Contextualized model for occupants' activities estimation in connected buildings

Huynh Phan*[1], Thomas Recht[1], Laurent Mora[1], Stéphane Ploix[2]

[1] **I2M Bordeaux, University of Bordeaux**
CNRS, Arts et Metiers Institute of Technology, Bordeaux INP, F-33400 Talence, France
[2] **G-SCOP, Grenoble Institute of Technology**
UMR CNRS 5272, 46 Avenue Felix Viallet, 38031 Grenoble Cedex 1
***huynh.phan@u-bordeaux.fr**

RÉSUMÉ. Récemment, de nombreuses approches se sont concentrées sur l'estimation des activités des occupants pour réduire l'écart significatif entre les impacts énergétiques simulés et réels. Cependant, la plupart d'entre elles prennent en compte des données statistiques d'occupation moyenne de nombreux ménages, ce qui est insuffisant pour un ménage spécifique dans un contexte particulier (caractéristiques, habitudes). Cette contribution propose une approche générale basée sur des graphes tenant compte du contexte particulier d'un ménage pour estimer les activités des occupants. Concrètement, un réseau bayésien basé sur les effets mesurés des activités est construit à partir de données de mesure collectées via divers capteurs (CO_2, température, humidité, consommation d'énergie d'appareils électriques, etc.) et de connaissances sur un bâtiment résidentiel spécifique. Les résultats mettent en évidence quatre variables principales pour estimer l'activité « cuisiner le petit déjeuner » : la consommation électrique la machine à café, celle du grille-pain, du micro-ondes, et le moment de la journée, le modèle final présentant un score F1 supérieur à 85 %.

MOTS-CLÉS. estimation d'activité, réseau bayésien, simulation.

ABSTRACT. Recently, there have been many approaches focusing on occupants' activities estimation to reduce the significant discrepancy between simulated and actual energy impacts. However, most of them consider general statistical data including average occupancy scenarios of many households that are not sufficient for a specific household in a particular context (characteristics, habits). This contribution proposes a general approach taking into account a particular context to estimate the occupants' activities in a specific household. Specifically, an activity consequences-based Bayesian Network is built based on measurement data collected from sensors (CO_2, temperature, humidity, the energy consumptions of electrical appliances, etc.) and the knowledge from a residential building. The final results show that the activity "cook breakfast" is mostly related to four variables (coffee machine, toaster and microwave energy consumption but also the time of the day) with a F1-score is higher than 85 %.

KEYWORDS. activity estimation, Bayesian Network, simulation.

1 Introduction

Recently, many researchers have focused on studying the factors, which are influencing energy performance in buildings. They are interested in exploring and understanding more accurately the major problems resulting in wasted energy in residential buildings. Zaraket (2014) addressed occupants' activities as factors having significant impacts on energy consumption in dwellings. In general, three major end-use groups are consuming the energy in residential buildings: space heating/cooling, electrical appliances, and hot water. They are strongly dependent on the activities of occupants. People use energy to satisfy their comfort and do daily life activities such as cooking, cleaning, washing clothes, etc. Therefore, the energy consumption of buildings is highly related to the profiles of occupants, their habits and their activities during their presence.

Consequently, occupants' activities are important factors for optimizing the energy performance of the building. Jia et al. (2017) presented three major benefits of the capability of detecting and modeling occupants' activities earlier for energy simulation: more realistic, support to building systems control measures, improve performance and services to users. Besides, occupants' activities are believed as the major element in explaining discrepancies between simulated and actual energy consumption. Better estimation in occupants' activities can improve the energy estimation in the building simulation software, which is essential for the Building Energy Verification. Additionally, the knowledge gains from the estimation of activities can help give the feedback or recommendation for occupants to enhance their energy usages. They are also used to build the profile of activities for the energy simulation in the design phase.

Many approaches have been proposed to estimate occupants' activities in residential buildings in recent years. The statistical model is a popular approach to extract the patterns of activities from survey data based on the characteristics (age, job, building type, etc.) of residential buildings. Due to the statistical nature of data, this approach considers average occupancy scenarios based on the context of many buildings and households with different habits. However, they are not sufficient to represent a specific building context, which includes determined settings (rooms, activities, appliances, etc.) and the particular members with corresponding characteristics and habits. Consequently, to improve the accuracy of occupants' activities estimation in a particular household, it is necessary to have a model taking into account the context in estimating the activities of members. Starting from the state of the art (see Literature Review section), this study proposed a general approach using a graphical model, which is based on not only measurements but also the context, to estimate the activities in a specific household. In this approach, the contexts information help locate the activities to their locations (rooms). For each activity, the relevant features are selected based on the collected labels from occupants and the measurements. Then, a Bayesian Network is built for each activity and estimate it independently from the others (see Methodology section). The case study corresponds to an individual house, and the shown results are about the estimation of the activity "cook breakfast" (see Case study: Cooking breakfast section).

2 Literature Review

In general, occupants' activity models can be divided into four main groups: profile-based, data-based, probabilistic and agent-based.

In **profile-based** approach, the patterns of the activities are scheduled and fixed for each thermal zone. They are extracted from standard conditions or statistical analysis of actual observations. However, Fabi et al. (2013) stated that these scenarios were repeatable for many households while their preferences were different, ambiguous and they played an important role in operating activities. Hence, predefined scenarios are not efficient to model occupants' activities.

Data-based approach aims to determine the patterns and systematic relationships between related variables using data mining, machine learning techniques such as Super Vector Machine

(SVM) and Decision Tree (DT). (Zhou et al., 2014) applied SVM to predict activities of using appliances and related electricity consumption. Alhamoud et al. (2015) applied DT to identify patterns of occupants' activities (eating, watching, etc.) for energy consumption estimation and energy saving recommendation systems. However, data-based is a deterministic black-box model, it does not provide human understandable outputs and does not concern the variety of the activities, which come from the uncertainty of preferences and habits.

Probabilistic approach is the most popular approach for activities estimation. The probability of occurrence is estimated based on statistical data. Wilke et al. (2013) proposed a probabilistic model based on French time-use survey data to simulate the sequences of 20 daily activities. The model was based on three types of time-dependent quantities and 41 candidate variables describing the characteristics of the individual (career, day, healthy, gender, job, etc.). Markov chain model was applied to estimate the transition probability between activities. Aerts (2015) derived realistic activity data in dwellings from the combined Belgian Time-Use Survey and Household Budget Survey. In this article, activities were divided into two groups: tasks and personal activities. Tasks were usually performed by only one of the household members in period of time. Personal activities can be performed independently from each member. These models concern the diversity of activities due to the investigation of both the attributes of occupants and the settings of building. However, the statistical data contains the information of many dwellings, individuals with different habits and characteristics, thus these models are hybrid and not entirely sufficient to estimate activities and evaluate energy feedback in a particular household.

Agent-based approach aims to investigate the interactions occupants-building and occupant-occupant through self-organization agents, which make the decision to satisfy their comfort. In (Kashif et al., 2013), the authors proposed an agent-based model to simulate reactive/deliberate group activities taking into account personal characteristics. BRAHMS (Business Re-design Agent-based Holistic Modeling System) language and 5W1H (who, when, where, what, why, how) context were used to learn activities patterns for energy control and management strategies. However, agent-based models become too complex when the number of agents increases because of the interactions network. The agent's comfort is also difficult to define because of the diversity of characteristics, personality and habits.

Besides the mentioned approaches, **Bayesian Network** (BN) is an approach to predict occupants' activities based not only on dataset but also on expert knowledge. It offers intuitive graphical representations of the uncertain relations among variables and their conditional dependencies *via* a directed edge. Hawarah et al. (2010) built a BN to predict the usage of oven in cooking based on the expert structure of causal nodes (hour, type of day, month) and effect nodes (duration, energy). Tijani et al. (2015) also proposed a Dynamic Bayesian Network (DBN) to predict the door movement in an office. The authors considered the periods of year, the periods of day, calendar, visitor presence, occupants' presence, season, CO_2 concentration and past door movement as influent variables. BN is flexible and friendly for the human to understand. It leads to the advantages for the people to explain and validate the model. However, all studies focus on simple activities such as the usage of appliances, the actions of windows/door or the presence of occupants in the building.

3 Methodology

Based on the Literature Review, this study aims to estimate daily human activities (cook breakfast, personal care, clothes-washing, sleep, absence, etc.) independently based on both measurement data and knowledge of the context in a specific household. This section presents the methodology for estimating a particular activity in a dwelling, with four main steps: 1. Data collection. 2. Data normalization. 3. Features selection. 4. Consequence-based Bayesian Network for activity estimation.

3.1 Definitions

In this approach, the contexts of a house include the activities with the labels provided by occupants, the list of rooms, the list of activities, and a "room-activity" relationship, which describes the room involved for a particular activity, and the available electrical appliances. Taking into account this information helps the estimation of the activities be more precise than the use of hybrid information from statistic data.

Different activities can occur in the same room while an activity can be done in many rooms. It is assumed that the impacts of an activity in different rooms are different. For example, the activity "cleaning" can take place in the living room, the kitchen, or the bedroom and their effects are distinct. To deal with this problem, we define a room-centered activity A as *(activity, room)*. It has a starting time, an ending time, an involved room, and several cause effects, impacts on sensors and appliances, and also possibly on electricity consumption. The time of day is divided into incremental steps. In each step, A is represented as a label, which could be 0 or 1 corresponding to the status inactive or active. A is detected as active in a step of time if either its starting time or its ending time is in that period. To be convenient, we will use the term "activity" for the term "room-centered activity" in this study. In general, an activity A in each time can be defined as the following:

$$A \overset{is}{\leftrightarrow} (activity, room) \overset{has}{\longrightarrow} label \overset{causes}{\longrightarrow} effects$$

The problem can be declared as a binary supervised problem. The study makes efforts to estimate the label of activity in each time of day. It leads to the necessity to collect labels of activities from occupants. Considering the duration of activities and the comfort of occupants for retrieving labels, 30-minutes is selected to be the step of time in this study.

3.2 Data collection

This paper focuses on the model of activities from both the measurements and the contexts of a household. A heterogeneous sensor environment has been installed to capture the information from a detached house. There are three groups of variables measured by the sensors: environment (temperature, humidity, CO_2, etc.), electricity consumption, and occupancy-related (motions, door/ windows contacts). Moreover, information such as time in the day extracted from data could be grouped into a fourth group: habit. The knowledge of household's context is also collected through the questionnaires. In addition, labels of activities for each time of day need to be collected from occupants to learn the model. Since a room-centered activity is connected to a specific room, only variables related to the room of that activity are considered.

3.3 Data normalization

To clean up and normalize the raw data of measurements, a 3-steps process is applied: Firstly, abnormal points of variables are detected and removed based on their natures or the physical properties of sensors. Secondly, measurement data is aggregated into a 30-minutes step. Different aggregation methods are applied for different groups of variables: a total of the energy consumption for electricity consumption, the average value for the others. Finally, filling methods (linear interpolation, forward fill, constant) are used to fill missing values. This process can be presented as the following:

$$\text{Raw data} \overset{apply}{\longrightarrow} \text{filter} \overset{apply}{\longrightarrow} \text{aggregate} \overset{apply}{\longrightarrow} \text{fill missing values} \overset{to}{\longrightarrow} \text{normalized data}$$

3.4 Features selection

The normalized values of each sensor is represented as a vector $V = [v_0, ..., v_i, ..., v_n]$, v_i - normalized value at time i, n - the number of samples. However, many useful features can be hidden

from the data and it is essential to extract them. Precisely, for each time i, for environment and occupancy-related variables, the difference $\Delta v = v_i - v_{i-1}$ is obtained while the total energy consumption v_i is selected for electrical variables. In addition, each feature in the Bayesian Network model should be categorical. Environmental variables are discretized manually by their nature. Values of occupancy-related features are divided into two values: 0 (inactive) or 1 (active). For the electrical appliances, we assume that electrical appliances have three levels of energy consumption: low, medium and high. In spite of extracted continuous value for each time i, features of electrical variables are thus represented by the level l_i, $\forall l_i \in [0, 1, 2]$ corresponding to these levels. These levels are detected by clustering method as k-means algorithm.

However, many features are redundant in the model of an activity. Removing them and keeping the most relevant features can help reduce the noise in the model and save the money for sensors. *Information Gain (IG)* is a popular method to estimate the importance of a feature with an activity. This method is based on *entropy (ENT)*, which measures the disorder of the data. *IG* estimates the reduction of impurity of labels for each feature, the higher information gain is, the more relevant feature is. With the vector of a feature $X = [x_0, ..., x_i, ..., x_n]$ with d possible values, and the labels of activity $Y = [y_0, ..., y_i, ..., y_k]$, k - the number of class of labels, n - the number of sample, $p(x)$- the percentage of x in the dataset. *ENT* and *IG* are defined as:

$$ENT(X) = -\sum_{i=1}^{d} p(x_i) \times \log p(x_i),$$

$$IG(X, Y) = ENT(Y) - \sum_{i=1}^{k} p(y_i) \times ENT(X|y = i)$$

Finally, m highest *Information Gain* features are selected to build the activity estimation model.

3.5 Consequences-based Bayesian Network

Bayesian Network is a graphical model including nodes and edges, which represents the cause-effect relationship $X \rightarrow Y$ between variables. Two elements in the graph have to be determined: the structure and the conditional probability table. The structure explains the relationships $X \rightarrow Y$ and the conditional probability table shows the conditional probabilities $p(Y|X)$ between variables. In this study, the structure of network is built based on expert knowledge and the conditional probability table is deduced from the data. Selected features are grouped into four categories: electricity consumption (C), environment (E), occupancy-related (O) and habits (H). It is assumed that the habits have influence in the decision of performing an activity and once an activity has been performed, it causes effects such as electricity consumption or changes on the environment. From this assumption and the groups of features, the consequences-based structure proposed for Bayesian Network is shown in figure 1. The joint probability of the network is:

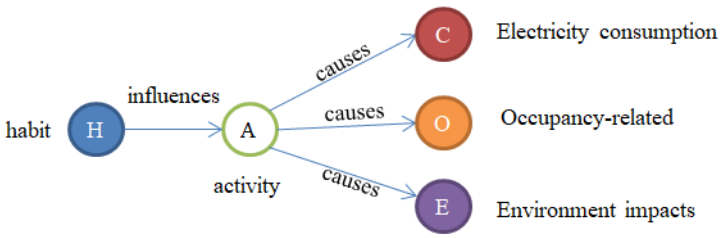

Figure 1. Consequences-based Bayesian Network

$$p(A, E, C, H, O) = p(A|H) \times p(E|A) \times p(C|A) \times p(O|A)$$

In each 30-minutes iterator, from the set of labels X, the activity state x^* is estimated by maximizing the joint probability as the formula 1.

$$x^* = \underset{x \in X}{\operatorname{argmax}}\, p(x|H) \times p(E|x) \times p(C|x) \times p(O|x) \tag{1}$$

4 Case study: Cooking breakfast

4.1 Testbed

The testbed is a detached house in France, which includes five household members. There are 11 rooms (four bedrooms, two bathrooms, one kitchen, one meal room, one living room, one mezzanine, and one cellar) and 14 daily activities are considered. In this case study, the activity *(cook breakfast, kitchen)* is studied, the concerned room is *kitchen*, where 11 sensors are installed: 2 motions, 2 door/windows, 6 electricity consumption (toaster, coffee machine, robot, yogurt maker, microwave, fridge), 1 for temperature, humidity, and luminosity. The dataset covers 54 weekdays (not weekend) from 01/02/2019 to 30/04/2019, except vacation days from 15/03/2019 to 19/03/2019. The selected step of time is 30-minute. The labels of *cook breakfast* in this period were collected manually with supports from the household members.

4.2 Results and discussion

The estimated distribution of collected labels over the time of day is shown in figure 2.

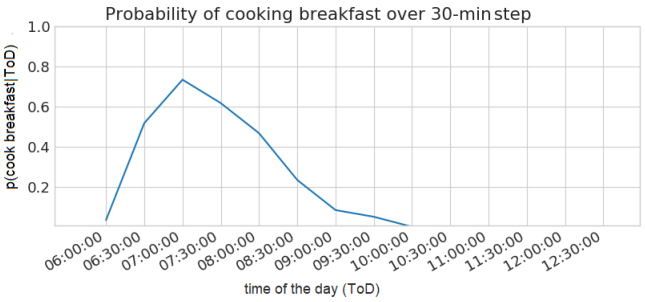

Figure 2. Distribution of collected labels over 30-min steps

It shows that occupants usually cook breakfast between 7 a.m and 8 a.m, it describes a contextualized habits of occupants in a particular case study. These habits are strongly related to the schedule, preferences of members in the house.

As mentioned in section 3, the useful features were the time of day (ToD), the energy consumption's level of appliances, the mean value, the difference of environmental variables (humidity, luminosity, temperature), and the state of occupancy-related variables (motions, door/windows) in the same 30-minute of the activity label. Figure 3 presents the information gains of these variables, which are estimated from the data set. The results show that besides the habit, the energy consumption levels of appliances involved frequently in the same time of the activity are the most relevant features for this activity. For instance, to estimate activity *cook breakfast* between 7:00 and 7:30, the levels of energy consumption of the coffee machine and the microwave at this time-step are important features. Based on the estimated information gain, we select 4 best features (ToD, energy consumption levels of the microwave, the coffee machine, and the toaster) to build a consequences-based Bayesian Network. The structure of this network is presented in figure 4.

K-fold cross-validation technique is used to validate the quality of the model. The dataset is divided into k subsets. For each fold, k-1 subsets are selected for training and the other one is for testing. This technique helps avoid the overfitting issue because, through many folds, most

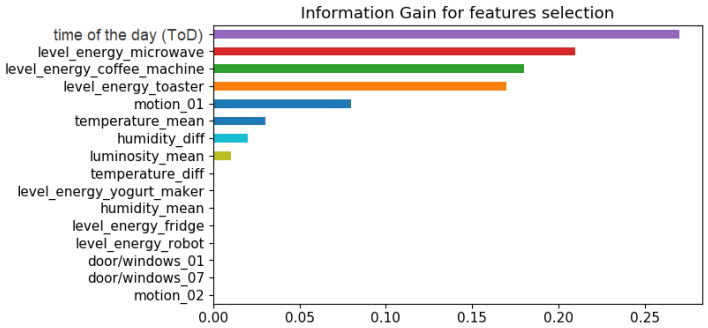

Figure 3. Information Gain for features in activity cook breakfast estimation

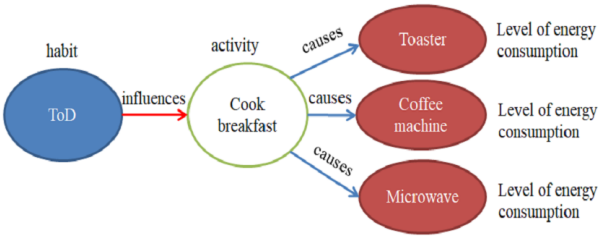

Figure 4. Consequences-based Bayesian Network for activity cook breakfast estimation

of the data is used for not only training but also testing. Besides, the F1-score is selected as the validation metric because this score concerns both the precision and the recall of the model. The higher F1-score is, the better model is. In general, F1-score can be estimated as the formula 2.

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{2}$$

In this case study, 5 folds cross-validation is applied and the obtained F1-score is approximately 85 %. However, the habit of occupants is not consistent all the time. They can change by internal aspects such as health issues, change of age, or external aspects such as the seasons and the potential retrofit of the building. It leads to the limitation of this model when the transition of the activity pattern needs to be considered. In addition, the task of collecting labels is a challenge because it is difficult for occupants to provide exactly labels for each time of day of the day, especially when we need information about many activities.

5 Conclusion

A supervised learning approach has been proposed in this paper to estimate activities in residential building. The proposed methodology helps to determine useful features and necessary sensors of an activity's estimation through the concept of *Information Gain*. Bayesian Network is used as an understandable model for training and predicting activity from relevant features. The structure of the network is based on the consequences of an activity, which is assumed to include environment, electricity consumption and occupancy-related effects. A case study on the *cook breakfast* activity is presented to explain the methodology in a particular context. Results show that *cook breakfast* is mainly linked to the energy consumption of frequently involved appliances (coffee machine, microwave, toaster) and the time in the day, which represents to occupants habit. Cross-validation is used and F1-score is approximately 85 %. Depending on not only measurements but also context's information (labels of activities, locations of activities, appliances, etc.), the approach improves the knowledge of activities and make it possible to reduce the discrepancies between the energy simulation and the measurements. However, the limitation

of collecting labels is an obstacle for modeling many activities. To deal with this problem, we have been developing a mobile application for getting labels. Occupants can use it to provide the information of activities every time and everywhere they feel comfortable. They can also see the patterns of activities in the past to improve the accuracy of their labels. In addition, this approach requires the sensors for measurement and the contexts' knowledge. Therefore, it is not flexible for the larger sample of households as the statistical approach.

6 Acknowledgements

This work was funded by the Regional Council of Nouvelle Aquitaine within the joint research lab GP2E between I2M laboratory and technical center Nobatek/INEF4.

References

Aerts, D. (2015). *Occupancy and Activity Modelling for Building Energy Demand Simulations, Comparative Feedback and Residential Electricity Demand Characterisation.* PhD thesis, Vrije Universiteit Brussel.

Alhamoud, A., Xu, P., Englert, F., Reinhardt, A., Scholl, P., Boehnstedt, D., et Steinmetz, R. (2015). Extracting Human Behavior Patterns from Appliance-level Power Consumption Data. In *Wireless Sensor Networks*, Lecture Notes in Computer Science, pages 52–67. Springer International Publishing.

Fabi, V., Andersen, R. V., et Corgnati, S. P. (2013). Influence of occupant's heating set-point preferences on indoor environmental quality and heating demand in residential buildings. *HVAC&R Research*, 19(5):635–645.

Hawarah, L., Ploix, S., et Jacomino, M. (2010). User Behavior Prediction in Energy Consumption in Housing Using Bayesian Networks. In *Artificial Intelligence and Soft Computing*, Lecture Notes in Computer Science, pages 372–379. Springer Berlin Heidelberg.

Jia, M., Srinivasan, R. S., et Raheem, A. A. (2017). From occupancy to occupant behavior: An analytical survey of data acquisition technologies, modeling methodologies and simulation coupling mechanisms for building energy efficiency. *Renewable and Sustainable Energy Reviews*, 68:525–540.

Kashif, A., Ploix, S., Dugdale, J., et Le, X. H. B. (2013). Simulating the dynamics of occupant behaviour for power management in residential buildings. *Energy and Buildings*, 56:85–93.

Tijani, K., Ngo, Q. D., Ploix, S., Haas, B., et Dugdale, J. (2015). Towards a General Framework for an Observation and Knowledge based Model of Occupant Behaviour in Office Buildings. *Energy Procedia*, 78:609–614.

Wilke, U., Haldi, F., Scartezzini, J.-L., et Robinson, D. (2013). A bottom-up stochastic model to predict building occupants' time-dependent activities. *Building and Environment*, 60:254–264.

Zaraket, T. (2014). *Stochastic activity-based approach of occupant-related energy consumption in residential buildings.* PhD thesis, Ecole Centrale Paris, Paris.

Zhou, S., Wu, Z., Li, J., et Zhang, X.-p. (2014). Real-time Energy Control Approach for Smart Home Energy Management System. *Electric Power Components and Systems*, 42(3-4):315–326.