Stjelja, Davor; Jokisalo, Juha; Kosonen, Risto

From electricity and water consumption data to information on office occupancy

Published in:
Sisäilmastoseminaari 2021. Verkkoseminaari 9.3.2021

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Stjelja, D., Jokisalo, J., & Kosonen, R. (2021). From electricity and water consumption data to information on office occupancy. In M. Ahola, & A. Merikari (Eds.), *Sisäilmastoseminaari 2021. Verkkoseminaari 9.3.2021* (pp. 113-118). (Sisäilmayhdistys raportti; No. 39). SIY Sisäilmatieto Oy.

FROM ELECTRICITY AND WATER CONSUMPTION DATA TO INFORMATION ON OFFICE OCCUPANCY: A SUPERVISED AND UNSUPERVISED DATA MINING APPROACH

Davor Stjelja[1,2] Juha Jokisalo[1] and Risto Kosonen [1,3]

[1] Department of Mechanical Engineering, Aalto University, Espoo, Finland
[2] Granlund Oy, Helsinki, Finland
[3] College of Urban Construction, Nanjing Tech University, China

ABSTRACT

Climate change and technology development are pushing buildings to become more sophisticated. The installation of modern building automation systems, smart meters and IoT devices is increasing the amount of available building operational data. The common term for this kind of building is a smart building but producing large amounts of raw data does not automatically offer intelligence which would offer new insights to the building's operation. Smart meters are mainly used only for tracking the energy or water consumption in the building. On the other hand, building occupancy is usually not monitored in the building at all, even though it is one of the main influencing factors of consumption and indoor climate parameters. This paper is bringing the true smart building closer to practice by using machine learning methods with sub-metered electricity and water consumptions to predict the building occupancy. In the first approach, the number of occupants was predicted in an office floor using a supervised data mining method Random Forest. The model performed the best with the use of all predictors available, while from individual predictors, the sub-metered electricity used for office equipment showed the best performance. Since the supervised approach requires the continuous long-term collection of ground truth reference data (between one to three months, by this study), an unsupervised data mining method k-means clustering was tested in the second approach. With the unsupervised method, this study was able to predict the level of occupancy in a day as zero, medium, or high in a case study office floor using the equipment electricity consumption.

INTRODUCTION

In order to promote smart ready technologies for the building sector, the EU has introduced a smart readiness indicator (SRI). The purpose of SRI is to determine the capability of buildings in using the information and communication technologies to adapt the building operation to the needs of the occupants and the grid while improving the overall performance of the building [1]. The outdoor and indoor environment, the building's energy consumption and the system operation are usually monitored with sensors and meters, but the occupancy pattern and behaviour are not. Collected data processed using data mining techniques can provide information about building occupancy.

Using building operational data with both supervised and unsupervised data mining methods to find more information on occupancy and its influence on building operation, has already been part of various studies, e.g. Kleiminger et al. [2], Mora et al. [3], Yang,

L. et al. [4]. Building operational data used in those studies was typically electricity usage, measured occupancy or indoor climate measurements. Information being discovered in those studies was ranging from grouping buildings or consumers by their energy usage, finding occupancy and consumption patterns or detecting and predicting the occupancy in buildings. This has all led to the development of this study, where the novelty is using a commonly measured variable such as sub-metered electricity or water consumptions to acquire the information on the occupancy from a larger area, such as office building floor. This is done with two different data mining approaches by using supervised and unsupervised methods. By using a supervised data mining method, the possibility of predicting the number of occupants on the office floor is assessed. For the training of the supervised methods, the long-term continuous and costly ground truth collection of the actual occupancy is needed. Following-on this, the unsupervised data mining method of consumption data to find daily occupancy level is evaluated.

METHODOLOGY

The building serving as a case study in this work is an office building of a consultancy company in Helsinki, Finland. The building with four floors above the ground was built in 1990. The total occupied floor area is 9672 m2. Focus in the study is the third floor, which has an area of 1900 m2. The electricity consumption data was collected with meters measuring the power consumed by the lighting system and the equipment connected via sockets in office. Water consumption is measured on an hourly basis by the central water meter for the building. For this study, the assumption is that water consumption in the office building is equally distributed on three floors since all floors are similar. The ground truth was collected using the people counting cameras installed on all four entrances to the floor.

Supervised method – Random Forest

The supervised method applied in this work was the Random Forest, a popular machine learning algorithm. The Random Forest was chosen because it has been already successfully applied in similar studies, where it has shown good performance, and it is easy to implement. The Random Forest was applied with Python module Scikit-learn, which is a general-purpose, high-level programming language for machine learning [5]. For results verification of the supervised method, the following error metrics have been used: explained variation, root mean square error (RMSE) and mean biased error (MBE).

Unsupervised method – the k-means clustering

The clustering method chosen for this work was k-means, which is one of the most popular clustering methods and it has been widely used for the clustering of raw datasets in energy and built environment field. The k-means algorithm used in this work is *tslearn*, which is a Python package that provides machine learning tools for time-series analysis [6].

RESULTS

Data gathering of the electricity and water consumption data was done by using Granlund Manager [7], a facility management software to which smart metered data was stored. The acquired dataset was for the period between the 1st of February and 31st of August 2017. Before analytics was performed, the dataset had to be cleaned.

First, the weekends were taken out of the dataset, since the floor was very rarely occupied during weekends. Second, since there were no occupants on the floor during night-time, it was decided to use only the data between 06:00 and 21:00.

Results of the Supervised Method

The dataset contained collected raw data and extracting additional features from the data could improve the prediction model. Extracted features used in model are first and second-order difference which captured temporal variations, while the moving average took into account the time delay between parameters.

In Table 1, the results of the supervised method for occupancy rate prediction for the test period (15th to 31st of August 2017) are presented. Results show that All predictors have the best result, when looking at the RMSE and explained variation, with slight underprediction, which is shown with a negative MBE value. Light and equipment are showing the second-best results, but with higher underprediction. Total floor electricity consumption shows the lowest absolute MBE value, while the value being positive means that the model using this predictor usually overpredicts. Surprisingly, Water consumption was shown to be a better predictor than Lighting consumption, which was shown as the worst. Regarding the explained variation, all results were above 90%, which meant that the models had a good prediction capability where even the model with the Lighting consumption predictor, still showed an average explained variation of 93%.

Comparison between ground truth occupancy and predicted occupancy using the best performing model (all predictors) is shown in Figure 1 for the test period. It is possible to see that predicted occupancy was following quite well with the ground truth occupancy and that the most considerable difference between the predicted and ground truth mostly happened during peak hours.

Table 1. Results of the supervised method for occupancy rate prediction shown through three error indices: RMSE, MBE and explained variation (EV).

Set of Predictors	RMSE (persons)	MBE (Persons)	EV (%)
Lighting	10.52	−1.6	93
Equipment	8.87	−2.34	95
Light and Equipment	8.14	−2.07	96
Total floor electricity consumption	8.8	0.41	95
Water consumption	9.67	−0.5	94
All predictors	7.88	−0.66	96

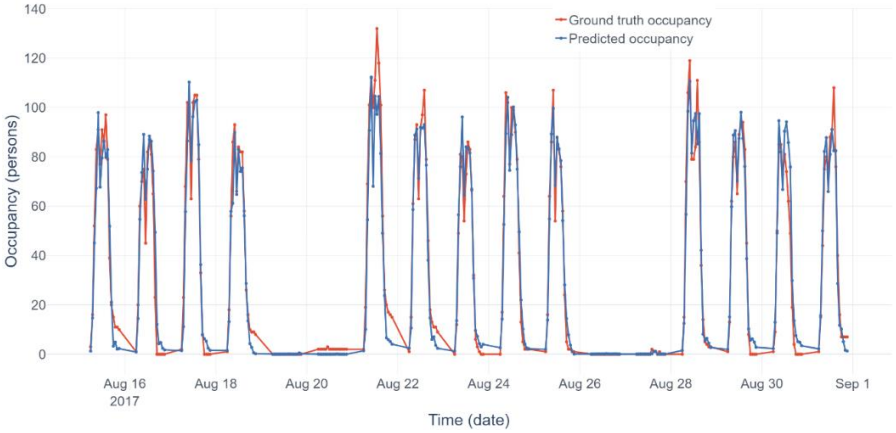

Figure 1. Measured and predicted occupancy during the test period. The predicted occupancy was calculated using all predictors

The results from the supervised method depicted that with the Random Forest model it was possible to predict the number of people present in the office and to create an occupancy profile that followed the measured one with reasonable accuracy. Generally, the higher number of occupancy correlated predictors used, the greater the occupancy prediction capability the model would have. Although, a model based on the highly correlated predictor such as office equipment consumption was in itself, giving a good result.

Results of the Unsupervised Method

The unsupervised clustering method k-means was used on the same dataset as in supervised method, the dataset was split into day profiles so that the days could be clustered. The time between 06:00 and 21:00 was used for the profile, which made the day profile size of 16 values.

In k-means clustering, most important parameter is number of clusters (k), in this work to select k the Elbow method was used [8]. Iterating between two and nine clusters, with Elbow method we concluded that for electricity equipment consumption, optimal number of clusters is three. Same analysis was done for lighting electricity consumption and water consumption, where the optimal number of clusters was chosen to be four and six, respectively.

In Figure 2, three clusters of equipment electricity consumption are presented, together with day profiles of consumption belonging to each cluster (black lines) and with the cluster centre (red line). Each of these clusters represents a typical equipment electricity consumption pattern for the case floor. Since the office equipment consumption was mostly influenced by occupancy (e.g., computers), the assumption was that by clustering this consumption, the occupancy patterns could be found. In Figure 3, the occupancy measured by people counting cameras is presented with each day profile coloured by the belonging cluster. The highest occupancy belonged to the first cluster (black) and the days with the lowest occupancy to the second cluster (red), while the days with occupancy in between belonged to the third cluster (blue). This figure shows that there was a good correlation between measured equipment consumption and office

occupancy. Correlation was most visible in the second cluster, where occupancy was nearly zero and consumption was low, compared to the other days. The first and third clusters were not clearly divided since they contained some border cases, which meant that certain days could potentially belong to either of those clusters. This was not the case with the second cluster, where belonging days with their consumption (and occupancy) had a clear membership. Looking at the dates belonging to each cluster, we could classify Cluster 1 as regular working days, Cluster 2 as public holidays and Cluster 3 as "semi-holidays". Semi-holidays being the periods when people usually take their vacation and "bridge" days, which are days between the public holiday and a weekend.

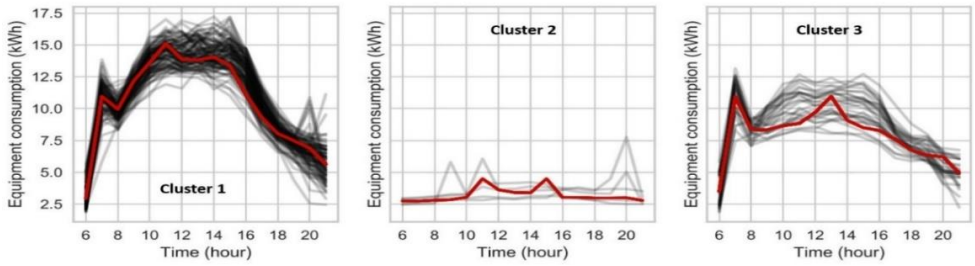

Figure 2. Equipment electricity consumption clusters. (Black) lines represent the day profiles of consumption, while (red) lines represent the cluster centre.

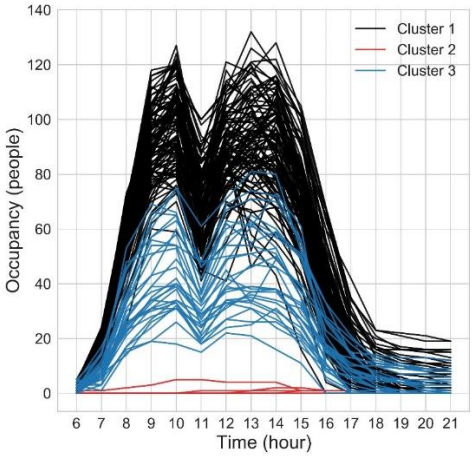

Figure 3. Occupancy day profiles coloured by clusters of equipment electricity consumption.

Following the equipment electricity consumption, the same method was performed on electricity consumption for office lighting and water consumption. Unfortunately, both lighting and water consumption, did not show good ability to group the days by occupancy into logical order (figures presenting this were omitted because of page limitation).

DISCUSSION & CONCLUSION

This study and its results have tried to make several contributions. First, was to find information on building occupancy, which is usually not available, by using the data which is more common in buildings. Secondly, to increase the value of energy and water consumption sub-metering in commercial buildings. Last, is to promote the smart building concept where one data stream can be used for multiple purposes. Further research is needed to reinforce these contributions, especially regarding scaling the methods to real-world applications.

Two methods of data mining studied in this work have been successful in discovering information about the office occupancy from the sub-metering. With the supervised method, this study proved that it is possible to predict the number of occupants in a larger office facility by using metered equipment, lighting electricity consumption or using water consumption. The unsupervised method developed in this work was able to cluster the days by the level of the office occupancy using equipment electricity consumption. In the case study office, the three levels of occupancy were found: zero (low), medium and high.

The way of working could change radically in the future and remote work will become more common. This means that there could be fewer regular working days and more irregular days. Having insight into which days and how many in a year are irregular can help to better manage buildings from the technical and the real-estate perspective.

REFERENCES

1. European Union Directive 2018/844 of the European Parliament and of the Council of 30 May 2018 amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency. Off. J. Eur. Union 2018, L 156/75-L 156/91.
2. Kleiminger, W.; Beckel, C.; Staake, T.; Santini, S. Occupancy Detection from Electricity Consumption Data. 2013, 1–8, doi:10.1145/2528282.2528295.
3. Mora, D.; Fajilla, G.; Austin, M.C.; De Simone, M. Occupancy patterns obtained by heuristic approaches: Cluster analysis and logical flowcharts. A case study in a university office. Energy Build. 2019, 186, 147–168, doi:10.1016/j.enbuild.2019.01.023.
4. Yang, L.; Ting, K.; Srivastava, M.B. Inferring occupancy from opportunistically available sensor data. 2014 IEEE Int. Conf. Pervasive Comput. Commun. PerCom 2014 2014, 60–68, doi:10.1109/PerCom.2014.6813945.
5. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.
6. Tavenard, R.; Faouzi, J.; Vandewiele, G.; Divo, F.; Androz, G.; Holtz, C.; Payne, M.; Yurchak, R.; Rußwurm, M.; Kolar, K.; et al. tslearn: A machine learning toolkit dedicated to time-series data 2017.
7. Granlund Granlund Manager Available online: https://www.granlundmanager.com/ (accessed on Mar 23, 2020).
8. Thorndike, R.L. Who belongs in the family. In Proceedings of the Psychometrika; Citeseer, 1953.