Occupancy Prediction from Electricity Consumption Data in Smart Homes

Davide Fialho Pereira
davide.pereira@tecnico.ulisboa.pt
Instituto Superior Técnico, Universidade de Lisboa, Portugal
June 2017

***Abstract*— The problem of occupancy detection and prediction plays a major role in the smart home era, since can provide benefits with regard to comfort, safety and energy savings to electricity consumers. Many authors have already explored occupancy monitoring and prediction systems, however, very few approached the occupancy detection and prediction by using smart meter data. In this work, we showed that an occupancy detection accuracy up to 92% can be achieved by using solely electricity consumption data.**

Also, we address the problem of generalizing a classification model, i.e., we analyze the possibility of using a single classification model to monitor occupancy in multiple households. We found that an occupancy detection accuracy of up to 82% was possible by using a generic classification model.

Regarding occupancy prediction, we showed that it is possible to predict occupancy in multiple households, with an accuracy of up to 78%, by using solely electricity consumption data. This result is not fantastic but we believe that better result can be obtained by: 1) exploring more features from the dataset 2) using a more sophisticated feature selection method and 3) using hybrid prediction approaches, since they combine schedule-based with context-aware approaches, e.g. combine a prediction timetable with information from mobile phones (GPS and wireless networks).

For both occupancy monitoring and prediction, we consider that households with low level of occupancy can benefit more from these systems.

***Index Terms*— Occupancy prediction, occupancy detection, electricity consumption, smart meter, opportunistic sensing**

INTRODUCTION

In the era of smart home, achieving a higher energy efficiency, comfort and safety at home are three important factors that households are interested in. With the advance in the statistics field and with the appearance of new fields such as the Internet of Things (IoT) and machine learning, energy utilities have now the possibility of providing these benefits to the customers.

One way of providing these benefits to the customers is by controlling automatically their electrical appliances depending if the house is occupied or not. According to [1], Heating, ventilation and cooling (HVAC) represents the largest source of residential energy consumption in the U.S., Canada and U.K. The same study refers that 20-30% of this energy could be saved by simply turning off the HVAC systems when residents are sleeping or away. However, the author states that these savings have been difficult to realize since households typically do not manually adjust the thermostat several times a day. Also, smart thermostats, such as the NEST thermostat, attempts to solve this problem by automatically programming itself based on occupancy patterns that it learns by a built-in motion sensor [2]. However, the relatively high cost represents a major disadvantage of the NEST thermostat.

According to [3], Non-Intrusive Occupancy Monitoring (NIOM) is possible by using smart meters and allows utilities to determine: 1) how much a programmable thermostat would benefit in each home and 2) to suggest an optimal customized thermostat schedule. In this work, we will not focus on analyzing the savings potential of smart thermostats, but on investigating how viable is it to monitor and predict occupancy by using solely smart meter data.

In addition to a higher energy efficiency, occupancy monitoring also provides more safety to the consumers. For example, if a high electricity consumption is verified in periods that are not supposed, occupancy monitoring systems can be used as an intruder's detector, by sending alarms in real time to the smartphones of the occupants. If we analyze occupancy at the room-level, occupancy monitoring systems can also be used for health monitoring applications. For example, by verifying the occupancy of the rooms (e.g. dispenser or kitchen), the algorithm can infer if a sick people missed his medication in the supposed time or if missed any meal during the day, and remind the caretaker. In this work, we monitor and predict occupancy at a binary level and not at the room-level, i.e., our objective is to detect and predict if a certain household is occupied or not.

Smart meters are already deployed in millions of households worldwide, representing an opportunity for occupancy monitoring systems without any additional cost, in contrast to conventional approaches, such as motion sensors.

The problem of occupancy detection trough smart meter data has not been yet widely investigated in the literature. Relevant works in this area have only explored the viability of detecting occupancy trough smart meter data individually for each household [4], [5], [6]. However, to the best of our knowledge, no single work analyzed the possibility of using a single and generic classification model to detect occupancy in multiple households. For electricity consumers, this represents a major interest, since they would benefit from occupancy monitoring applications (such as home automation) without the need of

monitoring the real occupancy through direct systems (e.g. motion sensors).
Occupancy prediction isn't also a mature topic in literature. Many works exist but no one, to the best of our knowledge, investigated the possibility of predicting occupancy by using solely the smart meter data.

1. RELATED WORK

We approach related works in two main areas: 1) occupancy monitoring and 2) occupancy prediction through electricity consumption data.

1.1. Occupancy monitoring

Occupancy monitoring systems can be divided into direct and indirect methods. Direct methods require physical devices to be installed in the buildings [7], [8] while indirect approaches use existent information to detect occupancy [3], [4], [5].

Detecting occupancy through smart meter data represents an indirect method and has not been yet very explored in the literature.

In [3], it was concluded that it is possible to monitor occupancy by using smart meter data. The authors instrumented two homes with smart meters and collected power consumption data in a summer week. Ground truth occupancy data (real occupancy) was collected through the interaction of the occupants with electrical loads (that imply occupancy) and via smartphones (through GPS) and was used to evaluate the occupancy detection accuracy. Then, a simple threshold-based method was applied to infer occupancy from the aggregate electricity consumption.

In [4], it was shown that it is possible to obtain an occupancy classification accuracy up to 80% by using common classification methods. In [5], the same authors performed a more detailed study. They showed that an occupancy classification accuracy of up to 94% is possible to obtain by using the data generated by electricity meters. In this study, 35 features were analyzed and two feature selection methods were tested.

1.2. Occupancy prediction

Three types of binary occupancy prediction algorithms exist: schedule-based, context-aware and hybrid algorithms. Schedule-based approaches predict occupancy by using solely the historical occupancy data of a building. Context-aware approaches use the information about the current position, activity and environmental factors (e.g. current traffic conditions) to predict the arrival time of each occupant. Hybrid approaches predict occupancy by combining schedule-based and context-aware algorithms.

Our work focus on predicting occupancy using schedule-based approaches, since we do not have data about the current context or activity of each occupant.

Schedule-based approaches can be divided in two categories. The first detect routines in the historical occupancy schedules and the second assumes that routines can be explained by daily or weekly timetable (i.e., depends with the day of the week and the time of the day). In [6], several state-of-art schedule-based algorithms were analyzed and the study concluded that, for their occupancy dataset, the Presence Probabilities (PP), Presence Probabilities Simplified (PPS) and PreHeat (PH) algorithms provided the best results.

In [9] it is presented the PreHeat algorithm, which is a schedule-based approach that predicts the future occupancy by analyzing the occupants' routines and finding the most similar historical patterns. In this study, five houses (3 in U.S. and 2 in U.K.) were used to analyze the benefit of controlling automatically home heating systems. Authors concluded that PreHeat algorithm allows to obtain a higher home heating efficiency (between 8% and 12% of savings in gas usage) while removing the necessity for users to program their thermostats.

Another schedule-based approach to predict occupancy is the Presence probabilities (PP) algorithm, presented in [10]. While PreHeat algorithm uses the current and historical occupancy data to detect routines and predict future occupancy, PP algorithm uses only historical occupancy data to build a 7-day timetable of occupancy.

In [6], authors obtained a median prediction accuracy of 85% by using the PP and PPS algorithms and 80% for the PreHeat algorithm. According to the same study, schedule-based algorithms for occupancy prediction are limited to the accuracy of 90%, since it relies only on the past occupancy data. To obtain a higher accuracy, the author refers that combining these algorithms with context-aware approaches could push the accuracy above the 90% limit, by providing information about the current context or activity of each occupant.

2. METHODOLOGY

The problem of occupancy monitoring and prediction can be approached through supervised classification algorithms and prediction algorithms, respectively.
In our experiments, we start by dividing the historical data (electricity consumption and occupancy data) of each household into 3 parts, as shown in Figure 1:

1. **Training set**: 40% of the data is used to train the models by performing cross-validations.

2. **Classification set**: 40% of the data is used to apply the classification models and to measure the accuracy of the occupancy detection. Accuracy is defined as balanced performance metric for classification problems, and is described further. By using the results from the occupancy classification, we create a probabilistic weekly timetable, containing the probability of a house being occupied in a certain time and weekday.

3. **Future set**: 20% of the data is used only to test the prediction accuracy using the probabilistic timetable, previously computed, and to measure its occupancy prediction accuracy. In reality, this dataset partition

corresponds to historical data but we consider as future data in order to simulate a real problem.

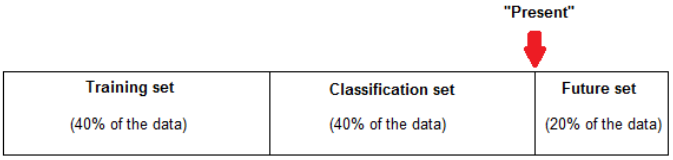

Figure 1: Partition of the historical electricity consumption and occupancy data into 3 subsets.

2.1. Occupancy classification

To detect occupancy from electricity data, we used three classification algorithms: neural networks, support vector machines and random forest (based on decision trees). We selected this three models since they are commonly used in similar problems and have proven to provide good results [11], [5], [12]. They are very different from each other. For example, neural networks may require more data to provide good results and takes more time to run while support vector machines runs faster and may provide better results with a smaller amount of data (comparing to neural network). Random forest is an ensemble algorithm that is based on decision trees to perform the classification. Typically, ensemble methods provide good results since they classify according to the contribution of multiple classifiers/decision trees, avoiding the overfitting.
For each classifier, we repeat a 10-fold cross-validation ten times over our training data in order to avoid specific allocation of data (overfitting) and to obtain the optimal feature set and model parameters.

2.2. Classification evaluation

Many evaluation metrics can be used to evaluate a classifier. In this work, we use the metrics accuracy and mathews correlation coefficient (MCC) as main criteria, since they complement with each other and were used in similar works [4], [5].
Accuracy is a simple metric to evaluate a classifier and can be computed by dividing the number of correct classifications (true positives and true negatives) by the total number of classifications, as shown in equation *(1)*.

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \qquad (1)$$

Because accuracy describes only partially the performance of a classifier, especially for data with unbalanced classes, the matthews correlation coefficient (MCC) is considered as a complementary evaluation metric. The value of MCC varies between -1 and 1. A value of -1 represents that no single instance was correctly classified. A value of 1 represents a perfect classification and a value of 0 indicates that the classification isn't better than a random guess. The MCC of a classifier is calculated as shown in *(2*) [4].

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \qquad (2)$$

We also consider the false positive and false negative rates as a complementary metric for classification evaluation.
The false positive rate (FPR) measures the percentage of unoccupied periods that were incorrectly classified as occupied (false positives or fp) and is given by equation *(3*).

$$FPR = \frac{fp}{fp + tn} \qquad (3)$$

The false negative rate (FNR) represents the percentage of occupied periods that were incorrectly classifies as unoccupied (false negatives or fn), as shown in equation *(4*).

$$FNR = \frac{fn}{fn + tp} \qquad (4)$$

In order to have a baseline for comparing and evaluating our classification algorithms, we introduce the Prior as a classifier that assumes that the household is always occupied or unoccupied (if the household is most of the times occupied or unoccupied, respectively)

2.3. Occupancy prediction

In order to predict occupancy, we use the Presence Probabilities Simplified (PPS) algorithm, which is a schedule-based approach for prediction. This algorithm was chosen since it only uses historical occupancy information and is compatible with our business objective. This algorithm creates a 7-day timetable that contains the probability of a households being occupied in every 15 minute's interval (p_{occ}), according to the hour of the day (time) and the day of the week (1 to 7), as shown in Table 1. By defining a threshold, these probabilities are converted into two classes (occupied or not) and this prediction table can be used to program a smart thermostat, for example.

Table 1: Probabilistic timetable created through the Presence Probabilistic Simplified (PPS) algorithm. This table contains, for each hour of the day (between 7:00h and 23:00h) and for each day of the week (1-7 represents Sunday-Saturday), the probability of presence in a certain household.

time	1	2	3	4	5	6	7
7:00h-7:15h	$p_{occ}(1,1)$	…	…	…	…	…	$p_{occ}(7,1)$
…	…	…	…	…	…	…	…
22:45h-23:00h	$p_{occ}(1,64)$	…	…	…	…	…	$p_{occ}(7,64)$

2.4. Prediction evaluation

The Receiver Operating Characteristic (ROC) curve is used to show how changing the threshold in the decision rule affects the true positive rate (TPR = 1-FPR) and FPR. For the same value of FPR, the higher is the TPR, the better. The ideal model would have a ROC curve passing through the point where TPR is 1 and FPR is 0. In Figure 2, it is possible to observe three ROC curves, for a bad, a good and a great prediction model. The dashed diagonal line represents a ROC curve for a random choice scenario, e.g., assuming that a household is always occupied, and the brown diagonal line represents the equal error rate line, where the false positive rate is equal to the false negative rate.

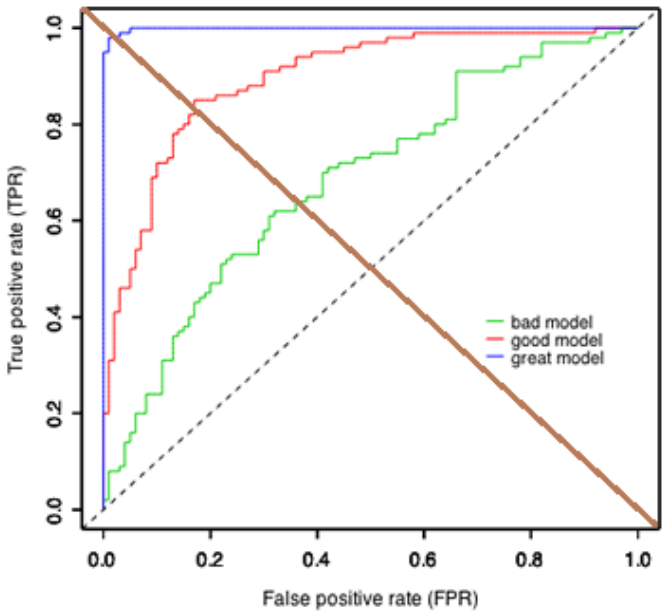

Figure 2: ROC curves for a bad, a good and a great classification model [13].

Before defining a threshold, it is important to know the goal of our prediction. In our work, because we do not intend to analyze the cost of a false positive and a false negative, we decided to use a threshold equal to 0.5 for our decision rule.
After defining a threshold equal to 0.5, we consider that, whenever the probability of occupancy, in a certain time period, is equal or higher than 0.5, the house is occupied in the respective period (and is unoccupied otherwise). Therefore, the probabilistic timetable is converted to a binary time table.
To evaluate the prediction performance of our algorithm, we use the metric accuracy, given by equation *(5)*, that we recall here for convenience.

$$Prediction\ Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \qquad (5)$$

3. Data Preparation

In this section we describe the process of collecting, pre-processing and extracting value from the data. Every step of this process is essential in order to obtain useful and valid data that is further used to monitor and predict household's occupancy based on the chosen machine learning algorithms.

3.1. Selection of households and data collection

To collect the necessary data for our study we installed a smart energy management system in five households. To obtain more information about each participant and to verify if they meet all the requirements for our analysis, we asked for each household to fill a questionnaire, containing multiple questions such as: the number of occupants, their age and occupation.

3.2. Measurement infrastructure

To obtain the aggregate energy consumption (kWh) and maximum power (kW) in every 15 minutes, we replaced the typical electricity meter by a smart meter, which was also used for billing purposes.
To collect the ground truth occupancy data, we gave each household a 7W LED lamp and instructed them to keep the light on whenever the house is occupied and turn it off whenever it's not occupied. The consumption of the LED lamp was obtained by using a smart plug.

3.3. Data cleaning and pre-processing

In this phase, we start by filtering missing and erroneous, that may occur due to bad communication between sensors or a weak internet signal. Also, we limited our analysis between the period of 7:00h-23:00h since we do not indent to analyze the occupancy during the sleeping period.
After the data clean and pre-processing, we obtain a different amount of days of data for each household, during the months of November, December, January, February March and April of 2017, as shown in the Table 2.

Table 2: Number of days available for each

Household	Number of Days
1	83
2	97
3	95
4	99
5	87

For our experiments it is important to have the same amount of samples for each household. For this reason, we limited the number of samples in the evaluation to 83 days for every household, which corresponds to the minimum number of days obtained.

3.4. Feature extraction and description

According to [5], occupancy may be correlated with the: 1) absolute value of the power consumption 2) viariability of the power consumption 3) time. Knowing this, we create 8 features based on expert knowledge and describe them in *Table 3*.

Table 3: Features extracted and used for occupancy classification.

Feature	Description
p_mean	Mean power excluding the "occupancy lamp" consumption.
p_mean_sd	Standard deviation of the mean power
p_mean_sad	Sum of absolute differences of the mean power
p_max	Maximum power verified in every 15 minutes
time	Time slot number (1-64)
workday	1 if it is a working day (Monday to Friday); 0 otherwise
before_workday	1 if the day precedes a working day (Sunday to Thursday); 0 otherwise
night_time	1 if it is dark outside (between sunset and sunrise); 0 otherwise

3.5. Feature scaling

After pre-processing and extracting relevant features from the raw data we obtained a set of heterogeneous features with different characteristics, such as different units, scale and range. We apply the standardization method to bring all the features to the same scale. Consequently, they have a similar contribution to the classification algorithm [8].
The standardization method ensures that all dimensions of the dataset (X) have zero mean (μ)and standard deviation (σ) equal to one, and is given by equation *(6)*.

$$X'_1 = \frac{X_1 - \mu(X_1)}{\sigma(X_1)} \qquad (6)$$

4. Experiences, Results and Discussion

In this section we perform the necessary experiments to answer our research questions.
In the first part, we analyze the performance of detecting occupancy through electricity consumption data. In this part, we also investigate the possibility of generalizing our classification models, i.e., the possibility of using a single and generic model do detect occupancy in multiple households, with a good classification performance.
In the second part, we investigate the possibility of predicting occupancy in multiple households based solely on historical electricity consumption data.

4.1. Occupancy detection from electricity consumption data

To analyze the viability of detecting occupancy from the electricity consumption data, we use the three classification algorithms already mentioned: neural network, support vector machines and random forest. In our classification analysis, we start by presenting the feature combinations that we considered relevant for our work. Then, we divide our experiments in two parts: *self-test* and *other's-test*.

Feature selection

The choice of the features used for our analysis was made based on expert knowledge, from similar works [4], [5], [6].
Table 4 presents five different feature combinations that we propose for our analysis.

Table 4: Proposed feature sets for our first experiment. All features are computed over a 15-minute interval and are described in Table 3.

Feature set	Features
1	p_mean, p_mean_sad, p_mean_sd, p_max
2	p_mean, p_mean_sad, p_mean_sd, p_max, time
3	p_mean, p_mean_sad, p_mean_sd, p_max, time, workday
4	p_mean, p_mean_sad, p_mean_sd, p_max, time, workday, before_workday
5	p_mean, p_mean_sad, p_mean_sd, p_max, time, workday, before_workday, night_time

4.1.1. Detecting household's occupancy (self-test)

This section describes the occupancy detection experiments, performed separately for each household (*self-test*). In the first part, we first try to obtain the highest classification accuracy possible, for each household (optimization by household). Then, in order to simplify our next experiments, we repeat the process by using equal classification parameters for every model and every household.

4.1.1.1. Optimization by household

In this experiment, we compute, for each model and household, the model parameters and the feature set that provide the highest accuracy in the training set, and the classification accuracy when applying the model in the classification set.
For each classification algorithm, we start by defining several combinations of parameters to test. These combinations result from the possible combinations that can be made by changing the feature set from 1 to 5 and by changing the parameters of the respective algorithm, within values defined by us. Then, for

each combination, we test the classification accuracy obtained in the classification set through cross-validations. The classification accuracy of each household represents the average of 100 runs.

Table 5 presents the classification accuracy obtained, in this experiment, according to the classification algorithm.

Table 5: Classification accuracy obtained by the three considered models, for each household.

Household	neural network	SVM	random forest	Prior accuracy (%)
1	89.72	**90.15**	89.96	90.15
2	85.41	87.05	**90.43**	69.31
3	77.90	75.97	**80.85**	75.22
4	82.87	85.31	**88.22**	50.59
5	92.68	92.12	**92.91**	66.82

In Table 6, we show the obtained Matthews correlation coefficient (MCC) for each household and classification algorithm.

Table 6: Matthews correlation coefficient (MCC) values obtained by the three considered models, for each household.

Household	neural network	SVM	random forest
1	0.05	0	**0.11**
2	0.65	0.69	**0.77**
3	0.38	0.32	**0.44**
4	0.49	0.71	**0.78**
5	0.83	0.82	**0.84**

It was possible to obtain higher classification accuracy than the Prior classifier in household 2,3,4 and 5. In terms of the matthews correlation coefficient (MCC) values obtained, it can be seen that the random forest model provided the highest performance in all households, confirming the results from the accuracy analysis (expect for household 1).

4.1.1.2. Assuming equal model parameters and feature set

We continue our next experiments with the neural network, SVM and random forest models. Regarding the SVM model, we use only the SVM model with the linear kernel in the next experiments, since it provided the best overall result, comparing to the other types of kernel.
To simplify our next experiments, we use only a single feature set and equal model parameters, for each algorithm.
To find the optimal model parameters for a certain algorithm, we choose the most frequent parameters verified in the top five results of the training phase, in the five households. For the neural network model, two hidden neurons were chosen. For the SVM model with linear kernel, a cost equal to 10 and a gamma equal to 1 was chosen. For the random forest model, 100 trees were selected.
A similar approach was used to select the best feature combination. Figure 3 shows the relative frequencies of the various feature combinations.

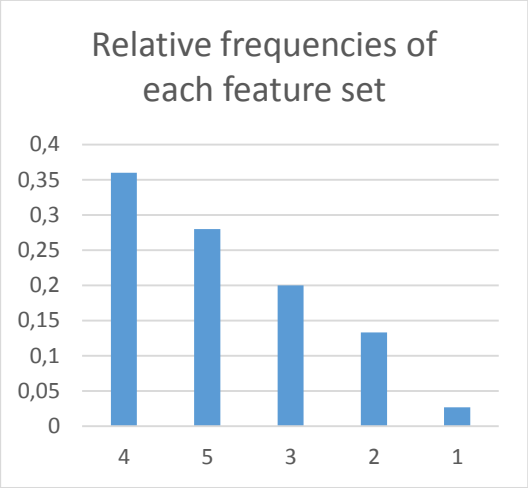

Figure 3: Relative frequencies of the choice of each feature set.

We can observe that the feature set 4 was chosen most frequently. Therefore, we continue our analysis considering this feature combination. Feature set 4 contains 7 features: *p_mean*, *p_mean_sad*, *p_mean_sd*, *p_max*, *time*, *workday* and *before_workday*. However, because we intend to generalize our models, i.e., to train a model using data from one household and apply it in a different household, we remove the feature *time* from the feature set 4, since this feature is highly dependent on the household (contains the occupancy patterns for a specific household according to the hour of the day). Thus, we obtain the feature set 6, defined as shown in Table 7.

Table 7: Feature set 6, used for the other's-test experiment.

Feature set	Features
6	p_mean, p_mean_sad, p_mean_sd, p_max, workday, before_workday

To analyze the effect of using equal conditions in all households, Figure 4 illustrates, for each household, a comparison between highest classification accuracy obtained in this experiment with the highest accuracy obtained in the *optimization by household* experiment. From this figure, we can verify that no significant classification accuracy differences are obtained by assuming equal conditions.

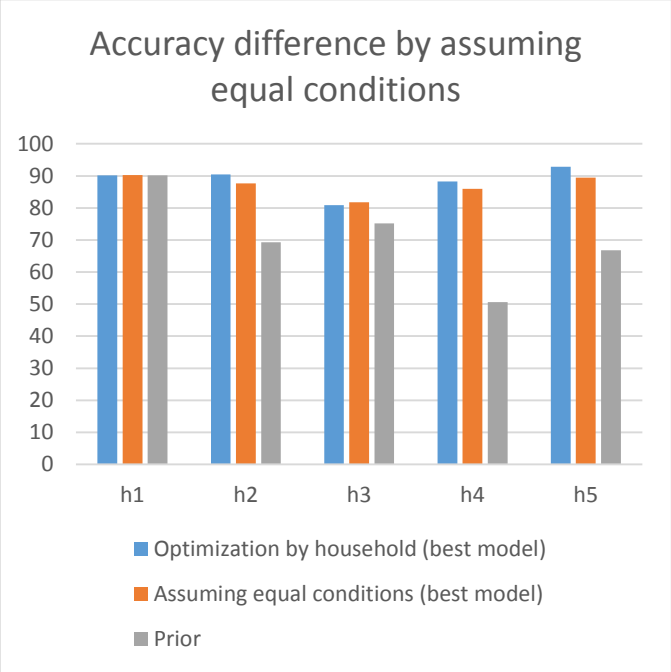

Figure 4: Comparison between the classification accuracies obtained in the optimization by household experiment and assuming equal conditions for the models.

4.1.2. Detecting household's occupancy in other households (other's-test)

To analyze the possibility of using a single and a generic classification model to detect occupancy in multiple households, we test, for each household, the classification models of the remaining households. For example, for testing in household 1, we first train the three algorithms in the training set of households 2,3,4 and 5 (separately). Then, we test the classification accuracy by applying these algorithms in the classification set of household 1.

To summarize the results in this experiment, Table 8 shows, for each household, the highest accuracy obtained by testing models trained in the remaining households. In the previous experiments (*self-test*), the random forest model provided the best overall results. In this experiment, the best results were obtained by the neural network model. This indicates that the random forest model may be better if we want to maximize the accuracy in a given household (*self-test*) and that the neural network model may be better if our goal is to use a single model to detect occupancy in multiple households.

Table 8: Best accuracy results obtained in the other's-test experiment.

Testing household	Best model(s)	Accuracy (%)	MCC	Prior accuracy (%)
1	ANN model of household 2 and 5; All models of household 4	90.15	0	90.15
2	ANN model of household 3	82.64	0.57	69.31
3	ANN model of household 4	80.10	0.39	75.22
4	ANN model of household 3	74.85	0.51	50.59
5	ANN model of household 3	72.27	0.49	66.82

The neural network model trained in household 3 provided the highest accuracy in the remaining households, except for household 1, with a small difference. Household 3 performed the best by using the neural network model from household 4

In households 2,3,4 and 5, the accuracies obtained are higher than the respective Prior accuracy and the lower value of MCC is 0.39, which clearly indicates that it is possible to use a single generic model to detect occupancy in multiple households. The best accuracy improvement, comparing to the Prior, was achieved in household 4 (48% improvement), which is in agreement with the highest value of MCC observed (0.51).

In Table 9, it is shown the false positive and false negative rates for each household and for each model correspondent in Table 8. The values in parentheses represent the number of minutes per day, in average, that are misclassified for the respective error type (false positive or false negative). For household 2, the table can be interpreted in the following manner: the neural network model of household 3 applied in the household 2, would misclassify in average and per day, 146 minutes of unoccupied periods as occupied (false positives) and 20 minutes of occupied periods as unoccupied (false negatives).

Table 9: False positive rate and false negative rate (in percentage) for best results of the other's-test experiment.

Household	FPR (%)	FNR (%)
1	100 (95 min)	0
2	49.70 (146 min)	3.41 (20 min)
3	64.58 (154 min)	5.2 (37 min)
4	39.03 (185 min)	11.60 (56 min)
5	34.90 (224 min)	13.30 (43 min)

From Table 9, we can observe that false positive errors are the most frequent errors verified in all households by using a classification model that was trained in other household. It means that the model has more difficult in detecting the unoccupied periods.

To summarize the results obtained in the occupancy detection experiments, Figure 5 illustrates the highest classification accuracy obtained in each of the three previous experiments:

1) Optimization by household (*self-test*): for each household, are chosen the optimal model parameters and feature set;
2) Optimization with assumed parameters (*self-test*): the same model parameters and feature set is applied in every household;
3) Using other household's models (*other's-test*): for each household, are applied the models of the remaining households and is selected the best model.

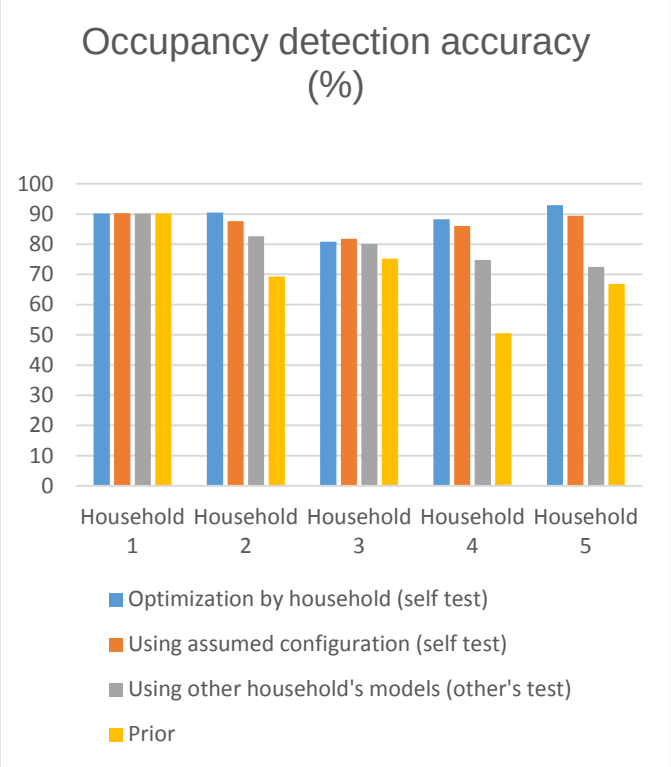

Figure 5: Comparison between the classification accuracies obtained in the optimization by household experiment and assuming equal conditions for the models.

4.2. Occupancy prediction

To analyze the possibility of predicting occupancy using solely electricity consumption data, we first use a classification algorithm to generate occupancy data from electricity consumption data. Then, we use the Presence Probabilities Simplified (PPS) to construct a prediction timetable based on the occupancy data that was generated by the classification algorithm. The classification algorithm is trained in the training set and is applied in the classification set. The PPS algorithm uses occupancy data with regard to the classification set and the prediction accuracy is calculated by applying the prediction timetable in the future set.

4.2.1. Using occupancy data generated by a classification model (type 1) and ground truth occupancy (type 2) (self-test)

To analyze the prediction accuracy separately in each household (*self-test*), we start by creating two types of prediction timetables. Then, we test the prediction accuracy of each type of timetable on the "future set".
The first type of prediction tables (type 1) is computed by using the occupancy data that was generated by the best classification algorithm in the previous experiment (random forest for household 1,2,4 and 5 and the neural network model for household 3). The second type of prediction table (type 2) represent the timetable that is constructed by using the ground truth occupancy data and is used as a baseline for our analysis. Obviously, we expect to obtain a higher prediction accuracy by using the prediction timetables type 2, since they do not incur on the error due to the classification process.

***Self-test* prediction summary**

Table 10 contains the results obtained, for each household, by applying the two types of prediction tables. In households 2,4 and 5, for both prediction tables type 1 and 2, the prediction accuracy was higher than the Prior accuracy and positive MCC values were obtained. This result indicate that our prediction algorithm allowed to predict occupancy in these households, but provided no value for households 1 and 3.

Table 10: Occupancy prediction results obtained by applying the prediction timetables type 1 and 2 on the future set of each household.

Household	PP table – type 1	PP table – type 2	Prior accuracy (%)
1	**88.74**	88.18	**88.74**
2	78.42	**80.02**	74.11
3	83.11	76.83	**85.27**
4	59.85	**66.89**	53.66
5	76.27	**80.77**	70.54

4.2.2. Using occupancy data generated by a classification model (type 1) from other household (other's-test)

Our final experiment consists on investigating the possibility of predicting occupancy in multiple households using solely their electricity consumption data. More specifically, our objective is to analyze the viability of predicting occupancy by using the prediction timetable type 1, but now, constructed with the best classification models from other households (*other's-test*).
For example, to predict occupancy in household 2, we first apply the neural network model of household 3 (the best) in household 2 to obtain occupancy data from electricity consumption data (classification). Then, we construct a

prediction timetable type 1 by using this occupancy data generated by the classification algorithm.
In section 4.1.2. (*other's-test* experiment), we verified that the neural network model of household 3 provided the highest occupancy detection accuracy in the remaining households. Thus, the prediction timetable type 1 of households 2, 4 and 5 are constructed by using occupancy data generated by this model.
After constructing the prediction timetables for household 2, 4 and 5, we test their occupancy prediction performance in the "future set". Table 11 summarizes the results obtained in this experiment.

Table 11 : Occupancy prediction accuracy obtained by using the prediction timetable type 1 (constructed by using occupancy data generated by the neural network model of household 3) and the respective MCC value.

Household	Prediction accuracy (%)	MCC	Prior accuracy (%) – future set
2	75.42	0.20	74.11
4	58.91	0.24	53.66
5	61.73	0.31	70.54

Despite of the not so high prediction accuracies, we can observe that a prediction accuracy higher than the Prior was obtained in households 2 and 4 and that all of the three households presented a positive MCC value, which indicates that the prediction algorithm performed better than a random choice in all households.

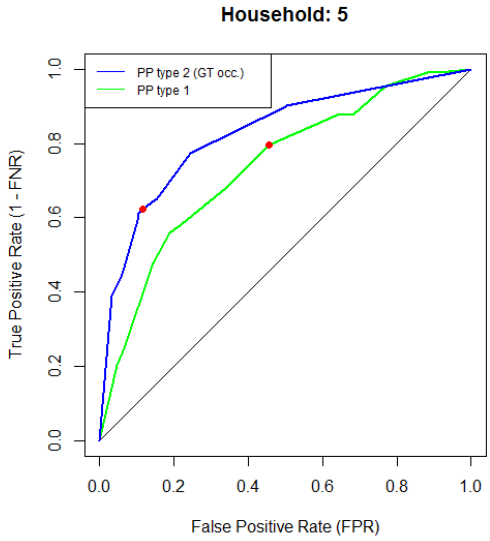

Figure 6: ROC curves for occupancy prediction in household 5. The green line corresponds to the prediction timetable constructed with data generated by the neural network model of household 3.

5. CONCLUSION

In this work, we verified that a classification accuracy of up to 92% is possible to be obtain from the electricity consumption data, in households with low occupancy level. Regarding the possibility of using a single classification model to monitor occupancy in multiple households, we observed that a single classification model allowed to detect occupancy in 4 out of 5 households and that a classification accuracy of up to 82% is possible to be obtained.
To predict occupancy through the electricity consumption data, we first used a classification model to obtain occupancy data from electricity consumption data. Then, we used the Presence Probabilities Simplified (PPS) algorithm to create a prediction timetable from the generated occupancy data.
We observed that, by using the respective classification model in each household, 3 out of 5 households presented a higher prediction accuracy than the Prior method, and that an accuracy of up to 78% was possible to obtain. By using the occupancy data generated by a single classification model, 2 out of 3 households provided a higher prediction accuracy than a Prior classifier and that a classification accuracy of up to 75% was possible to obtain. This result is not magnificent and may not be sufficient to justify the use of our approach to predict occupancy, however, it indicates that it is possible to predict occupancy in multiple household by using a single classification model. We consider that better results may be obtained by using a hybrid approach to predict occupancy, i.e. by combining context-aware and schedule-based approaches.

6. ACKNOWLEDGMENT

I would first like to thank EDP for providing me the opportunity for developing this thesis in the company. It was a pleasure to contribute for the interests of the company. I wish to acknowledge my two supervisors, Prof. Rui Castro (from Instituto Superior Técnico) and Pedro Adão (from EDP) for their support and guidance during the period of this work. Without them, this work would not be possible.
I would also to thank Prof. Alexandre Bernardino, for the availability in clarifying many doubts that I had with regard to specific topics in the statistics field.
In this work, we collected the necessary data from five households. A special thanks to these five participants who made themselves available in helping me during a period of about five months.
Last but not least, I would like to thank my family and my friends for the support and encouragement that they provided to me during this work, especially in the more difficult periods.

7. REFERENCES

[1] J. Lu, T. Sookoor, V. Srinivasan, G. Gao and B. Holben, "The Smart Thermostat: Using Occupancy Sensors," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, Zürich, 2010.

[2] R. Yang and M. W. Newman, "Learning from a learning thermostat: lessons for intelligent systems for the home," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013.

[3] D. Chen, S. Barker, A. Subbaswamy, D. Irwin and P. Shenoy, "Non-Intrusive Occupancy Monitoring using Smart Meters," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, 2013.

[4] W. Kleiminger and C. Beckel, "Occupancy Detection from Electricity Consumption Data," in *Proceedings of the 5th ACM Workshop on Embedded Systems For Energy-Efficient Buildings*, 2013.

[5] K. Wilhelm, B. Christian and S. Silvia, "Household Occupancy Monitoring Using Electricity Meters," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 975-986, 2015.

[6] K. WILHELM, "Occupancy Sensing and Prediction for Automated Energy Savings," 2015.

[7] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei and T. Weng, "Occupancy-driven energy management for smart building automation," in *Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*, 2010.

[8] A. Akbar, M. Nati and F. Carrez, "Contextual occupancy detection for smart office by pattern recognition of electricity consumption data," in *2015 IEEE International Conference on Communications (ICC)*, 2015.

[9] J. Scott, A. J. Brush, J. Krumm, B. Meyers, M. Hazas, S. Hodges and N. Villar, "PreHeat: Controlling Home Heating Using Occupancy Prediction," in *Proceedings of the 13th international conference on Ubiquitous computing*, 2011.

[10] J. Krumm and A. J. Bernheim Brush, "Learning Time-Based Presence Probabilities," in *Pervasive Computing*, 2011.

[11] M. C. Mozer, L. Vidmar and R. H. Dodier, "The Neurothermostat: Predictive Optimal Control of Residential Heating Systems," *Advances in neural information processing systems*, 1997.

[12] L. Yang, K. Ting and M. Srivastava, "Inferring occupancy from opportunistically available sensor data," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, 2014.

[13] J. Weiss, "Lecture 22—Wednesday, November 10, 2010," [Online]. Available: https://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture22.htm. [Accessed 02 04 2017].

[14] S. Raschka, "Machine Learning FAQ," [Online]. Available: https://sebastianraschka.com/faq/docs/evaluate-a-model.html. [Accessed 04 04 2017].