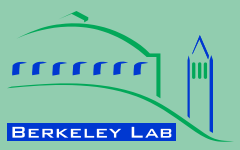
LBNL-xxxxx



ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Occupancy schedules learning process through a data mining framework

Simona D'Oca[1,2], Tianzhen Hong[1]

[1] Building Technology and Urban Systems Division Energy Technologies Area

[2] Polytechnic of Turin, Italy

May 2015

This is an article published in the journal of Energy and Buildings, February 2015.

Disclaimer

Please cite this report as follows:
D'Oca, S., Hong, T. Occupancy schedules learning process through a data mining framework. Energy and Buildings, 88: 395-408, 2015. LBNL-xxxxx.

Acknowledgement

This work was sponsored by the United States Department of Energy (Contract No. DE-AC02-05CH11231) and the China Ministry of Housing and Urban - Rural Development and the Ministry of Science & Technology (Grant No. 2010DFA72740-02) under the U.S.-China Clean Energy Research Center for Building Energy Efficiency. It is also part of the research of Annex 66, Definition and Simulation of Occupant Behavior in Buildings, under the International Energy Agency Energy in Buildings and Communities Program.

Occupancy schedules learning process through a data mining framework

Simona D'Oca[a,b], Tianzhen Hong[a,*]

[a]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[b]Polytechnic of Turin, Corso Duca degli Abruzzi 24, 10129, Torino, Italy

[*]Corresponding author. T. Hong. Tel: +1(510)4867082; Fax: +1(510)4864089
Email address: thong@lbl.gov

Occupancy schedules learning process through a data mining framework

Abstract
Building occupancy is a paramount factor in building energy simulations. Specifically, lighting, plug loads, HVAC equipment utilization, fresh air requirements and internal heat gain or loss greatly depends on the level of occupancy within a building. Developing the appropriate methodologies to describe and reproduce the intricate network responsible for human-building interactions are needed. Extrapolation of patterns from big data streams is a powerful analysis technique which will allow for a better understanding of energy usage in buildings. A three-step data mining framework is applied to discover occupancy patterns in office spaces. First, a data set of 16 offices with 10 minute interval occupancy data, over a two year period is mined through a decision tree model which predicts the occupancy presence. Then a rule induction algorithm is used to learn a pruned set of rules on the results from the decision tree model. Finally, a cluster analysis is employed in order to obtain consistent patterns of occupancy schedules. The identified occupancy rules and schedules are representative as four archetypal working profiles that can be used as input to current building energy modeling programs, such as EnergyPlus or IDA-ICE, to investigate impact of occupant presence on design, operation and energy use in office buildings.

Keywords
Occupant Behavior, Data mining, Occupancy schedule, Behavioral Pattern, Office Building, Building Simulation

1. Introduction

One of the paramount efforts engineers, architects and policymakers are currently facing is the need to deliver highly efficient buildings. In the roadmap toward net-zero energy buildings, office buildings play an important role as they represent approximately 17% of the energy used in the U.S. commercial building sector [1].
Several efforts have been made to accelerate the uptake of energy efficiency technologies in office buildings. While the driving factors of building energy performance such as climate, building envelope and building equipment are well recognized, the description of factors such as operation and maintenance, occupant behavior, and indoor environmental conditions are still oversimplified. Often building occupancy schedules are based upon generalized assumptions that hinge on standards, energy codes or rely on the experience of energy modelers. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) Standard 90.1-2004 [2] provides standardized occupancy factors for different building types which can be used to design occupancy when actual schedules are unknown (Fig 1). A daily profile, handled differently for weekend and weekdays, is composed of hourly values, each of which corresponds to a fraction of the occupancy peak load.

Fig. 1. Recommended office building occupancy factors [%] by day type, ASHRAE Standard 90.1-2004

Nevertheless the stochastic nature of occupant behavior, the number of people that occupy a space and the duration occupied, is a non-trivial aspect to characterize. Literature studies have focused on the impact of occupancy presence scenarios on energy use in office buildings, with Gunay et al. [3] providing a comprehensive and up-to-date critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices. In 2013 a study conducted by Duarte et al. [4] analyzed the

occupancy sensors of a large commercial multi-tenant office building and showed up to 46% variation in occupancy patterns for the time of day, day of the week, holidays and months, when compared with the standardized occupancy schedules in ASHRAE Standard 90.1-2004 [2]. The discrepancy presented by Duarte et al. [4] may lead to the incorrect design of office building equipment and to system inefficiencies. Chang and Hong [5] demonstrated the stochastic nature of occupancy profiles was one of the driving factors behind the discrepancy between the measured and simulated energy consumption in buildings. Based on statistical analysis of measured lighting-switch data, Chang and Hong [5] proved the frequencies of occupants leaving their cubicles and the corresponding durations of absence had significant impact on the total energy use and operational controls of the office building. Results from EnergyPlus simulations to evaluate the impact of occupant behavior on energy use of private offices with single occupancy [6], demonstrated that occupants with wasteful work-style consumed up to 90% more energy than standard users, while austerity work-style occupants used half of the energy of the standard occupants. Moreover, real-time estimation of occupancy in commercial buildings is largely treated with the aim to achieve better dynamic modeling results. However, it is a challenging task to develop reliable mathematical models of occupant presence due to the *stochastic* nature of human behavior [7].

Some stochastic models of the occupancy level of single offices have been proposed in the last decade within the scientific community [8-10]. In 2005 D. Wang et al. [8] examined the statistical properties of occupancy in single person offices of a large office building in San Francisco and found that, while vacancy intervals could be treated as a constant over the day, occupancy intervals were more complex due to their varied distribution in time. Tabak and de Vries [9] proposed a model to predict the occurrence and the frequency of intermediate break activities during an office working day (i.e. walking to a printer/mailbox or using the bathroom). For each intermediate activity, a probabilistic formula was presented for use in office occupancy schedule designs. More recently, Sun et al. [10] developed a stochastic model, using a binomial distribution to represent the total number of occupants working overtime and an exponential distribution to represent the duration of overtime periods. Moreover, Stoppel and Leite [11] presented a probabilistic occupancy model simulating annual building occupancy rates based on frequency, duration and seasonality of occupants' long vacancy activities that can be further implemented into a building simulation model.

Additionally, there has been a gowning interest in agent-based models (ABM) to simulate patterns of human individual action and presence at the building level. Most notably, the Markov Chain method, provides a simulation approach to capture the movement process per occupant in the time and space dimensions of building models. The earliest ABM of occupant presence using a Markov Chain was proposed in 2005 by Yamaguchi et al. [12] in the development of a district energy system simulation model. The working state of each occupant of a group of commercial was simulated based on appliances energy consumption data, where the times of arrival, lunch break and departure were selected on a 5 minute interval with a random distribution by using the inverse function method (IFM).

One of the first agent-based models of occupancy in single office was provided in 2008 by Page et al. [13]. The model predictions used a Markov Chain to create random occupancy profiles (i.e. time of arrival and departure, periods of intermediate absence and presence, as well as periods of long absence from the space) based on and validated by sensor data, and were later used as an input to occupant behavior models within building simulation tools [13]. C. Wang et al. [14] handled occupancy as the straightforward result of occupant movement processes which occurred among the spaces inside and outside a

building. By using the Markov Chain method, the model generated the location for each occupant and the zone-level occupancy for a whole office building type. Additionally, Virote and Nueves-Silva [15] used the Markov Chain method to relate behavior in an office space catalogued by data logger measurements to occupant presence in the office building.

More recently in 2014, Dong and Lam [16] developed a real-time predictive control model for building heating and cooling systems based upon the occupancy behavior pattern detection in coordination with local weather forecasting, using advanced machine learning methods including Adaptive Gaussian Process, Hidden Markov Model, Episode Discovery and Semi-Markov Model.

Currently, more granular real-time measurements of the occupant presence, movement and interaction with system controls (thermostats, lighting) and building envelope action (windows, shades) are streamed. Sensor networks enable multidisciplinary and integrated layers of big data source collection, providing reliable information on occupancy recognition and scheduling, in addition to building performance and operation.

State-of-the-art data mining methods provide a powerful analysis technique to extrapolate useful and understandable occupancy patterns from big data streams.

For clarity, data mining is defined in 2001 by Hand et al. [17] as: “The analysis of large observation datasets to find unsuspected relationships and to summarize the data in novel ways so that owners can fully understand and make use of the data.” Cabena et al. [18] provided another definition as: “An interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases and visualization to address the issue of information extraction from large databases.” In many applications, it is difficult to extrapolate useful information from monitored building data due to large data scattering. Instead patterns of data discovered through data mining techniques may present applicable solutions at high levels of abstraction. Data mining of frequent patterns has been a focused theme in data mining research for over a decade with a comprehensive review provided by Han et al. [19]. Although, data mining techniques are largely applied to research fields such as marketing, medicine, biology, engineering, medicine, and social sciences, the application of a data mining framework to building energy consumption and operational data, is still in elementary phases. One highly effective technique of data mining for obtaining information on human-building interaction is the use of patterns correlating repetitive behaviors and actions to typical user profiles [20-22]. In this context, between 2011 and 2012 Yu et al. [23-26] tested several systematic data mining methodologies for identifying and improving occupant behavior in buildings. The results showed that the analysis methodology was powerful in providing insights into energy patterns related to the occupant behavior, facilitating evaluation of building saving potential by improving users’ energy profiles as well as driving building energy policy formulation [23-26].

2. Methodology

Traditional methods of turning data into useful knowledge require data cleaning, analysis and interpretation. However, such manual data analysis often becomes impractical, slow and expensive as data volume grows exponentially. In view of these facts, researchers in the field of machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition and data visualization, have focused their attention on the Knowledge Discovery in Databases (KDD), advancing beyond traditional methods [27]. KDD is the broad process of knowledge extraction in big data streams and involves the application of the following six steps (Fig. 2):

1. *Data selection*
   Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.
2. *Data cleaning and preprocessing*
   Removal of noise or outliers, strategies for handling missing data fields.
3. *Data transformation*
   Finding useful features to represent the data depending on the goal of the task.
4. *Data mining*
   Matching a particular data mining method for searching patterns in the data.
5. *Data interpretation and evaluation*
   Deciding which parameters may be appropriate and interpreting mined patterns.
6. *Knowledge extraction*
   Consolidating discovered knowledge that can be used for further analysis.

Fig. 2. Graphical representation of the Knowledge Discovery in Databases (KDD) process

This study uses the KDD data mining process to extrapolate information on occupancy schedule patterns from measured building big data streams (Fig. 3). A three-step data mining schedule method is applied to a data set to provide insight into patterns of occupancy in office buildings. In step 1, a data set of 16 offices with 10 minute interval occupancy data over a 2 year period is mined though a decision tree model that predicts the occupancy presence. In step 2, a rule induction algorithm is used to mine a pruned set of rules on the results from the decision tree model. In step 3, cluster analysis is employed to obtain consistent patterns of occupancy schedules representative of typical single office working user profiles.

Fig. 3. Proposed occupancy schedule learning framework

The data mining algorithms are employed along with the open source data mining program Rapid Miner 6 [28] to perform the analysis. Rapid Miner is a free open source visual environment for predictive analytics and data mining. Rapid Miner is based on an XML internal process structure, it has an intuitive graphical user interface and no programming is required. For these reasons, it one of the best open source data mining tools both in terms of technology and applicability.

2.1 The Data Set

An office building located in Frankfurt am Main [29] is used as the case study (Table 1).

Table 1. Building Characteristics

Frankfurt is located in central Germany with a temperate-oceanic climate with relatively cold winters and warm summers. The building combines a high energy standard with high occupant comfort. The building is naturally ventilated and cooled in summer and equipped with a night-time mechanical ventilation. Moreover, the monitored office building shows very strict design criteria in terms of energy efficiency and energy optimization for heating, cooling, ventilation and lighting. With an average U-value of 0.54 W/m²K (façade including windows), the building exceeds the requirements of the German 2002 Energy Saving Standards by approximately 30%.
In this study, we use the following dataset (Fig. 4 and 5) with:

a) 16 private offices with single or dual occupancy (Table 2). E01 to E11 are eleven offices facing the east while W01 to W05 are five offices facing the west.

b) 10-minute occupancy interval data over two complete years

Fig 4. Two-part sun protection enables glare-free use of daylight

Fig 5. Offices with operable windows and sun protection, allowing natural ventilation and lighting

Table 2. Dataset Characteristics

2.2 Decision Tree Model

A decision tree is a branched flowchart graphical classification model. This representation of the data has the advantage of being easy to interpret. Decision tree models segregate a set of data into various predefined classes and provide description, categorization and generalization of a given dataset. The goal of a decision tree is to create a classification model (Fig. 6) that predicts the value of a target attribute (*label attribute*) based on several input attributes (*predictor attribute*). Each interior node (*leaf node*) of tree corresponds to one of the predictor attributes. The number of edges (*branche*s) of a nominal interior node is equal to the number of possible values of the corresponding predictor attribute. Each leaf node represents a value of the label attribute represented by the path from the root tree (*root node*) to the final leaf (*possible answers*).

Fig. 6. Graphical representation of the flowchart tree-like graph decision tree model

Decision tree model generation is a two-step process, namely learning and classification, as shown in Fig. 7.

Fig. 7. The decision tree model generation process

In the *learning process*, records in the dataset are automatically and randomly divided into two subsets: a *training* dataset and a *test* dataset. Then, a decision tree algorithm *generates* a decision tree. In this study, we employ the C4.5 algorithm, along with the open-source data mining software RapidMiner.

The C4.5 algorithm was first introduced by Quinlan [30] for inducing decision trees classification models from data. In building a decision tree, the C4.5 algorithm deals with training sets that have records with unknown attribute values by evaluating the “Gain” (also called “Gain-ratio”), for an attribute by considering only the records where that attribute is defined. Gain is defined in Eq. 1:

$$Gain(\vec{y}, j) = Entropy(\vec{y} - Entropy(j|\vec{y})) \quad (1)$$

Where $Entropy(y) = -\sum_{j=1}^{n} \frac{y_j}{y} log \frac{y_j}{y}$ and $Entropy(j|y) = \frac{y_j}{y} log \frac{y_j}{y}$

This process uses entropy as a measure of the disorder of the data. The final aim is to maximize the Gain by dividing by the overall Entropy due to split argument *y* by value *j*.

In the *classification process*, the accuracy of the obtained decision tree is *validated* by cross-validation in order to estimate the statistical performance of a learning process. In the cross-validation process, the data set is partitioned into k subsets of equal size. Of the k subsets, a single subset is retained as the testing data set and the remaining k − 1 subsets are used as training data set. The cross-validation process is then repeated k times, with each of the k subsets used exactly once as the testing data. The k results from the k iterations then can be averaged (or otherwise combined) to produce a single estimation. The value k is adjusted in this study using a k=10 number of validations. If the accuracy is considered acceptable, the decision tree can be applied to new datasets for classification and prediction.

2.3 Rule Induction

The Rule Induction is a classification data mining technique which generates sets of rules in big data sets. Rules have the advantage of being easy to understand, representable in first order logic and prior knowledge can be easily added. A variety of rule-induction algorithms are applied in the field of machine learning and applied data mining literature (Quinlan 1992 [30]; Breiman et al. 1984 [31]). Such algorithms are likelihood-based model evaluation methods, which are typically used for predictive modeling, both for classification and regression, although they can also be applied to descriptive modeling in big data (Agrawal et al. 1996 [32]). Pruning in decision trees is a technique in which leaf nodes that do not add useful information to the classification of the decision tree are removed. This is done by converting an over-specific or over-fitted tree to a more general form in order to enhance its predictive power on unseen datasets. In the prune phase, for each rule any final sequences of the antecedents is pruned with the pruning metric p/(p+n).

In this study, *information gain* has been used as criterion parameter for selecting attributes and numerical splits of the rule induction. Similarly for the decision tree model, the entropy (Eq. 1) of all the attributes is calculated, and the attribute with minimum entropy is selected for split. Accordingly, the Rule Induction operator algorithm is applied to the given data set to iteratively grow and prune rules until there are no positive examples left or the error rate is greater than 50%.

2.4 Cluster Analysis

Cluster analysis is the process of merging data into different clusters, so that (i) instances in the same cluster have high similarity and, (ii) instances in different clusters have low similarity (Fig. 7). The similarity between clusters is normally computed based on the distance between the clusters. The most popular distance measure is described by the Euclidian distance shown in Eq. 2:

$$d(\mathrm{a},\mathrm{b}) = d(\mathrm{b},\mathrm{a}) = \sqrt{(b_1 - a_1)^2 + (b_2 - a_2)^2 + \cdots + (b_n - a_n)^2} \quad (2)$$

Where $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_n)$ are two points in an Euclidean n-space.

The *k-means algorithm* is a method of vector quantization for cluster analysis in data mining. Given the simple nature of the algorithm, it is one of the widely used classification technique. Assumed a data set D, containing a number n of records (instances), the number of clusters *k* must be specified. Each cluster is associated with a centroid (center point) representing the mean of the points in the cluster and each point is assigned to the cluster with the closest centroid.

The performance of the cluster models is evaluated by means a Cluster Distance Performance operator. In this study, the Davies–Bouldin index (DBI) is used for performance evaluation. This index is defined in Eq. 3 (3) as “the ratio of the sum of average distance inside clusters to distance between clusters” [33].

$$DBI = \frac{1}{n}\sum_{i=1}^{n} \max_{i \neq j} \left[\frac{R_i + R_j}{M_{i,j}}\right] \quad (3)$$

Where:

n: number of clusters:

R_i; R_j : average distance inside cluster *i* and cluster *j* by averaging the distance between each cluster object and the cluster centre;

M_{ij}*:distance betwwen the cluster centres*

Consequently, a smaller DBI indicates a better performance of the clustering algorithm. The k=n algorithm that produces clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low

Davies–Bouldin index, and will be considered the $k=n_{opt}$ cluster algorithm for the specific data set (Fig. 8).

Fig. 8. Graphical representation of data clusters

3. Results

In this study, a data mining learning framework was applied in an effort to extrapolate valid and understandable occupancy schedules patterns from the given measured building data set. The main outcomes are summarized as follows:

3.1 Data Transformation

Transformation methods are applied to the given data set with the aim of finding useful predictor attributes to classify the data depending on the office occupancy rates. Raw data at each 10 minute time step are transformed into more significant pre-processed data representing invariant *predictor and label attributes* of the data set, as follows:

1. *Season* (Summer, Spring, Autumn, Spring)
2. *Day of the week* (Monday to Sunday)
3. *Time of the day* (Early Morning 6-9am, Morning 9am-12pm, Noon 12-3pm, Afternoon 3-6pm, Evening 6-9pm, Night 9pm-6am)
4. *Window Change* (if occupancy state $t_{n-1} = t_n$ then = 0, otherwise = 1)
5. *Office Occupancy State* (0=vacant, 1=occupied)

The pre-processed data set is then used for the three-step schedule learning via data mining.

3.2 Step 1: Decision Tree Induction

The goal of this step is to create a decision tree model that predicts the value of a label attribute (occupancy) based on several input attributes (predictor attribute) of the data set. In this study, we employ the C4.5 algorithm, along with the open-source data mining software RapidMiner to generate a decision tree. Gain ratio is the criterion on which attributes (occupied/vacant) are selected for splitting. Gain ratio calculates the entropy of all the attributes and selects for the split the attribute with minimum entropy. According to the gain ratio criterion, the range and uniformity of the attribute splits is assured with a minimum confidence of 50%. Fig. 9 shows the decision tree for the classification of the 16 offices occupancy state. The classification model predicts the value of the label attribute (*Office Occupancy State vacant/occupied)* based on the predictor attributes (*Time of the day, day of the week, season of the year, Window Change*). The tree-like graph presented above must be read from top to bottom. The predicted answers of the model represent the probability of the office being vacant/occupied based upon the path from the root-to-final leaf of the tree.

Fig. 9. Decision tree for the classification of the 16 offices occupancy state

The decision tree is validated by cross-validation along with the open-source data mining software RapidMiner to estimate statistical performance of the learning process. As shown in Table 3, 90.53% of all the training records are correctly classified as vacant or occupied. This indicates a good accuracy of the decision tree model which can be further applied to new datasets for classification and prediction.

Table 3. Multiclass classification performance of the decision tree model

3.3 Step 2: Rule Induction

Based on the decision tree, decision rules are induced by traversing the tree model from the root node to a leaf node. Since each leaf node produces a decision rule, the complete set of decision rules, which is equivalent to the decision tree, is derived. Accordingly, generated decision tree is converted to a set of decision rules, as show in Table 4. For example, a decision rule can be generated from root node to node 4 in above decision tree as follows: if Time of the day = Morning and Window Change = 1 then the office is OCCUPIED with a probability equal to 26% (1058 records over 4118 assigned).

3.4 Step 3: Cluster Analysis

The goal of this step is to disaggregate the occupant presence during working days into valid working user profile schedules. The k-means algorithm is employed, along with the open-source data mining software RapidMiner, to generate clusters of occupancy patterns in 16 single occupancy offices of the same building. The value $2>k<10$ is adjusted in this study in order to find the k_{opt} by using Cluster Distance Performance operator. In this study, the Davies–Bouldin index is used for performance evaluation. As shown in Fig. 10, the k=4 algorithm has the lowest Davies–Bouldin index (-1.31). For this reason, a k=4 is chosen as the k_{opt} cluster algorithm for the specific data set.

Fig. 10. Cluster Distance Performance Analysis with the Davies–Bouldin index

The cluster centroids of the k_{opt} =4 algorithm are plotted in order to provide a visualization of the emerged occupancy patterns. Significantly, the algorithm highlights four different office occupancy patterns, as A, B, C, and D (Fig. 11). A variation up to 60% occurred in the hourly occupancy rate, among the four patterns of occupancy, with noticeable variation happening during times of arriving (8am) and leaving (5pm) the office.

Fig. 11. Occupancy patterns (Monday-Friday) emerged by applying the k_{opt}=4 cluster algorithm

Patterns of occupancy presence clustered in the data set are leading to four typical working occupancy rates, from Monday to Friday (Fig 12).

- Pattern A presents the highest occupancy rate Monday through Friday
- Pattern B presents a medium occupancy rate Monday through Friday
- Pattern C characterizes the most variable occupancy rate. This pattern presents a medium occupancy rate on Monday, Tuesday and Thursday and medium-high occupancy on Friday (before-after the lunch vacancy). On Wednesday, user's vacancy/presence state in the office space varies with high frequency.
- Pattern D characterizes the lowest occupancy rate Monday through Friday.

Fig 12. Occupancy rate patterns (Monday-Friday) emerged by applying the k_{opt}=4 cluster algorithm

Each of the 16 offices is assigned to an occupancy behavioral cluster for every day of the working week. Results in Fig. 13 demonstrate single occupancy offices are characterized by dissimilar patterns of occupancy during working weekdays. This means that the working profile of a singular office may vary broadly on different working days. The distribution of the office working profiles over the working days (Monday-Friday) is presented in Fig. 13, showing a predominance of occupancy Pattern A on Friday (44% offices), Pattern D on Monday (38% offices) and Pattern C on Wednesday and Thursday (38% offices). Occupancy patterns are distributed uniformly over the 16 offices on Tuesday.

Fig. 13. Distribution of the occupancy patterns in 16 offices (Monday-Friday)

In order to understand the implication of the occurrence and the frequency of a single occupancy office being occupied/vacant over the 24 hour period, analysis of the correlation among occupancy patterns during the same days of the week, is conducted. Fig. 15 illustrates a scatter plot matrix of the four occupancy patterns, sorted by day of the week, from Monday to Friday. R-values of the linear correlation of occupancy patterns coupled two by two (Pattern A – Pattern B, Pattern A – Pattern C, Pattern A – Pattern D, Pattern B – Pattern C, Pattern B – Pattern D, Pattern C – Pattern D) show a good correlation of occupancy patterns during the same days of the week (R-value > 0.8). This means office users tend to occupy the office space with a similar pattern over the working hours but that occupancy may vary based on the frequency an occurring event (i.e. arriving or leaving the office space). On the contrary, a small correlation is found in between occupancy Pattern C and Pattern D (R-value < 0.7), indicating that the event the office user arriving or leaving the office space occurs with dissimilar frequency and is also strongly shifted in the 24 hour time schedule.

Fig. 14. Correlation among occupancy patterns in the scatter plot matrix

Fig. 15. Scatter plot matrix of the four occupancy patterns during same day of the week

3.4.1. From occupancy patterns to working profiles

The Knowledge Discovery in Database (KDD) process, patterns extraction from the data base and cannot be considered the final step of the data mining occupancy learning process. In order to extrapolate useful, valid and further applicable knowledge on the occupancy of the case study office building, the mined 24 hour occupancy patterns must be transformed into working user profiles. For this, time dependent description of typical working activity, presence and intermediate absence in a singular office space is required. Fig. 16 provides a graphical visualization of main typical working activities for the four mined patterns:

- going to work: increase in global occupancy curve
- working: stable global occupancy curve
- lunch/breakfast: one valley decrease/increase in global occupancy curve
- going off work: decrease in global occupancy curve

Fig. 16. Graphical visualization of main working typical activities for the four mined patterns

The patterns of occurrence of repetitive, typical activities occurring during a working day in a single occupancy office, in the 24 hour time schedule, include: (i) time of arriving/leaving the office, (ii) period of stable work from the office and (iii) period of intermediate absences. These activities are characterized in four emerging working-user profiles, as shown in Fig. 17 and described as follows:

- Pattern A working-user profile arrives at work around 6-9am, works stable from the office in the morning from 9am-12pm and in the afternoon from 1:30-4pm, going for lunch around 12-1:30pm, leaves work around 4-7pm.
- Pattern B working-user profile arrives at work around 8-9:30am, works stable from the office in the morning from 9:30am-12:30pm and in the afternoon from 2-5:30pm, going for lunch around 12:30-2pm, leaves work around 5:30-10:30pm.
- Pattern C working-user profile arrives at work around 5:30-6:30am, leaves office in between 6:30-7am, works stable from the office in the morning from 7-

11:45am and in the afternoon from 1-5pm, going for lunch around 12-1pm, leaves work around 5-7pm.

- Pattern D working-user profile arrives at work around 5:30-6am, leaves office in between 6-8am, returns to office around 8-9:30am, works stable from the office in the morning from 9:30am-12pm and in the afternoon from 2-4pm, going for lunch around 12-2pm, leaves work around 4-6:30pm.

Fig. 17. 24-hour typical office activities for Patterns A, B, C and D working user profiles

Fig. 18. Breakdown of occupancy hours for Patterns A, B, C and D working user profiles

Pattern B working-user profile tends to arrive later in the office and typically works beyond normal working hours (Fig. 17). Fig. 18 shows the average total breakdown of hours spent at work during the work week, with the following breakdown of occurrence for Pattern B working-user: (i) working stable from the office (26%), (ii) moving from the office (26%) and, (iii) taking breaks (7%). Therefore, Pattern B working-user is just slightly above (60%), whereas the breakdown for the other working profiles is less (49% Pattern A, 48% Pattern C and 38% Pattern D).
Pattern C is characterized for a more stable working-user profile, whom tends to spend more time working from office (35%) than moving from the office (13%).
Pattern D working-user profile tends to spend less time in the working space, with an average of about 20% considered working stable and an almost equivalent amount (18%) arriving/leaving the working position. Additionally, more work day time (17%) is spent away from the office.

3.4 Occupancy Schedule

Final goal of the proposed method is to identify archetypal user profiles for which different energy conservation strategies, as well as building design recommendations, may be appropriate. For this aim, the identified occupancy patterns are transformed into four typical working profile schedules of occupancy (Fig. 19).

Fig. 19. Example of 24 hour schedules of occupancy for Patterns A, B, C and D

Such schedules characterize the probability of an office being occupied at a specific time of the day and day of the week (Monday-Friday). Schedules do not represent the percentage of full occupancy as represented by traditional occupant schedules. Fig. 19 shows the occupancy level in similar offices of the same building may vary largely, based on different working-user profiles. This finding greatly impacts appliance, lighting, plug-load, system controls and therefore on total energy consumption of the building.

4. Discussion

In the last decades, the evolution to more granular measurements of the occupancy presence, movement and interaction with system controls (thermostats, lighting) and building envelope (windows, shades) have transpired into building development. Sensor networks make available multidisciplinary and integrated layers of big data source, providing reliable information on occupancy recognition and scheduling besides of actual building performance and operation. The aim of researchers to improve the accuracy of occupant presence in simulation models and some stochastic models and the topic of real-time estimation of occupancy treated largely in commercial buildings, all feeds into the larger objective of building better dynamic modeling for design, daily energy management and energy conservation measures assessments.

Nonetheless, previous research has highlighted due to the stochastic nature of human behavior, evolving randomly with time as one of the major shortcomings in obtaining reliable mathematical models.
Moreover in many applications it is difficult to extrapolate useful information on the occupant movement and presence from monitored building data due to the data scattering at this level. Instead patterns of data discovered through data mining techniques may highlight commonsense knowledge applicable to solutions at high levels of abstraction. In this context, data mining techniques have been shown as able to automatically extrapolate valid, novel, potential useful and understandable occupancy patterns from big data streams, highlighting expressions describing typical and repetitive working user profiles in office spaces or in office buildings.
Useful information can be extracted from the proposed data mining occupancy schedules learning process to improve energy building simulations. The decision tree model can help understand repetitive rules in occupancy patterns in order to optimize appliances, plug loads, lighting use, HVAC control systems, fresh air requirements, internal heat gain and building design plans in office buildings.
One of the limitation of this study is that the mined working user profiles and patterns of occupancy are circumstantial to the given data set.
The application of the proposed data mining framework on different data sets will enhance the robustness of individual as well as group energy related behavioral patterns description and prediction in office buildings.
The implementation of the emerged occupancy rules and schedules into current building energy modelling programs, would support the implementation of more efficient energy measures through more accurate investigation on the impact of typical working user profiles on appliances, plug loads, lighting use, HVAC control systems, fresh air requirements and components heat gain, both at the single office and at the office building level.
Moreover, it has to be underlined that the proposed methodology is not intended to substitute or contrast the agent-based stochastic models already developed for the integration of occupants' presence into building energy simulations.
More likely, the knowledge discovered related to occupancy rules and working user profiles schedules aims to support operators, building designers, auditors and managers decision making by providing solutions with fast legibility, high replication potential and low capital investment to direct specific operation and maintenance strategies at a building level as well as future energy-saving policy in the commercial building sector.
Scalability of solutions is probably one of the most critical point in pattern mining. The mined schedules and occupancy rules are circumstantial to the case study building and do not represent the complete set of patterns that can be derived within a comprehensive dataset of 10-min data in two full years. Nevertheless, they characterize the most compact, physical meaningful and high quality set of patterns that can be derived with satisfactory performance. Interesting results may emerge from further analysis, by applying the proposed data mining framework to a seasonal or one year behavioral data set, providing solutions to direct specific operation and maintenance strategies at a building level that may be appropriate for each of the distinct working user profiles in different periods of time. Moreover, is generic enough to be possibly

5. Conclusion

Using the Knowledge Discovery in Database (KDD) process, a data mining learning process was proposed to extrapolate office occupancy patterns and working user profiles from big data streams. A three-step data mining schedule learning method was applied to a data set along with the open source data mining program RapidMiner to provide insights into patterns of occupancy in 16 offices located in Frankfurt, Germany.
The transformation methods were applied to a given data set with 10-minute interval occupancy data over two complete year periods. Raw data were transformed into more significant pre-processed data representing invariant attributes of the data set. The pre-processed data were mined though a decision tree model with the goal to predict the value of a label attribute (occupancy) based on several predictor attributes (Season, Day of the week, Time of the day, Occupancy State and Window Change) of the data set. The results demonstrated that the C4.5 is a suitable algorithm for learning the occupancy presence in offices. This was verified with a 90.3% accuracy rate of all training records correctly classified as vacant or occupied. The predicted answers of the tree-like graph model are probabilities of the office being vacant/occupied based on the condition defined by the path from the root tree to the final leaf.
Secondly, by traversing the tree model from the root node to a leaf node, a complete set of 45 rules was derived. The proposed tree and rule models can be used to understand repetitive occupancy patterns in order to optimize operation, maintenance and energy performance both at the single office and at the office building level.
Thirdly, a cluster analysis was performed in order to disaggregate the occupancy presence into valid working user profile schedules for every working day (Monday-Friday) and for the whole working week. In this study, we employed the k-means algorithm, to generate an optimal k=4 number of clusters. The results showed the occupancy patterns in single offices were not assigned to the same behavioral cluster every day, meaning that working profiles may vary broadly during different working days. A conspicuous variation up to 60% in the hourly occupancy rate was noticeable among patterns of occupancy especially during office arriving or leaving time (8am to 5pm). Furthermore, the clustering of typical office working activities such as working stable from the office, moving (arriving or leaving) from the office and taking breaks, such as having breakfast or lunch, highlighted that the average breakdown of hours spent every day by the monitored users in the single office space may differ broadly, as well as a significant shifting in the time the occupancy pattern activity/presence/ intermediate absence was occurring. Also, the occurrence occupants arrive early in the morning before 5.30am and depart after 10.30pm (overtime work) emerged as phenomenon occurring frequently, nonetheless such patterns are omitted by using fixed deterministic occupancy schedules.
Finally, the proposed methods in this study identified rules of occupancy and archetypal user profiles for which different energy conservation strategies, as well as building design recommendations, may be appropriate. Characterization of the probability of an office being occupied at a specific season of the year, day of the week and time of the day will enable the more accurate development of building energy models. The results supported the assumption that occupant stochastic behavior and presence cannot easily described by means deterministic 24 hour schedules. Instead more accurate profiles having the same patterns of occupancy of real building users are required to close the gap between predicted and actual building performance. The future applications of the proposed method to discern occupancy schedules and their implementation into a building energy modelling programs, like EnergyPlus or IDA-ICE, would strongly support the investigation of the impact of typical working occupancy patterns on design and operation

of office appliances and control systems. In this context, further investigations are suggested to uncover cost-effective, applicable and reliable best practices and solutions supporting energy efficiency policies and decision makers on how to incorporate patterns of human movement and actions into behavioral models, with the aim of bridging the gap between actual and predicted energy performance in buildings.

Acknowledgement

This work was sponsored by the U.S. Department of Energy (Contract No. DE-AC02-05CH11231) under the U.S.-China Clean Energy Research Center for Building Energy Efficiency. The authors very appreciated Marcel Schweiker and Andreas Wagner of Karlsruhe Institute of Technology, Germany for sharing the dataset and answering our questions. This work is also part of the research activities of the International Energy Agency Energy in Buildings and Communities Program Annex 66, Definition and Simulation of Occupant Behavior in Buildings.

References

[1] J. Laustsen, Energy Efficiency Requirements in Building Codes, Energy Efficiency Policies for New Buildings, 2008, OECD/IEA International Energy Agency
[2] American Society of Heating, Refrigerating and Air-Conditioning Engineers ASHRAE Standard 90.1-2004, Energy Standard for Buildings except Low-Rise Residential Buildings, 2004, ISSN 1041-2336
[3] H. Burak Gunay, W. O'Brien, I. Beausoleil-Morrison, A critical review of observation studies, modeling, and simulation of adaptive occupant behaviors in offices, Building and Environment 70 (2013) 31-47
[4] C. Duartea, K.Van Den Wymelenberga, C. Riegerb, Revealing occupancy patterns in an office building through the use of occupancy sensor data, Energy and Buildings 67 (2013) 587–595
[5] W. Chang., T. Hong, Statistical Analysis and Modeling of Occupancy Patterns in Open-Plan Offices using Measured Lighting-Switch Data. Building Simulation 6 (2013) 23-32
[6] T. Hong T, H. Lin, Occupant Behavior: Impacts on Energy Use of Private Offices. ASim 2012 - 1st Asia conference of International Building Performance Simulation Association, Shanghai, China (2013)
[7] C. Liao, P. Barooah, An integrated approach to occupancy modeling and estimation in commercial buildings. ACC American Control Conference (2010) 3130-3135
[8] D. Wang, C. C. Federspiel, and F. Rubinstein, Modeling occupancy in single person offices, Energy and Buildings 37 (2005) 121–126
[9] V. Tabak, B. de Vries, Methods for the prediction of intermediate activities by office occupants, Building and Environment 45 (2010) 1366–1372
[10] K. Sun, D. Yana, T. Hong, S. Guo, Stochastic modeling of overtime occupancy and its application in building energy simulation and calibration, Building and Environment 79 (2014) 1-12
[11] C. M. Stoppel, F, Leite, Integrating probabilistic methods for describing occupant presence with building energy simulation models, Energy and Buildings 68 (2014) 99–107
[12] Y. Yamaguchi, Y. Shimoda, M. Mizuno, Development of district energy system simulation model based on detailed energy demand model, Proceedings of the Eighth International IBPSA Conference Eindhoven, Netherlands (2003)

[13] J. Page, D. Robinson, N. Morel, and J.-L. Scartezzini, A generalized stochastic model for the simulation of occupant presence, Energy and Buildings 40 (2008) 83 – 98
[14] C. Wang, D. Yan, Y. Jiang, A novel approach for building occupancy simulation, Building Simulation 4 (2011) 149–167
[15] J. Virote, R. Neves-Silva, Stochastic models for building energy prediction based on occupant behavior assessment, Energy and Buildings 53 (2012) 183–193.
[16] B. Dong, K.P. Lam, A real-time model predictive control for building heating and cooling systems based on the occupancy behavior pattern detection and local weather forecasting, Building Simulation 7 (2014) 89-106
[17] D. J. Hand, H. Mannila, P. Smyth, Principles of Data Mining: MIT Press (2001)
[18] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, A. Zanasi, Discovering data mining: From Concept to Implementation, Prentice-Hall, Inc. (1998)
[19] J. Han, H. Cheng, D. Xin, X. Yan, Frequent pattern mining: current status and future directions, Data Mining Knowledge Discovery 15 (2007) 55–86
[20] O.G. Santin, Behavioural Patterns and User Profiles related to energy consumption for heating, Energy and Buildings 43 (2011) 2662-2672.
[21] V.D.K. Wymelenberg, Patterns of occupant interaction with window blinds: A literature review, Energy and Buildings 51 (2012) 165-176.
[22] W. F. Van Raaij, T.M. Verhallen, Patterns of residential energy behavior. Journal of Economic Psychology 4 (1983) 85-106.
[23] Z. Yu, F. Haghighat, B.C.M. Fung, H. Yoshino, A decision tree method for building energy demand modeling, Energy and Buildings 42 (2010) 1637–1646.
[24] Z. Yu, B.C.M. Fung, F. Haghighat, H. Yoshino, E. Morofsky, A systematic procedure to study the influence of occupant behavior on building energy consumption, Energy and Buildings 43 (2011) 1409–1417.
[25] Z. Yu, F. Haghighat, B.C.M. Fung, L. Zhou, A novel methodology for knowledge discovery through mining associations between building operational data, Energy and Buildings 47 (2012) 430–440
[26] J. Zhao; R. Yun; B. Lasternas, H. Wang; K.P. Lam; A. Aziz; V. Loftness, Occupant behavior and schedule prediction based on office appliance energy consumption data mining, Proceedings on CISBAT 2013, Lausanne, Switzerland (2013)
[27] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From Data Mining to Knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, (1996) 1-34
[28] RapidMiner Studio,V 5.3 http://rapid-i.com/content/view/181/190/.
[29] IEA Annex 53 Task Force. Final report, Total energy use in residential buildings – the modeling and simulation of occupant behavior (2012)
[30] J.R Quinlan, Simplifying decision trees. International Journal of Man-Machine Studies 27 (1987) 221-234
[31] Breiman, L.; Friedman, J. H.; Olshen, R. A.; and Stone, C. J. 1984. Classification and Regression Trees. Belmont, Calif.: Wadsworth.
[32] Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; and Verkamo, I. 1996. Fast Discovery of Association Rules. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 307–328. Menlo Park, Calif.: AAAI Press.
[33] Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224–227