learn and tune the parameters of the earlier layers in the network. And it became even worse when the number of hidden layers are increases. To overcome that problem in 1997 by Sepp Hochreiter, model called Long Short Term Memory was introduced (Yuhuang Hu, 2018) (Graves, Supervised Sequence Labelling with Recurrent Neural Networks, 2012).

LSTM can learn to bridge time intervals in excess of 1000 steps even in case of noisy, incompressible input sequences, without loss of short time lag capacities and this is accomplished by an effective, gradient based algorithm for a design implementing consistent error flow through internal states of special units (Sepp Hochreiter, 1997).

A common LSTM architecture contains memory cells with an input gate, an output gate and a forget gate. Forget gate layer will decide what parameter should throw away. The process is looking at previous time step output, y_{t-1}, and current input $X_{t,}$ then produce value between [0 1] for each number in state cells, if the calculated value equal to one means the value should be kept otherwise throw it away or forget that value. The next step is to decide what new values should be memories, this happens through two steps starts by input gate layer which make a decision about updating values. Next goes to tanh layer which creates a vector of that updated values. Outcomes from input gate layer and tanh later combined to update the old cell state. Finally, the output will be produced by using these states.

The LSTM in this project are developed using Keras library. Model consist of 4 hidden layer while each layer contains 16 neurons. After each iteration 20% of neurons will be disabled to avoid overfitting. At each iteration of the training data some neurons are randomly disabled to prevent them from being dependent to each other when they learn the correlation. Therefore, by writing these neurons the model learns several independent correlations in the data because each time there is not the same configuration of the same neurons. The fact that we get these independent correlations of the data, then neurons work more independent that prevents them from learning too much thus that can overcome overfitting.

In study done by (Vinyals, 2014) they also show the effect of drop out at RNN. We will use drop out functionality to disable that portion of neurons after each iteration. The algorithm improved through experiments to find the best match, model with long history makes better estimation over test and validation datasets. The first model was construction of four hidden layer, with 16 neurons at each layer and 20% dropout after each layer. Model was developed to retrieve every 30 minutes' prior history, validation set selected to be room with same capacity as train and test dataset because we aim to make general model. Table (4-4) illustrate the performance of different execution.

Feature vector	Room capacity	Train RMSE	Test RMSE	Validation RMSE
CO_2, temperature, humidity and noise	10-12	0.142	0.146	0.120
	4-6	0.104	0.086	0.124
CO_2, temperature, humidity	10-12	0.135	0.124	0.113
	4-6	0.098	0.086	0.118
CO_2, temperature, humidity and motion	10-12	0.130	0.131	0.102
	4-6	0.092	0.080	0.112
CO_2, motion	10-12	0.141	0.128	0.108
	4-6	0.112	0.102	0.126
CO_2, temperature and motion	10-12	0.148	0.136	0.113
	4-6	0.109	0.101	0.127

Table 4-4 Performance of LSTM with 30 minutes time step and 4 hidden layer

It is observable that feature vector which contain CO_2, temperature, humidity and motion data perform quite better than others. Model was trained by sensor number 3 and validate over all other sensor with this feature vector show acceptable results.

One of strategies to improve accuracy is to expand hidden layer and number of iteration which in Keras model it is called 'epoch'. Therefore, next model consists of six hidden layers and sixteen neurons along with 320 number of epochs. Table (4-5) represent the summary of different tests run with this architecture.

Feature vector	Room capacity	Train RMSE	Test RMSE	Validation RMSE
CO_2, temperature, humidity and noise	10-12	0.155	0.170	0.136
	4-6	0.116	0.099	0.153
CO_2, temperature, humidity and motion	10-12	0.153	0.164	0.158
	4-6	0.103	0.112	0.150

Table 4-5 Performance of LSTM with 30 minutes time step and 6 hidden layer

Comparing results of LSTM with four hidden layer with other model which contains six hidden layer it is obvious again that expanding hidden layer cannot be solution to improve accuracy in current project.

Third LSTM model was developed to use CO_2, temperature, humidity, weekday as input feature vector. Model is provided to keep 30 minutes, 60 minutes, 90 minutes and 120 minutes' prior time steps. Thus input vector is [$CO_{2,\ t\text{-}30}$, $CO_{2,\ t\text{-}60}$, $CO_{2,\ t\text{-}90}$, $CO_{2,\ t\text{-}120}$, $Humidity_{t\text{-}30}$, …, $Temperature_{t\text{-}30}$, …, Weekday and number of people]. The result demonstrates better accuracy when we introduce more history in the model. Also weekday parameter used to distinguish between working days and weekends.

Feature vector	Room capacity	Train RMSE	Test RMSE	Validation RMSE
CO_2, temperature, humidity and weekday	10-12	0.136	0.146	0.145
	4-6	0.097	0.094	0.153

Table 4-6 Performance of LSTM with more history and 4 hidden layer

Table (4-6) shows slightly improvement on model accuracies, especially on bigger rooms.

Since we are dealing with variety of rooms capacities, and we suffer from great amount of training data, an important test is to train model using dataset which contain observation of different rooms. Three different dataset was selected; rooms' capacity was introduced to distinguish between rooms. At first step LSTM with four hidden layer and 30 minutes' history was trained to check for improvement.

Feature vector	Room capacity	Train RMSE	Test RMSE	Validation RMSE
CO_2, temperature, humidity, weekday and capacity	10-12	0.094	0.099	0.151
	4-6			0.137

Table 4-7 LSTM trained with multiple rooms dataset and 30 minutes time step

Table (4-7) illustrates improvement on all datasets, that shows whenever LSTM model trained more, it can perform better. Because the model can adjust weight on parameters.

Last LSTM model constructed to use CO_2, temperature, humidity, weekday and capacity as input feature vector. Model is provided to keep 30 minutes, 60 minutes, 90 minutes and 120 minutes' prior time steps. Thus input vector is [$CO_{2,\ t-30}$, $CO_{2,\ t-60}$, $CO_{2,\ t-90}$, $CO_{2,\ t-120}$, Humidity$_{t-30}$, …, Temperature$_{t-30}$, …, Weekday, room Capacity and number of people]. This model trained using one rooms and combination of different rooms data.

Training files	Room capacity	Train RMSE	Test RMSE	Validation RMSE
Single room data	10-12	0.135	0.137	0.138
	4-6	0.106	0.090	0.147
Multiple rooms data	10-12	0.083	0.086	0.123
	4-6			0.089

Table 4-8 LSTM with long history

Table (4-8) shows when we train model with multiple datasets, model accuracy improves and it can be used as general model for all rooms with different capacities. Thus one of study questions is to make generalization, here we can claim that this model is answer for that statement.

Final model trained with multiple rooms' dataset made predictions on sensor number 16, which is room with capacity of six people, figure (4-5) shows result of prediction for one day based on sensor data gathered from sensor number 16.

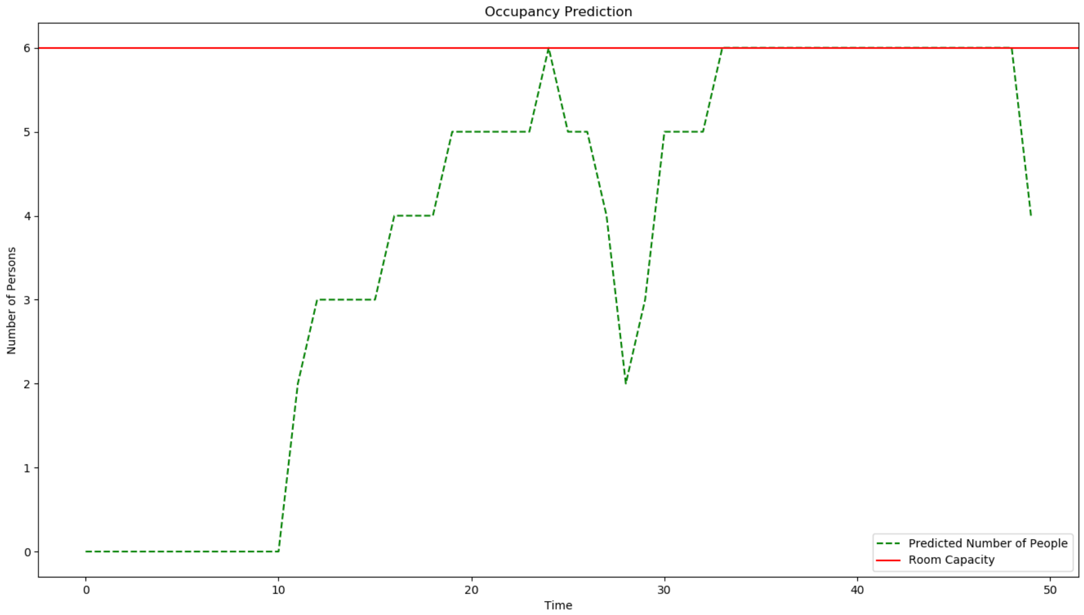


Figure 4-5 Prediction of people using sensor 16 data, model trained with multiple rooms' dataset

4.4 Movement flow

In this research study the minor objective is to obtain transition flow within open office space. One of the options to capture movement flow is to utilize cameras, but the privacy is a big challenge here and we couldn't use camera for this project. Thus we decided to proceed on sensor data then make estimation on number of people to find main transition flow at this area. To answer this question, we start to acquire similarity between CO_2 concentration in various sensors. The trend displays sensor which are located close to each other, and located in same direction has similar CO_2 level.

Figure (4-6) is a proof for that statement. It is observable that CO_2 level has delay, this can arise two possibilities, first sensors captured data with delay which we know sensor has 10 minutes' delay, second theory is related to people transitions. Since CO_2 concentration is main indicator of people presence. That might be also indicator for people transitions in a way that they move from one sensor to another.

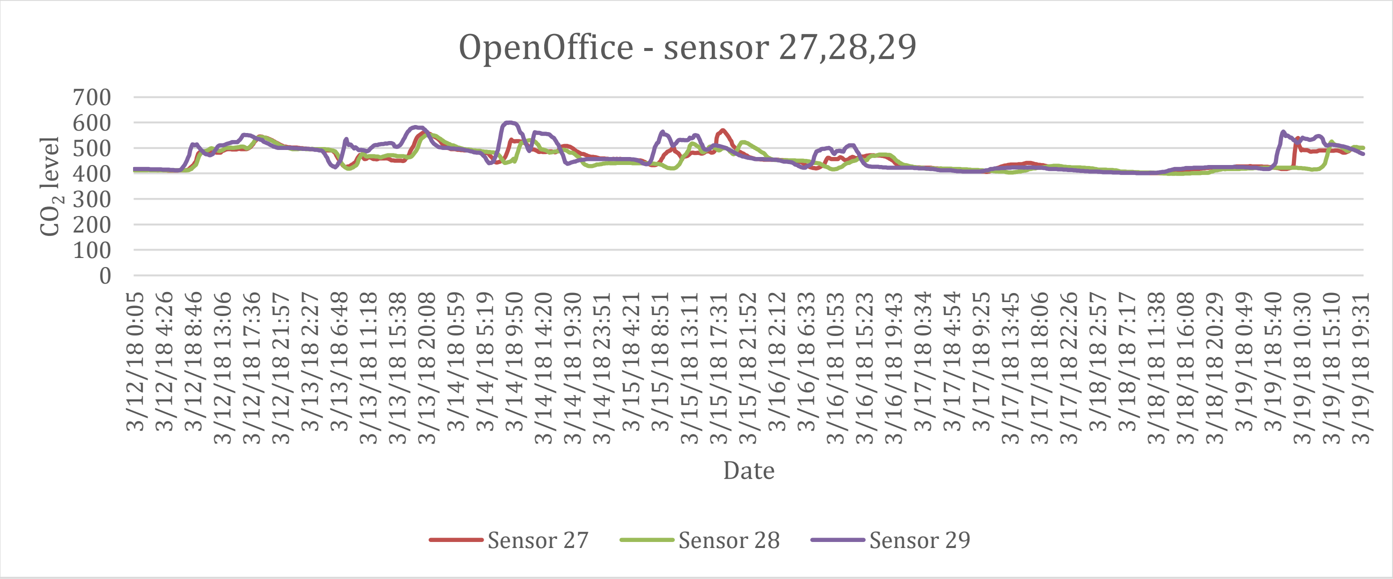


Figure 4-6 Comparison between multiple sensors CO_2 level

Evaluating PCA on open office space illustrates gradient tendency between CO_2 and humidity which can be indicator of certain interference between people sitting at either corner of the office area. To make proof on this results below steps have been performed:

1. No People in the room at all (evaluate CO_2, humidity and noise Level)
2. A number of people gathered at one end of the room only (close to one sensor).
3. The same number of People gathered at the opposite end of the room.
4. Divide the same number of People at either side of the room.
5. Repeat steps 3 to 4 for other sides of the room too.

Figure (4-7) displays result of above calculation, we can see whenever number of people raises in each end the covariance gets higher.

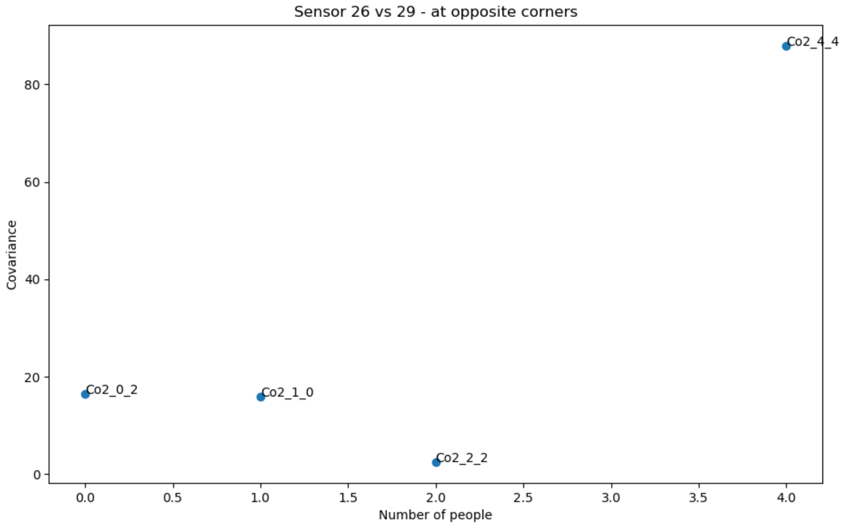


Figure 4-7 Comparing covariance of CO_2 level between two sensors located at each corner of open space, various number of people at same time is selected to check CO_2 gradient

Next step is to calculate covariance using different parameters, the covariance is to understand the relation between distance of sensor and variance, to estimate this indicator one of sensors selected as index to compare with other sensors.

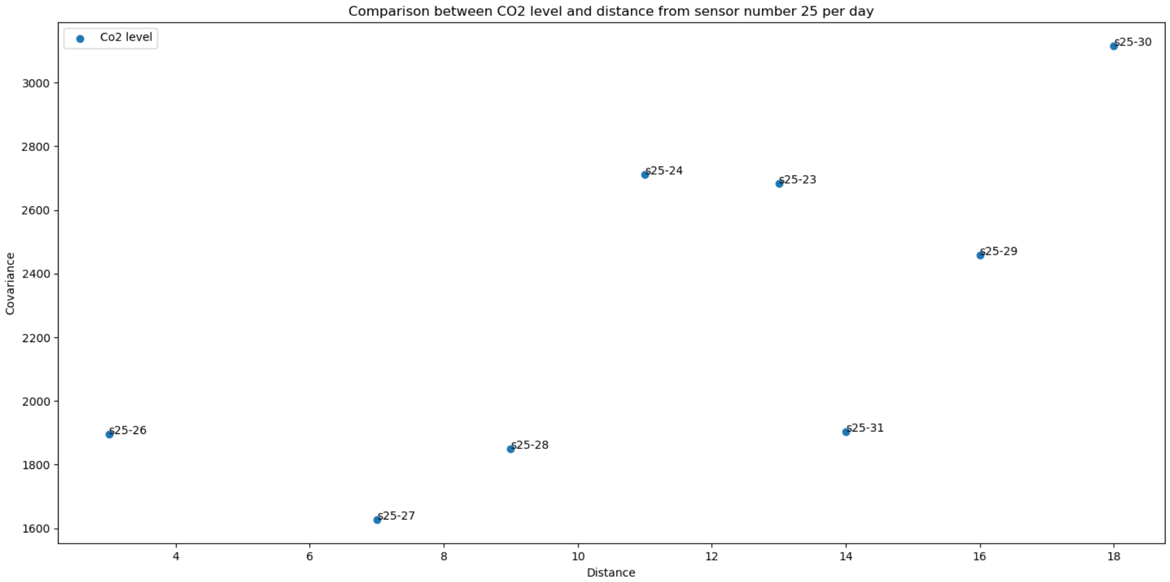


Figure 4-8 CO_2 level Covariance from multiple sensor in compare with sensor number 25 per one day

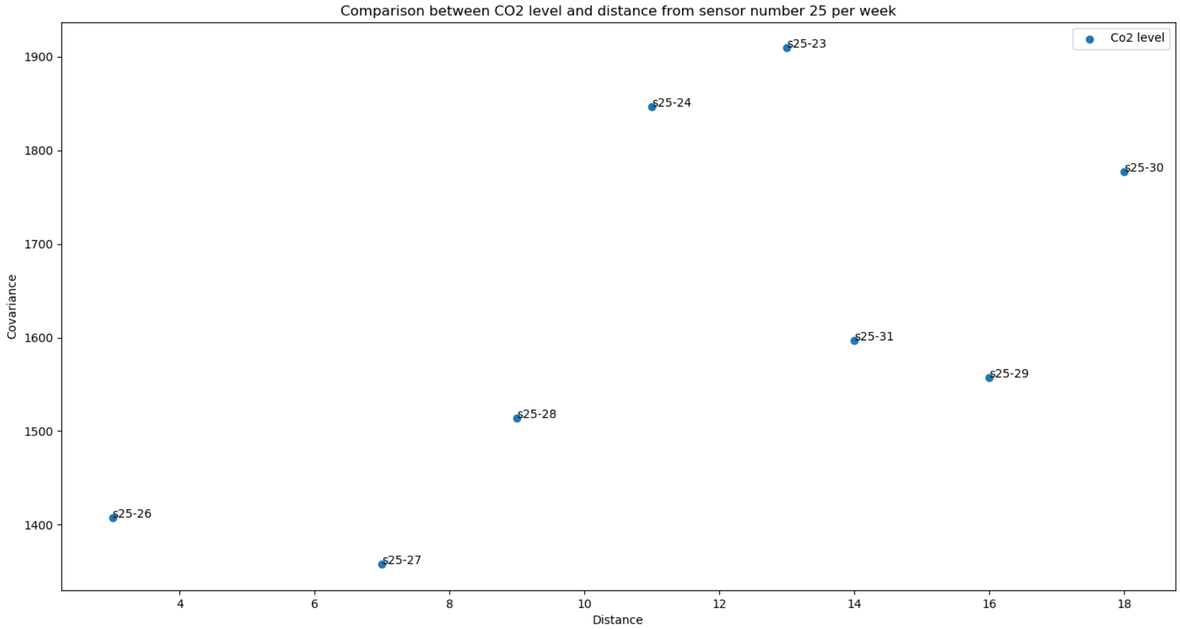


Figure 4-9 CO$_2$ level Covariance from multiple sensor in compare with sensor number 25 per one week

Figures (4-8) and (4-9) illustrate CO$_2$ covariance calculation using sensor number 25 as index sensor in compare with other sensors to find the relationship between sensors data. The correlation between distance and sensor is detectable, whenever two sensors are located far from each other their covariance's are higher. That might cause intersection between sensors. Using final LSTM model prediction over a week of data was performed to predict the number of people.

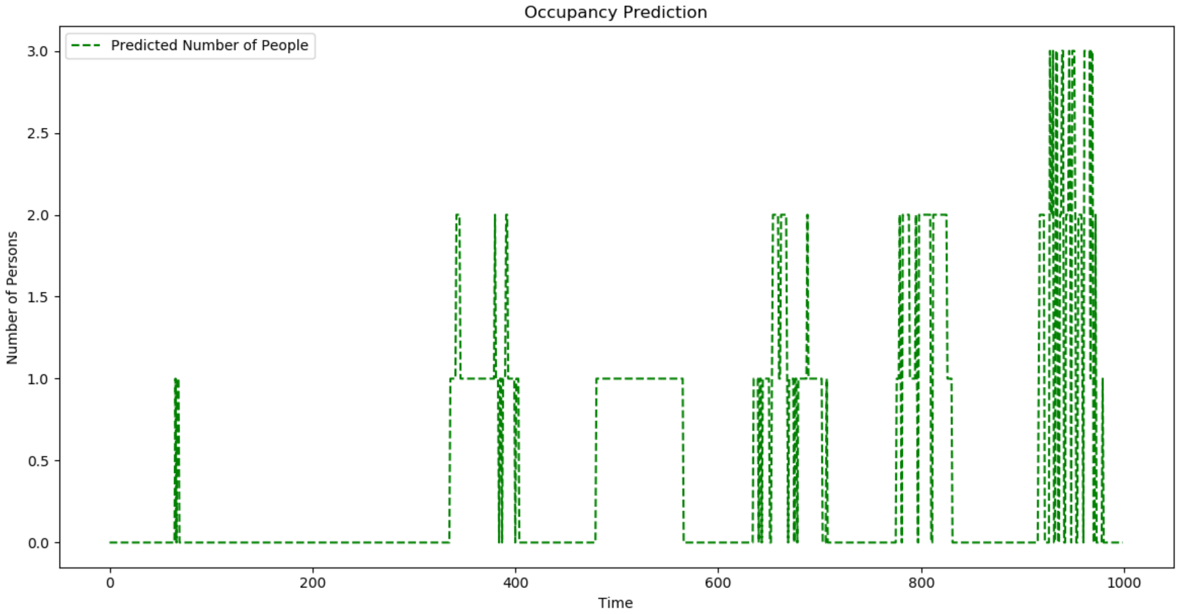


Figure 4-10 The number of people prediction based on sensor number 25

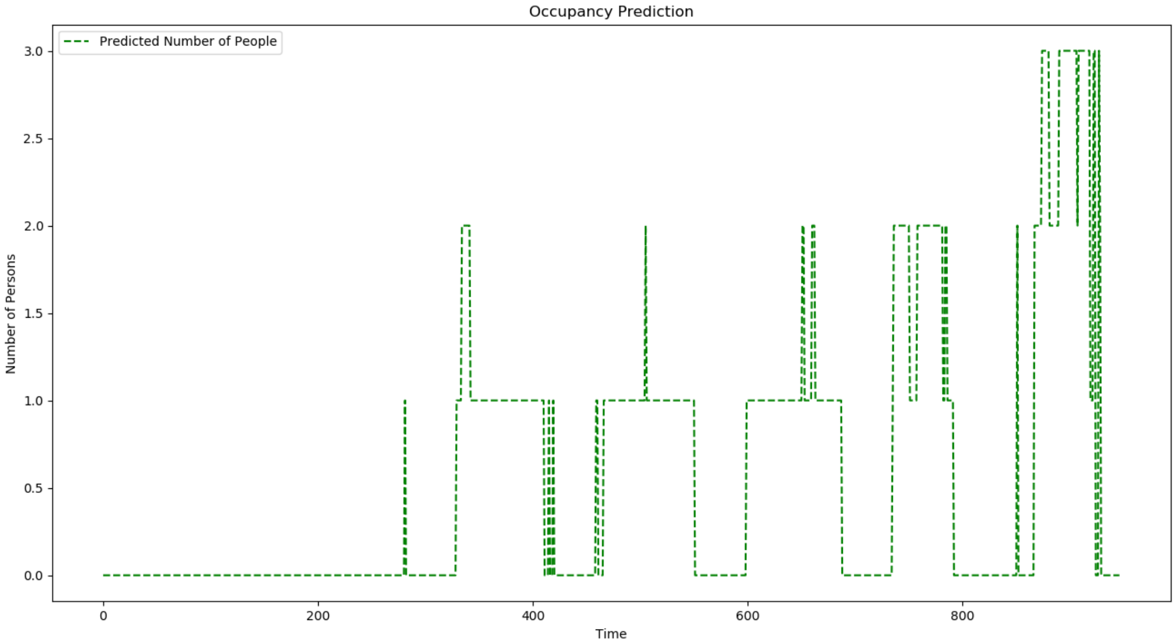


Figure 4-11 The number of people prediction based on sensor number 26

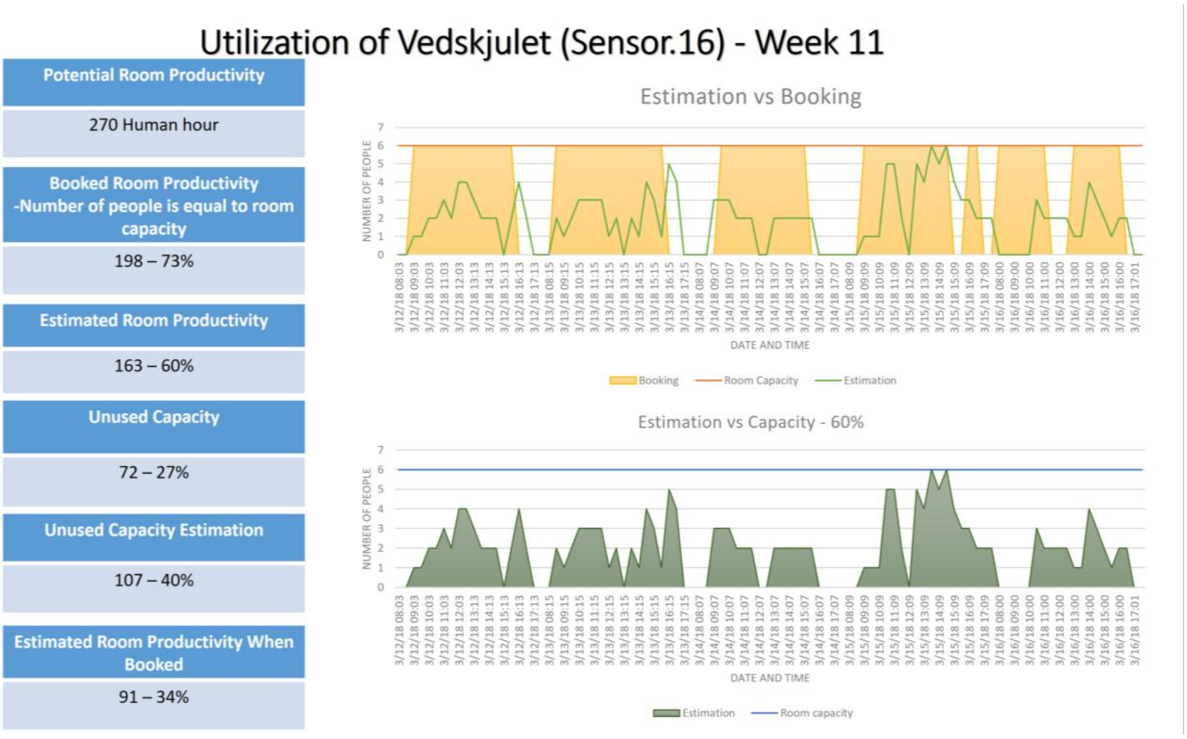
Figure (4-10) illustrates prediction on the number of people based on sensor number 25, we can see during weekend model was able to predict one person. And based on real observation for that day, two persons was present at that day. But the sensor was lost CO_2 level for beginning of that day.

Figure (4-11) demonstrates prediction on the number of people based on sensor number 26. Accumulating the predicted number of people based on these two sensors might be indicator for total number of people close to them.

5 Further experiments

5.1 Sample dashboard

In this section we briefly discussed different reports which produced during research. Since target business is highly interested in their meetings rooms usage. Based on sensor data and developed model a simple dashboard was presented to show the usage and effect of mounted sensors data and ML model. These data are useful source on building strategy planning. The number of people estimation shows their usage is lower than planned, also contradiction between booking system and room usage is observable. Most of calculation in this phase was performed on Microsoft Excel. Figures (5-1) (5-2) and (5-3) are three dashboards that created based on target building data.



Utilization of Vedskjulet (Sensor.16) - Week 11

Potential Room Productivity	270 Human hour
Booked Room Productivity -Number of people is equal to room capacity	198 – 73%
Estimated Room Productivity	163 – 60%
Unused Capacity	72 – 27%
Unused Capacity Estimation	107 – 40%
Estimated Room Productivity When Booked	91 – 34%

Figure 5-1 Dashboard created to compare room productivity and booking

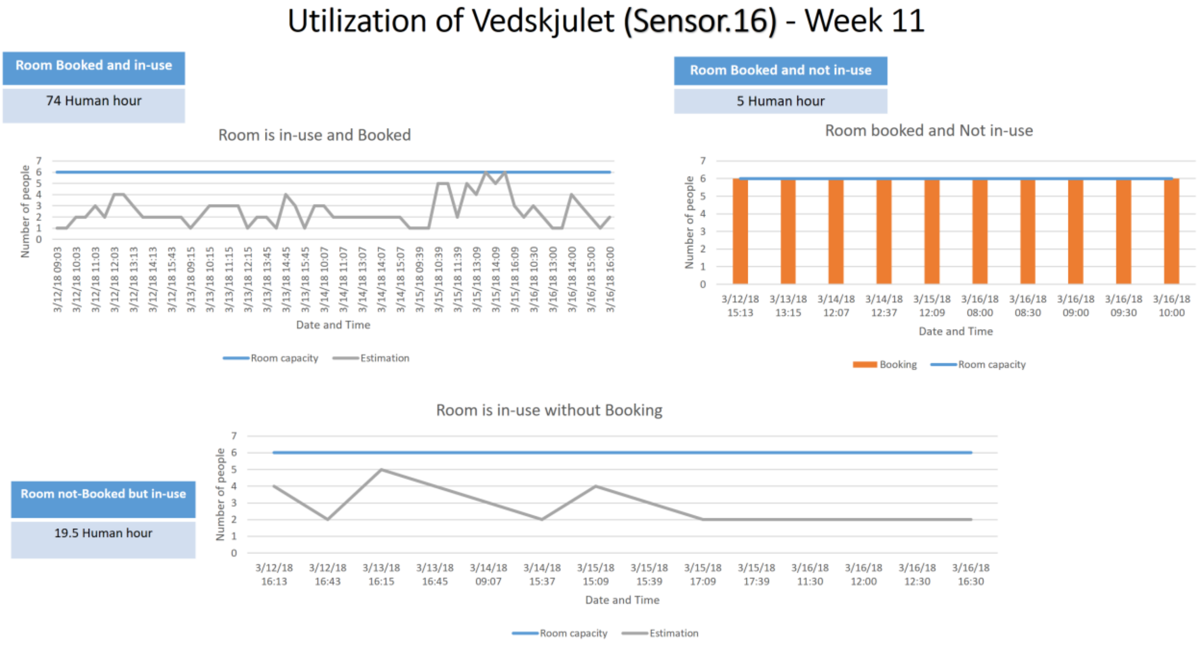
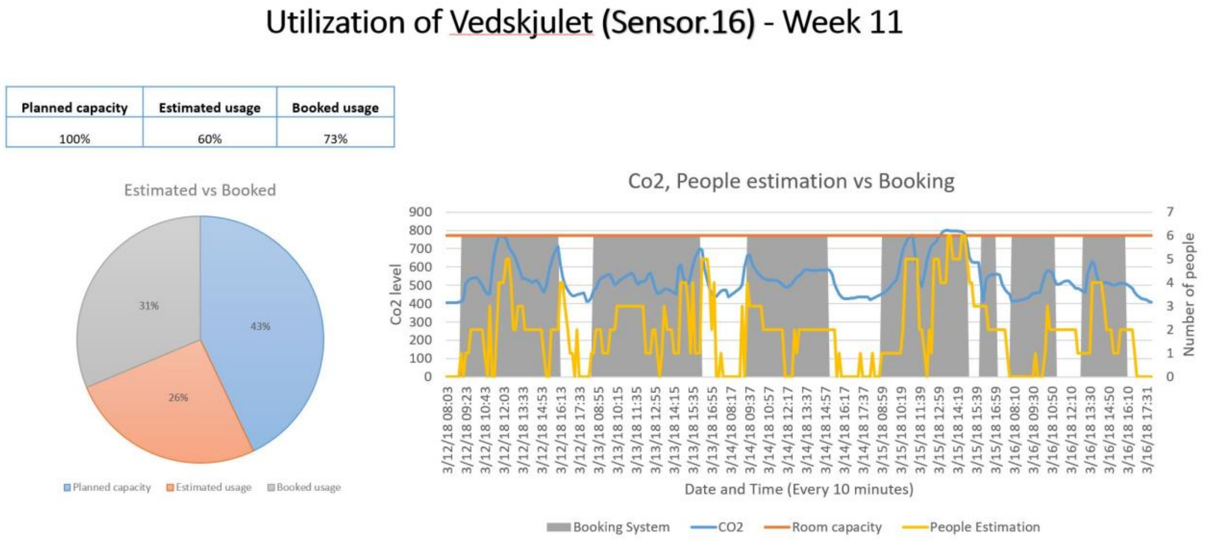


Figure 5-2 comparison between booking system and room usage



Planned capacity	Estimated usage	Booked usage
100%	60%	73%

Figure 5-3 CO$_2$ level, estimated number of people based on LSTM model

5.2 Experiments

During data gathering week at target building – Gjensidige, some valuable observations was captured which one can help to improve sensor functionality. One of them is, if the sensor was faced in front of door and rooms’ door is open, every outside movement detected as a motion, which can introduce outlier for model performance. Opening door is important event because some of their human sensitive sensors which can turn on lights that are located close to the door. Hence outside movement detected and consequently lights turn on which again can cause outliers in data analysis process.

Lights are turned off after 30 minutes if sensors do not recognize another motion. The building facade has high glass windows which can observe daily lights, these also can be resulting to detect light by sensors.

To see the effect of building type and age specially on CO_2 level, a sensor was mounted in an apartment with 11 years' old. Figure (5-4) shows CO_2 concentration on the old building when we know just two people are living there.

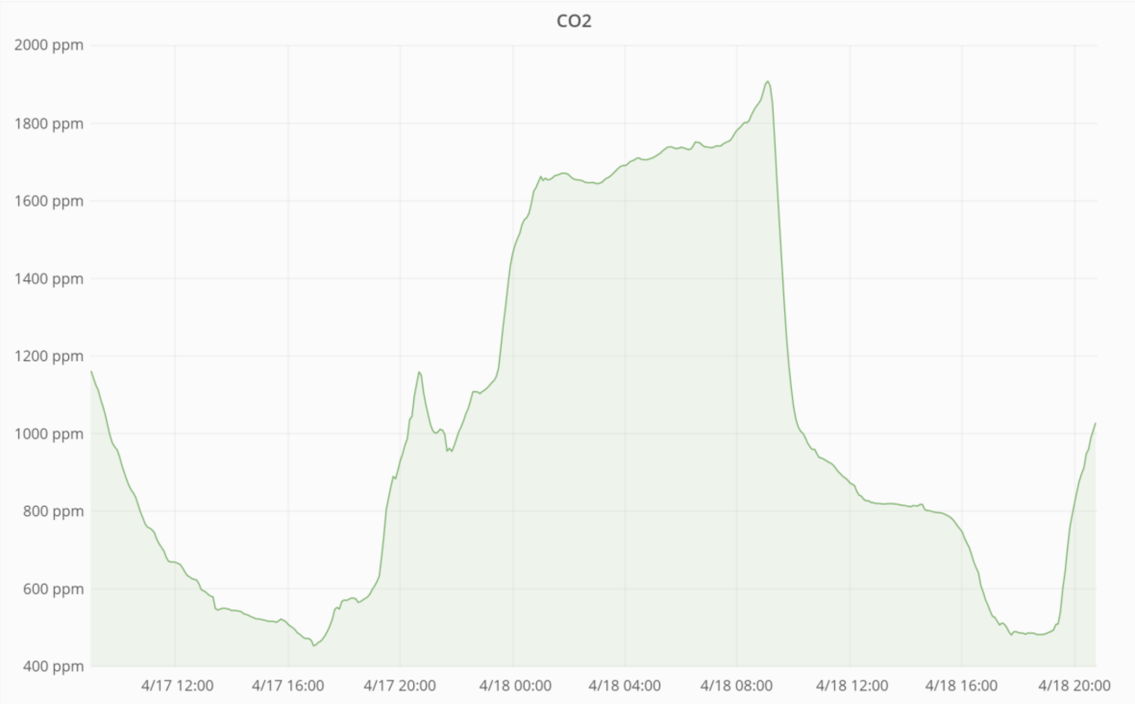


Figure 5-4 CO_2 concentration at apartment with 11 years' old

As it has shown on figure (5-4) the CO_2 concentration raises to 1900 ppm and it has minimum level 457 ppm when no one is present at that time.

Figure (5-5) shows CO_2 level and prediction of people at target building as same day as in apartment, while the maximum level of CO_2 is 595 ppm and the minimum level is 432 ppm, based on sensor data LSTM model can predict up to three persons.

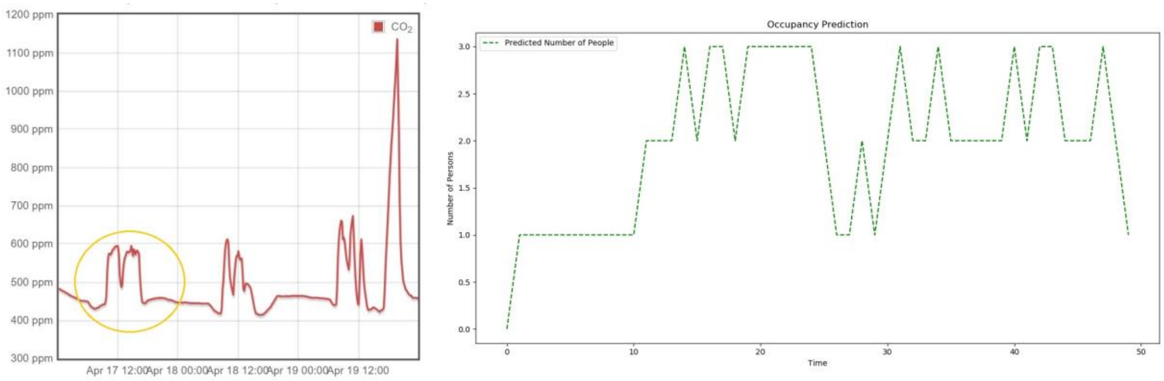


Figure 5-5 CO_2 concentration and number of people estimation

An experiment in apartment has been performed to check the effect of door situation on CO_2 concentration, as it has shown on figure (5-6) when the door is closed CO_2 level goes up to reach at 2584ppm which is approximately 1700ppm more than the moment when door was open.

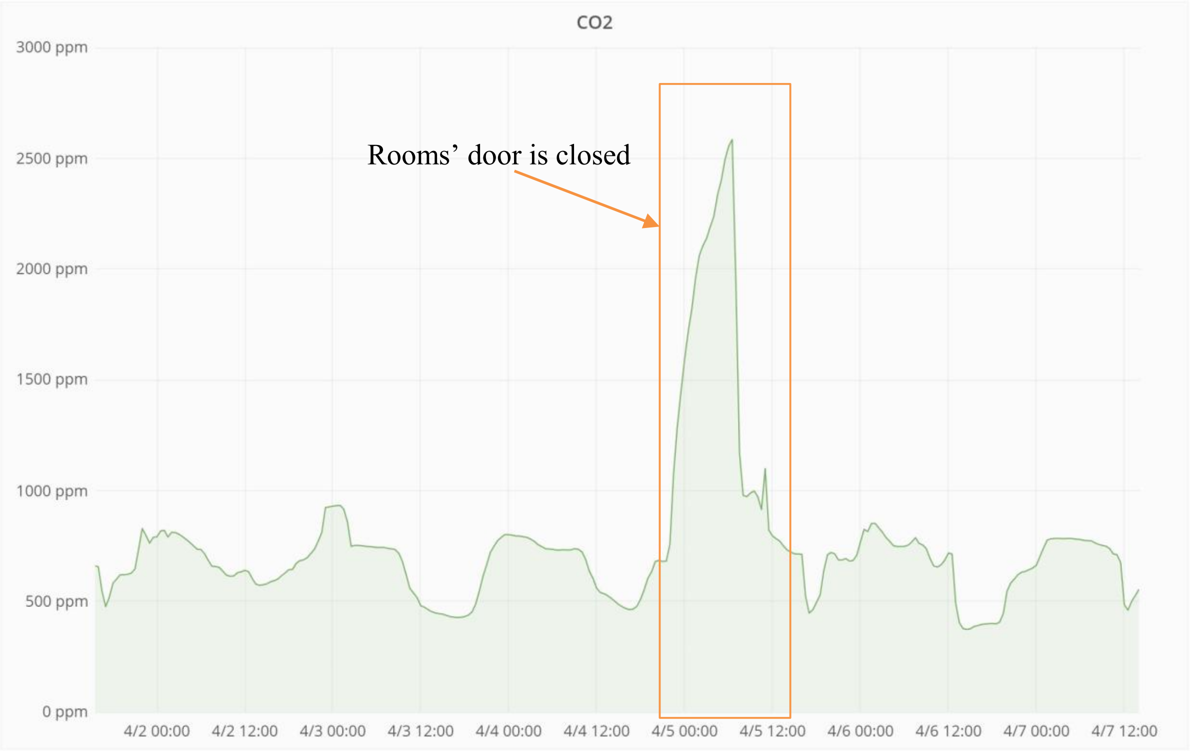


Figure 5-6 The effect of door on CO2 concentration

6 Discussion

In this work four different models were developed to estimate number of people within office spaces based on IAQ parameters. IAQ parameters was recorded using special sensor which is manufactured by Serinus technology. In this study, office area is divided in two main categories, close and open area. Close area is meeting rooms while open area is a floor with big common space, during development we found that handling close space data is easier than open office, because in open area there is more room for air flow which has effect on CO_2 and humidity.

The first research question is about to identify Parameters that make an impact on the perceived indoor climate, to answer this question we start with raw data analysis to gain insight about data then proceed with correlation coefficient method and PCA. Also among previous studies, one that done by (I. B. A. Ang, 2016) shows that CO_2 is the best ambient sensor predictor for detecting human presence, this can be starting point to use CO_2 as one of main parameters. Beside that another reason to use CO_2 as human indicator is coming from the human body metabolism and breathing process. To make proof on these claim, different methods was established, correlation coefficient which is used to show the strength of values in compare to each other. The result of that presents strong correlation between CO_2 and the number of people. Next step we proceed to make cluster map which again highlighted that correlation between those two parameters. Thus CO_2 has been chosen as first input parameter.

To discover another input parameters, a study which is about Building Level Energy Management Systems (BLEMS) project and done by (Z. Yang, 2012) gives motivation to take temperature and humidity in consideration. That study tries to use a combination of sensors (light, sound, motion, CO_2, temperature and humidity sensor) to create a simulation model to estimate the human occupancy. Also referring to the nature of human body inertia and cluster map presented in section 3.2.3 we decide to utilize humidity and temperature as second and third parameters.

Second or main question is to estimate number of people at each space, different machine learning algorithms was developed to predict respective human occupancy. The models run in two different type of spaces, close and open space. Each space has their own characteristic and the target building is modern structure with eco-light certificate which makes problem even more challenging. In section 4 we discussed each developed model in more details and we ended up to use LSTM with long history. While LSTM with long history gives better accuracy than other models, this model is more complex and time consuming. other models perform better on smaller spaces with capacity lower or equal to 6 and low level of CO_2 concentration.

As LSTM uses sequences to make prediction, it is better to train and test using sequential data, but in this research we trained and tested model using random values, one reason for that is lack of enough training data, another is we need to make general model which can be used by several rooms type. LSTM proposed as general model in this research.

Third research question is to identify movement flow within open space, despites the fact that this question is remained open due to privacy concerns, because we could not use camera, but some experiments was performed to address solution. At first attempt covariance of different parameters was calculated to show the relationship between each sensors. Using developed model shows there is intersection between each close sensors. However, the model can predict the number of people but identifying the intersection between sensors to estimate with high accuracy in open office is not easy. I suggest to use raw data analysis and image processing to get better result.

Based on models' performance LSTM was selected as general algorithm. The input feature vector requires weekday and capacity of room, because these two values help validation to find relevant data. Thus input vector is containing:

[$CO_{2,\ t-30}$, $CO_{2,\ t-60}$, $CO_{2,\ t-90}$, $CO_{2,\ t-120}$, Humidity_{t-30}, …, $\text{Temperature}_{t-30}$, …, Weekday, room Capacity and number of people]

The best accuracy is obtained when model was trained using multiple datasets, means data from rooms with different capacities was merged to make common dataset. A reason for that is neural networks generally need a lot of data to get the weights right.

7 Conclusion

In this work IAQ parameters of Gjensidige, a leading Nordic insurance group, in Oslo is analyzed. This work represents case study and it aims to estimate number of people using ML models for captured IAQ data. The main domain of time-series forecasting has been statistical analysis for a long time until machine learning models became more popular. So the first research question is about identifying perceived air parameters. As turned out CO_2, humidity and temperature had impact on perceived indoor air, because correlation coefficient shows the strong relation between CO_2, temperature and number of people, and we selected humidity during experimenting with different input vectors.

There is also moderate correlation between number of people and motion, but this parameter cannot be used because during experimental phase, we observed that if sensor was faced to the door then can detect outside motion and that means a noisy data because outside movement does not indicate the human presence at the close space.

Then to answer second which is main question different model was developed and it was showed that LSTM outperformed other three models, but the model needs to be trained using multiple rooms data set. Also the model is trained using random values to increase accuracy. Afterward, during training the model learned to capture best matches. Third research question is to identify movement flow within open office, based on this sensor, number of people can be estimated but we cannot identify intersection between sensors, so to answer this question we need to use other sensors like camera, which in this study it is not possible to use because of privacy issues.

The indoor air parameter captured by sensor can be used in BMS systems. Identifying events related to each trend is really important for these data. An example is the effect of door situation on CO2 concentration because sometimes we observe slightly decrease in CO2 level while we know the number of people does not change at the time. These type of even can help to boost the accuracy of model and make better analysis over data.

Works Cited

(BSD), s.-l. d. (2007 - 2017). *scikit-learn*. Retrieved from http://scikit-learn.org/stable/index.html

A.C.K. Lai a, K. M. (2009). An evaluation model for indoor environmental quality (IEQ) acceptance in residential buildings. *Elsevier*.

Arief-Ang, I. B. (2017). CD-HOC: Indoor Human Occupancy Counting using Carbon Dioxide Sensor Data. *CoRR*, 24. Retrieved from http://arxiv.org/abs/1706.05286

Ba, D. P. (2014). Adam: A Method for Stochastic Optimization. *CoRR*.

Chollet, F. (2014-2017). *Keras*. Retrieved from https://keras.io/

developers, s.-l. (2007 - 2017). *Importance of Feature Scaling*. Retrieved from scikit-learn: http://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html

developers, s.-l. (2007-2017). *Principal component analysis (PCA)*. Retrieved from scikit-learn: http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Graves, A. (2012). Neural Networks. In *Supervised Sequence Labelling with Recurrent Neural Networks* (p. 141). Springer.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks.* Berlin, Heidelberg: Springer. doi:https://doi.org/10.1007/978-3-642-24797-2

I. B. A. Ang, F. D. (2016). Human occupancy recognition with multivariate ambient sensors. *IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 1-6.

Irvan B. Arief-Ang, F. D. (2017). Indoor Human Occupancy Counting using Carbon Dioxide Sensor Data. *CoRR*.

J.Smola, A. (2004). A tutorial on support vector regression. 24.

John Joseph Valletta, C. T. (2017). Applications of machine learning in animal behaviour studies. *elsevier*, 203-220.

Klepeis NE, N. W. (2001). The National Human Activity Pattern Survey (NHAPS): a resource forassessing exposure to environmental pollutants. *Exposur Analysis and Environmental Epidemiology*, 231-252. Retrieved from https://www.ncbi.nlm.nih.gov

Luis Pedro Coelho, W. R. (2015). *Building Machine Learning Systems with Python Second Edition.* Birmingham: Packt Publishing Ltd.

Mamidi, S. a.-H. (2012). Improving building energy efficiency with a network of sensing, learning and prediction agents. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (pp. 45--52). Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems.

Mitchell, T. M. (1997). Machine Learning. *McGraw Hill*.

Sepp Hochreiter, J. S. (1997). LONG SHORT-TERM MEMORY. *Neural Computation*, p. 32.

Suthaharan, S. (2016). *Machine Learning Models and Algorithms for Big Data Classification.* Springer.

V. L. Erickson, Y. L. (2009). Energy efficient building environment control strategies using real-time occupancy measurements. *ACM*, 19-24.

Vinyals, W. Z. (2014). Recurrent Neural Network Regularization. *CoRR*, 8. Retrieved from http://arxiv.org/abs/1409.2329

Yuhuang Hu, A. H.-C. (2018). OVERCOMING THE VANISHING GRADIENT PROBLEM IN PLAIN RECURRENT NETWORKS. *CoRR*.

Z. Yang, N. L.-G. (2012). A multi-sensor based occupancy estimation model for supporting demand driven hvac operations. In *Proceedings of the 2012 Symposium on Simulation for Architecture and Urban Design* (pp. 2:1--2:8). San Diego, CA, USA: Society for Computer Simulation International.

Appendix

Appendix 1 - Source code

The complete source code uploaded at https://source.uit.no/fhe017/Thesis-SorceCodes/tree/master

Appendix 2 – Project description

Faculty of Engineering Science and Technology
Department of Computer Science and Computational Engineering
UiT The Arctic University of Norway

Study of area use and floor space occupation in office buildings

A machine learning approach

Fatemeh Heidari

Thesis for Master of Science in Computer Science

Problem description

The MSc-thesis will investigate and document area use and floor space occupation in office buildings. Transient and permanent behavior are both of interest. For this data feeds and machine learning is required.

The property and office market needs to know more on how patterns of floor space and office areas are used. This has important economic implications. It is essential for how comfortable users feel during the day and for collaboration as well as for production efficiency. The basic questions to be asked are how people's working experience can be enhanced and how more cost effectiveness can be achieved.

For this purpose, sensors are deployed that measure different aspects of indoor climate such as CO2, humidity, light, noise level and temperature. The student will capture time series from different sensors providing this type of data and use LTSM networks, multi-regression methods and/or similar techniques to determine the following:

1. Parameters that make an impact on the perceived indoor climate and how they are correlated
2. Determine the number of people in a room or different office spaces based on these parameters
3. Determine movements of people within an office area (transient use)

Two data sets will be made available. The primary data will be sourced from the company View-Serinus. It will consist of a training set and a test set. A supporting data set will be sourced from UiT and will serve as validation case for the other.

In addition to the three tasks specified above the student should attempt a generalization. By using data from multiple rooms and the two different sites create a generic empirical model that allows no or less supervised training. This implies that sensors can be placed in rooms and after calibration produce reliable results with not additional training.

In addition, the student can use and request additional data from other sources. This could be floor space area, calendar, room booking systems, state information such as doors open/windows open, weather data, specifications on technical systems such as lights and ventilation. Data that suggests the number of people in a building at any time can also be used. This could be the number of cars in the parking area, traffic on the outside of the building and similar. The idea is to synthesize this, determine dependencies and specify what matters regarding the four tasks specified above.

It is also possible to explore the use of additional sensors to support the tasks specified to improve accuracy. One example could be to augment the set up with an IR camera with low resolution that does not violate privacy regulations. This has been applied in Asbjørn Danielsen's PhD work.

The effort will be carried out in cooperation with View-Serinus and Chronos on behalf of Cushman & Wakefield. The results are considered confidential, but it is the aim of UiT to support the candidate and allow external publishing of the results at an international conference on AI. This will depend on the quality of the result and the conclusions drawn.

Location of work: Oslo

Prerequisites: A signed non-disclosure agreement

Dates

Date of distributing the task:	<16.01.2018>
Date for submission (deadline):	<01.06.2018>

Contact information

Candidate	Fatemeh Heidari fahaydari@gmail.com
Advisor at UiT-IVT	Bernt A. Bremdal bernt.a.bremdal@uit.no
Advisor at UiT-IVT	Asbjørn Danielsen asbjorn.danielsen@uit.no
External advisor (optional)	Rolv-Møll Nilsen, Serinus rmn@serinustechnology.com

General information

This master thesis should include:

- ❄ Preliminary work/literature study related to actual topic
  - A state-of-the-art investigation
  - An analysis of requirement specifications, definitions, design requirements, given standards or norms, guidelines and practical experience etc.
  - Description concerning limitations and size of the task/project
  - Estimated time schedule for the project/ thesis
- ❄ Selection & investigation of actual materials

- Development (creating a model or model concept)
- Experimental work (planned in the preliminary work/literature study part)
- Suggestion for future work/development

Preliminary work/literature study

After the task description has been distributed to the candidate a preliminary study should be completed within 3 weeks. It should include bullet points 1 and 2 in “The work shall include”, and a plan of the progress. The preliminary study may be submitted as a separate report or “natural” incorporated in the main thesis report. A plan of progress and a deviation report (gap report) can be added as an appendix to the thesis.

In any case the preliminary study report/part must be accepted by the supervisor before the student can continue with the rest of the master thesis. In the evaluation of this thesis, emphasis will be placed on the thorough documentation of the work performed.

Reporting requirements

The thesis should be submitted as a research report and could include the following parts; Abstract, Introduction, Material & Methods, Results & Discussion, Conclusions, Acknowledgements, Bibliography, References and Appendices. Choices should be well documented with evidence, references, or logical arguments.

The candidate should in this thesis strive to make the report survey-able, testable, accessible, well written, and documented.

Materials which are developed during the project (thesis) such as software / source code or physical equipment are considered to be a part of this paper (thesis). Documentation for correct use of such information should be added, as far as possible, to this paper (thesis).

The text for this task should be added as an appendix to the report (thesis).

General project requirements

If the tasks or the problems are performed in close cooperation with an external company, the candidate should follow the guidelines or other directives given by the management of the company.

The candidate does not have the authority to enter or access external companies’ information system, production equipment or likewise. If such should be necessary for solving the task in a satisfactory way a detailed permission should be given by the management in the company before any action are made.

Any travel cost, printing and phone cost must be covered by the candidate themselves, if and only if, this is not covered by an agreement between the candidate and the management in the enterprises.

If the candidate enters some unexpected problems or challenges during the work with the tasks and these will cause changes to the work plan, it should be addressed to the supervisor at the UiT or the person which is responsible, without any delay in time.

Submission requirements

This thesis should result in a final report with an electronic copy of the report including appendices and necessary software, source code, simulations and calculations. The final report with its appendices will be the basis for the evaluation and grading of the thesis. The report with all materials should be delivered according to the current faculty regulation. If there is an external company that needs a copy of the thesis, the candidate must arrange this. A standard front page, which can be found on the UiT internet site, should be used. Otherwise, refer to the "General guidelines for thesis" and the subject description for master thesis.

The advisor(s) should receive a copy of the the thesis prior to submission of the final report. The final report with its appendices should be submitted no later than the decided final date.