MINIPROJECT: Occupancy of a Smart Home

Smart Systems - 5EU9SSY1

Authors
Michel Farah
Paula Alejandra Pedraza Aguirre

Guided by
Jerome Ferrari

Grenoble INP - ENSE3
SEM - 202

Contents

1 Data Collection **3**
1.1 Expe-smarthouse 3
1.2 Data selection 4

2 Data Analysis **4**
2.1 Power demand 5
2.2 Noise 6
2.3 CO_2 emissions 6

3 Clustering **7**
3.1 Clustering method (K-mean) 7

4 Conclusions **9**

Abstract

This paper analyses the correlation between four types of variables (CO2 emissions, total power, total power of a TV and noise) using the machine learning method known as K-mean clustering to estimate the occupancy of a room. The analysed data are obtained from the study of a household through the Expe-smarthouse project, which provides the opportunity to access in real time through a Grafana portal with Influxdb database. Additionally, the selected sensors and data are studied individually to better understand their behaviour. The methodology used corresponds to the collection of data through the installation and correct execution of the different softwares, the subsequent selection of relevant data and its management by means of appropriate statistical techniques to obtain the same step time of analysis. Finally, the application of the clustering method that allows the variables to be related and to conclude about the occupation.

Key words: Occupancy, sensors, clustering, power, statistical techniques.

Introduction

The advancement of smart buildings responds to the constant progress of electronics, currently being a wide range of solutions and proposals that aim to save energy, save on maintenance and to contribute to the environment. According to Fortune Business Insights, the Internet of Things (IoT) market is forecast to grow at a 25.4 percent compound annual growth rate (CAGR), zooming from $381 billion in 2021 to $1.8 trillion to 2028 [1]. Through low-cost computing, cloud, big data analysis and mobile technologies, physical things can share and collect data, enabling a hyper-connected society. The physical and digital worlds go hand in hand and cooperate with each other, especially thanks to new technologies that have been developed such as access to low-cost and low-power sensors, machine learning and analytics, the rise of network protocols for the internet, among others.

Therefore, for the design of a smart home, a digitisation plan must be taken into account with a robust infrastructure, a convergence of all systems to the same network and an adequate control relationship between people and the variables of the environment.Different types of techniques and methods to estimate the load of the building and consequently its needs are used. One of the main tools consists of machine learning and artificial intelligence in order to estimate the number of room occupancy, which takes data from different types of installed sensors and tries to analyze the occupancy behavior using a different data set than the one that tries to estimate the correct occupancy in the future.

As an approach to the complex development of this type of building, this project seeks to estimate the occupancy of a smart house through the analysis of the data collected in the "Expe-smarthouse" project, data which are accessible in real time using the free software Grafana and the Influxdb database. Estimating the occupancy by applying the decision tree method and comparing it to other techniques such as the random forest and the K-mean clustering method.

1 Data Collection

As stated above, the data are taken from an open-access database from a project developed in 2018, which is described in more detail below.

1.1 Expe-smarthouse

"The *Expe-smarthouse* project was initiated in order to provides data from a 120 m^2 household where a 5 people family lives" [2]. Using open source hardware and software there is access in real time to 340 measuring points.

The sensors include the following information:

1. ***Device Consumption***: Electricity, gas and water.
2. ***Environmental conditions***: Temperature, air analysis, humidity and brightness of each common room. Outdoor weather conditions.
3. ***Others***: Opening position of each door and window. Motion sensors.

1.2 Data selection

The data that was collected from the Grafana Website taking into account the process shown in the figure 1.

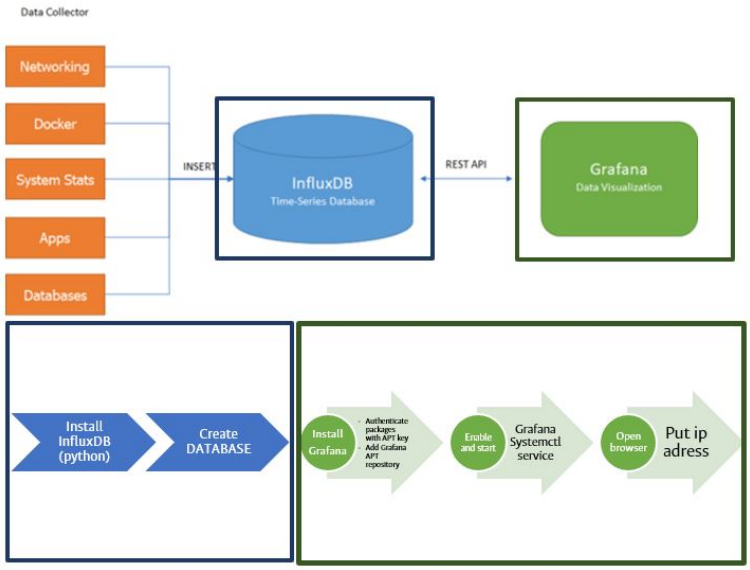

Figure 1: Process for accessing to the data generated by Expe-smarthouse platform.

Using our Python code each set of data as a csv file format was extracted. The Data chosen are the power demand for different equipment in the house, like the television, the hotplate, the dishwasher and the washing machine. Futhermore, the noise sensor and the CO_2 concentration sensor in the salon were chosen, to be able to understand more the behaviour of the inhabitants and estimate the occupancy.

Each sensor collects data differently, some sensors collect data daily or hourly and some even with a random frequency. To solve this problem was necessary to interpolate our data so that all the data collected are in the same frequency, An hourly mean value was chosen, as follows:

	Time	CO2_Emissions	Bruit	TV_Power_Consumed	Total_Power	Lave Linge Consumption [W]	Lave vaisselle consumption [W]	Plaque Cuisson Consumption [W]
0	12/6/21 13:00	413	37	0.00	321.9760	53	21	0
1	12/6/21 14:00	411	37	0.00	400.8976	53	21	0
2	12/6/21 15:00	404	37	0.00	375.9136	0	202	0
3	12/6/21 16:00	390	53	0.00	413.5552	0	202	0
4	12/6/21 17:00	454	47	15.07	564.9072	137	2192	0
...	...	...	...	...	...	...	...	...
91	12/10/21 8:00	570	43	0.00	420.8064	75	2135	1180
92	12/10/21 9:00	554	41	0.00	337.6960	63	69	899
93	12/10/21 10:00	580	43	45.86	315.8400	63	69	0
94	12/10/21 11:00	610	43	0.00	384.7376	83	68	1558
95	12/10/21 12:00	600	45	0.00	1030.9824	83	68	1577

96 rows × 8 columns

2 Data Analysis

In this section, we will analyse the different data collected from the sensors and try to understand the behaviour of the inhabitants and finally estimate the occupancy specifically in the salon.

2.1 Power demand

Sensors have been connected to several electronic devices in the house, including the Television, the hotplate, the dishwasher and the washing machine. The data has an hourly frequency and so now we can observe the power demand behavior of the user from the figures below.

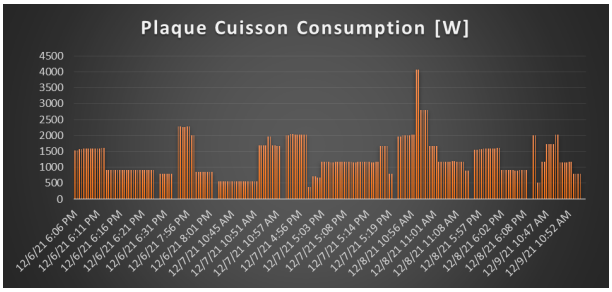

Figure 2: Hourly Hotplate power consumption [W].

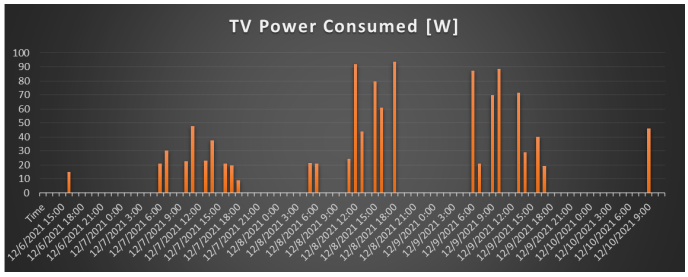

Figure 3: Hourly TV power consumption [W].

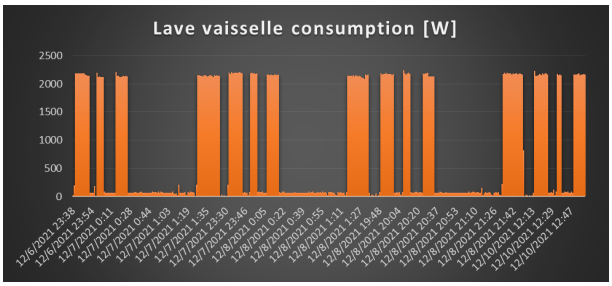

Figure 4: Hourly dishwashing power consumption [W].

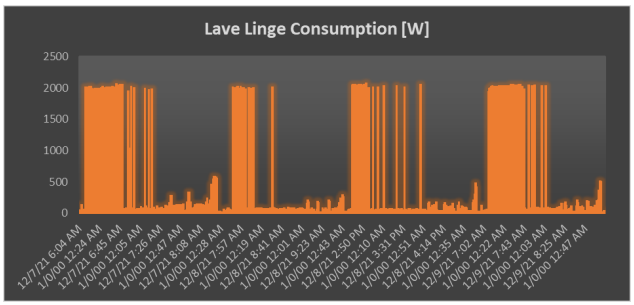

Figure 5: Hourly washing machine power consumption [W].

As we can see, every device has a different behaviour from the other. The TV is used at noon or at Night mostly and not every day. So we can assume that when on monday and friday the user is working in the company and on tuesday, wednesday and thursday is working remotely. The hotplate is mostly used around 12 PM or at 6 PM, meaning that the user is cooking lunch when working remotely or dinner after coming back home from work. For the dishwasher, we observe that is mostly done at night every day while the washing machine is used in the morning most of the time. We can clearly see that most of the power demand is for the dishwasher, washing machine and hotplate but they are not used constantly.

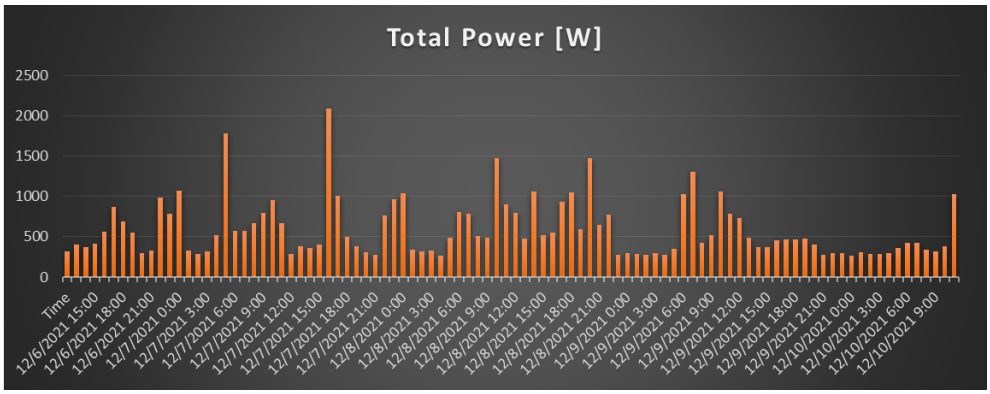

Figure 6: Hourly Total power consumption [W].

The total power demand can be seen in the figure above, we have a base load of around 500 W and many peaks of demand depending on the equipment used. We notice that on monday and friday we have lower demands than during the other weekdays.

2.2 Noise

Next we will analyse the noise behaviour in the salon,so mainly it is when the user is watching TV or just sitting relaxing.

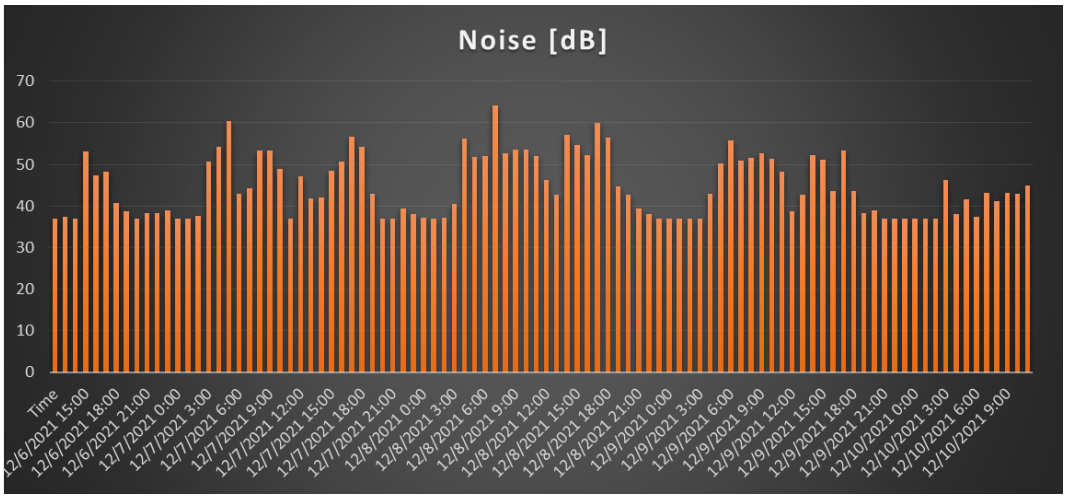

Figure 7: Hourly Noise disturbance [dB].

As we can see, on monday, we go above the average around 18 PM,meaning it is when the user returned home. On the other hand, during the other days except friday, we notice that the user did not leave the house so he must be working remotely and it is convenient because from figure 1 of the TV power demand, we saw that the TV was used mostly during tuesday, wednesday and thursday. This helps us to predict the occupancy at any given time.

2.3 CO_2 emissions

Our final observation will be on the CO2 concentration of the salon, to see if the results with other sensors are compatible and help us have more accurate predictions.

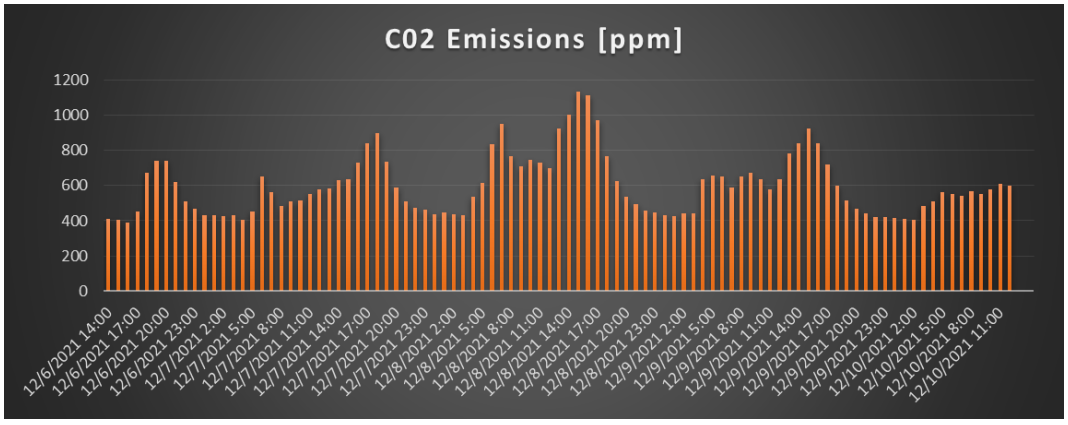

Figure 8: Hourly CO2 concentration [ppm].

The results from figure 7 indicates that their is a presence around 17-18 PM on monday in the salon, while during the other days we have more presence during the day in the salon. The average CO2 concentration is 400 ppm sp when it increases to 600 or 800, we can assume there is human presence in the room. The CO2 concentration behaviour confirms even more our predictions, the user during the week from 6 december to 10 december 2021 was at work on monday and friday, and was working remotely the other days.

3 Clustering

Taking into account that it is unlabeled information, machine learning methods focused on unsupervised data are used. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. There are a number of clustering methods but in this case, just the K-mean method is used.

3.1 Clustering method (K-mean)

This method requires the completion of several steps, which will be described below. First, the optimal number of clusters is determined using the "Elbow Method". The elbow method uses the *Within Cluster Sum of Squares (WCSS)* concept to draw the plot by plotting *WCSS* values on the Y-axis and the number of clusters on the X-axis. This is computed as

$$WCSS = \sum_{i \in n} (x_i - y_i)^2$$

where y_i is centroid for observation x_i. By definition, this is geared towards maximizing the number of clusters, and in the limiting case, each data point becomes its own cluster centroid. Therefore, the value for WCSS for different k values ranging from 1 to 10 was calculated and the desired k number of **4** was obtained as a result. This bend indicates that additional clusters beyond the fourth have little value.

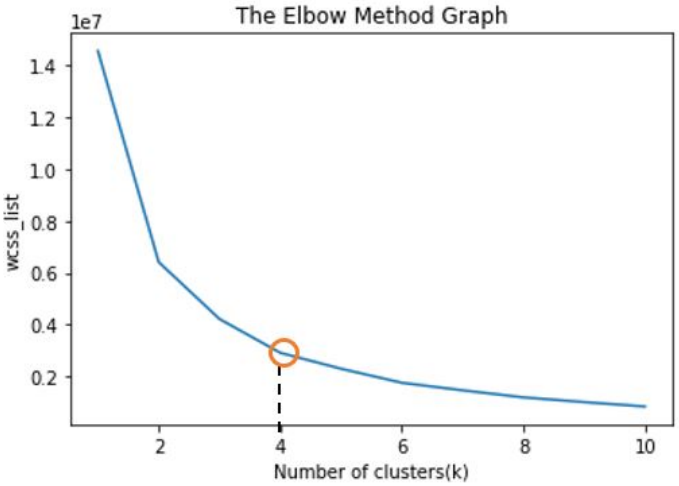

Figure 9: Elbow method for optimal value of k in KMeans.

Since the number of clusters is known, it is selecting random centroids for each cluster. Two clusters were selected and then select random observations from the data as the centroids. In Figure 10, the red dots represent the four centroids for each cluster. Because the points were chosen randomly and hence every time the code is run, different centroids are gotten.

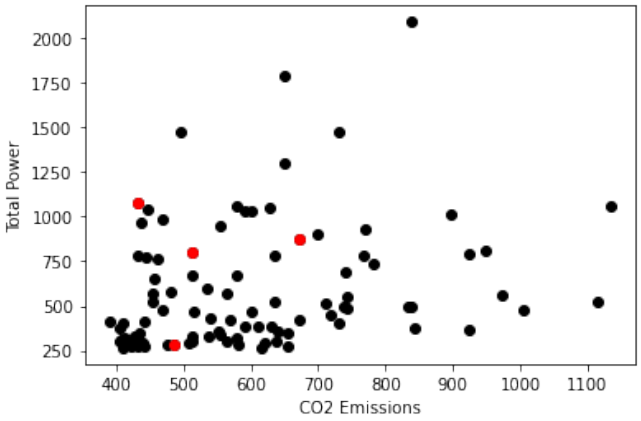

Figure 10: Random selection of centroids.

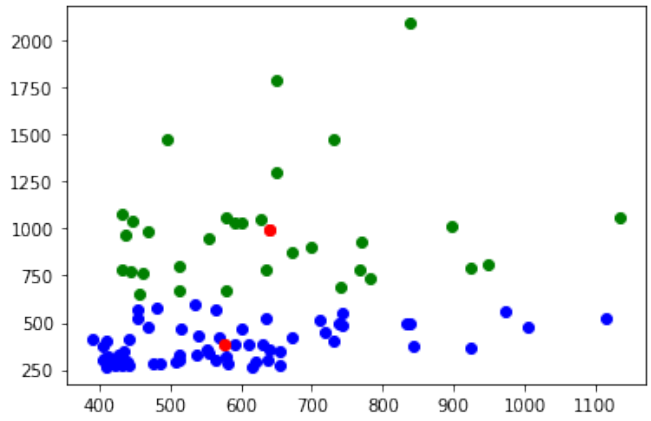

Figure 11: CO_2 Emissions versus Total Power based on k-means algorithm.

After executing all the required steps, the clustering method is applied to obtain the data separation shown in Figure 11, which just take two centroids based on running with the lowest total within-cluster sum of square as the final clustering solution. Another way to carry out this method is to apply it directly without identifying the ideal number of clusters. The following results are obtained.

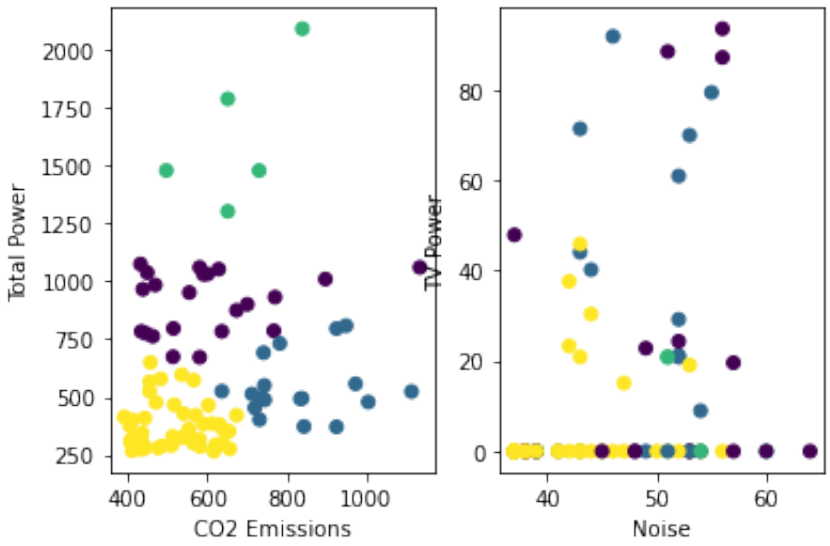

Figure 12: Analysis of correlation between sensors data based on k-means algorithm.

The output image is showing the four different clusters with different colors. The clusters are formed between two parameters of the dataset in each case. A basic expected pattern can be determined at this point:

- At the bottom-left of the left graph (yellow points), the lowest CO^2 emissions correspond to a lower total power.
- Upper-middle of the right graph (blue points) reflects more noise than usual for higher power consumption values of the TV.

4 Conclusions

In this Analysis we were able to understand and predict the occupancy in the house, and is was all because of the smart use of sensors installed and IoT devices to communicate data. We conclude that smart houses are the future and are very essential to understand the user behaviour to be able to reduce the power consumption in a better way for a more sustainable future. We need alsoto mention the importance of machine learning and AI that makes it easier for us to do our estimations and predictions that are harder for us by just looking at data.

Regarding the clustering method, it was possible to identify the most important relationships between the selected variables. Low power consumption means low CO_2 emissions and thus, an increase in total power means an increase in emissions. However, this increase is not always proportional, because other emission factors not related to power can have an influence, such as an increase in the number of people in the room, emergencies, etc. On the other hand, and as expected, the use of a device such as a television represents a significant and direct variable for the presence of noise. As in the previous case, noise variation also responds to other environmental and occupancy factors.

From these results it is possible to estimate room occupancy by noting a non-proportional increase in both CO_2 emissions and the presence of noise other than the usual consumption and duration of an appliance such as a television. All this considering that the method used is sensitive to inappropriate sample values and varies randomly with respect to its centroids.

References

[1] Fatehpour, Y., 2021. 5 IoT Trends in 2022. [online] eWEEK. Available at: <https://www.eweek.com/networking/iot-trends/>.

[2] n.d. EXPE-SMARTHOUSE – The full connected living house. [online] Available at: <http://expe-smarthouse.org/index.php/en/project/>.