Prediction model of energy use in a house.

J. A. Celis Gil

Abstract

We present and discuss data-driven predictive models for the energy use of appliances. We discuss data filtering to remove non-predictive parameters and feature ranking. A model was done using vector auto-regression with repeated cross validation. We were able to predict until 24 hours ahead with an accuracy of 9%. For longer periods of time the accuracy of the model decreases. The data from the parents' room, teenager room and ironing room were ranked the highest in importance for the energy prediction, probably because of the devices in those places and the time spent there. Finally we remark the importance of monitoring the energy consumed in order to reduce the expenses and improve the efficiency in the energy use.

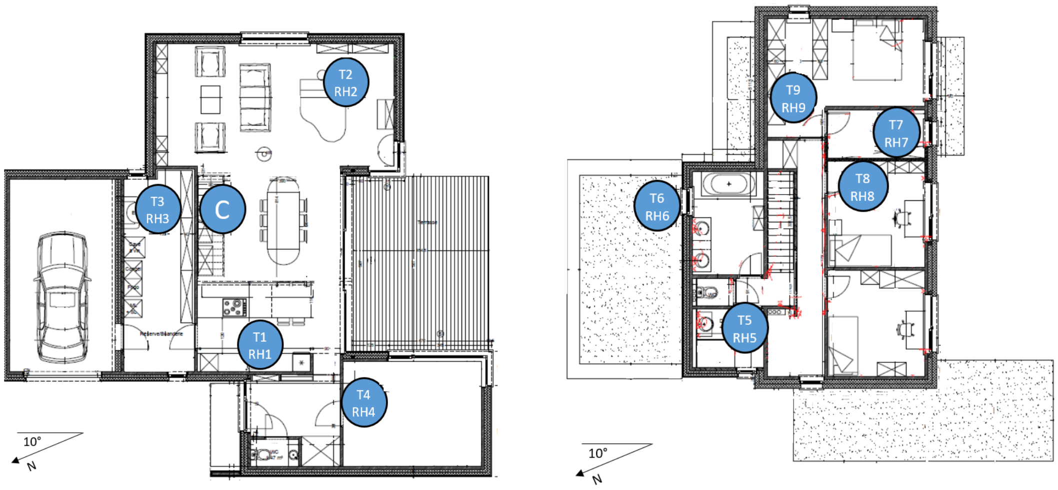
1. Introduction

Nowadays, one of the most common topics in society is the efficiently energy use. But, what is the meaning of this term? Energy efficiency is the goal to reduce the amount of energy required to provide products and services. In other words, energy efficiency is the idea to obtain the same products but producing less impact over the environment.

We can translate this concept to the daily life, more specifically to the house appliances. Using our appliances in a more efficient way will be reflected in shorter energy and gas bills, which means extra money.

Let us see one simple example:

Peter is a normal guy; he lives with his family, a beautiful woman and his two sons (11 and 16 years old) in a normal city. Usually they get up at 7:00 in the morning and prepare themselves for work and school. One of the main goals for Peter is to travel and meet as many countries as he can. For this, every month he saves some money from his salary, but recently he realized that he would need more money if he wants to visit an especial place for his wedding anniversary. He was wondering if it would be possible to reduce the amount of energy that he and his family spend at home without sacrificing their comfort.



First floor

Second floor

Figure 1. House plans. First floor (left) and second floor (right)

During the morning the house is empty and usually they return home around 18:00. They hired a company that helped them monitoring the energy consumption. Peter wants to detect if all the devices are working properly, but he does not know that he can do more than that. With the data he can also predict the energy consumed by the appliances in a normal day and then create a model to reduce his expenses.

The present work mostly deals with the problem of aggregate appliances energy use prediction, which would help to control and predict the energy consumption.

In this work, we include environmental parameters like temperature and humidity. Measurements were done by sensors from a wireless network installed a house located in Stambruges, which is about 24 km from the City of Mons in Belgium. Weather from a nearby airport station and recorded energy use of appliances and lighting fixtures. Measurements were done at 10 min for about 4.5 months. The house temperature and humidity conditions were monitored. Weather from the nearest airport weather station (Chievres Airport, Belgium) is included in the data set [1]. This data is available at this link http://archive.ics.uci.edu/ml/machine-learning-databases/00374/

Using statistical models applied to time series and multivariate time dependent regression models have been tested to predict the energy consumption in the house a few days after the data set record. The purpose of this work is to understand the relationships between appliances energy consumption and different predictors.

2. Data exploratory analysis

Figure 2 shows the comparison between the energy consumed by the appliances and the lights in the complete data set.

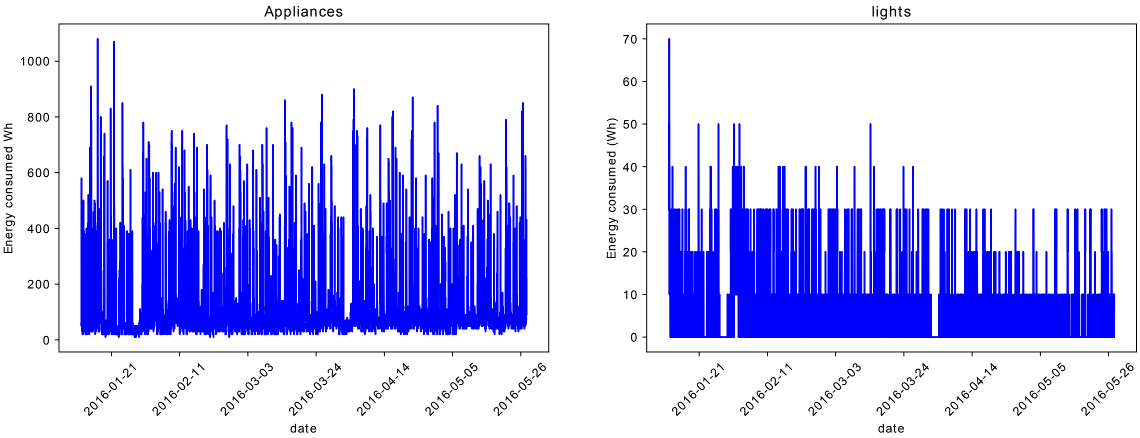


Figure 2. Energy consumed in the house by appliances (left) and lights (right).

The energy use in the house is mainly due to the appliances, however, there is a correlation between the energy use by appliances and lights. This implies the presence of people and the consumption, discarding malfunctioning devices.

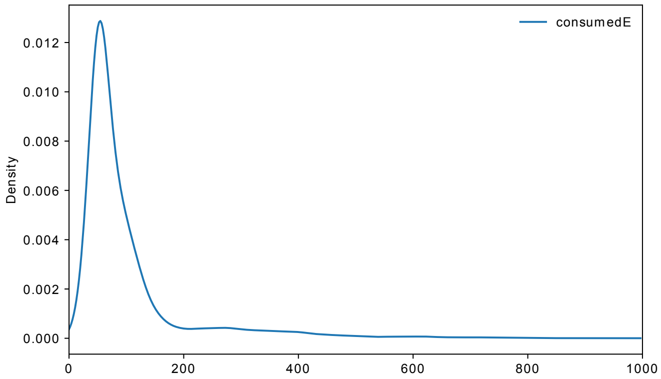


Figure 3. Energy consumption distribution. KDE plot.

2.1. Feature reduction

The data set contains 30 columns; many of them will not present a significant contribution to the variable that I am predicting. Those ones that are poorly correlated with the consumed Energy can be removed. Mutual information measures the dependency between two variables [2]. These methods can capture any kind of statistical dependency. In the next table we can appreciate the mutual information coefficient normalized between different variables and the consumed energy in the house.

Variable	MI coefficient
T1	0.705631
RH_1	0.506109
T3	0.802691
RH_3	0.570058
T4	0.774469
RH_4	0.466110
T5	0.830087
RH_6	0.665144
T7	0.889468
T8	0.780161
T9	1.000000
RH_9	0.507065

Table 1. Normalized mutual information coefficients between the variables in the left columns and the consumed energy.

It can be seen that the variables that are most correlated with the consumed energy are the T9, T7 and T8. These variables correspond to the parent's room, the teenager room and the ironing room respectively. (In the appendix it can be found the pairs plot.)

3. Model

In time series, predictions are made for new data when the actual outcome may not be known until some future date. The future is being predicted, but all prior observations are almost always treated equally. Methods like vector auto-regression (VAR) [3] capture the linear interdependencies among multiple time series, allowing calculating the evolution of more than one variable.

3.1. Dickey Fuller test

At first we test the stationarity of the time series that we are analyzing, more specifically, the consumed energy. To do that, we make use of the Dickey-Fuller test [4-6]. In figure 4 we can see the rolling mean and the rolling standard deviation of the logarithm of the consumed energy. From this plot the Dickey-Fuller test gives

```
Test Statistic                    -8.257370e+00
p-value                            5.193399e-13
#Lags Used                         2.700000e+01
```

```
Number of Observations Used     3.236000e+03
Critical Value (1%)            -3.432372e+00
Critical Value (5%)            -2.862434e+00
Critical Value (10%)           -2.567246e+00
```

Which means that the time series is stationary. These results do not change drastically if we use different values for the number of lags used.

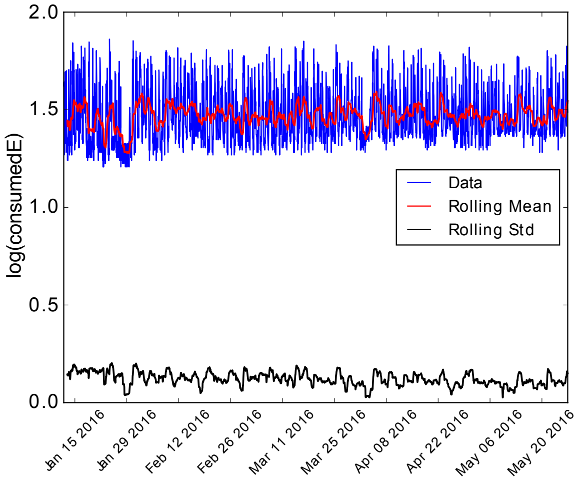


Figure 4. Rolling mean and rolling standard deviation for the consumed energy.

3.2. Cross validation

In order to optimize our model, before we can do the analysis, we need to determine the number of lag terms to use. We know that VAR is a very good tool but we also know the limitations of this method. For example given the length of 4 month of our data, with this method it is possible to predict a few steps in the future and not an entire month. At first, we choose a granularity of hours and then we make use of the cross-validation [7]. We split the data set in training and test. We train our model and we check its performance in a test data set of 24 hours. In figure 5 we show the mean absolute percentage error obtained for the forecast.

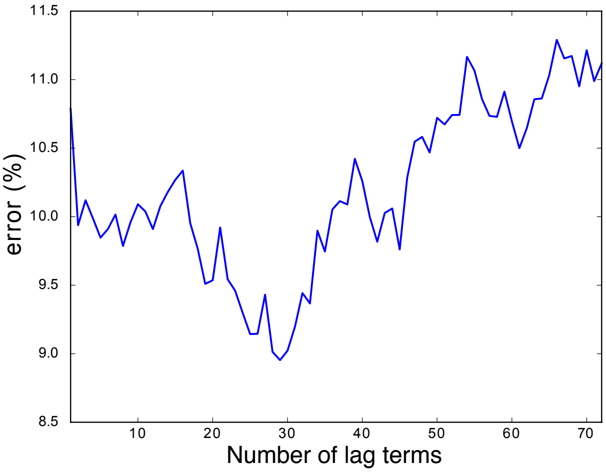


Figure 5. Mean absolute percentage error obtained for the forecast of the test data set.

It can be seen that the optimal number of lag terms is 29 and shows an error close to 9%.

3.3. Forecast

Now that we have selected the optimal number of lag terms to use, we can build a model with our data set and then predict some days in the future. From figure 5 we know that a 24 hours forecast will show an error close to 9%. The error will increase if we do the forecast for more hours.

In figure 6 we show the 4 days forecast of the energy consumed. It can be seen that the trend becomes smoother for longer periods of time, this means that the model do not capture completely the behavior, so our system predict in good agreement until one day in the future, which is a great progress given that these data can be collected daily.

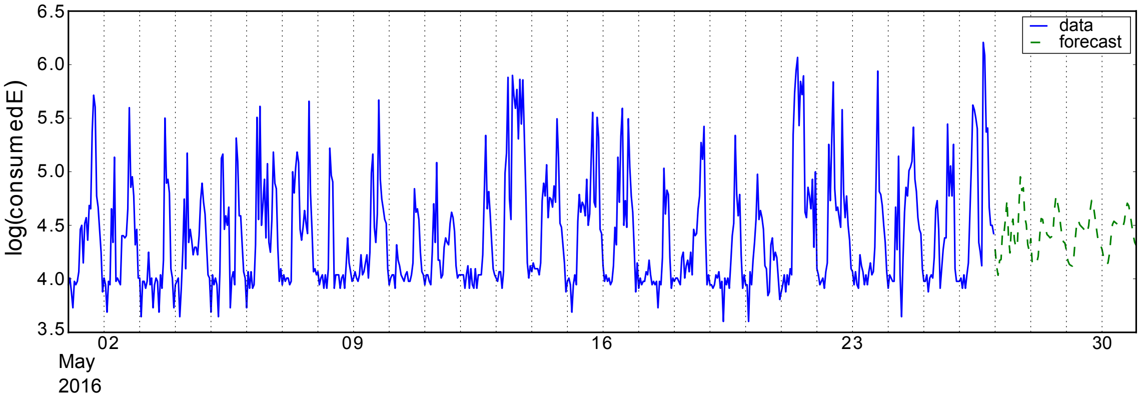


Figure 6. 4 days forecast of the consume energy in the house. The blue line shows the values in the data set while the green line show the predicted values.

Conclusions

The prediction of appliances' consumption with data from the wireless network indicates that it can help to locate where in a building the main appliances' energy consumption contributions are found. We find that the main consumption is done in the parents' room, the teenager room and the ironing room, probably due to the equipment present in those places.

In conclusion we encourage people to monitor their energy consumptions at home to improve their efficiency energy use and predict their expenses.

Bibliography

[1] Luis M. Candanedo, Véronique Feldheim, Dominique Deramaix. Data driven prediction models of energy use of appliances in a low-energy house. Energy and buildings, 140, 81-97,1 April 2017.

[2] Mohamed Bennasar, Yulia Hicks, Rossitza Setchi. Feature selection using Joint Mutual Information Maximisation. In Expert Systems with Applications. 42, Issue 22, 2015, Pages 8520-8532.

[3] Duo Qin. Rise Of VAR Modelling Approach. Journal of Economic Surveys. 25. 1467-6419. 2011.

[4] David A. Dickey. Stationarity Issues in Time Series Models. Statistics and Data Analysis. 192-30

[5] Dickey, D. A. and Fuller, W. A. (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". Journal of the American Statistical Association, 74, p. 427-431.

[6] Dickey, D. A. and Fuller, W. A. (1981). "Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root". Econometrica 49, 1057-1072.

[7] Picard, Richard; Cook, Dennis (1984). "Cross-Validation of Regression Models". *Journal of the American Statistical Association*. **79** (387): 575–583.

Appendix A

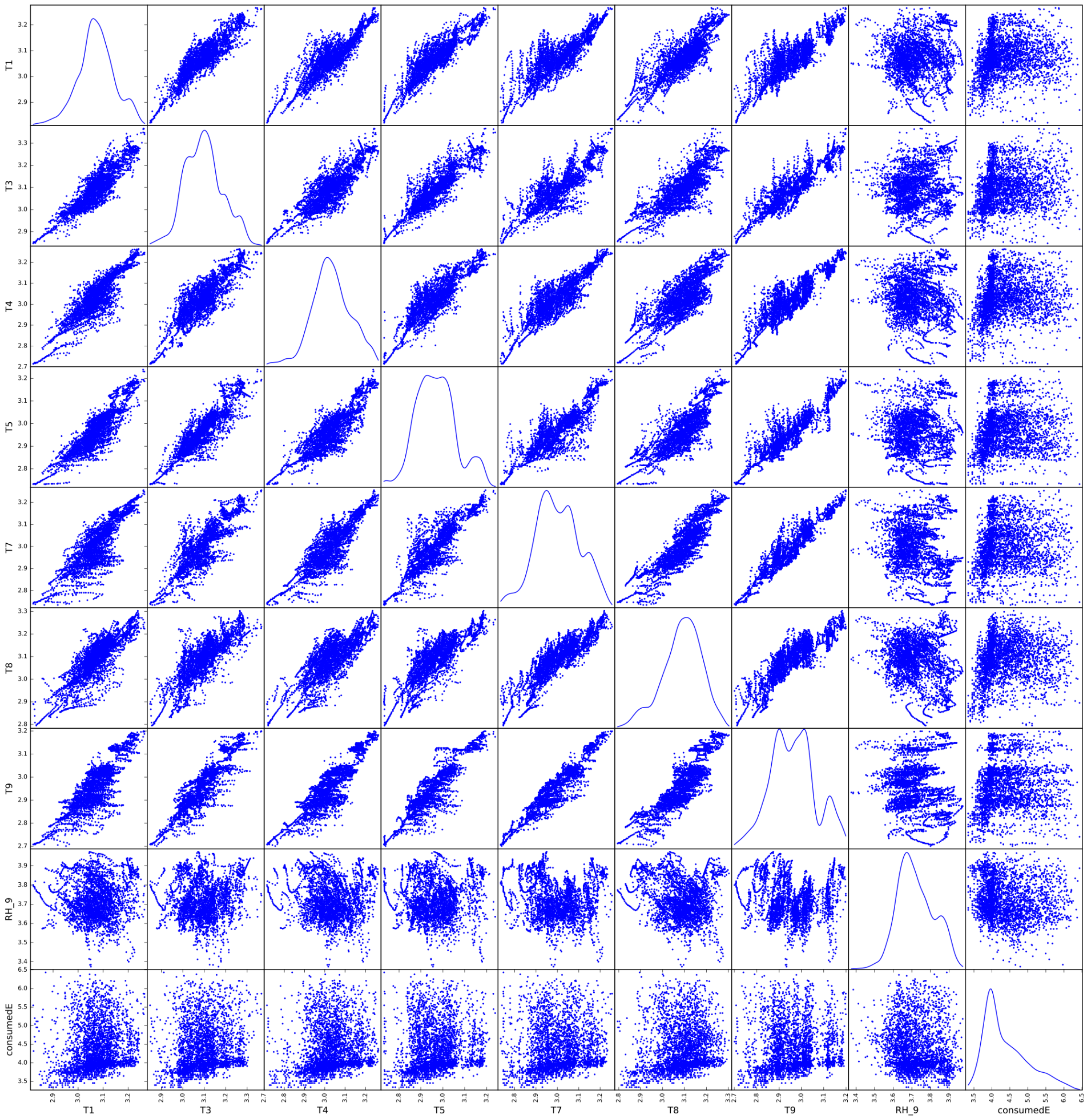


Figure A1. Pairs plot of the variables that are most correlated with the energy consumption in the house. The normalized mutual information coefficient is shown in table 1.