JÖNKÖPING UNIVERSITY
School of Engineering

Categorizing conference room climate using K-means

PAPER WITHIN Computer Science
AUTHORS: *Jin Asp, Saga Bergdahl*
TUTOR: *Niklas Lavesson*
JÖNKÖPING 2019 June

This exam work has been carried out at the School of Engineering in Jönköping in the subject area computer science. The work is a part of the three-year Bachelor of Science in Engineering program. The authors take full responsibility for opinions, conclusions, and findings presented.

Examiner: Ulf Johansson
Supervisor: Niklas Lavesson
Scope: 15 credits
Date: 2019-06-26

Postadress:	Besöksadress:	Telefon:
Box 1026	Gjuterigatan 5	036-10 10 00 (vx)
551 11 Jönköping		

Abstract

Smart environments are increasingly common. By utilizing sensor data from the indoor environment and applying methods like machine learning, they can autonomously control and increase productivity, comfort, and well-being of occupants.

The aim of this thesis was to model indoor climate in conference rooms and use K-means clustering to determine quality levels. Together, they enable categorization of conference room quality level during meetings. Theoretically, by alerts to the user, this may enhance occupant productivity, comfort, and well-being. Moreover, the objective was to determine which features and which *k* would produce the highest quality clusters given chosen evaluation measures.

To do this, a quasi-experiment was used. CO_2, temperature, and humidity sensors were placed in four conference rooms and were sampled continuously. K-means clustering was then used to generate clusters with 10 days of sensor data. To evaluate which feature combination and which *k* created optimal clusters, we used Silhouette, Davis Bouldin, and the Elbow method.

The resulting model, using three clusters to represent quality levels, enabled categorization of the quality of specific meetings. Additionally, all three methods indicated that a feature combination of CO_2 and humidity, with $k = 2$ or $k = 3$, was suitable.

Keywords

Ubiquitous computing, model development and analysis, unsupervised learning, smart environment, indoor climate.

Postadress:	Besöksadress:	Telefon:
Box 1026	Gjuterigatan 5	036-10 10 00 (vx)
551 11 Jönköping		

Acknowledgment

We would like to thank our supervisor Professor Niklas Lavesson for his expertise, ideas, feedback, time, and encouragement during the course of our thesis.

Another thanks to ROL Ergo and our contacts there, whom we are grateful to for providing us with daily updated data, assistance with sensors, insight in their business, and access to their facility.

This thesis was written within the scope of the Mining Actionable Patterns from complex Physical Environments (MAPPE) project. It is a three-year research project at Jönköping University in collaboration with ROL Ergo and Saab AB, which is financed by the Knowledge Foundation.

At last, a special thanks to our classmates for their continuous cooperation and time devoted to providing us with feedback.

1 Introduction

Weiser (1999) proposed a vision of the future where computers would become integrated in our everyday lives to the point where we would cease to notice them. To explain this, he coined the term ubiquitous computing, also known as pervasive computing. This trend has emerged in different areas, smart environments being one of them (Satyanarayanan, 2001; Davies & Clinch, 2017).

A smart environment can be described as "one that is able to acquire and apply knowledge about the environment and its inhabitants in order to improve their experience in that environment" (Cook & Das, 2007, p. 54). A common setting for this type of environment is in office buildings.

There is a distinct relationship between human behavior and indoor air quality (Lin et al., 2017). It has also been concluded that occupancy can be determined by measuring CO_2, temperature, and humidity (Candanedo & Feldheim, 2016; Szczurek, Maciejewska & Pietrucha, 2017). There are approaches that use machine learning to control HVAC (Heating, Ventilation, Air Conditioning) by learning from the behavior of the users and adapting toward their needs (Peng, Rysanek, Nagy & Schlüter, 2018). Machine learning can be defined as "the systematic study of algorithms and systems that improve their knowledge or performance with experience" (Flach, 2012, p. 3).

By analyzing indoor climate (air quality, temperature, space, acoustics, lighting) with statistical models or machine learning to enable autonomous smart environments, it is possible to increase well-being and productivity of occupants and decrease energy consumption (Cook & Krishnan, 2014; Mozer et al. 1995; Peng et al., 2018). In order to implement a system that applies machine learning effectively, it is necessary to model the indoor environment to determine which factors are relevant and suitable to consider.

This thesis was in collaboration with ROL Ergo, a company specialized in smart office solutions and activity-based workplaces (ABW's). The purpose of ABW's is to reduce costs, increase flexibility and save space (Rolfö, Eklund & Jahncke, 2018). ROL Ergo is a part of the Mining Actionable Patterns from complex Physical Environments (MAPPE) research project at Jönköping University, which is about creating machine learning solutions that can provide understandable explanations of predictions. According to ROL Ergo, what is needed is technology that goes further than describing the data collected. Users desire technology that can automatically predict and explain e.g. office and resource use, productivity, or user satisfaction based on sensor data. A possible use of this technology could be an algorithm that predicts room quality and alerts the user about current room quality status.

As literature shows, there are approaches to how smart environments can be used to control indoor climate, and research about which factors can be controlled to enhance productivity. However, as far as we are aware, there is no research about how indoor climate in conference

rooms can be modeled to categorize room quality level. Therefore, this thesis aimed to contribute to the knowledge base by creating a model of indoor climate in conference rooms and using K-means clustering to determine quality levels, which enables categorization of quality level of meetings. Moreover, the objective was to determine which features and which *k* produces the highest quality clusters given chosen evaluation measures.

2 Background

The outline of this chapter is as follows. First, smart environments are reviewed followed by indoor climate. The potential relationship between indoor climate, occupant productivity, and well-being is explained. Thereafter, the possible connection between smart environments and indoor climate, and implementation of smart environments using machine learning and data mining is described. Lastly, the research problem is addressed.

2.1 Smart environments

A smart environment continuously perceives the environment and based on the goals and outcomes makes automated decisions about what actions to take to change the state of the environment. It can be defined as "one that is able to acquire and apply knowledge about the environment and its inhabitants in order to improve their experience in that environment" (Cook & Das, 2007). The components range from physical sensors and actuators, software interfaces and sensor networks, methods for data mining and predictions, to decision making.

Different domains where smart environments are applied encompass traditional infrastructure, industry sectors, and personalized digital services (Curry & Sheth, 2018). Liang, Cao, Liu and Liang (2016) envision a smart world, where advanced techniques like advanced networks, ubiquitous sensing, and collaborative computation are used to enable a more productive, safe, efficient, and connected world.

A smart world contains smart cities, which are cities that through investments in human and social capital and communication infrastructure promote a high quality of life and economic growth, while managing natural resources wisely (Caragliu, Del Bo & Nijkamp, 2011).

Smart buildings and indoor environments like homes and offices are a part of smart cities. They use multiple sensors and actuators to react to the occupants and/or the utilities without the occupants' need to intervene with the system (Torunski, Othman, Orozco & El Saddik, 2012). According to De Silva, Morikawa and Petra (2012), there are three main categories of smart homes in the literature; ones that support the well-being of inhabitants by detecting behavior, ones that store and retrieve multimedia captured within the home, and ones that deal with surveillance by detecting information that can raise alarms.

2.2 Indoor climate

Indoor climate is important not only for energy consumption saving, but it is indicated to affect the occupants' comfort, health, and productivity. Indoor climate can be defined as; thermal

environment, air quality, acoustics, and lighting (European Standard, 2006). These factors are indicated in the literature to affect productivity in indoor environments. Additional factors are office layout, biophilia and views, look and feel, and location and amenities (Al Horr et al., 2016).

Productivity, defined as ratio of output to input (Al Horr et al., 2016), is not just about speed of work, it is about mental and physical health (Browning, 2012). It is indicated that humans experience a higher level of happiness in a natural environment. Biophilia is a hypothesis that implies that there is an instinctive bond between humans and their attraction to nature (Wilson, 1984). Therefore, biophilic design, meaning built environment integrated with nature, is often seen as a luxury for employers. In workplaces, this encompasses plant life, water, breezes, sounds, scents, and other natural elements (Browning, 2012).

Indoor thermal comfort is a major factor that affects productivity of occupants. An office environment that is satisfactory for occupants may reduce complaints and absence while increasing productivity. Comfort has positive effects on well-being and can be defined as the absence of unpleasant sensations. It constitutes the physical environment (air, climate), functional environment (disturbances, resources), and psychological environment (privacy, territory) (Al Horr et al., 2016).

A high CO_2 concentration in the air is connected to physiological changes, which lead to a decrease in the user's functional ability. To improve productivity, there needs to be appropriate ventilation in relation to the number of occupants in a building (Vehviläinen et al., 2016). Human metabolism is the main source of CO_2 in indoor environments. No toxic effects or cognitive performance losses are expected below 10 000 ppm (Zhang, Wargocki & Lian, 2016) but at levels above 10 000 ppm there are measurable effects on the respiratory system (Maresh et al., 1997) as well as heart rate and systolic blood pressure (Bailey, Argyropoulos, Kendrick & Nutt, 2005). Arbetsmiljöverket (Swedish Working Environment Agency) (2009) state that the ambition should be to keep CO_2 below 1000 ppm. They also specify that indoor air temperature should be kept at 20-24 °C in winter and 20-26 °C in summer. It has been indicated that CO_2, temperature, and humidity correlate in indoor environments, since they are exuded by humans (Lazovic, Stevanovic, Jovasevic-Stojanovic, Zivkovic & Banjac, 2016).

Very low or very high indoor humidity can cause discomfort since the humidity affects the perception of air temperature. Humidity below 40% is considered low, and can cause eye, skin, and mucous membrane irritation. At 30% and below, skin irritation and static electricity are the biggest concerns (Derby & Pasch, 2017).

2.3 Intelligent analysis of indoor climate for increased well-being

A smart environment may control indoor environment factors to optimize energy consumption as well as productivity and well-being. This often involves control of HVAC in buildings. By control and reduction of the consumption of energy, an optimal indoor climate can be maintained for both environment and cost purposes. For example, by detection of occupancy,

the HVAC system and lighting in the building can be optimized (Candanedo & Feldheim, 2016).

To make a smart environment further understand the context of the environment and make automated decisions, intelligent analysis is necessary. Automated support can be provided by analysis of indoor climate data from sensors with algorithms and machine learning techniques like data mining (Cook & Krishnan, 2014).

Drawing from the pervasive computing vision of Weiser (1999) and other research motivated by it, Davies and Clinch (2017) identify a new research area called pervasive data science. It is characterized by "a focus on the collection, analysis (inference) and use of data (actuation) in pursuit of the vision of ubiquitous computing" (Davies & Clinch, 2017, p. 1). They consider smart environments to be one of the most obvious applications of pervasive data science and include algorithms for processing pervasive sensor data as a topic of research. In practice, this could involve smart environments which utilize sensor data to understand the context of the environment and make autonomous decisions regarding its climate.

There are several approaches to control an indoor climate in a smart environment. Merabet, Essaaidi, Benhaddou, Khalil and Chilela (2018) developed a model, which with environment sensor data as well as information about the occupants, could predict their thermal comfort and automatically adjust the environment accordingly. With the same purpose in mind, Peng et al. (2018) used supervised and unsupervised learning to develop a control strategy which responded to occupant behavior and successfully controlled the cooling system in an office.

2.4 Problem description

The problem is associated to how the indoor environment can be modeled with relevant data to enable use of machine learning e.g. autonomous decisions, pattern recognition or predictions.

In order to utilize sensor data from the environment for machine learning solutions, it is necessary to model the specific environmental factors. Hence, to model indoor climate in an office environment, suitable factors need to be considered. Humans prefer different types of indoor climate and it may be difficult to create statistical models with subjective data based on perception. Useful connections can be drawn between objective data that is indicated by research to affect the brain capacity e.g. CO_2 and temperature (Vehviläinen et al., 2016), as well as humidity. There are many different factors that can be studied and modeled to draw connections, and one can use statistical models to determine which relationships are relevant.

Specifically, this thesis aims to contribute to the knowledge base by modeling conference room environments, using clustering to categorize meeting room quality, and evaluating the clustering performance. This knowledge can be used in a smart office environment to enhance the comfort and productivity. Furthermore, it can provide means to create more energy efficient environments by establishing awareness of the indoor climate.

3 Related work

The literature shows different ways in which smart environments are used to monitor human activity, analyze human behavior's impact on indoor climate, and decrease energy consumption. It also shows how environmental data can be analyzed with K-means clustering.

Mozer et al. (1995) conducted research on home automation using machine learning. They saw potential in a system that operates the home and adapts to the behavior of the occupants, both to meet their needs and the energy consumption goals of the home. They explored neural network reinforcement learning and prediction techniques as tools to use adaptive control in a residence that was equipped with multiple sensors and actuators.

To study how human behavior affects indoor air quality, Lin et al. (2017) collected behavior data from sensors and chemical indoor air quality measurements in two smart home environments. They used machine learning algorithms to see what indoor air quality factors were impacted by smart home features. The results showed that there is a strong relationship between human behavior and indoor air quality. This result is useful knowledge to us when we select factors to measure and analyze.

Candanedo and Feldheim (2016) measured the accuracy of predictions regarding occupancy in a room using light, temperature, CO_2, and humidity data. Szczurek et al. (2017) did a similar study with a time series of CO_2 concentration, temperature, and humidity to determine occupancy during a 60-minute period as well as determine duration of occupancy periods. The experiment was successful, and it was concluded that certain measured amounts of CO_2, temperature, and humidity could be used to determine occupancy. This supports our decision to study temperature, CO_2 and humidity, since this indicates that they are affected by utilization.

To decrease HVAC energy consumption in an office building, Peng et al. (2018) used a control strategy based on supervised and unsupervised learning that responded to the occupants' behavior. It was used to control an office cooling system and succeeded to improve energy savings by between 7% and 52% compared to a conventional cooling system.

Merabet et al. (2018) conducted an experiment where a sensor network measured the thermal comfort of users and adjusted the environment accordingly. They concluded that biometric information about the occupants, like height and waist size, could be used to create a logistic regression predictive model of thermal comfort. They said their results can pave the way to systems that can predict comfort of occupants and automatically adjust the environment.

Li, Logenthiran, Phan and Woo (2018) proposed a power alert system for a smart home with an energy management system. Using the K-means clustering algorithm, they classified power consumption into three levels (high, average, low). They evaluated the clustering performance and concluded that the system succeeded to perform reliable clustering and created energy consumption awareness for residents. Since K-means clustering was successfully used to

cluster power consumption levels in a smart home, we consider it a suitable method to cluster indoor environmental factors since it is a related area with similar data.

Evaluation of clustering performance can be used to verify that the clustering algorithm that is used achieves a satisfactory result. Celestino et al. (2018) used several evaluation measures, e.g. Silhouette, to see if K-means clustering with a reduced dimension of features was preferable to using a high dimension of features. The conclusion was that the method that reduced the number of features improved clustering performance. Since we do not know which *k* is suitable, we too aim to use Silhouette, among other methods, to verify that we get a fair result.

The literature contains a substantial amount of research in these areas, but there was no research found in our literature search that specifically addresses how indoor climate can be modeled to analyze conference room quality with K-means clustering. Neither was there any research about categorization of conference room quality during meetings to provide feedback to the occupants.

4 Aim and Scope

The purpose of this thesis is:
To model indoor climate factors in conference rooms and use K-means clustering to categorize quality levels, use measures to evaluate which feature combination and which k that produces the best results, and to evaluate the model by fitting meetings into it.

This is done with a quasi-experiment in smart office conference rooms to measure and analyze CO_2, temperature, and humidity sensor data in connection to room quality. The K-means clustering algorithm is used to cluster the data which enables categorization of conference room quality during meetings. The clustering is then evaluated to determine the best clustering performance depending on feature selection and *k*, number of clusters. To evaluate the model further, meetings are fitted into it to determine meeting quality.

The purpose is supported by the following research question:

- How can CO_2 level, temperature, and humidity in a conference room be modeled to categorize room quality?

5 Method

The objective of the study was to determine how indoor climate factors best could be modeled to categorize room quality. This included feature engineering, design and application of the experiment, and use of evaluation measures to analyze the results. The literature review led us to the selection of indoor climate factors that may be suitable to study. Our study had an empirical quantitative approach and used a quasi-experiment as research method.

5.1 Experiment design

The aim of the quasi-experiment was to answer the research question. A quasi-experiment has a similar purpose as true experiments; testing a hypothesis about manipulable causes. The biggest difference is that a quasi-experiment lacks random assignment and does not necessarily indicate a true causal relationship (Shadish, Cook & Campbell, 2002). It was not reasonable to assign employees who used the conference rooms randomly since they all work at ROL Ergo and we wanted to explore a natural setting. Additionally, we were not looking to answer a research question about a true causal relationship. Hence, we chose a quasi-experiment. The quasi-experiment was structured into different phases seen in Figure 1.

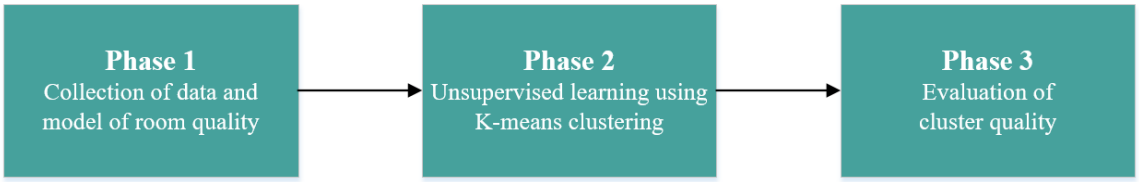


Figure 1. Description of phases of the quasi-experiment.

Phase one involved creating a model of room quality, which meant measuring and collecting data from the selected indoor climate factors using sensors. We chose to measure temperature, CO_2, and humidity based on our literature search, which confirmed that these factors are exuded by humans, hence human occupancy impacts indoor climate (Lazovic et al., 2016; Lin et al., 2017; Candanedo & Feldheim, 2016). It also indicated that human productivity, comfort, and well-being is affected by these factors (Al Horr et al., 2016).

Temperature, CO_2, and humidity sensors were installed in four conference rooms at ROL Ergo and streaming data was collected for ten working days. Booking data was also collected to determine when meetings occurred and number of participants in each meeting. We assumed that all participants that were invited and had not declined, attended the meetings. We also assumed that extreme outliers and values outside the range of the sensors were false, therefore they were removed.

In phase two, the data was prepared and used with K-means clustering. In phase three, we examined which features and which number of K that resulted in high-quality clusters. This was done by using Silhouette, Davies Boulding Index, and the Elbow method. To test the found clusters, we labeled the clusters to create qualitative descriptions of what each cluster represented. After this we took data points from specific meetings and categorized them.

5.2 Data collection

The sensors used to measure CO_2, temperature, and humidity can be seen in Table 1. They use an I^2C protocol to communicate and were sampled every minute. One of each sensor was placed in every room at 1.2 m above the floor, which is at occupant height approximately half-way from floor to ceiling (Yun & Kim, 2013). Two rooms were of larger size (s958 and s959) and the remaining two rooms were of smaller size (s960 and s962).

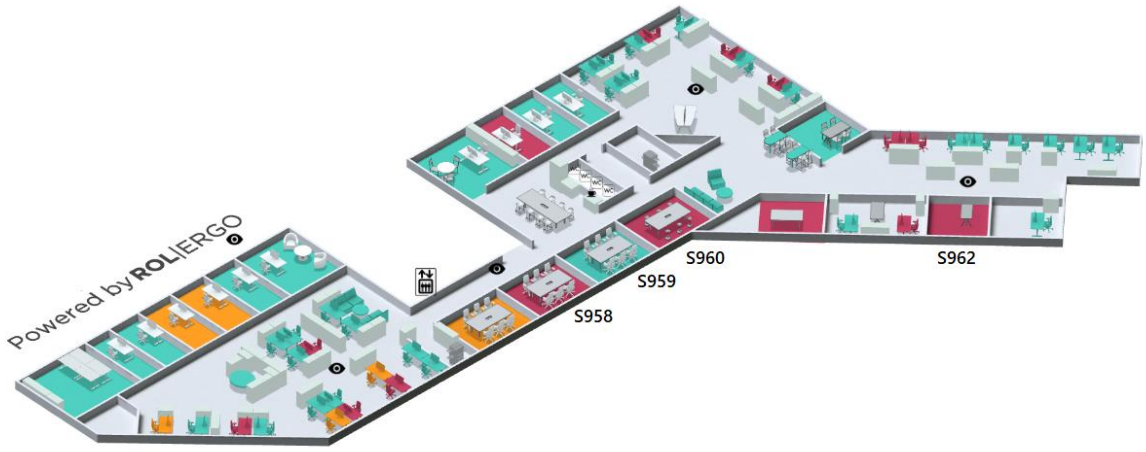


Figure 2. Floor plan of ROL Ergo (ROL Ergo, personal communication, May 5, 2019).

The sampled data from the sensors, as well as booking data, was collected working days during working hours, 07.00-18.00, in separate json-files. We chose not to collect data outside of these hours since the HVAC was turned off and few occupants were present. The data was extracted with a Python script, which organized it into arrays; CO_2, temperature, bookings, and timestamps. We added criterions in the script, for example all CO_2 values below 400 were recognized as false data and therefore not used. The limit was chosen based on interval limits of the CCS811 sensor (see Table 1). We also removed all temperature values below 18 °C since temperatures this low were uncommon and when they occurred it was likely an error, based on temperatures surrounding these values.

Table 1
Table listing the sensors used in the experiment.

Name	Manufacturer	Type	Unit	Interval	Format
BME280	Bosch	Temperature	Celsius	0 - 850	Integer
		Humidity	Percent	0 - 100	Integer
CCS811	AMS	CO_2	Ppm	400 - 29206	Integer

10 days of data from temperature, CO_2, and humidity sensors was combined into one data set. One data point in this set consists of one measurement of temperature, CO_2, and humidity each, at a certain point in time. The purpose of this set was to use it to generate the clusters with K-means clustering. Since we wanted as much data as possible to create a comprehensive model of the indoor climate, both during meetings and not, we chose to include both.

In the case of meetings, we extracted the start and finish time of each meeting and calculated the mean CO_2 level, temperature, and humidity during each specific meeting. This included conversion of integers to floats. After this, each meeting, with start and finish time, was

connected to its average CO_2, temperature, and humidity level, as well as number of occupants. All booking data was censored to protect the identity of the occupants.

5.3 Feature Engineering

The features we chose to use in the experiment were CO_2 level, temperature, and humidity. They are quantitative, meaning they have a real numerical scale (Flach, 2012). In addition, for the meeting data sets, number of occupants, timestamps (year-month-day-hour-minute), mean CO_2 level, mean temperature, and mean humidity were used to create few data points representative of whole meetings.

5.4 Clustering

The K-means clustering algorithm, also known as "Lloyd's Algorithm", was introduced in 1967 as a solution to difficult classification problems. The algorithm is an unsupervised learning algorithm that classifies different groups of data based on pre-determined *k* number of clusters. Unsupervised learning uses data without labels, i.e. previous information about the data, to learn (Flach, 2012). We chose K-means clustering since it is a popular clustering algorithm and we found in the literature that it was used in a related area by Li et al. (2018). They described the algorithm as follows:

1. Initialize *k* centroids
2. Assign every data point to nearest centroid
3. Recalculate the *k* centroids with the same data points
4. Repeat step 2 and 3 until the centroids no longer shift their position

See Figure 3 to see our flow chart of the K-means clustering algorithm based on the previous description of the algorithm.

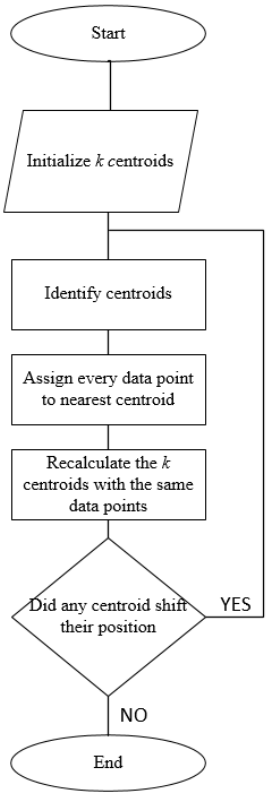


Figure 3. Flow chart of K-means algorithm.

In our case, the first step in the K-means clustering algorithm was to plot our data and visualize it in a diagram. Then we randomly picked three points somewhere in the diagram and marked them as cluster centers, also known as centroids. The next step was to identify the closest centroid for each point by calculating the Euclidean Distance (d),

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

We assigned all points to a centroid, updated the centroids based on the points, and repeated the process until no centroid shifted their position. After that we marked all the points with different colors to represent different clusters. In our case we used teal, blue, and green. The colors had no other meaning than which cluster it represented.

5.5 Evaluation

Since clustering, unlike supervised techniques, uses no previous information about existing partitions of the data, it is not possible to know if the cluster quality is dependent on the structure of the data or the performance of the algorithm (Menardi, 2010). Hence, because K-means clustering requires the input of number of *k*'s, evaluation of the clustering performance and structure of the data is necessary to determine which *k* produces the best cluster quality. For this purpose, we used internal evaluation measures; Silhouette analysis, Davis Bouldin Index, and the Elbow method. Three measures were chosen since it is recommended that several evaluation measures are used considering they might perform differently with different sets of

data. Additionally, no measure is dominant in all contexts (Hämäläinen, Jauhiainen & Kärkkäinen, 2017). Since there is no prior information of the datasets, only internal measures were chosen. External measures are available, but require prior information about the dataset (Rendón, Abundez, Arizmendi & Quiroz, 2011).

Silhouette analysis is based on the tightness and separation of clusters. It uses a measurement of how close each data point is to the allocated cluster, compared to a measure of the distance from the closest alternative cluster (Menardi, 2010). To determine the Silhouette, the partitions from a clustering algorithm and the proximities between the data points are needed (Rousseeuw, 1986). The Silhouette for the elements s_i is defined as

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

Which results in a coefficient between –1 and 1, where 1 is a positive result. Results with s_i near 1 are considered good, while s_i near 0 indicates that the observation lies between clusters, and observations with s_i below 0 indicates they are placed in the wrong cluster (Menardi, 2010). An average s_i can be determined either per cluster, or from the entire data set (Lleti, 2004). We used code from Scikit-learn to calculate and visualize the Silhouette (Pedregosa et al., 2011).

Additionally, we used Davies Bouldin Index. It is a measure of the computing quality of clustering. Similar to Silhouette, clusters that are less dispersed and further apart will result in a better score. The Davies Bouldin Index formula is defined as

$$DBI = \frac{1}{K}\sum_{i=1}^{K} \underset{i;j\neq i}{max} \frac{s_i + s_j}{d_{i,j}}$$

Where $s_i = \frac{1}{|C_i|}\sum x_j \in C_i \,||\, x_j - v_i \,||$ is a measure of scatter within the cluster i, k is the number of clusters, x_i is an n dimensional feature vector assigned to cluster i, v_i is the center of the cluster i, C_i represents the cluster i. The output known as the Davies Bouldin score is a float and the minimum score is zero with lower values indicates better clustering (Xiao, Lu & Li, 2017).

Furthermore, we used the Elbow method with Scikit-learn (Pedregosa et al., 2011). The Elbow method calculates Sum of Square Distances (SSD) for a chosen k. SSD is the sum of the average Euclidean distance of each point to the centroid. It is calculated as follows,

$$SSD = \sum_{i,j=1}^{n} ||x_i - x_j||^2$$

(Witsenhausen, 1974). This is plotted with, for example, k from 1-10 on the x-axis and SSD on the y-axis. The idea is that the graph levels out when adding another k does not significantly

reduce SSD, which is where the optimal *k* is located. This becomes visually clear as the graph forms an “elbow” (Soni,, 2012).

When the suitable *k* has been determined, labels can be assigned to the clusters. The labels are based on the temperature, CO_2, and humidity limits from Arbetsmiljöverket (2009) and Derby and Pasch (2017). This means that clusters with temperature within the 20 - 26 °C range, CO_2 below 1000 ppm, or humidity above 20% are considered good quality. Clusters outside these limits are considered bad. If more than two clusters are used, the labels will range from bad, medium, and good quality.

6 Results

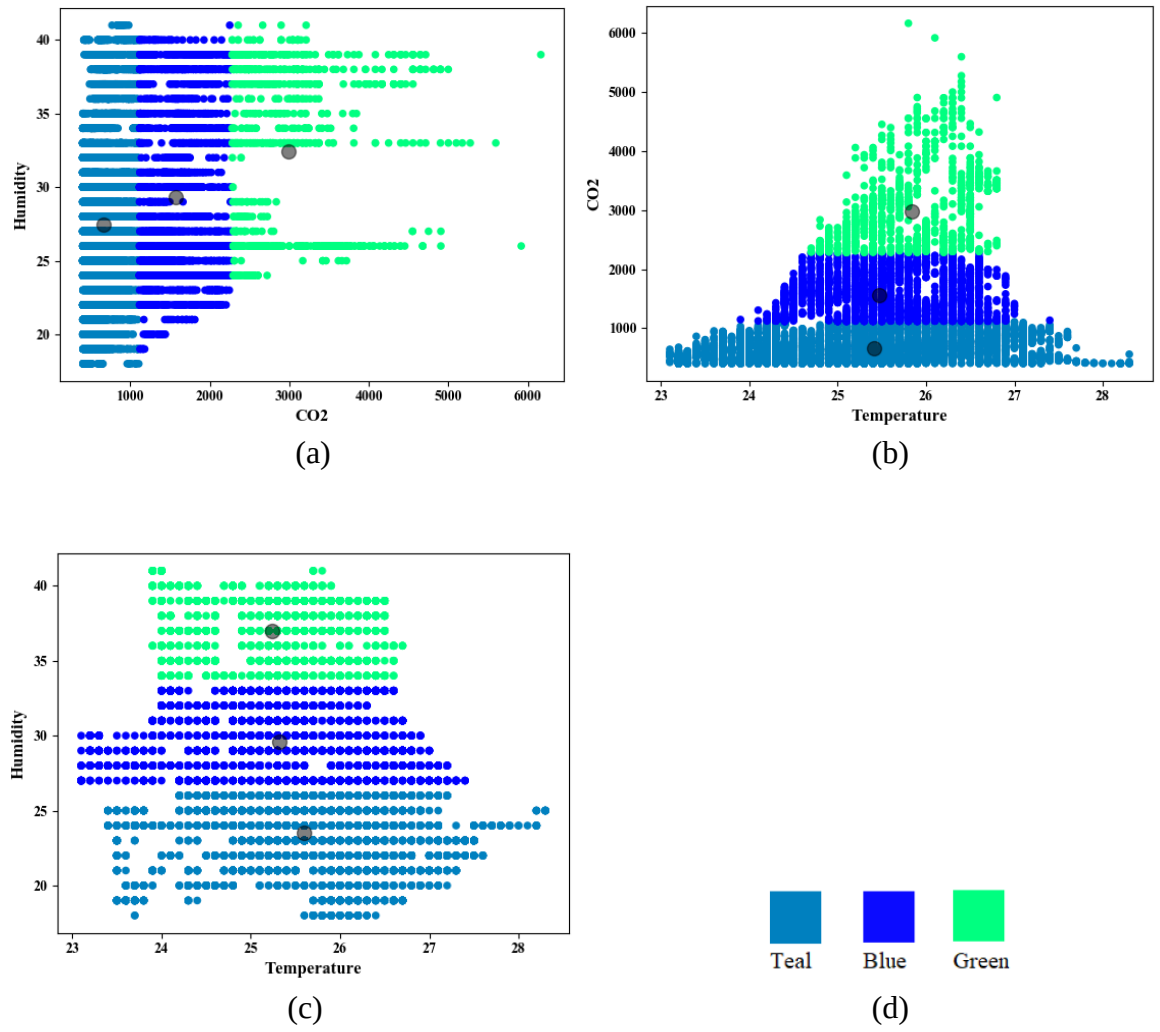


Figure 4. Visualization of the result from the K-means clustering algorithm; (a) CO2 in relation to humidity, (b) temperature in relation to CO2, (c) temperature in relation to humidity, (d) Names of the colors that we are referring to.

We plotted the results of the K-means clustering algorithm including the centroids so that we could receive a visual result. As shown in Figure 4a above, CO_2 is placed on the x-axis and

humidity on the y-axis. We see that the teal colored cluster, with low CO_2 values, covers the whole humidity axis' range, while the blue and the green clusters are gathered in the middle and upper range of the humidity's axis. When looking at the CO_2 we can see a distinct border between the teal, blue and green cluster. We can see that humidity increases along with the CO_2.

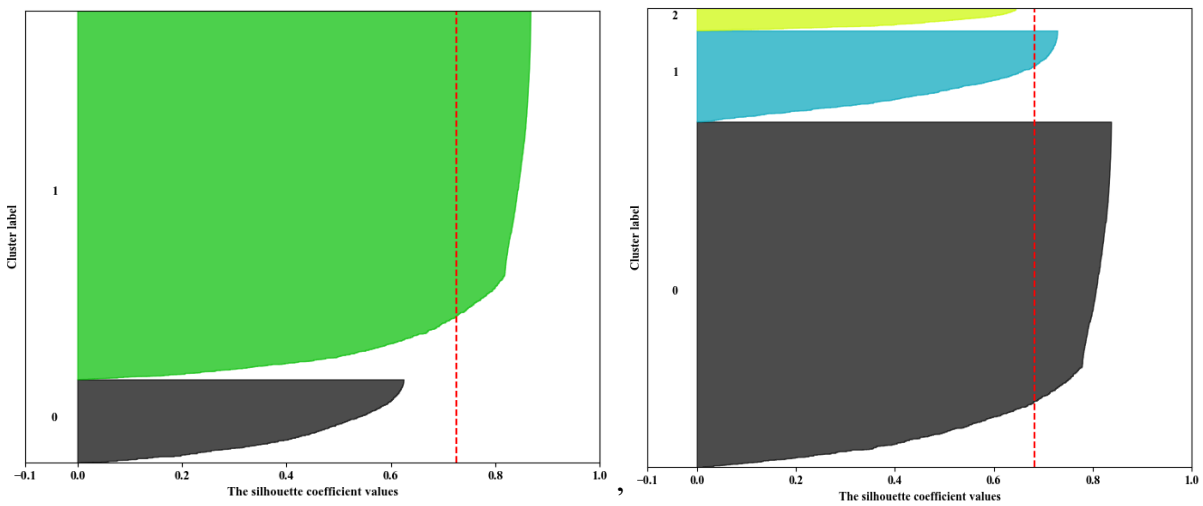
In Figure 4b the temperature is placed on the x-axis and CO_2 on the y-axis. Here we see that the teal cluster with the lowest CO_2 covers the whole temperature axis. The blue and green clusters are spread out through the middle axis of the temperature and the middle and high range of the CO_2 axis.

In Figure 4c, the temperature is placed on the x-axis and the humidity on the y-axis. We can see that the teal and the blue clusters cover the whole temperature axis and the green covers the middle temperature axis. When looking at the humidity, the teal covers the lower area, the blue covers the middle area and the green covers the upper area.

Table 2
Silhouette score for all respective feature combinations.

k	CO_2, Humidity, Temp	CO_2, Humidity	CO_2, Temp	Humidity, Temp
2	0.73	0.72	0.73	0.61
3	0.68	0.68	0.68	0.50
4	0.61	0.61	0.61	0.49
5	0.59	0.60	0.60	0.48
6	0.57	0.57	0.56	0.46
7	0.57	0.57	0.57	0.45
8	0.57	0.57	0.57	0.42
9	0.56	0.56	0.57	0.41
10	0.55	0.55	0.56	0.43

Looking at the results from the Silhouette analysis, it can be seen in Table 2 than the Silhouette scores are similar except for the feature combination humidity and temperature, which is lower. $k = 2$ and $k = 3$ resulted in the highest Silhouette scores, see Table 2. Their average Silhouette scores are illustrated by the red vertical lines shown in Figure 5a and 5b.



(a) (b)

Figure 5. Silhouette score for *k*; (a) *k* = 2, (b) *k* = 3.

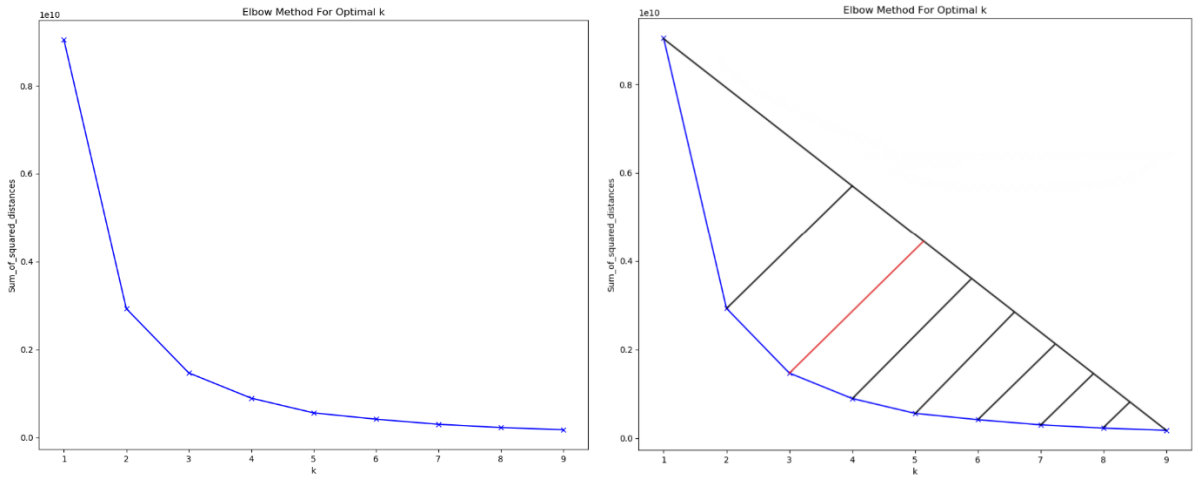
The average Silhouette was calculated based on the entire data set. The size of each cluster was also illustrated by the vertical size of the Silhouettes. We can see that Figure 5a has a higher average Silhouette index, while Figure 5b has a slightly lower one. However, Figure 5b displays a result where each cluster has a similar Silhouette score, meaning that the clusters have a higher Silhouette score independent of each other.

The Davies Bouldin index analysis indicated *k* = 2 as preferred, since it has the lowest score. See Table 3.

Table 3
Result from the Davies Bouldin Index analysis for each k.

Number of *k*	Davies Bouldin
2	2.00
3	6.14
4	16.73
5	28.77
6	55.58
7	82.36
8	99.45
9	126.32
10	142.75

From the Elbow method we used temperature, CO_2, and humidity with the combinations: CO_2 and temperature, CO_2 and humidity, temperature and humidity and all three of them together. The result showed no significant difference between the different combinations. Therefore, we only show the graph containing all three of them. If we look at the graph, we can see how that the marginal gain drops and creates an angle in the graph where $2 < k < 3$.



(a) (b)

Figure 6. Elbow method result with CO_2, humidity and temperature; (a) Results from the Elbow method, (b) same graph as graph (a) but with help lines added.

With a small margin, by looking at the slope where it forms an Elbow, it is inclined that $k = 3$ is slightly better than $k = 2$ for our type of dataset. It is also possible to see this when drawing a help line between the start and the finish point in the Elbow diagram and then from each point to the line using Euclidean distance, see Figure 6b. The point with the longest distance to the help line indicates the optimal k (Soni, 2012)

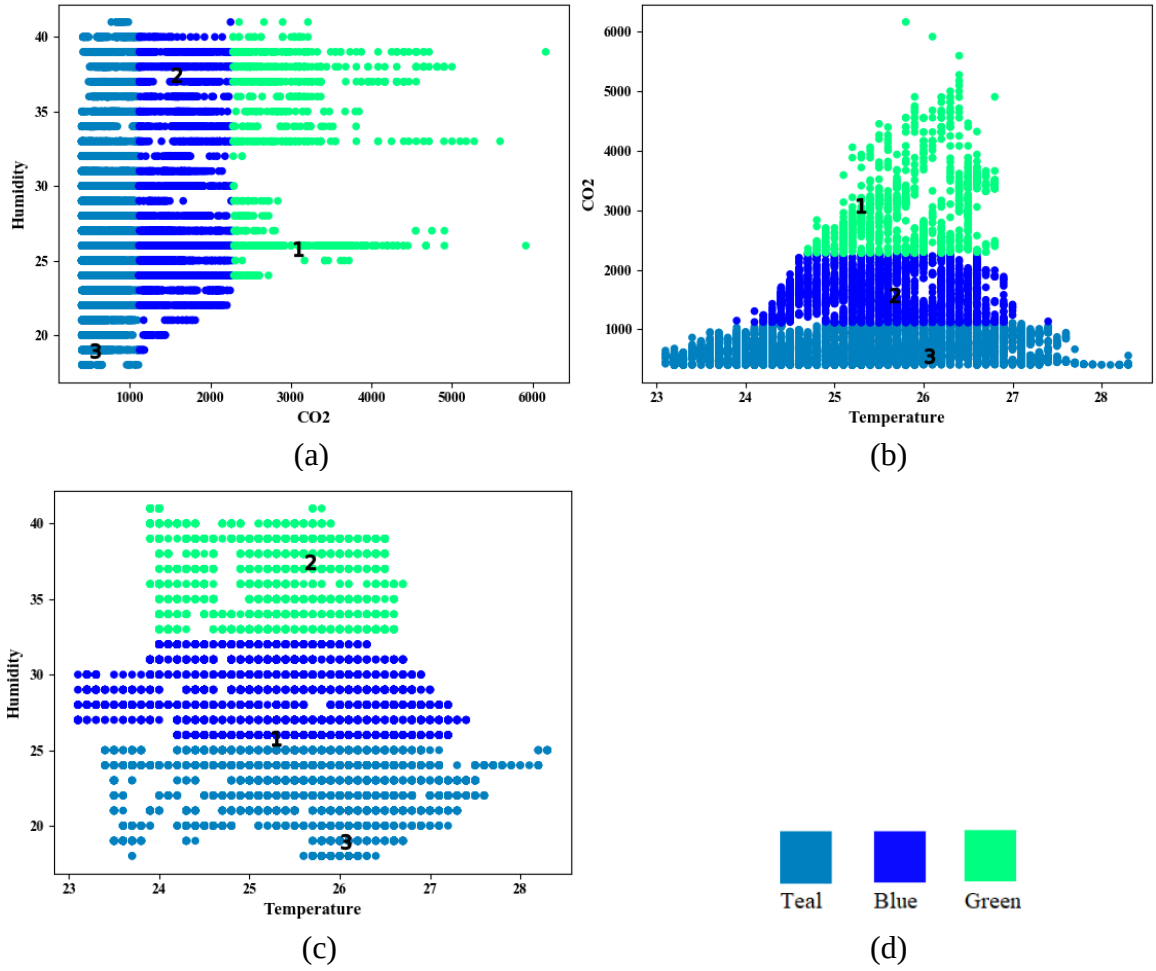


Figure 7. Determination of which cluster each meeting (1, 2, 3) belongs to; (a) based on Figure 4a, (b) based on Figure 4b, (c) based on Figure 4c. (d) Names of the colors that we are referring to.

We added meetings to see which clusters they would be categorized into. To visualize the result, we show three meetings where each meeting is marked by a number. For information about the data from the meeting, see Table 4.

Table 4

Description of meetings and their respective features.

Meeting	Occupants	Date	Duration	Temp	CO_2	Humidity	Room
1	8	24/4/19	70 min.	25.3	3079.9	25.8	s962
2	3	26/4/19	90 min.	25.7	1577.1	37.5	s960
3	2	3/5/19	60 min.	26.1	555.3	18.9	s960

Meeting 1 was categorized into the green cluster, meeting 2 was categorized into the blue cluster and meeting 3 was categorized into the teal cluster.

7 Discussion

7.1 K-means results

What each cluster represents is different levels of temperature, CO_2, and humidity, which can be labeled using limits for indoor climate. When the K-means clustering algorithm was executed with $k = 3$, labels were assigned to the clusters. Looking at Figure 4, teal was labeled 'good', blue 'medium', and green 'bad' quality.

After we visualized the result, we could see an indication of CO_2 increasing with humidity (see Figure 4a). This suggests a relationship between CO_2 and humidity, which is in line with findings by Lazovic et al. (2016). The clusters in Figure 4a also indicate the quality levels of CO_2, with distinct cluster limits at approximately 1000 ppm and 2500 ppm. This correlates well to the CO_2 limit recommendations found in the literature where <1000 ppm is considered good (Arbetsmiljöverket, 2009). When looking at temperature and CO_2, it was indicated that high CO_2 levels occurred at temperatures between 24 and 27. However, this may have occurred since 94% of all temperatures in the data set were between these values. Furthermore, temperature and humidity did not indicate any visible correlation but were quite evenly distributed regardless of their respective levels. Temperature and humidity showed no correlation even though they often covary (Lazovic et al., 2016). This could be because the correlation was not strong enough to be clearly visible, or because temperature could have been controlled more efficiently by the HVAC than CO_2.

7.2 Evaluation results

The results from the Silhouette and the Elbow method indicated that both $k = 2$ and $k = 3$ were suitable, while the Davis Bouldin Index indicated that $k = 2$ was suitable. The result from $k = 3$ (Figure 5b) showed that each independent Silhouette had a similar, high score compared to each other, while $k = 2$ (Figure 5a) had a total average better Silhouette but larger difference between independent Silhouettes. This means there is less difference between cluster quality when $k = 3$, than when $k = 2$. Therefore, when choosing between $k = 2$ and $k = 3$, there is a trade-off between higher average Silhouette, and more defined independent Silhouettes. The results also showed that the Silhouettes were similar when using humidity and CO_2, or temperature and CO_2 for clustering, as when using all three features. However, temperature and humidity used together returned slightly lower Silhouette scores (see Table 2). This relates to our analysis of the visualization of the clusters where temperature and humidity showed no correlation.

These evaluation measures give us insight into the quality of clusters depending on number of k, but one should also consider the purpose of the clusters when determining which k is suitable. Since we wanted to show quality levels, $k = 2$ or $k = 3$ were suitable options because higher k's, more levels, would be difficult to comprehend and therefore not useful for this purpose.

7.3 Categorization of meetings

When data from specific meetings was analyzed by categorizing the meetings to clusters, it was possible to tell that CO_2 levels were higher, and humidity levels slightly higher, when more occupants attended the meetings. This correlates to the causal relationship between human presence, CO_2 and humidity (Zhang, Wargocki & Lian, 2016; Lazovic et al., 2016). However, temperature showed little correlation to occupancy. In Figure 7, each meeting from Table 4 is visualized and one can tell that each meeting was categorized to the appropriate cluster when $k = 3$, depending on its temperature, CO_2, and humidity values. However, it was not clear that either $k = 2$ or $k = 3$ could explain other patterns, like if one of the clusters encompassed all collected meeting data and another cluster only non-meeting data.

7.4 Validity threats

We verified that we measured relevant data since CO_2, temperature, and humidity are exuded by humans and are generally accepted in the literature to be important factors to maintain a productive work environment and can be used to measure occupancy. Since we did not alter the features or indoor environment, there was no risk of threatening the treatment validity before or during measurement. However, there was a small risk of the employees at ROL unconsciously altering their behavior since they knew that data was being collected. We did not consider this a significant threat since the employees at ROL are used to sensors in the office environment.

Furthermore, poorly functioning sensors could result in inaccurate data. This was noticeable in the raw data, where missing or invalid data from sensors occurred. We counteracted this threat by removing all data from days with errors and removing any extreme or unpredictable outliers. However, this also posed a risk that we removed data that we thought had errors but did not. The best solution to counteract these threats is to use high quality sensors and verify that they are always functioning.

In the case of meetings, we did not take into consideration if there had been a meeting prior to a meeting that we collected measurements from. This means the previous meeting could have impacted the measurements, which matters when analyzing the impact of number of occupants. This is something to bear in mind for future research where, for example, a baseline level for CO_2, temperature, and humidity in the specific rooms could be determined, and only meetings with this start level could be used. Furthermore, the data of number of occupants during the meetings was not necessarily correct. This data was collected from a booking system, so there is a possibility that some occupants did not actually participate in a meeting they were invited to. However, the booking system showed if the meeting had occurred at all. To ensure accurate data, there is a need for a more effective method that determines number of occupants automatically.

Our results are generalizable to other environments since they follow generally accepted relationships between occupancy and indoor climate. However, since the clusters are based on

data from a specific environment, new measurements and cluster calculations should be done in the said environment to be able to utilize the categorization effectively. In an automated system, data can be measured continuously.

We consider the results valid in terms of how levels of room quality are determined, since we use generally accepted limits for CO_2, temperature, and humidity to analyze the results. Additionally, we used three evaluation measures to confirm that the K-means clustering algorithm produced satisfactory clusters.

8 Conclusions

After clusters were defined using the K-means clustering algorithm, it was possible to assign three meeting to specific clusters depending on their features. This meant it was possible to categorize the quality level of the respective meetings. Since this was the main purpose of the thesis it was an important result. We were also able to determine the optimal number of *k*, $k = 2$ or $k = 3$, using Silhouette, Davies Bouldin Index, and the Elbow method. Our results confirmed that CO_2 and humidity correlated in response to usage of conference rooms, while no discernible patterns could be found using temperature. A limitation that may have affected the validity was the occurrence of incorrect sensor data, and uncertain data of the number of occupants during meetings. In future research, we suggest that sensors are checked regularly and that a more sophisticated method for determining number of occupants is applied.

We suggest that future research explores how qualitative aspects can be included in the model of room quality. This enables study of subjective data such as personal preference of look and feel, biophilia and views, and distance and amenities. In a working product, this could entail that the user is able to not only get recommendations for conference rooms based on level of quality from objective sensor data, but on subjective data about their personal preference and previous ratings of rooms.

The contribution of this thesis to the knowledge base is a model of indoor climate in conference rooms, to enable categorization of room quality during meetings. It is also an evaluation of which features, and number of *k* is suitable for determining levels of room quality with K-means clustering. Moreover, it contributes as a preceding study to the MAPPE research project, and to ROL Ergo, with increased knowledge about how sensor data can be utilized in conference rooms. At last, it can provide means to create more energy efficient environments by establishing awareness of the indoor climate.

References

Al Horr, Y., Arif, M., Katafygiotou, M., Mazroei, A., Kaushnik, A., & Elsarrag, E. (2016). Impact of indoor environmental quality on occupant well-being and comfort: a review of the literature. *International Journal of Sustainable Built Environment, 5*(1), pp. 1-11. doi: 10.1016/j.buildenv.2016.06.001

Arbetsmiljöverket. AFS 2009:2 - *Arbetsplatsens utformning - Arbetsmiljöverkets föreskrifter om arbetsplatsens utformning samt allmänna råd om tillämpningen av föreskrifterna.* Retrieved June 26, 2019: https://www.av.se/arbetsmiljoarbete-och-inspektioner/publikationer/foreskrifter/arbetsplatsens-utformning-afs-20092-foreskrifter/

Bailey, J., E., Argyropoulos, S., V., Kendrick, A., H., & Nutt, D., J. (2005). Behavioral and cardiovascular effects of 7.5% CO_2 in human volunteers. *Depression and anxiety, 21*(1), pp. 18-25. doi: 10.1002/da.20048

Browning, B. (2015). Healthier workplaces, happier employees: incorporating nature into the built environment. *People & Strategy, 38*(3), pp. 14-17.

Candanedo, L. M., & Feldheim, V. (2016). Accurate occupancy detection of an office room from light, temperature, humidity and CO_2 measurements using statistical learning models. *Energy and Buildings, 112*, pp. 28-39. doi: 10.1016/j.enbuild.2015.11.071

Caragliu, A., Del Bo, C., & Nijkamp, P. (2011). Smart cities in Europe. *Journal of Urban Technology, 18*(2), pp. 65-82. doi: 10.1080/10630732.2011.601117

Celestino Marín, A., E., Martínez Cruz, D., A., Otazo Sánchez, E., M., Reyes, F., G., & Vásquez Soto, D. (2018). Groundwater quality assessment: an improved approach to K-means clustering, principal component analysis and spatial analysis: a case study. *Water*. doi: 10.3390/w10040437

Cook, D. J., & Das, S. K. (2007). How smart are our environments? An updated look at the state of the art. *Pervasive and Mobile Computing*, *3*(2), pp. 53-73. doi: 10.1016/j.pmcj.2006.12.001

Cook. D. J., & Krishnan, N. (2014). Mining the home environment. *Journal of Intelligent Information Systems, 43*(3), pp. 503-519. doi: 10.1007/s10844-014-0341-4

Curry, E., & Sheth, A. (2018). Next-generation smart environments: from system of systems to data ecosystems. *IEEE Intelligent Systems*, *33*(3), pp. 69-76. doi: 10.1109/MIS.2018.033001418

Davies, N., Clinch, S. (2017). Pervasive data science. *IEEE pervasive computing, 16*(3), pp. 50-58. doi: 10.1109/MPRV.2017.2940956

De Silva, L., Morikawa, C., & Petra, I. M. (2012). State of the art of smart homes. *Engineering Applications of Artificial Intelligence, 25*(7), pp. 1313-1321. doi: 10.1016/j.engappai.2012.05.002

Derby, M., M., & Pasch, R., M. (2017). Effects of low humidity on health, comfort & IEQ. *ASHRAE Journal, 59*(9), pp. 44-58.

European standard. (2006). ICS 91.140.01 - *Indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics.*

Flach, P. (2012). *Machine learning – the art and science of algorithms that make sense of data.* Cambridge, England: Cambridge University Press.

Hämäläinen, J., Jauhiainen, S., & Kärkkäinen, T. (2017). Comparison of internal clustering validation indices for prorotype-based clustering. *Algorithms, 10*(3). doi: 10.3390/a10030105

Lazovic, I., M., Stevanovic, Z., M., Jovasevic-Stojanovic, M., V., Zivkovic, M., M., & Banjac., M., J. (2016). Impact of CO_2 concentration on indoor air quality and correlation with relative humidity and indoor air temperature in school buildings in Serbia. *Thermal science, 20*(1), pp. 297-307. doi: 10.2298/TSCI150831173L

Li, W., Logenthiran, T., Phan, V., & Woo, W., L. (2018). Power alert system using K-means for smart home. *IEEE innovative smart grid technologies – Asia,* pp. 722-727. doi: 10.1109/isgt-asia.2018.8467949

Liang, G., Cao, J., Liu, X., & Liang, J. (2016). Smart world: a better world. *Science china – information sciences, 59*(4). doi: 10.1007/s11432-016-5518-8

Lin, B., Huangfu, Y., Lima, N., Jobson, B., Kirk, M., O´Keefe, P., …, & Cook, D. J. (2017). Analyzing the relationship between human behavior and indoor air quality. *Journal of Sensor and Actuator Networks, 6*(3). doi: 10.3390/jsan6030013

Lleti, R., Ortiz, M., C., Sarabia, L., A., & Sánchez, M., S. (2004). Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. *Analytica chimica acta, 515*(1), pp. 87-100. doi: 10.1016/j.aca.2003.12.020

Maresh, C. M., Armstrong, L. E., Kavouras, S. A., Allen, G. J., Casa, D. J., Whittlesey, & M., LaGasse, K. E. (1997). Physiological and psychological effects associated with high carbon dioxide levels in healthy men. *Aviation Space and Environmental Medicine, 68*(1).

Menardi, G. (2010). Density-based Silhouette diagnostics for clustering methods. *Statistics and computing, 21*(3), pp. 295-308. doi: 10.1007/s11222-010-9169-0

Merabet, G. H., Essaaidi, M., & Benhaddou, D. (2018). Measuring human comfort for smart building application: experimental set-up using WSN. *ICSDE'18,* pp. 56-63. Rabat, Morocco. doi: 10.1145/3289100.3289110

Mozer, M. C., Dodier, R. H., Anderson, M., Vidmar, L., Cruickshank III, R. F., & Miller, D. (1995). The neural network house: an overview. *Current trends in connectionism*, pp. 371-380. Hillsdale, NJ, USA.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ..., & Duchesnay, É. (2011). Scikit-learn: machine learning in Python. *JMLR*, 12, pp. 2825-2839.

Peng, Y., Rysanek, A., Nagy, Z., & Schlüter, A. (2018). Using machine learning techniques for occupancy-prediction-based cooling control in office buildings. *Applied Energy, 211*(1), pp. 1343 - 1358. doi: 10.1016/j.apenergy.2017.12.002

Rendón, E., Abundez, I., Arizmendi, A., & Quiroz, M, E. (2011). Internal versus external cluster validation indexes. *International journal of computers and communications, 1*(5).

Rolfö, L., Eklund, J., Jahncke, H. (2018). Perceptions of performance and satisfaction after relocation to an activity-based office. *Ergonomics, 61*(5), pp. 644-657. doi: 10.1080/00140139.2017.1398844.

Rousseeuw, P., J. (1986). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics, 20*, pp. 53-65. doi: 10.1016/0377-0427(87)90125-7

Satyanarayanan, M. (2001). Pervasive computing: vision and challenges. *IEEE Personal Communications, 8*(4), pp. 10-17. doi: 10.1109/98.943998

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Soni Madhulatha, T. (2012). An overview of clustering methods. *IOSR journal of engineering, 2*(4), pp. 719-725.

Szczurek, A., Maciejewska, M., & Pietrucha, T. (2017). Occupancy determination based on time series of CO_2 concentration, temperature and relative humidity. *Energy Buildings, 147*, pp. 142-154. doi: 10.1016/j.enbuild.2017.04.080

Torunski, E., Othman, R., Orozco, M., & El Saddik, A. (2012). A review of smart environments for energy savings. *Procedia Computer Science, 10*, pp. 205-214. doi: 10.1016/j.procs.2012.06.029

Vehviläinen, T., Lindholm, H., Rintamäki, H., Pääkkönen, R., Hirvonen, A., Niemi, O., & Vinha, J. (2016). High indoor CO_2 concentrations in an office environment increases the transcutaneous CO_2 level and sleepiness during cognitive work. *Journal of Occupational and Environmental Hygiene, 13*(1), pp. 19-29. doi*:* 10.1080/15459624.2015.1076160

Weiser, M. (1999). The computer for the 21st century. *SIGMOBILE ACM, 3*(3), pp. 3-11. doi: 10.1145/329124.329126

Wilson, E. O. (1984). *Biophilia.* Cambridge, MA: Harvard University Press.

Witsenhausen, H., S. (1974). On the maximum of the sum of squared distances under a diameter constraint. *The American Mathematical Monthly, 81*(10), pp. 1100-1101. doi: 10.2307/2319046

Xiao, J., Lu, J., Li, X. (2017). Davies Bouldin Index based hierarchical initialization K-means. *Intelligent data analysis, 21*(6), pp. 1327-1338. doi: 10.3233/IDA-163129

Yun, J., Kim, J. (2013). Deployment support for sensor networks in indoor climate monitoring. *International Journal of Distributed Sensor Networks*, *9*(9). doi: 10.1155/2013/875802

Zhang, X., Wargocki, P., & Lian, Z. (2016). Human responses to carbon dioxide, a follow-up study at recommended exposure limits in non-industrial environments. *Building and Environment, 100*(1), pp. 162-171. doi: 10.1016/j.buildenv.2016.02.014