Semester Project: Residential consumption forecasting in the Smart Grid

Samuel Humeau

***Abstract*—The recent development of smart meters has allowed the analysis of electric consumption at the level of the house in real time. It makes possible a prediction of the electric consumption at a very low scale. It could increase the efficiency of distribution networks and the pricing of the energy. In this work, we address the problem of the prediction of consumption at the level of one house. This is done using various machine learning techniques such as SVR and MLP. Moreover we exploit statistical relations between houses in order to improve prediction, by finding leaders among the time series, and by grouping houses with similar consumption.**

I. Introduction

THE management of energy have currently become a central matter worldwide. Most electricity distribution networks that are in place nowadays have not been designed to face current and future demand. Congestion and atypical power flows threaten to overwhelm the system while demand increases for higher reliability and better security and protection [1]. In this context, anticipation and load forecasting are more and more important, while the geographical scale of the predictions is lower and lower, stepping from a country in the 90s to a district those last years. Techniques for prediction have been inspired by research on machine learning, and have passed from linear regression [13] and ARMA model [11] to neural networks [7], boosting [2] or SVM regressions [4], [8], [12] . According to literature this last technique stands for a state of the art method for load forecasting. Those methods have been used successfully for consumption prediction at the scale of a country. However, this report deals with the possibility to use those tools and particularly SVM in the prediction of load consumption at the scale of one house. Moreover, we explore some techniques to exploit the recent smart meters installation to establish prediction one hour and 24 hours ahead.

II. Related work

The main goal of this project was forecasting the consumption of residential houses. In this sense, it is similar to what have been done by PJM in its elaboration and comparisons of consumer baselines [5]. The techniques used, such as Support Vector Machine for Regression, are greatly inspired from the article of Chen, Chang and Lin [4] who used SVR to win the EUNITE competition in 2001 (large scale electricity load forecast). Discretizing data, as it has been made in Section V is an approach that is directly inspired from one piece of work of HP labs [9], and also from the work of Wijaya, Eberle and Aberer [14] which deals with anonymity of collected data through discretization. Finally, Section VI where we look for leaders in time series in order to establish prediction is similar to the work of Yu, Ke and Wu [16] who established a method to use leaders for predicting exchange rates.

III. Preliminaries

A. Dataset

The Smart Metering Electricity Customer Behavior Trials (CBTs) took place during 2009 and 2010 with over 5,000 Irish homes and businesses participating. Its goal was to study the behavior of consumers exposed to different pricing policies of electricity. The dataset consists in a measure of the consumption of each house (in kWh) every half an hour. To conduct this study, only the control test of residential houses has been considered, which reduces the dataset to 782 houses. The owners of those houses were not aware of any pricing policy and keep consuming like they usually do. In the following paper, data has been aggregated in order to get a measure every hour.

For every results that are presented, we considered the first year as the training set (8460 hours) and the remaining 6 months as the test set (4317 hours).

This gives us a formidable tool to check if relations between houses could be used to obtain a better forecast of the load. In section IV we concentrate in predicting the consumption of each house separately, one hour and 24 hours ahead, without using our knowledge of the consumption of the other houses. Section V is an extension of this trial and try to discretize data in order to use other ways of classifying. In section VI, we use the consumption of every houses in order to forecast the load for each one, by finding leader houses in the set. Finally in Section VII we study the impact of the scale on the accuracy of the prediction. Moreover we link regroup similar houses into clusters in order to improve the forecast of the overall consumption.

B. Evaluation

In the literature the two well-known methods to evaluate forecasted values are MAPE (Mean of Average Percentage Error) and RMSE (Root Mean Square Error). In this report we mainly make predictions at the level of a house. For many houses, there are periods where the consumption is 0, or extremely low. In those cases, the MAPE is infinite which makes its evaluation not convenient. The RMSE does not have this problem. The RMSE on a time series $S : (s_1, ..., s_N)$ of its estimation $\tilde{S} : (\tilde{s}_1, ..., \tilde{s}_N)$ is defined by:

$$RMSE(S, \tilde{S}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \tilde{s}_i)^2}$$

We can notice that the RMSE is strongly influenced by the order of magnitude of the data. In our dataset, we dispose of multiple houses that average consumptions are comprised between 0.05kWh/h and 3.83kWh/h. In order to reduce this influence, in all the report we use the following normalization for the RMSE:

$$NRMSE(S,\tilde{S}) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(s_i - \tilde{s}_i)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N} s_i^2}}$$

In fact we normalize by the L_2 norm of the series. We will each time use the notation NRMSE to refer to the normalized RMSE.

C. Implementation details

All the programs realized for this study have been made using JAVA and Matlab, and the codes have been made public. We used the Weka framework, developed by university of Waikato [6], [15] to compute the predictions.

The Irish smart meter training set offers measures on a time span of 1 year and a half on 782 houses. Since computation of MLP and SVM are time consuming, the evaluation of each method of prediction has been made on a panel of 25 random houses, the same for each new experiment. This implies a standard deviation on the mean that we provide in our results.

To refer to an algorithm that realizes predictions, the correct name would be predictor. However, since the techniques that we use are extensions of classification methods, the term classifier will also be used.

IV. Predict without taking advantage of massive installation of smart meters

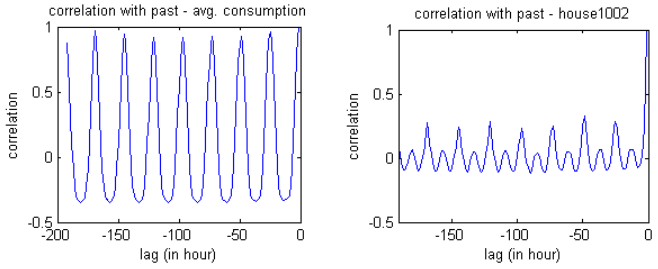
In order to be able to use efficiently the consumption of every house in the computation of predictions, it is necessary to see how far we can go without it. In this part we try to establish a prediction of the consumption of one house, both one hour ahead and 24 hours ahead, without using our knowledge of the consumption of other houses, but only considering the past of each house and the temperature.

A. Prediction made using historical consumption

1) Self correlation: The idea of using previous consumption values to predict the new one is very intuitive. Even for a house, the consumption follows a certain rhythmic, and is very low at night to reach a maximum around 20h. The common sense tells us that the consumption at a time t must be similar to the one 24 hours before. However, due to the high variability of the consumption of one house, the accuracy of this intuitive rule is low at this scale. The Figure 1 illustrates this fact, by showing the auto-correlation of the aggregated consumption over all houses. The auto-correlation of $S : (s_1..., s_N)$ for a certain lag τ is calculated as below:

$$\frac{\mathbf{E}((s_i - \mathbf{E}(s)) \times (s_{i-\tau} - \mathbf{E}(s)))}{\sigma(s)^2}$$

where $\mathbf{E}$ is the average and σ the standard deviation. Figure 1 shows clearly that the consumption at time t is highly related to the consumption at the same hour the day before, especially 7 days before. Figure 1 shows that this self-correlation is lower in the case of an individual house. The consumption is highly correlated with consumption up to 3 hours before, which will motivate to consider those features in the case of the one hour ahead prediction.



(a) (b)

Figure 1. a: self correlation of the aggregated consumption over the 782 houses. b: self correlation of the consumption of individual houses (id 1002).

2) Prediction using previous consumption: Figure 2 shows the average consumption of the houses along one week. We can clearly distinguish each day (5 similar days, then Saturday and Sunday that follow a different pattern). It is noticeable that consumption depends on the hour of the day, but also on the day of the week. This leads us to add them in our set of feature as nominal features. Table I shows the set of features that have been chosen for this first prediction. Given that recent literature describes SVM based Regression methods (SVR) as the most effective one to predict future consumption [4], [8], we have used an SVR classifier. It uses an RBF kernel, and parameters of this model have been optimized, (full procedure is explained in Section VIII). In order to compare our results, we have also considered 2 other predictors:

- a simple linear regression on all features considered
- a Multi-Layer Perceptron, composed of one sigmoid hidden layer and one output. The setting of this classifier is also described in Section VIII.
- Moreover we compare our results with two predictors that would simply output the consumption 24 hours ago and 1 hour ago.

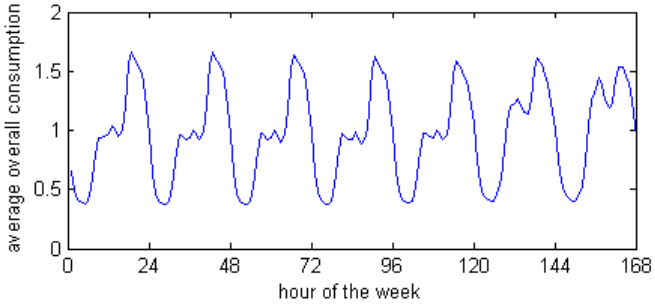


Figure 2. Average consumption (over time and houses) function of the hour of the week. Notice that we distinguish clearly each day, and that patterns for week-end strongly differ.

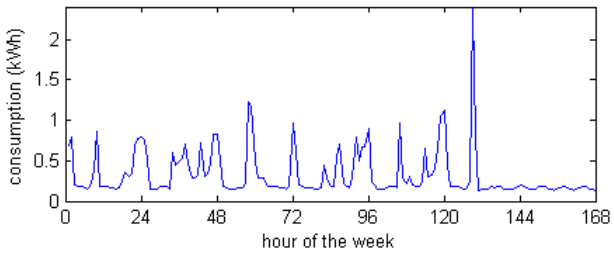


Figure 3. Consumption of house with id 1002 the week from 2009/09/07 to 2009/09/13

Table I
FEATURES USED FOR PREDICTION IN SECTION IV-A

1 hour ahead	24 hour ahead
$cons.t - 1h$	$cons.t - 1 \times 24h$
$cons.t - 2h$	$cons.t - 2 \times 24h$
$cons.t - 3h$	$cons.t - 3 \times 24h$
$cons.t - 1 \times 24h$	$cons.t - 4 \times 24h$
$cons.t - 2 \times 24h$	$cons.t - 5 \times 24h$
$cons.t - 3 \times 24h$	$cons.t - 6 \times 24h$
$cons.t - 7 \times 24h$	$cons.t - 7 \times 24h$
hour of the day	hour of the day
day of the week	day of the week

Results can be seen Table II. To our great surprise, the linear regression has been the most effective on our test bench, both for one and 24 hours ahead predictions. This is curious since SVR and MLP give better results (by far) when considering aggregated consumptions (see sub-section VII-A). It could come from the fact that our data are not structured enough, and that to a same set of historical consumptions corresponds different loads. Indeed, Figure 3 shows the consumption of one house during one week, and even with human eye it is hard to recognize structures in it. It implies that other features should be necessary in order to do a more discriminative prediction.

B. Usage of temperature

An improvement that is often seen in the literature consists in adding the temperature as a feature. Figure 4 shows that there is a relation between the temperature and the consumption. We used the Wunderground database in order to obtain the temperatures in Dublin during the period of the establishment of the dataset. Real temperatures have been used, and not the one that could have been predicted. It is non-realistic, since in real conditions, only predicted values

Table II
A. NRMSE OF 4 CLASSIFIERS B. RESULTS FOR 4 CLASSIFIER ON THE TEST SET FOR ONE HOUR AHEAD PREDICTION WITH FEATURES OF SECTION IV-A.

	1 hour ahead	h-24	h-1	Lin. Reg.	SVR.	MLP
A.	Average NRMSE	0.80	0.71	0.56	0.57	0.59
	σon the average	0.030	0.038	0.023	0.025	0.027

	24 hour ahead	h-24	Linear Regression	SVR	MLP
B.	Average NRMSE	0.80	0.61	0.64	0.70
	σon the average	0.030	0.019	0.023	0.038

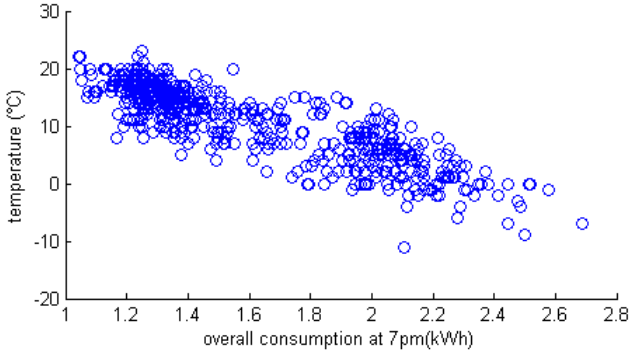


Figure 4. Relation between the averaged consumption at 19h, and the temperature (averaged over all houses)

of temperature could be accessed. However, it is used here to check whether adding temperatures could help or not. The results, shown in Table III can be directly compared with Table II. We have found no improvements when using temperatures, in both one hour ahead and 24 hours ahead cases.

Table III
A. NRMSE OF 4 CLASSIFIERS B. RESULTS FOR 4 CLASSIFIER ON THE TEST SET FOR ONE HOUR AHEAD PREDICTION WITH FEATURES OF SECTION IV-B.

	1 hour ahead	h-24	h-1	Lin. Reg.	SVR.	MLP
A.	Average NRMSE	0.80	0.71	0.56	0.57	0.59
	σon the average	0.030	0.038	0.023	0.025	0.027

	24 hour ahead	h-24	Linear Regression	SVR	MLP
B.	Average NRMSE	0.80	0.61	0.64	0.70
	σon the average	0.030	0.019	0.023	0.038

C. Existing baselines

The idea of predicting the consumption for each house individually is not new. In its "Empirical analysis of response baseline method"[5], PJM has compared 12 different baselines methods. Those methods give predictions for the next day consumption, and are currently use to establish contracts between clients and electricity provider. Those methods are mainly used on big electricity consumer (>100kWh/day) like small company, which have a regular consumption. On those 12 methods, we have selected the three that were, according to the analysis of PJM, both efficient for regular and variable consumptions.

- Middle 4 of 6: is a x out of y method. It separates weekdays and week-end. For each kind of day, it takes into consideration the 6 previous days of the same type. For each hour, it drops the highest and lowest kWh day of the considered six, and takes the average of the remaining 4.
- PJM Eco: differentiates also weekdays and weekend. For weekday events, the baseline consists of the average hourly loads of the 5 most recent weekdays preceding the event. For weekend events, the baseline consists of the average hourly loads of the 2 highest kWh days out of the 3 most recent week-ends.

- ISONE: is a weighted average of the preceding days. It can be summarized as:
  - the baseline for the first predicted day is an average of the 5 previous days
  - current day baseline = 0.9×previous day baseline +0.1×current day metered load

We have implemented those baselines, and Table IV shows their results on the test set. The comparison with our results of Section IV-B for 24 hours ahead predictions shows that ISONE in particular realizes nearly the best possible linear regression with the previous consumption.

Table IV
PERFORMANCE OF THE BASELINES (EVALUATED ON 24 HOURS AHEAD PREDICTION)

	middle 4 of 6	PJM Economic	ISONE
NRMSE	0.63	0.69	0.61
σon mean	0.020	0.022	0.019

D. Baselines as features

Baselines of previous Section are simple predictions based on previous day's consumptions. But they produce nonlinear operations on them (for example middle 4 of 6 suppresses the lowest and highest values of the 6 preceding days). From there rises the idea to use them as features. In addition to the features used in Section IV-E, we added the prediction of those 3 baselines. Results can be seen on Table V. It is noticeable that the error on prediction with linear regression has slightly decreased by adding the baselines, which proves that this feature can help a little.

Table V
A. NRMSE OF 4 CLASSIFIERS B. RESULTS FOR 4 CLASSIFIER ON THE TEST SET FOR ONE HOUR AHEAD PREDICTION WITH FEATURES OF SECTION IV-B

A.

1 hour ahead	h-24	h-1	Lin. Reg.	SVR.	MLP
Average NRMSE	0.80	0.71	0.54	0.56	0.57
σon the average	0.030	0.038	0.022	0.023	0.023

B.

24 hour ahead	h-24	Linear Regression	SVR	MLP
Average NRMSE	0.80	0.60	0.63	0.67
σon the average	0.030	0.019	0.022	0.039

E. Including derivatives

Other features often added are gradients and Lagrangian (Second derivative) of the series in features. This means that classifiers will not only deal with the absolute value of the consumption at every hour, but also take into account their evolution. For a signal $S : (s_1, .., s_N)$, the gradient at time i is given by $\frac{\partial S}{\partial t}_i = s_i - s_{i-1}$, and the second derivative is given by $\frac{\partial^2 S}{\partial t^2}_i = s_i - 2 \times s_{i-1} + s_{i-2}$. Notice that those 2 operations are linear, and thus the linear regression is naturally taking account of those two features (we provide it consumption at time t-1, t-2 and t-3). Since linear regression seems to perform well at the scale of one house, the idea of giving these features to the other classifiers rises. The other features considered were the same as previous Section (past consumption, temperatures, and baselines predictions). The results are shown in Table VI. Notice that results are only provided for one hour ahead prediction, since we only add the derivatives of the recent past consumption. Unfortunately, we have not noticed any statistically significant improvement of the quality of the prediction for SVR and MLP.

Table VI
A. PERFORMANCE OF 4 CLASSIFIERS ON THE TEST SET FOR ONE HOUR AHEAD PREDICTION WITH FEATURES OF SECTION IV-E.

1 hour ahead	h-24	h-1	Lin. Reg.	SVR.	MLP
Average NRMSE	0.80	0.71	0.54	0.56	0.57
σon the average	0.030	0.038	0.023	0.023	0.022

F. Separating night and day

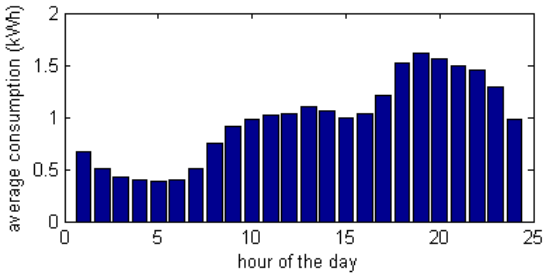


Figure 5. Averaged consumption (over 782 houses and 1 year) for each hour of the day

Figure 5 gives the average consumption of houses for each hour of the day. We can notice that days are split in two logical parts, day and night. Between 1am and 7am the consumption is extremely low, compared to the rest of the day. A perfect prediction of the night is not expected in the context of load forecasting, while an accurate forecast during heavy hours in mandatory. This has led to train our predictors separately on day and night (which starts from 1am to 7am included). An improvement of the SVR prediction was not expected with this operation: since we give the hour of the day as a feature, SVR is supposed to discriminate those by cases by its own. The idea was to improve the performance of linear regression and MLP. The features used were the same as in Section IV-E. As usual, results for one hour prediction are presented in Table VII. Once again we do not observe any statistically significant improvement.

Table VII
PERFORMANCE OF 4 CLASSIFIERS ON THE TEST SET FOR ONE HOUR AHEAD PREDICTION WITH FEATURES OF SECTION IV-F.

1 hour ahead	h-24	h-1	Lin. Reg.	SVR.	MLP
Average NRMSE	0.80	0.71	0.54	0.56	0.59
σon the average	0.030	0.038	0.023	0.023	0.024

G. Including last hour consumptions for 24 hour ahead prediction

Despite the fact that we were unable to find features that greatly improve the average quality of the prediction of the

consumption for one house, we could have notice that the forecast one hour ahead is better than 24 hours ahead. This seems common sense, and is due to the fact that the correlation between the consumption at time $t-1$ and t is extremely high. A lot of electric consuming tasks (watching TV for example) take more than one hour. Then it leads us to consider the prediction that we established for the time $t-1$, $t-2$ and $t-3$ when predicting of the consumption at time t. During the training, real values of the previous consumption are used. This technique has been used without distinguishing night and day, and considering only the features use in Section IV-B for 24 hours ahead prediction, plus the consumption at time $t-1$,$t-2$ and $t-3$ (Table VIII summarizes the set of features used in this Section). Since only those 3 features are added, Table IX can be directly compared with Table III to determine the effect of adding the last hour consumption. The comparison revealed that results are worse for all three classifiers, particularly for the MLP. It can illustrate that classifiers "rely" strongly on the last hour's consumption. If the values for these features are false, it induces errors.

Table VIII
FEATURES USED IN SECTION IV-G.

$cons.t-1h$	$cons.t-4\times 24h$
$cons.t-2h$	$cons.t-5\times 24h$
$cons.t-3h$	$cons.t-6\times 24h$
$cons.t-1\times 24h$	$cons.t-7\times 24h$
$cons.t-2\times 24h$	temperatures
$cons.t-3\times 24h$	day of the week
hour of the day	

Table IX
PERFORMANCE OF 4 CLASSIFIERS ON THE TEST SET FOR 24 HOURS AHEAD PREDICTION WITH FEATURES OF SECTION IV-G.

24 hours ahead	h-24	Linear Regression	SVR.	MLP
Average NRMSE	0.80	0.62	0.66	0.72
σon the average	0.030	0.019	0.022	0.041

H. Propagation of error through time

As said before, it noticeable that prediction one hour ahead gives a better result than the three consumer baselines of Section IV-D. Notice that if we are able to predict the consumption 1 hour in advance, then by iterating the process we can obtain a prediction of the consumption 2 or more hours ahead. Suppose the consumption is $S:(s_1...s_N)$:

$$\begin{cases} \tilde{s_i} = F(s_{i-1}, \quad features) \\ \tilde{s}_{i+1} = F(\tilde{s}_i \quad features) \\ ... \end{cases}$$

The question is: how far can we go while keeping the error sufficiently low. The results of the computation can be seen on Figure 6. To obtain those results, only a linear regression (that seems the best that we have so far for this scale) has been considered. The results are averaged on the 782 houses of the dataset. We notice that after 2 iterations of the technique, the error induced by this technique is above these of ISONE baseline. This shows a high rate of propagation of the error.

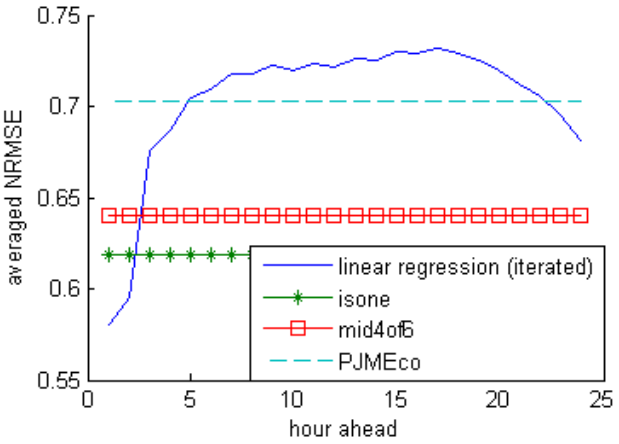


Figure 6. Propagation of the error when iterating the linear classifier of Section IV-A.

V. DISCRETIZATION

Many possibilities are allowed by discretizing data. Among them, there is the possibility to use nominal algorithms in order to find relations through the series of the consumptions of the houses.

A. CTTM discretization

A good discretization should approximate the real time series at best with the lowest number of bits (the number of categories). Two methods are mostly used to discretize data.

- The uniform method, which splits the value space into M equal intervals (where M is chosen by user) between the lowest and the highest values. The value assigned to the interval is usually its center. As its name suggests, it is particularly appropriate when values are distributed uniformly on a segment. In our case, the series are positive, with some rare high pikes and most frequent very low values. Then uniform discretization gives too much precision on the rare pikes, but not enough for the low values, which will be regrouped in the same category.
- The quantile method, which splits the value space into M quantiles in order to fill them with the same number of datapoints. The value assigned to each interval is usually the average value on the interval. In our cases, since most of values are very low, the quantile method brings a lot of precisions in the discretization of those values. However, we are mostly interested in predict well the consumption at time where it has a big influence on the average consumption.

In order to find a more appropriate trade-off between those two methods, we introduce the third method, that we called CTTM as Contribution To The Mean. As its name suggests, it discretizes the values according to its importance in the average consumption.

Let $S:(s_1,...,s_N)$ be a time series composed of positive scalar values, that follow a probability density function p.

$$\int_{s=-\infty}^{+\infty} p(s)ds = 1$$

The average consumption is:

$$\bar{s} = \int s \cdot p(s) ds$$

Let's define $\tau_0, ..., \tau_M$ the boundaries, defined by the following relationships:

$$\begin{cases} \tau_0 = -\infty \\ \tau_M = +\infty \\ \int_{t=-\infty}^{\tau_j} s \cdot p(s) ds = j \frac{\bar{s}}{M} \end{cases}, \ 0 < j < M$$

Then given a datapoint s_k, we discretize it by finding in which interval it belongs, i.e. find the integer i such that $s_k \in [\tau_i, \tau_{i+1}]$. Then we assign it to the average of the values in this interval. The Discretization operator D_{CTTM} can therefore be defined as:

$$D_{CTTM} \ : \ s_k \rightarrow \frac{\int_{\tau_i}^{\tau_{i+1}} s \cdot p(s) ds}{\int_{\tau_i}^{\tau_{i+1}} p(s) ds}, \qquad s_k \in [\tau_i, \tau_{i+1}]$$

Figure 7 gives a clear view of the strength of this method, while Table X compare the RMSE induced by the compression, obtained by those three methods in average on the Irish smart meter dataset. CTTM gives a precise discretization using only a few bits, therefore we will use it for any discretization of our time series.

Table X
AVERAGE (OTHER ALL HOUSES OF THE IRISH SMART METER DATASET) OF THE ERROR CAUSED BY COMPRESSION, FOR EACH METHODS. BOUNDS OF DISCRETIZATION HAVE BEEN EACH TIME ESTABLISHED ON THE TRAINING SET (FIRST YEAR).

Average RMSE				
	training set		test set	
	8 bits	16 bits	8 bits	16 bits
Uniform	0.35	0.16	0.35	0.17
Quantile	0.41	0.28	0.43	0.30
CTTM	0.20	0.12	0.22	0.13

B. Results of discretization

We can now apply some well-known nominal classifiers. We chose J48 and Random Forests (tree based classifiers), the naïve Bayes classifier, and a SVM. The subsequent idea is that those algorithms could be more appropriate to find patterns in the series and establish rules. Table XI shows the performances of each of those 4 nominal classifiers, with the feature set of Section IV-B. The first conclusion is that those classifiers produce a performance far worse than the regression based classifiers. In fact, they do not even reach the level of the baselines. A second conclusion is that CTTM distribution has led to better results than the 8 bits uniform discretization which is also shown on Table XI for comparison. Results are similar while using an 8 bits or 16 bits discretization. This has led us to discredit nominal classification for load forecasting. However, discretization will be used in Section VI to determine the mutual information between the consumption of each house.

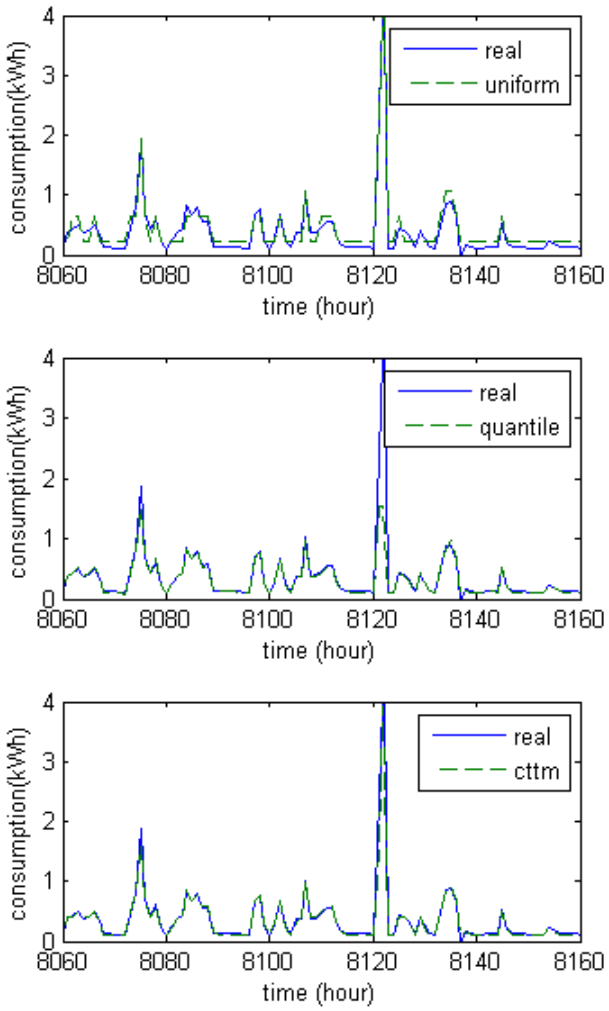


Figure 7. Comparison between the three ways of discretizing. In each case, the discretization is made with 16 bits. We can see clearly that the uniform method lacks of precision for low values, while the quantile method is inappropriate for large values. CTTM offers a trade-off between those two methods. (Extract of the consumption of house 1002 of the Irish dataset).

Table XI
PERFORMANCE OF 4 NOMINAL CLASSIFIER ON THE SET OF SECTION IV-B WITH TWO METHODS OF DISCRETIZATION, CTTM AND UNIFORM DISCRETIZATION.

Using 8 bits CTTM	J48	Naive Bayes	Rand. Forest	SVM
Average NRMSE	0.67	0.68	0.71	0.65
σon the average	0.029	0.026	0.025	0.023

Using 8 bits Unif	J48	Naive Bayes	Rand. Forest	SVM
Average NRMSE	0.69	0.75	0.74	0.71
σon the average	0.022	0.044	0.022	0.026

C. Using rule based algorithms

Having discretizing the data also allows us to use rule based algorithm such as the "A priori" algorithm. Rules are not sufficient to establish prediction: there might be some case where no rules can be applied. That is why rules can only be applied as a post processing. The idea is to discover in the dataset rules that characterize the consumption at time t with a good precision. In the case where a classifier would

produce a prediction that does not respect this rule, the post process could then correct it. In practice, we discovered that on the discretized dataset (with the set of features of Section IV-D, and an 8 bits CTTM discretization) 100% of rules with a precision above 70%, and also the 1000 rules with the highest precision (for a support above 0.01) concern the minimum value of consumption. One typical example of the generated rules would be: "if the consumption at 3am is minimal, then consumption at 4am is minimal". The application of those rules then produces only a minimal change, concerning periods where the consumption is low and prediction accurate. Thus, it has not been possible to exploit those rules.

VI. Taking advantage of massive installation of smart meters

Every method that has been presented so far only used the consumption of the house on which we want to make prediction. The Irish smart meter dataset gives us the possibility to establish relations between houses, and exploit them. This has been made in two different ways. The first is finding houses that lead the consumption of others, to try to improve one hour ahead prediction. The second, which is developed in Section VII is to assemble houses that have similar consumption, in order to create clusters of similar houses, and use them to establish the overall consumption with accuracy.

A. Finding leaders in time series

The initial idea of this part was the hypothesis that some houses play the role of leaders in terms of consumption. This approach has already been exploited by Wu [16], in the domain of financial trends forecasting. If we find some houses that seem to lead consumption, then we can exploit them for short term prediction. This part is exclusively dedicated to one hour prediction.

1) Finding leaders using cross correlation: The task is to determine which house leads the consumption of another house. In order to determine it, we interpret the time series of the consumption of the houses as signals in order to use the cross correlation. The cross-correlation between two signals $S^{(1)} = (s_1^{(1)}, ..., s_N^{(1)})$ and $S^{(2)} = (s_1^{(2)}, ..., s_N^{(2)})$ for a certain lag τ is given by:

$$CC(S^{(1)}, S^{(2)}, \tau) = \frac{\mathbf{E}((s_i^{(1)} - \mathbf{E}(s^{(1)})) \times (s_{i-\tau}^{(2)} - \mathbf{E}(s^{(2)})))}{\sigma(s^{(1)})\sigma(s^{(2)})}$$

Where $\mathbf{E}$ is the average and σ is the standard deviation. Since the goal is to predict at short term, we can restrict the delay τ to be small, but not 0. In our case, for each pair of houses (h_1,h_2), we compute the cross correlation between the consumption of those two houses for τ comprised between -4 and 4, and different from 0. Then we determine the lag $\tau_{h_1h_2}$ for which the correlation between the two signals is maximal. A positive $\tau_{h_1h_2}$ means that house h_2 leads the consumption of house h_1. We can therefore store the obtained result in two symmetric matrix, MCC and LCC. MCC(h_1,h_2) gives the maximum correlation between house $h_1$and h_2 while LCC(h_1,h_2) gives $\tau_{h_1h_2}$. On the diagonal we observe the correlation of each house with itself (lagged). In most of the cases, the consumption of houses was more correlated with their past consumption than with the past consumption of any other houses.

For each house we can now extract the k best leaders for each house.

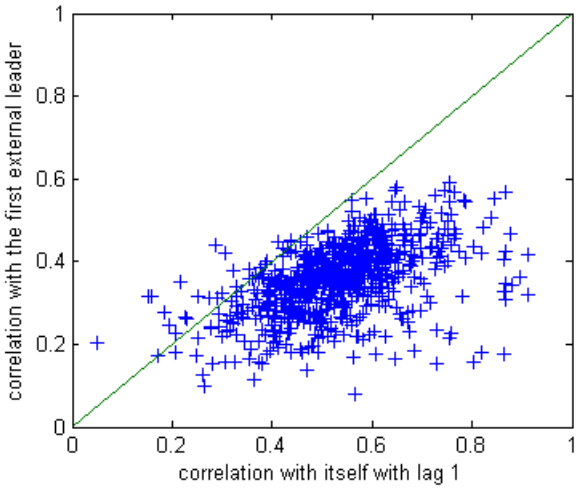


Figure 8. Correlation between the first external leader function of the self correlation with lag 1. Notice the consumption of some houses is more correlated with the one of external leaders than with their own.

MCC

id house	1002	1014	1018	1022	1027	...	6817	...
1002	0.59	0.20	0.11	0.21	0.14	...	0.22	
1014		0.58	0.16	0.20	0.23	...	0.27	
1018			0.67	0.18	0.16	...	0.22	
1022				0.33	0.22	...	0.18	
1027					0.46	...	0.36	
...						...	...	
6817							0.45	

LCC

house	1002	1014	1018	1022	1027	...	6817	...
1002	1	1	1	1	1	...	1	
1014		1	1	1	2	...	1	
1018			1	1	1	...	1	
1022				1	1	...	-1	
1027					1	...	-1	
...						...	...	
6817							1	

Figure 9. Illustration of the matrix MCC and LCC. Here the house with id 1027 seems to "lead" the house with the id 6817, because the correlation between the consumption of the first at time $t-1$ with the consumption of house 6817 is 0.36, which is close to its self correlation with lag 1 (0.45)

2) Leaders determined using mutual entropy: Determine leaders using cross correlations means that we are looking for houses that have a very similar consumption, but delayed. However, in the objective to make prediction using sophisticated machine learning techniques, we are interested in every type of relation between the consumption series of two

houses. In this case, the mutual entropy is more appropriate. This concept comes from information theory and expresses the quantity of information that links two random variables. Let X, Y be two discrete random variables. Then the mutual information between X and Y is expressed as:

$$I(X,Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$$

Where $H(X)$ is the entropy of the variable X.

$$H(X) = -\sum_x P(X=x)\log_2(P(X=x))$$

$$H(X,Y) = -\sum_{x,y} P(X=x, Y=y)\log_2(P(X=x, Y=y))$$

We now consider our consumption series as random variables, discretized using CTTM 8 bits. To be precise each datapoint (one for each hour) is considered as an evaluation of this random values. As for the cross correlation method, we try to maximize the mutual entropy between the consumption series of two houses, after introducing a delay τ in one of the series. Let $\tilde{S}_{h_1} = (\tilde{s}_1^{(h_1)}, ..., \tilde{s}_N^{(h_1)})$and $\tilde{S}_{h_2} = (\tilde{s}_1^{(h_2)}, ..., \tilde{s}_N^{(h_2)})$ be the discretized series of the consumption of house h_1 and h_2. $\tilde{s}_i^{(h_j)}$ is an integer comprised between 1 and 8. Then the mutual information between the consumptions of two houses h_1 and $h_2$for a certain lag τ is:

$$I(\tilde{S}_{h_1}, \tilde{S}_{h_2}, \tau) = -\sum_k P(\tilde{s}_i^{(h_1)} = k)\log_2(P(\tilde{s}_i^{(h_1)} = k)) - ...$$

$$\sum_k P(\tilde{s}_{i-\tau}^{(h_2)} = k)\log_2(P(\tilde{s}_{i-\tau}^{(h_2)} = k)) + ...$$

$$\sum_{k_1,k_2} P(\tilde{s}_i^{(h_1)} = k_1, \tilde{s}_{i-\tau}^{(h_2)} = k_2)\log_2(\tilde{s}_i^{(h_1)} = k_1, \tilde{s}_{i-\tau}^{(h_2)} = k_2)$$

The probabilities are estimated on the series (if a value k appears n_0 times, then its estimated probability of appearance is $\frac{n_0}{N}$). In the same manner as what we have done for cross correlation we can extract the k best leaders for each house.

B. Taking advantage of leaders for one hour ahead prediction

In order to evaluate whether the introduction of leaders can really improve the very short term prediction (one hour ahead), we applied our predictors on two different sets:

- The first set of features, $Set_{without}$ is the same as used in Section IV-B
- The second, $Set_{with} = Set_{without} + Set_{leaders}$ where $Set_{leaders}$is composed of the consumption of the 5 best leaders with the corresponding lag.

The experiment has been repeated for the two kinds of leaders and the results can be read on Table XII, and can be compared directly with the results of Section IV-B. Unfortunately, it seems that the information added by the consumption of the leaders do not bring more accuracy in the forecast. This result is true for the two types of leaders, determined by cross-correlation and with mutual entropy.

Table XII
A. NRMSE OF 4 CLASSIFIERS, ON THE SET OF FEATURE INCLUDING THE LEADER DETERMINED USING CROSS CORRELATION. B. USING MUTUAL INFORMATION.

	Cross Correlation	h-24	Linear Regression	SVR.	MLP
A.	Average NRMSE	0.80	0.55	0.57	0.58
	σon the average	0.030	0.023	0.024	0.024

	Mutual Information	h-24	Linear Regression	SVR	MLP
B.	Average NRMSE	0.80	0.55	0.57	0.58
	σon the average	0.030	0.023	0.024	0.027

VII. AGGREGATIONS

A. Classification on aggregated value

The result we obtained for the prediction for each house is as we could have guessed less accurate that what we can find in the literature for aggregated consumption. Let's consider the aggregation of the predictions realized for each houses. Since the best classifier that we have obtained so far for individual house consumption prediction is a linear classifier, which is extremely fast to compute, we can predict the consumption of the 782 houses of our dataset, using features of Section IV-E. Then we aggregate the predictions, in order to obtain a forecast of the group of houses.

Table XIII
PERFORMANCE OF THE AGGREGATION OF THE PREDICTION ON EACH HOUSES, COMPARED WITH 3 FORECAST TRAINED ON AGGREGATED DATA. (ONE HOUR AHEAD PREDICTION)

RMSE (kWh)			
Aggregation	Lin. Reg.	MLP	SVR
65.0	58.3	60.1	37.4

MAPE			
Aggregation	Lin. Reg.	MLP	SVR
5.7%	5.2%	6.0%	3.4%

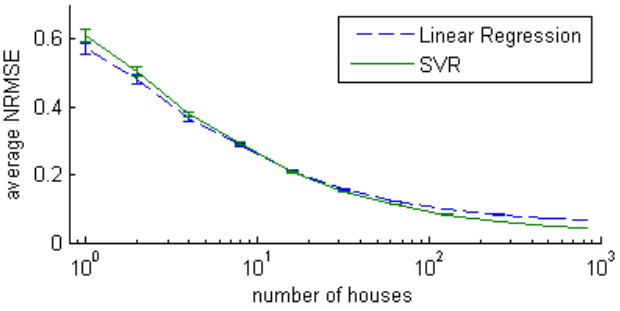
We compared this forecast with 3 others, created by using directly the aggregated consumption. The three predictors were once again linear regression, SVR and MLP. The features selected were composed of *consumption at time t-1, t-2, t-3, t-24, t-2x24, t-3x24, t-7x24, the hour of the day, the day of the week, the derivatives at time t-1, t-2, t-3 and the temperature*. Since this dataset is different from what have been considered so far, we optimized the SVR by using a validation set. The complete setup is described in Section VIII. Table XIII shows the results we obtain in terms of RMSE and MAPE. It shows clearly that when considering the prediction of the overall consumption, it is better to run regression on aggregated data. In our case, SVR has been the most efficient regression, with a MAPE of 3.4% for one hour prediction.

Exploiting aggregation for individual house prediction.: This leads to the idea of exploiting this accuracy to help predicting the consumption for each house. We have considered estimating the difference between the prediction of the overall consumption (obtained with SVR) and the aggregation, and redistributing the difference to every house. The obtain result (0.54 NRMSE) shows no improvements compared with Section IV-B, and the reason seems very intuitive. Errors on

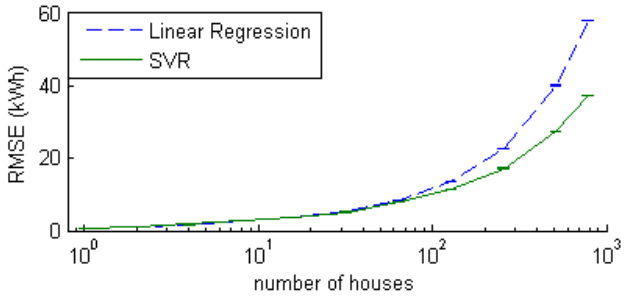
the prediction of houses consumptions come from random high peaks of consumption from a small number of houses. Applying a top down approach and redistribute the error of the aggregation to each houses, forces us to do it in a uniform way. Then the prediction of each house is slightly modified, while a correct but unfeasible way to correct it would be to do the redistribution on the small number of houses which consumption knows a peak.

B. Subscale

The previous Section and Section IV have highlighted two paradoxical results. To predict the consumption of a large number of houses, SVR is much better, but concerning the prediction for one house only, the linear regression is better. The question is now, from which number of houses that we consider does it inverse. To determine it, we have consider groups of $1, 2, 2^2, ..., 2^9$ and 782 houses (the whole set). Let K be the number of houses considered, we take randomly K houses among the dataset, and aggregate their consumption. Then, using the features of the previous Section, we train a linear regression and a SVR regression (using parameters that are described in Section VIII). We redo it $N_{redundancy}$ times (we chose $N_{redundancy} = 36$), in order to obtain a representative average. The result is shown Figure 10, in term of RMSE and NRMSE. It shows that the transition is made for K=16. This means that for a group of houses of more than 16 houses, using SVR starts increasing the quality of prediction.



(a)



(b)

Figure 10. Performance of linear and SVR regression for one hour ahead prediction. (a) expressed with NRMSE, described in Section III-B (b) expressed with RMSE (average over 36 redundancy).

C. Exploiting clustering

Previous sections have shown that predicting the overall consumption is more accurate when considering the aggregation of consumption than considering each house independently. A question that can be raised is to know whether there is a subscale that outperforms both, i.e., if we can group houses into clusters and forecast efficiently the consumption of every cluster, in order to obtain a better prediction on the whole set of houses. The subsequent idea is to regroup houses that have a similar consumption into clusters.

To define the "profile" of a house, we have considered the average consumption for each hour. Then each house is defined by a 24-dimensional vector. In order to find clusters among the 782 houses that composed our dataset, we applied a k-means algorithm. We repeated the operations for 1,2,4,8,16,32,48 and 64 clusters. Figure 11 shows that houses are well distributed along the clusters.

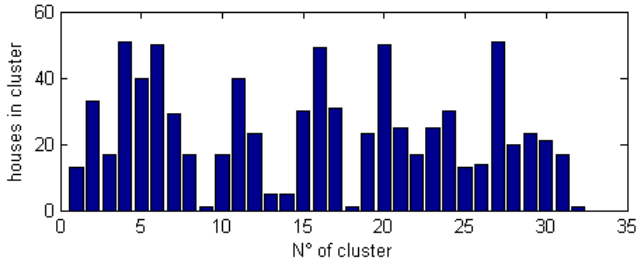


Figure 11. Number of houses in each cluster, when splitting the houses in 32 clusters with kmeans.

Finally, we forecast one hour ahead each cluster, and aggregate the consumption. Figure shows the evolution of the error on the overall consumption for different number of clusters, and different classifiers. It appears that for MLP and (but not on the same degree) for SVR, a minimum of the error is obtained for 4 clusters. Results are shown on Figure 12. This is a proof that in some cases, use of a subscale can lead to better results. Further work on this part would be to consider the best classifier for each cluster (that could be determined on a validation set).

VIII. About the optimization of the classifiers

Let apart linear regression, two classifiers have been mainly used through this work. A Multi-Linear Perceptron and a Support Vector machine for Regression.

A. Configuration of the MLP

The implemented MLP features only one sigmoidal hidden layer. Therefore the output is $Y = W_2 \times \sigma(W_1 \cdot X)$ where X is the input, Y is the output, W_1, W_2 are matrices that are optimized by the MLP.$\sigma(T)$ where T is a vector applies the sigmoid opera to each component of T. Notice that a constant dimension is added to the input in order to feature a bias. The input is also normalized (removing the mean and dividing by the standard deviation). In order not to overfit the MLP with the training set, a validation set is considered, taken randomly and which size is 30% of the train set. The optimization of the

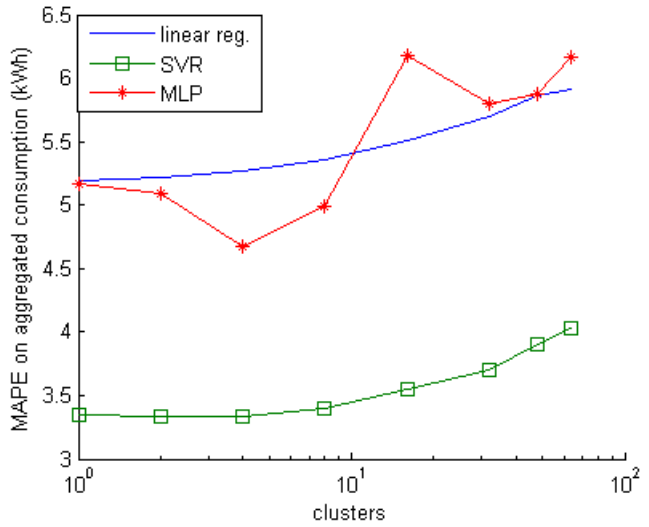


Figure 12. Evolution of MAPE on aggregated consumption, for different number of clusters, with different classifiers

C \ γ	0	0.01	0.1	1
1.E+00	0.69	0.62	0.59	0.57
1.E+01	0.62	0.59	0.58	0.57
1.E+02	0.59	**0.58**	0.57	0.58
1.E+03	0.58	0.58	0.57	0.63
1.E+04	0.58	0.57	0.58	0.82
1.E+05	0.57	0.57		

(a)

C \ γ	0	0.01	0.1	1
1.E+00	0.02	0.03	0.02	0.02
1.E+01	0.03	0.02	0.02	0.20
1.E+02	0.02	**0.02**	0.11	0.02
1.E+03	0.02	0.02	0.02	0.02
1.E+04	0.02	0.02	0.02	0.04
1.E+05	0.02	0.11		

(b)

C \ γ	0	0.01	0.1	1
1.E+00	2.1	2.7	3.2	4.6
1.E+01	2.3	2.6	3.1	3.2
1.E+02	2.3	**2.4**	4.0	8.1
1.E+03	2.7	3.0	8.9	45
1.E+04	4.0	8.2	50	222
1.E+05	7.2	52		

(c)

Figure 13. (a) Average NRMSE on the validation set (evaluated over 25 houses) depending on the values of C and gamma. (b) Standard deviation on the average. (c) time of computing on 25 houses (in mn). Due to the small differences of performance between each set of γ and C, the couple ($C = 100$ and $\gamma = 0.01$) which realizes top performances in minimal time, has been chosen.

MLP (gradient descent, without momentum and with learning rate =0.3) is stopped whenever the error on the validation set, calculated each epochs, has increased 10 times in a row.

B. Configuration of SVR

SVR is a regression method based on Support Vector Machines that have been invented in 1996 by Vapnik and is fully described in Smola and Schölkopf's tutorial[10]. During this work, we have used the implementation of SVR by the library LIBSVM [3], developed by Chang and Lin.

When implementing a SVR regression, there are parameters to choose. Among them the type of kernel, and the cost of the SVM error, usually noted C. In all our experiment, we used a RBF kernel : $k(x, y) = e^{-\gamma\|x-y\|^2}$, which introduces a parameter γ. This Section deals about the optimization of the used SVR classifier in terms of C and γ.

1) Case of the prediction for each house independently: In order to find the suitable C and γ, the following methodology has been applied. We considered 25 houses taken randomly. For each of those houses, we split the training set in two parts: a sub-training set and a validation set. For each houses, we trained multiple SVR classifier with different values of C and γ on the sub-training set and evaluated the NRMSE on the validation set. Results are presented on Figure 13. It is noticeable that the set (C, γ) does not strongly influence the performance in terms of NRMSE. However, they strongly change the computation time, which explodes then $C > 1000$ and $\gamma > 0.1$. In whole Section IV, the couple ($C = 100$ and $\gamma = 0.01$) has been used.

2) Case of the prediction on aggregated consumption: In Section VII, we establish prediction on the aggregated consumption. Once again, our optimization has been based on the separation between a sub-training set and a validation set. The set of features of Section VII-A has been considered. Results can be seen Figure 14. On the contrary to the previous Section, this time changing the set (C, γ) has a great importance, and the error can be divided by 10 with a good choice. We found that the couple $(C = 1000, \gamma = 1)$ outperformed all the other tested, and we use those values in Section VII.

C \ γ	0.001	0.01	0.1	1	10
1.E-01	0.418	0.415	0.395	0.356	0.318
1.E+00	0.415	0.392	0.226	0.160	0.201
1.E+01	0.391	0.200	0.074	0.070	0.082
1.E+02	0.197	0.073	0.064	0.052	0.082
1.E+03	0.073	0.066	0.059	**0.045**	0.065
1.E+04	0.066	0.065	0.054	**0.045**	0.072
1.E+05	0.066	0.061	0.050	0.050	
1.E+06	0.065	0.057	0.047	0.063	
1.E+07	0.061	0.054	0.046	0.087	
1.E+08	0.063	0.110	0.188	0.149	

Figure 14. NRMSE on the validation set in the case of prediction

IX. Conclusion

The most impressive recent results in electric load forecasting come from the use of SVR for regression. Those results are obtained when dealing with the prediction of the consumption on a large scale, like a region or a country. We have verified that applied on a small district (782 houses), SVR seems effectively the best method that we tested to establish the overall consumption. However, in this report, we mainly challenged a different problem, which is a prediction of the consumption at the scale of a house, one hour and 24 hours ahead. Our work has concluded to the following points :

- Concerning the prediction at the level of one house, the method that produced the lowest error was the simple

linear regression. Both MLP and SVR have revealed to be deceptive at this scale.

- We have shown that SVR starts to bring more precision than Linear Regression when considering the aggregated consumption of more than 16 houses of the Irish Smart Meter Dataset.
- About the idea of taking into account of relations of leadership between houses in order to help the short term prediction, our experiments have not concluded to any relevant results that can valid the hypothesis. The accuracy of the forecast, using and without using the leaders information were the same.
- We presented a way to discretize data that is more efficient than the classic uniforms and quantile methods. However, we showed that discretization has lowered the efficiency of the predictor we put in place. In the case of electric load prediction, regression methods seem the most accurate.
- Concerning the problem of predict the overall consumption, it seems that for MLP and (for a little degree) SVR predictions can be improved by regrouping houses that have similar consumption into clusters can lead to an improvement of the results. However, aggregate the prediction made on each single house independently leads to a poor forecast on the overall consumption.

Concerning this last point further work can consist into selecting the best classifier for each cluster, in order to improve the global quality of the overall prediction.

X. Acknowledgments

I would like to thank Tri Kurniawan Wijaya, my supervisor for this project. First for giving me the opportunity to work on it, then for his availability, his patience and his incredibly constant good mood through this semester, as well as his essential help.

I am also very grateful towards Matteo Vasirani for his time and his advices.

References

[1] S.M. Amin and B.F. Wollenberg. Toward a smart grid: power delivery for the 21st century. *Power and Energy Magazine, IEEE*, 3(5), 2005.

[2] Souhaib Ben Taieb and Rob J Hyndman. A gradient boosting approach to the kaggle load forecasting competition. 2013.

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] Bo-Juen Chen, Ming-Wei Chang, and Chih-Jen Lin. Load forecasting using support vector machines: a study on eunite competition 2001. *Power Systems, IEEE Transactions on*, 19(4), 2004.

[5] PJM Load Management Task Force. Pjm empirical analysis of demand response baseline methods. 2011.

[6] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software : An update. *SIGBDD Explorations*, 11(1), 2009.

[7] H.S. Hippert. Neural networks for short-term load forecasting : a review and evaluation. *Power Systems, IEEE Transactions on*, February 2001.

[8] NI Sapankevych and Ravi Sankar. Time series prediction using support vector machines : A survey. *Computational Intelligence Magazine, IEEE*, May 2009.

[9] Mehmet Sayal. Detecting time correlations in time-series data streams. *Hewlett-Packard Company*, 2004.

[10] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. September], 2003.

[11] James W Taylor. Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research*, July 2010.

[12] Liang Tian and Noore Afzel. A novel approach for short-term load forecasting using support vector machines. *International Journal of Neural Systems*, 34(9), 2004.

[13] Bianco Vincenzo and manca Oronzio. Electricity consumption forecasting in italy using linear regression models. *Energy*, 34(9), 2009.

[14] Tri Kurniawan Wijaya, Julien Eberle, and Karl Aberer. Symbolic representation of smart meter data. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops*, EDBT '13, 2013.

[15] Ian H. Witten, Eibe Frank, and A. Matk Hall. *Data Mining Practical Machine Learning Tools and Techniques third edition*. Morgan Kaufman, 2011.

[16] Di Wu, Yiping Ke, Jeffrey Xu Yu, Philip Yu, and Lei Chen. Detecting leaders from correlated time series. *Database Systems for Advanced Applications*, 2010.