Predicting hourly energy consumption. Can regression modeling improve on an autoregressive baseline?

Pierre Dagnely[1], Tom Ruette[1], Tom Tourwé[1], Elena Tsiporkova[1], and Clara Verhelst[2]

[1] Sirris - Software Engineering & ICT Group, Brussels, Belgium
{Pierre.Dagnely,Tom.Ruette,Tom.Tourwe,Elena.Tsiporkova}@sirris.be
[2] 3E - iLab, Brussels, Belgium
Clara.Verhelst@3e.eu

Abstract. According to the Third Industrial Revolution, peer-to-peer electricity exchange combined with optimized local storage is the future of our electricity landscape, creating the so-called "smart grid". Such a grid not only has to rely on predicting electricity production, but also its consumption. A growing body of literature exists on the topic of energy consumption and demand forecasting. Many contributions consist of presenting a methodology, and showing its accuracy. This paper goes beyond this common practice on two levels: first, by comparing two regression techniques to a univariate autoregressive baseline and second, by evaluating the models in term of industrial applicability, in close collaboration with domain experts. It appears that the computationally costly regression models fail to significantly beat the baseline.

Keywords: Energy consumption prediction · Ordinary least squares regression · Support vector regression · Autoregressive models

1 Introduction

According to the Third Industrial Revolution defined by Rifkin [26], peer-to-peer electricity exchange combined with optimized local storage is the future of our electricity landscape. Therefore, a "smart grid" not only has to rely on predicting electricity production, but also consumption. Moreover, the prediction of hourly electricity consumption is of great economical interest for players on the global electricity market, since an accurate prediction of the consumption is needed to obtain the best prices on the day-ahead market and to avoid purchasing on the more expensive real-time spot market. This study[3] focuses on the prediction of the (hourly) electricity consumption of a medium-scale office building. Through its Software-as-a-Service platform *SynaptiQ*, 3E offers an interface to monitor (green and grey) electricity production and consumption, and is now additionally interested in predicting electricity consumption.

Energy consumption prediction is a typical forecasting challenge, where a univariate continuous variable (hourly total energy consumption in a building) needs to be predicted by means of a model that may contain multiple exogeneous variables, i.e. independent variables that affect a model without being affected by it, and indogeneous

[3] This study takes place in the context of the Artemis project Arrowhead (http://www.arrowhead.eu), in which 3E and Sirris collaborate.

variables, i.e. variables that affect a model but are also affected by it. For the current paper, regression models are investigated as candidate methods for the forecasting challenge. Concretely, the research question is formulated as "which regression model is most successful in predicting as precisely as possible the total energy consumption in a building". Thus, two regression techniques, Ordinary Least Squares and Support Vector Regression, are compared against a naive autoregressive baseline. It appears that it is (for the given realistic dataset) impossible to significantly improve on that autoregressive baseline with the two established regression models, despite their increased computational cost.

The paper first offers a literature review of energy consumption forecasting work, and some references for a better understanding of regression models (Section 2). Then, the data and the considered predictors are discussed in detail (Section 3). In the following section, the baseline against which the results of the regression models will be compared and the two regression models (Section 4) are presented. Then, an evaluation of these models against a held-out test dataset and an investigation of the improvement over the baseline is presented (Section 5). The penultimate section discusses these results (Section 6). Finally, Section 7 concludes the paper and offers a list of topics for further research.

2 Literature review

Predicting the electricity consumption of buildings has been a topic in the literature for at least fifty years [31], during which mainly three approaches have been explored: the engineering approach, the statistical approach, and the machine learning approach.

Engineering methods use physical models that are based on building characteristics, such as the wall materials or the HVAC[4] characteristics of each room, and external parameters, such as the weather, to predict electricity consumption. These methods range from simple manual estimations to detailed computational simulations [4].

An example of a simple engineering method is the bin or temperature frequency method. This method starts by defining outdoor temperature bins of usually 2 or 3 $^{\circ}$C. For each bin, the electricity consumption of the building is estimated, assuming that the consumption depends on the outdoor temperature. By multiplying the estimated electricity consumption for each temperature bin with the amount of times that this frequency bin occurs, the total electricity consumption is estimated. To have a more accurate prediction, there can be multiple electricity consumption estimates for a temperature bin, e.g. an estimate for electricity consumption during business hours and an estimate outside of the business hours [4].

On the other side of the spectrum, a detailed computational simulation would estimate the electricity consumption based on a finely grained picture of all the building components and characteristics by means of an existing tool. A list of such tools is maintained by the US department of energy [1]. These tools usually follow four steps [4]:

- Estimate the cooling and heating consumption for each zone of the building

[4] HVAC stands for Heating, Ventilation and Air-Conditioning

- Estimate the required energy flows of the various equipments to obtain this cooling and heating consumption
- Determine the amount of electricity that is consumed to generate these energy flows
- Estimate the cost of this electricity consumption (optional)

Two strategies exists to tackle these steps:

- Sequential strategy: all steps are handled one after the other, using the result of the previous step as input.
- Integrated strategy: all steps are performed together, with feedback loops between them. This strategy is more complex but takes into account the interactions between the steps.

An example of a computational simulation tool is *EnergyPlus*, an integrated simulation system sponsored by the US Department of Energy and first released in 2001. EnergyPlus is composed of three basic components [10]. The first component is the simulation manager, which manages all the processes. The second component is the heat and mass balance simulation module, which computes the building thermal zones and their interactions. This module can use parameters such as room surface, heat conduction of walls (based on their sizes and materials) or daylight illumination. The third component is the building system simulation module, which simulates the HVAC and electrical systems, equipments and components to update the zone-air condition and estimate the electricity consumption.

Statistical approaches are based on methods coming from the classical statistician toolbox. Typically, historical data that represents the past behaviour of a building, e.g. past consumption, weather or sensor data such as occupancy, is used to fit a model, e.g a regression model. However, predicting electricity consumption can be a time series problem, where data points at time t depend on data points at a time $t - i$. Therefore, methods such as Auto Regressive Integrated Moving Average (ARIMA) are widely used, as to account for the interdependency of data points. These ARIMA methods consist basically of two models. First, an autoregressive model assumes that a value at time t is linearly depending on values at times $t - i$ and a noise term. Second, a moving average model assumes a stationary mean, and future values in the time series are functions of this mean, altered by weighted noise terms. In [24], Newsham et al. evaluated an ARIMAX model, i.e. an ARIMA model with eXogenous inputs that influence the noise terms, for predicting the electricity consumption of a three-story building of 5800 m^2, comprising laboratories and 81 individual work space. Due to the presence of laboratories, this building does not represent a conventional office building. As predictors, they used past consumption data, weather data and occupancy data (through sensor monitoring logins but not logoffs). Adding the login predictor improved slightly the accuracy of the model. Nevertheless, they conclude that measuring logoffs or better sensors, such as a camera, may have a positive impact on the accuracy in a more conventional office building.

Regular regression analyses have also shown good accuracy, despite the potential violation of the assumption of independence among data points. In [6], Ansari et al. computed the cooling consumption of a building by means of a linear regression function with the temperature difference between inside and outside as a predictor. Cho

et al. studied the impact of the training dataset size when applying linear regression to predict the energy consumption of a building in [9]. They used the average outdoor temperature as a predictor for the energy consumption of the heating system. With one day of hourly measurement data, they obtained a very inaccurate system with an error range from 8 to 117%. A training dataset of one week of daily data gave an error range of 3 to 32%, while for three weeks of training data the errors ranged from 9 to 26%. They concluded that in a training set shorter than one month the outdoor temperature variability is a more important cause of error than the length of the training set. In contrast, with a training set of more than one month of daily data, the length of the measurement period strongly influences the accuracy. As an example, with three months of daily data, their model was able to to make electricity consumption predictions with errors ranging from only 1 to 6%.

Machine learning approaches are, like statistical approaches, data-driven, but they use techniques coming from the field of Artificial Intelligence. Two of the most used methods in energy consumption prediction are Artificial Neural Networks [28] and Support Vector Machines [11]. Artificial Neural Networks create a network of input nodes, in-between nodes, and output nodes, all connected by weighted links. The output nodes are thus a function of the input nodes, but their relationship can be obfuscated by hidden in-between nodes, and the relationships created by the links. Support Vector Machines are typically used for classification problems. To resolve non-linear classification problems, Support Vector Machines transpose the data points to a higher dimensional space by means of a kernel. In this higher dimensional space, the data points may have a linear separation. The linear separation is then found by fitting a hyperplane between the classes. The hyperplane has to maximize the margins, i.e. the largest distance to the nearest training data point from the classes. Support Vector Regression [28] is an extension of a Support Vector Machine adapted to regression problems. Here, the hyperplane has to minimize the error cost of the data points outside of the margin ε, while the error cost of the data points inside the margin are considered as null. It prioritize the minimisation of the upper bound of the training error rather than the (global) training error.

In [13], Ekonomou used Artificial Neural Network to predict the (yearly) Greek long-term energy consumption (2005-2015) using the past thirteen years as training data (from 1992 to 2004). The inputs for fitting the model were the yearly ambient temperature, the installed power capacity, the yearly per resident electricity consumption and the gross domestic production. He tested various models, using the combination of 5 back-propagation learning algorithms, 5 transfer functions and 1 to 5 hidden layers with 2 to 100 nodes in each hidden layer. The best model had a compact structure and a fast training procedure with 2 hidden layers of 20 and 17 nodes, using the Levenberg-Marquardt back-propagation learning algorithm and logarithmic sigmoid transfer function. Gonzalez et al. [14] used an Auto-associative feed-forward Neural Network to predict the electricity consumption of two buildings, with input variables such as the temperature difference $\triangle T = T_{k+1} - T_k$, the hour of the day, the day of the week and the previous consumption. Gonzalez et al. showed that the previous consumption reflects other parameters such as the occupancy level: their models performed rela-

tively accurately during recurring holidays, although this information was not directly encoded as predictor.

For the past ten years, Support Vector Machines are on the rise and many studies have been performed. One of the earliest applications of a Support Vector Machine to the (monthly) forecast of building energy consumption is [12], where Dong et al. successfully applied it to four commercial buildings with a better accuracy than other machine learning methods. They emphasized on the fact that Support Vector Machines only require the tuning of a few parameters: ε, C and the kernel parameters (such as γ for a Radial Basis Function kernel). In a more recent study, Jain et al. [17] examined the impact of various spatial and temporal variables to predict the energy consumption of a multi-family residential building. They applied a Support Vector Machine with the following predictors: the past consumption, temperature, solar flux, weekend/holidays and hour of the day. They tested the impact of spatial (i.e. by apartment, by floor or for the whole building) and temporal (i.e. every 10 minutes, hourly or daily) granularity of the data. The model that was based on data per floor and per hour was found to be significantly more accurate than the other possible combinations.

Over the years, some studies that compare these three types of approaches have been performed. In [23], Neto et al. compared the aforementioned *EnergyPlus* tool to a machine learning approach based on Feed-forward neural networks. The result of this comparison revealed that both approaches are similar in term of prediction accuracy. However, the Feed-forward neural network approach turned out to be more straightforward, as it only relied on 17 months of past consumption and weather data. In contrast, the engineering model was cumbersome to construct, as it depended on the availability of domain experts and a precise knowledge of all the building characteristics.

Azadeh et al. [7] compared Artificial Neural network models using a supervised multi layer perceptron network to a conventional regression technique for predicting the monthly electricity consumption in Iran. The more accurate Artificial Neural Network model outperformed the conventional regression significantly. For the specific case of ARIMA models, studies shown that they are usually easy to estimate, but lack in accuracy in comparison with machine learning methods [31]. However, Amjady [5] tested a modified ARIMA taking into account the domain knowledge of experts (e.g. to define manually a starting point for the parameter tuning). Artificial Neural Network and ARIMA where similar in term of accuracy but the modified ARIMA succeeded to outperform slightly both models.

From a mathematical perspective, and in terms of optimization, Support Vector Machines present some advantage in comparison with Artificial Neural Network [12]. Support Vector Machine methods always lead to a unique and globally optimal solution, whereas (back propagation) Artificial Neural Network methods may lead to a locally optimal solution. From an application perspective, both Artificial Neural Networks and Support Vector Machines produce accurate results, but need a sufficient amount of representative historical data to train the model. Usually, however, Support Vector Machine models are slightly more accurate, as shown in [2].

Many other machine learning approaches have been compared. Tso et al. [29] applied decision trees, which have the advantage of being easily interpretable (in contrast to Artificial Neural Networks and Support Vector Machines). Huo et al. [16] experi-

mented with genetic programming, which mimic the natural evolution to make evolutionary programs by recombination, mutation and selection to find good solutions. Yang et al. [30] studied the use of fuzzy rules, through a combination of Wang-Mendel method and a modified Particle Swarm Optimization algorithm.

However, a promising approach could be the use of a hybrid method, i.e. a combination of different methods. Hybrid methods may be able to go beyond the accuracy of a Support Vector Machine or Artificial Neural Network, as shown in [18]. In that study, Kaytez et al. combined Support Vector Machine with Least Squares as a loss function to construct the optimization problem. Nie et al. [25] investigated the combination of ARIMA, to predict the linear part of the consumption and Support Vector Machine to predict its non-linear part. Li et al. [20] tested a General Regression Neural Network, i.e. an Artificial Neural Network where the hidden layer consists of Radial Basis Functions.

From this extensive body of literature, three aspects are missing, which we try to address in the current paper. First, there has been, up to now, no comparison of ordinary least squares regression versus support vector regression. Second, none of the presented studies compare the more advanced techniques to the simplest possible baseline. Third, all methods have only been evaluated in terms of accuracy, but not in terms of industrial applicability, i.e. complexity to deploy and maintenance of the system. This last evaluation step has been performed in close collaboration with the domain experts of 3E.

3 Response and predictor variables

For this study, a 72 hours-ahead prediction of the hourly electricity consumption of the 3E office in Brussels was requested by 3E. They provided five years of past electricity consumption data (from 2010 to 2014). In addition, this study was performed in close collaboration with experts at 3E, who are not only experienced with energy consumption metrics, but also know the context and activity of the 3E office. This proved to be very useful to identify certain predictive patterns in the data that emerged during explorative visualizations.

This chapter starts by an explorative analysis of the received data. Then the response variable, i.e. the variable to predict, is explained more deeply. Finally, the predictors tested during this study are described in detail, with a focus on how they have been implemented.

The following building specific patterns have been found:

- A drop in energy consumption was visible between 13:00 and 14:00, which, after consultation with a 3E employee, appeared to align with the typical lunch break moment in the building. Since the effect was relatively small, the lunch break was not modelled as a predictive variable.
- Between 2010 and 2013, energy consumption on Saturdays was higher than on Sundays, and in 2014 this pattern was not visible. It appeared that between 2010 and 2013, the cleaning company worked in the building on Saturdays, and in 2014 this policy changed. Because of this very clear policy change, 2014 has been left out from this analysis.

– In 2013, a new electrical heating system was installed, which increased the electricity consumption to a certain extent. Since this effect was relatively small, 2013 has not been discarded from the analysis.

3.1 Response variable

The response variable will be the hourly energy consumption. The histogram in Figure 1 shows the distribution of the hourly energy consumption measurements at a granularity of 100W, which is the measurement granularity.

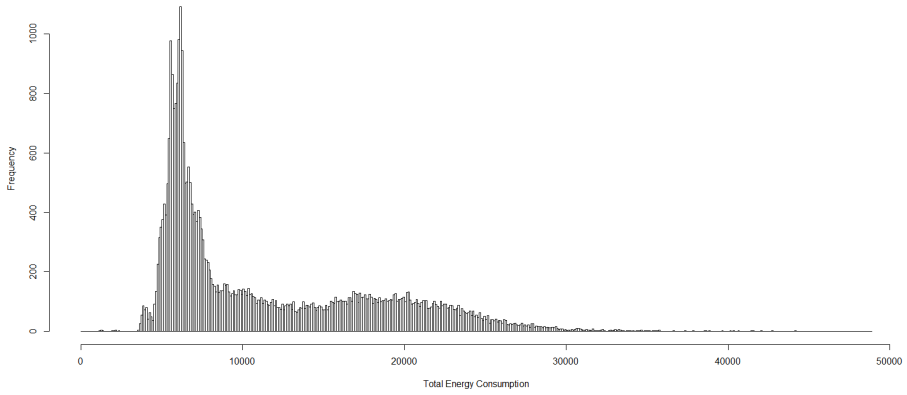

Fig. 1. Histogram of total energy consumption at 100W granularity.

3.2 Predictors

Some potentially relevant predictors have been selected, based on the analysis of the data and a literature review on the predictors. In the following subsections, these predictors and how they have been incorporated in the benchmark data set, during the preprocessing step, is discussed.

Recency An autocorrelation analysis of the electricity consumption shows a clear weekly pattern. In addition, an autocorrelation with the previous day and all the $t-(\alpha \times 7)$ previous days with a decreasing impact over times has also been observed. Therefore, three autoregressive attributes have been created: RECENCY1 that merely contains the logged energy consumption value of the previous day, RECENCY7 that contains the logged energy consumption value of exactly one week ago and RECENCY14 that contains the value of exactly two weeks ago. These recency attributes also correspond to the autoregressive baselines used to evaluate the added value of the more advanced regression models.

Temporal predictors As defined in [19], the first temporal attribute considered is the typical OCCUPANCY of the building. For the 3E office building, standard business hours range from 08:00 until 19:00. It is obvious that in this period, the total energy consumption will be higher than outside of this period. In addition, during weekend days, the facilities are typically not used. This information has been encoded as a binary attribute. In principle, the OCCUPANCY attribute may also contain planned holidays, or even foreseeable late night events, or general off-site meetings. However, this additional information has not been taken into account since it was not readily available.

The second temporal attribute is the day of the week. Visual inspection has revealed that there are recurring patterns of reduced/increased energy consumption that relate to the specific day of the week. This can be explained by habits of employees, who have typical days for teleworking. This information is encoded as the categorical WEEKDAY variable that contains simply the name of the day of the week. The encoding of this information is based on the visual inspection of data, but can also be based on the outcome of a model that can be inspected, such as a decision tree.

Meteorological predictors As shown in [8], two meteorological attributes can be sufficient to integrate the weather into a predictive model. The first meteorological attribute is (ambient) TEMPERATURE. The outside temperature has an influence on total energy consumption, since on cold days, additional electrical heating may be used. On warm days, the air conditioning might be responsible for increased energy consumption. 3E provided the hourly measured ambient temperature in Brussels (Uccle), where their office building is located. The TEMPERATURE has been encoded as a continuous variable.

The second meteorological attribute is IRRADIANCE. The irradiance, which is the amount of sunlight that reaches the earth, may influence total energy consumption. A lower irradiance may indicate darkness or cloudiness, which increases the need for artificial lighting. Also, there is an influence on the heating. 3E provided the hourly measured irradiation in Brussels (Uccle). The IRRADIANCE has also been encoded as a continuous variable. As expected, there is a strong correlation to the temperature predictor.

4 Modeling

As mentioned in the introduction to this paper, two regression approaches (Ordinary Least Squares and Support Vector Regression) have been compared to a naive autoregressive baseline. It is expected that the Ordinary Least Squares regression and Support Vector Regression models will predict the hourly electricity consumption more accurately than an autoregressive baseline, because they take the temporal and meteorological factors into account, in addition to the recency attributes.

The data set has been divided between a test dataset and a training dataset. The training dataset covers the two years 2011 and 2012. The year 2010 has been discarded, because there was no meteorological data available. The test dataset, which is never used during the training phase, consists of data from the year 2013.

4.1 Autoregressive baseline

To evaluate the quality of the regression models, three autoregressive models have been evaluated and the most accurate one has been used as a baseline. These three models predict the energy consumption at time t by taking the energy consumption at respectively times $t-1$ day, $t-7$ days and $t-14$ days, because these are the three strongest autocorrelation lags. Note that these components have also been used as the recency attributes in the regression models.

4.2 Regression

Based on a review of the literature, two regression methods have been selected. The first regression method considered is Ordinary Least Squares regression [15], a statistical method. The optimal linear combination of predictors is found by fitting a hyperplane between the data points. This hyperplane has to minimize the sum of the squares of the distances from the points to this hyperplane. This regression method is straightforward to implement, free of parameters and can be run without making any configuration decisions.

The second regression method considered is Support Vector Regression [28], a machine learning method. Unfortunately, an exact method to obtain the optimal parameters of a Support Vector Machine does not exist. Therefore, a search algorithm must be applied. Three types of search algorithm approaches exist:

1. Grid search, where a set of possible parameter values is tested, i.e. for each parameter, a range of possible values is assessed and all combinations are tested. Such an approach is used by Akay in [3]. This approach produces good result, but it usually is a time consuming method.
2. Local search type methods, such as the pattern search applied by Momma et al. [22], where locally optimal Support Vector Machine models are created and then bagged or averaged to produce the final model.
3. A more recent approach uses machine learning methods to estimate the parameters. For example, Salcedo-Sanz et al. [27] compared Evolutionary Programming and Particle Swarm Optimization to find the parameters of a Support Vector Regression model for a wind speed forecasting problem. Both methods had very good performance, but such methods are more complex to deploy than a simple grid search.

For this study, the grid search approach has been used. Since the Radial Basis Function kernel has been chosen — it was successfully applied by Dong et al. [12] — three parameters have to be tuned (on the training set): C and ε, which are Support Vector Machine related parameters, and γ, which is a kernel related parameter. For the model using all predictors, the grid search suggested as parameters: kernel = Radial Basis Function (RBF), $C = 100000$, $\gamma = 0.03$ and $\varepsilon = 0.00005$.

For both methods, 7 different combinations of predictors have been generated to find out which interactions yield the most accurate results. The first four models only take a single predictor, respectively OCCUPANCY, TEMPERATURE, IRRADIANCE and

RECENCY7. The fifth model considers the two temporal attributes and the two meteorological variables in interaction (OCCUPANCY, WEEKDAY, TEMPERATURE and IRRADIANCE). The sixth model uses the three recency attributes in interaction (RECENCY 1, RECENCY 7, RECENCY 14). The seventh model considers all attributes (RECENCY 1, RECENCY 7, RECENCY 14, OCCUPANCY, WEEKDAY, TEMPERATURE and IRRADIANCE) in interaction.

5 Evaluation

Figure 2 illustrates nicely how well the predicted values estimate the observed values, and how close to one another the predictions of the baseline, Ordinary Least Squares regression and Support Vector Regression are. The figure depicts a seven days-ahead forecast based on one year of training data and using all the predictors aforementioned for the Ordinary Least Squares regression and Support Vector Regression models.

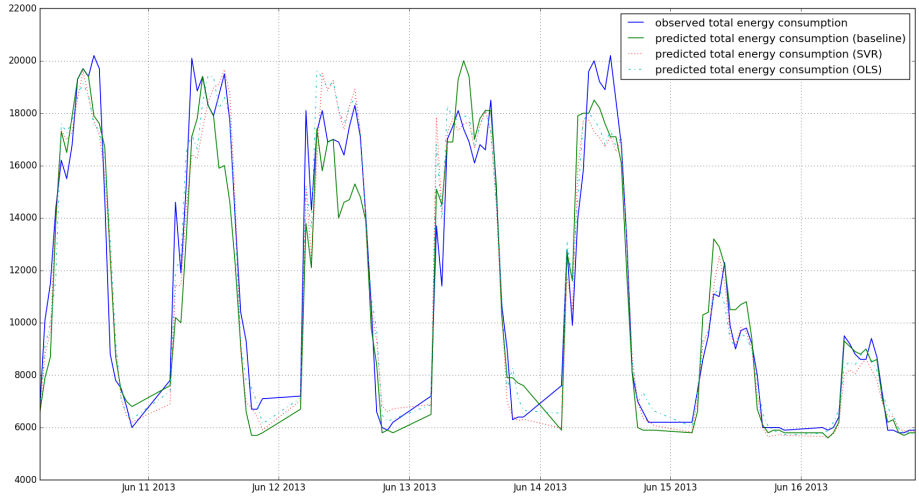

Fig. 2. Showcase of prediction performance.

The predictive power of the models have been evaluated by letting them forecast the hourly (total) energy consumption for the test dataset. For each day of the test dataset a forecast of the three next days have been made (72-hours ahead prediction is requested by 3E) and the prediction accuracy of this forecast has been calculated by using the Mean Absolute Error (MAE). Note that these predictions rely on actual meteorological data, instead of relying on predicted temperature and irradiance values, so that the results tend to be over-optimistic. Then, the mean and standard deviation of all these 72-hour ahead prediction accuracies across the whole year have been computed. The MAEs of the 7 models from both the Ordinary Least Squares regression and the Support Vector Regression are presented in Table 1. The same scores have been computed for

the three autoregressive models. The model using the energy consumption at $t-7$ days appeared as the more accurate and has been chosen as a baseline.

Table 1. Mean absolute errors (MAE) of the seven models for each regression method (with their standard deviation) and the baseline

Model	predictor(s)	MAE scores
OLS	occupancy	3710 ($\pm$ 1195)
	temperature	5538 ($\pm$ 1290)
	irradiance	4967 ($\pm$ 1227)
	recency 7 days	2227 ($\pm$ 817)
	occupancy * weekday * temperature * irradiance	3343 ($\pm$ 1025)
	recency 1 days * recency 7 days * recency 14 days	1971 ($\pm$ 699)
	recency 1 days * recency 7 days * recency 14 days * occupancy * weekday * temperature * irradiance	1914 ($\pm$ 757)
SVR	occupancy	3657 ($\pm$ 1251)
	temperature	5424 ($\pm$ 1402)
	irradiance	4630 ($\pm$ 1561)
	recency 7 days	2125 ($\pm$ 837)
	occupancy * weekday * temperature * irradiance	3219 ($\pm$ 1071)
	recency 1 days * recency 7 days * recency 14 days	1914 ($\pm$ 691)
	recency 1 days * recency 7 days * recency 14 days * occupancy * weekday * temperature * irradiance	1719 ($\pm$ 596)
Baseline	recency 1 days	3242 ($\pm$ 1287)
	recency 7 days	2189 ($\pm$ 871)
	recency 14 days	2391 ($\pm$ 1016)

The MAEs only provide a global view on the errors in the prediction. To analyze the errors, three checks have been undertaken (1) a check of the distribution of the errors, (2) an analysis of errors to identify common patterns that can be addressed in the model, (3) a comparison of the MAE distribution for the main models.

First, the distribution of the (absolute) errors of the prediction have been inspected – based on the Support Vector Regression model with all predictors (the errors of the Ordinary Least Squares regression prediction are similarly distributed) – by means of a histogram, represented in Figure 3. Taking into account that the errors are similarly distributed in the baseline, the Ordinary Least Squares regression model and the Support Vector Regression model, the MAE is a decent metric for comparison.

Second, the significant deviations between observed and predicted values have been explored to identify recurring patterns, by computing the MAE scores of the main models for various specific time periods, as shown in Table 2. For both regression flavors, it can be observed that the prediction errors appear (to be expected) mainly during the working hours, and not during the night. A similar distinction can be observed between the workweek and the weekend in the errors. However, as the electricity consumption of the night and the weekend is significantly lower, the percentage of the error is higher for these periods, e.g. a Support Vector Regression model using all predictors has an error of 16.6% for the weekend and 10.4% for the workweek. Further inspection points

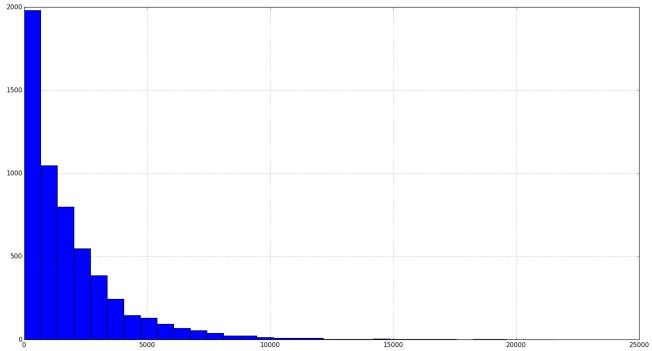

Fig. 3. Histogram of prediction errors of SVR.

to a slight increase in errors during the spring and the winter periods for both regression models whereas the prediction error of the autoregressive baseline is similar for all seasons. This difference of accuracy in spring can be attributed to increased energy consumption during more dynamic, extreme warm or cold, weather. For winter, 5 weeks of data are missing, which could impact the accuracy of the models and lead to this result.

Third, a comparison of the distribution of MAE values for the main models have been conducted, as shown in Table 4. It can be observed that models only based on meteorological and temporal predictors are less accurate than the others. There is no significant difference (in the sense that confidence intervals overlap) between the Ordinary Least Squares regression, Support Vector Regression models and the baseline.

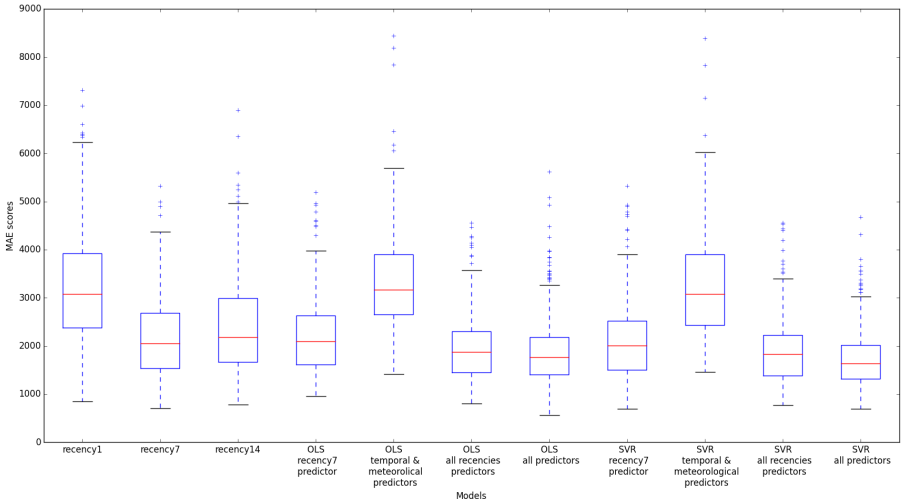

Fig. 4. Distribution of MAE values for the main models

Table 2. Mean absolute errors (MAE) of the main models and the baseline for forecast of specific periods in time

	OLS all recencies predictors	OLS all predictors	SVR all recencies predictors	SVR all predictors	Recency 7 baseline	Mean of the electricity consumption
Workweek	2145 (± 951)	2112 (± 1028)	2090 (± 960)	1884 (± 877)	2415 (± 1165)	18.031 (± 5403)
Weekend	1578 (± 765)	1480 (± 693)	1488 (± 754)	1390 (± 781)	1680 (± 1095)	8343 (± 3225)
Working hours	2476 (± 1362)	2354 (± 1444)	2407 (± 1363)	2140 (± 1243)	2799 (± 1672)	14.938 (± 6504)
Night and weekend	1570 (± 753)	1459 (± 612)	1535 (± 749)	1410 (± 678)	1616 (± 903)	8094 (± 3295)
Spring	2095 (± 726)	2122 (± 959)	2092 (± 813)	1911 (± 786)	2258 (± 1036)	12.618 (± 5877)
Summer	1888 (± 794)	1739 (± 656)	1779 (± 680)	1572 (± 536)	2190 (± 958)	13.478 (± 7310)
Autumn	1829 (± 627)	1645 (± 474)	1758 (± 573)	1590 (± 475)	2071 (± 713)	12.634 (± 5992)
Winter	2030 (± 506)	2147 (± 610)	1998 (± 514)	1885 (± 519)	2166 (± 526)	13.766 (± 6732)

For the calculation of the models, the following Python libraries have been used: Statsmodels for the Ordinary Least Squares regression, SciKit-Learn for the Support Vector Regression and Pandas for the data manipulation.

6 Discussion

The evaluation of the models (see above) points out that the following two aspects have an impact on the prediction results: (1) the selection of the predictors for regression models, (2) the selection of a modeling method.

For the analysis of the impact of the predictors, it appears that:

- The recency attributes already convey most of the information needed to make an accurate prediction. A Support Vector Regression model only based on the three recency attributes achieves a MAE score of 1914 whereas a model using in addition the temporal and meteorological attributes only reduces the MAE score with 195, to reach 1719. These reduction are probably not statistically significant as their confidence intervals overlap. This observation also stands for Ordinary Least Square.
- The explorative part of the study has shown that the two meteorological attributes are important long-term predictors, but have less impact on the short-term, e.g. hourly prediction and they relate well to the meteorological seasons. However, for short-term prediction, their impact is negligible. We assume that this is caused by the

thermal inertia of the building, which has a smoothing effect on the electricity consumption.
- Both Ordinary Least Squares regression and Support Vector Regression rely on meteorological predictors. For this analysis, the actual values of temperature and irradiance have been used. However, in a real application, only predictions of these values will be available. This might decrease slightly the efficiency of both regression methods (but has no influence on the baseline).

Based on the above findings, we would like to argue that a regression model can be successfully applied, even if the previous consumption is the only available data. This result is particularly interesting for companies working at the grid level as the recency data is the only one that they directly have access to. They could therefore do without purchasing expensive weather forecasts or to study the occupancy of the buildings monitored.

For the comparison of regression models with an autoregressive method, it appears that:

- The naive baseline already produces good results, with a mean absolute error of 2189. Given a mean hourly energy consumption of 13099W at the 3E building, this is an error of 16.7%.
- Support Vector Regression and Ordinary Least Square Regression manage to improve slightly on this baseline. They respectively reach a mean absolute error of 1719 (13,1%) and 1914 (14,6%). However, the confidence intervals for these mean absolute error scores overlap, which suggests that improvement will not be statistically significant.
- Both Ordinary Least Squares regression and Support Vector Regression are slow in constructing the model on the basis of the training data. The baseline does not need this intensive computation step, and will return predictions much faster.

These points are also particularly interesting from an industrial point of view as they imply that a very simple autoregressive method already gives good results. In addition to the obvious simplicity of this method in comparison to more complex regression models, this method has the advantage to be easily deployable and maintainable by people without specific data science expertise, as it only relies on a simple preprocessing of the data.

7 Conclusion

In this study a comparison of Ordinary Least Square and Support Vector Machine models with an autoregressive method have been performed to make a 72 hours-ahead forecast of the electricity consumption of an office building. We conclude that both the autoregressive baseline as the more advanced regression models are able to predict the hourly energy consumption fairly well. However, the advanced regression models do not significantly outperform the autoregressive baseline. Given the computational cost of the advanced regression models, an autoregressive model is the most effective methodology at present to do energy consumption prediction. The temporal and meteorological predictors do not improve the accuracy significantly in our tests.

7.1 Further research

This study is a reflection of our first explorations of the prediction of hourly energy consumption. Next, we want to further explore the application of Support Vector Regression by testing other kernels, by estimating the parameters more dynamically, i.e. recalibrating the model when concept drift occurs, or by using preprocessing methods, e.g. binning (a first investigation did not show much improvement) or wavelets.

In addition, we want to explore more advanced autoregressive models, using a weighted average of past values. The intuition behind the weighted average is that the recency lags do not have the same importance. It can be expected that the electricity consumption of 7 days ago is closer to the future consumption than the one of 14 days ago. The consumption of one day ago has also less importance than might be assumed at first: the presence of daily patterns tends to minimize it. As an example, Friday is usually a day where employees tend to telework, which decreases the electricity consumption. Therefore, basing the forecast on the consumption of Thursday could lead to an overestimation. The problem is more obvious for the prediction of the electricity consumption of Mondays based on the consumption of Sundays. An initial inspection with a weighted average of the three recency attributes has already produced encouraging results, with a MAE score of 1984 ($\pm$ 740). This MAE score is close to the one of the Ordinary Least Square model only using recency attributes as their weights are relatively similar and as the interaction between the predictors do not have a big impact in the Ordinary Least Square model.

We also want to test if other methods could significantly out-perform the autoregressive models. One such method is k Nearest Neighbour, a method as simple as the autoregressive method. To predict the electricity consumption of day_{i+1}, this method takes the $days_{\{i-j,i\}} = H$, find k sequences of j days that resemble H (the so-called k nearest neighbours, i.e. the k sequences of days with the closest electricity consumption behaviour to H that are present in the historical data) and then take, for each sequence, the subsequent day, this day would correspond to a potential prediction for day_{i+1}. The forecast is made with a weighted average of these following day of the k nearest neighbours, using weights based on the similarity of these neighbours with H. This method was successfully applied by Lora et al. [21], who used it to make a 24 hours-ahead forecast of the spanish electricity demand.

Acknowledgements

This work was subsidised by the Region of Bruxelles-Capitale - Innoviris.

References

1. Building Technologies Office: Building Energy Software Tools Directory.
2. A.S. Ahmad, M.Y. Hassan, M.P. Abdullah, H.A. Rahman, F. Hussin, H. Abdullah, and R. Saidur. A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews*, 33:102–109, May 2014.
3. Mehmet Fatih Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36(2):3240–3247, March 2009.

4. Mohammad Saad Al-Homoud. Computer-aided building energy analysis techniques. *Building and Environment*, 36(4):421–433, 2001.
5. Nima Amjady. Short-term hourly load forecasting using time-series modeling with peak load estimation capability. *Power Systems, IEEE Transactions on*, 16(3):498–505, 2001.
6. F. A. Ansari, A. S. Mokhtar, K. A. Abbas, and N. M. Adam. A simple approach for building cooling load estimation. *American Journal of Environmental Sciences*, 1(3):209–212, 2005.
7. Ali Azadeh, S. F. Ghaderi, and S. Sohrabkhani. Forecasting electrical consumption by integration of neural network, time series and ANOVA. *Applied Mathematics and Computation*, 186(2):1753–1761, 2007.
8. M. Beccali, M. Cellura, V. Lo Brano, and A. Marvuglia. Short-term prediction of household electricity consumption: Assessing weather sensitivity in a Mediterranean area. *Renewable and Sustainable Energy Reviews*, 12(8):2040–2065, October 2008.
9. Sung-Hwan Cho, Won-Tae Kim, Choon-Soeb Tae, and M Zaheeruddin. Effect of length of measurement period on accuracy of predicted annual heating energy consumption of buildings. *Energy Conversion and Management*, 45(18-19):2867–2878, November 2004.
10. Drury B. Crawley, Linda K. Lawrie, Curtis O. Pedersen, Frederick C. Winkelmann, Michael J. Witte, Richard K. Strand, Richard J. Liesen, Walter F. Buhl, Yu Joe Huang, Robert H. Henninger, and others. EnergyPlus: New, capable, and linked. *Journal of Architectural and Planning Research*, pages 292–302, 2004.
11. Nello Cristianini and John Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
12. Bing Dong, Cheng Cao, and Siew Eang Lee. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings*, 37(5):545–553, May 2005.
13. L. Ekonomou. Greek long-term energy consumption prediction using artificial neural networks. *Energy*, 35(2):512–517, 2010.
14. Pedro A. González and Jesús M. Zamarreño. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. *Energy and Buildings*, 37(6):595–601, June 2005.
15. James Douglas Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.
16. Limin Huo, Xinqiao Fan, Yunfang Xie, and Jinliang Yin. Short-term load forecasting based on the method of genetic programming. In *Mechatronics and Automation, 2007. ICMA 2007. International Conference on*, pages 839–843. IEEE, 2007.
17. Rishee K. Jain, Kevin M. Smith, Patricia J. Culligan, and John E. Taylor. Forecasting energy consumption of multi-family residential buildings using support vector regression: Investigating the impact of temporal and spatial monitoring granularity on performance accuracy. *Applied Energy*, 123:168–178, June 2014.
18. Fazil Kaytez, M. Cengiz Taplamacioglu, Ertugrul Cam, and Firat Hardalac. Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines. *International Journal of Electrical Power & Energy Systems*, 67:431–438, May 2015.
19. Simon S.K. Kwok and Eric W.M. Lee. A study of the importance of occupancy to building cooling load in prediction by intelligent approach. *Energy Conversion and Management*, 52(7):2555–2564, July 2011.
20. Qiong Li, Peng Ren, and Qinglin Meng. Prediction model of annual energy consumption of residential buildings. In *Advances in Energy Engineering (ICAEE), 2010 International Conference on*, pages 223–226. IEEE, 2010.
21. Alicia Troncoso Lora, Jesús Manuel Riquelme Santos, José Cristóbal Riquelme, Antonio Gómez Expósito, and José Luís Martínez Ramos. Time-series prediction: Application

to the short-term electric energy demand. In *Current Topics in Artificial Intelligence*, pages 577–586. Springer, 2004.
22. Michinari Momma and Kristin P. Bennett. A Pattern Search Method for Model Selection of Support Vector Regression. In *SDM*, pages 261–274. SIAM, 2002.
23. Alberto Hernandez Neto and Flávio Augusto Sanzovo Fiorelli. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. *Energy and Buildings*, 40(12):2169–2176, January 2008.
24. Guy R. Newsham and Benjamin J. Birt. Building-level occupancy data to improve ARIMA-based electricity use forecasts. In *Proceedings of the 2nd ACM workshop on embedded sensing systems for energy-efficiency in building*, pages 13–18. ACM, 2010.
25. Hongzhan Nie, Guohui Liu, Xiaoman Liu, and Yong Wang. Hybrid of ARIMA and SVMs for Short-Term Load Forecasting. *Energy Procedia*, 16:1455–1460, 2012.
26. Jeremy Rifkin. The third industrial revolution: How the internet, green electricity, and 3-d printing are ushering in a sustainable era of distributed capitalism. *World Financial Review*, 1, 2012.
27. Sancho Salcedo-Sanz, Emilio G. Ortiz-Garcı'a, Ángel M. Pérez-Bellido, Antonio Portilla-Figueras, and Luis Prieto. Short term wind speed prediction based on evolutionary support vector regression algorithms. *Expert Systems with Applications*, 38(4):4052–4057, April 2011.
28. Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.
29. Geoffrey K.F. Tso and Kelvin K.W. Yau. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, 32(9):1761–1768, September 2007.
30. Xueming Yang, Jiangye Yuan, Jinsha Yuan, and Huina Mao. An improved WM method based on PSO for electric load forecasting. *Expert Systems with Applications*, 37(12):8036–8041, December 2010.
31. Hai-xiang Zhao and Frédéric Magoulès. A review on the prediction of building energy consumption. *Renewable and Sustainable Energy Reviews*, 16(6):3586–3592, August 2012.