Prediction model of energy use in a house.

J. A. Celis Gil

Abstract

I present and discuss data-driven predictive models for the energy use in a house. I discuss data filtering to remove non-predictive parameters and feature ranking. Two models were done, one using vector auto-regression (VAR) with repeated cross validation and the other using an iterative random forest (RF) procedure. Using VAR, I was able to predict until 24 hours ahead with an accuracy of 9%, while with RF model I predicted the same period with an error of 6%. For longer periods of time the accuracy of the models decrease. 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 that is spent there. Finally I remark the importance of monitoring the energy consumed in order to reduce the expenses and improve the efficiency in the energy use.

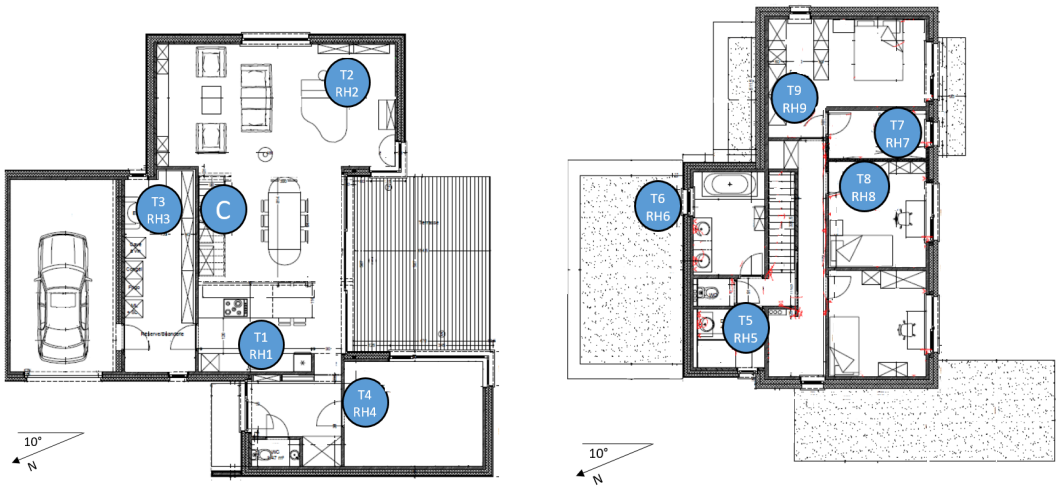
1. Introduction

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 reducing the cost and 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 and save some money.



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 17:00. They hired a company that helps them monitoring the energy consumption. Peter wants to detect if all the devices are working properly, but he does not know that with the data 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 try to modify their customs to reduce the 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, I 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. The data consist of weather from a nearby airport station (Chievres Airport, Belgium) 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 [1]. This data is available at this link http://archive.ics.uci.edu/ml/machine-learning-databases/00374/

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

2. Data exploratory analysis

I am mainly interested to analyze the energy consumed in the house. Figure 2 shows the comparison between the energy consumed by the appliances and lights in the complete data set.

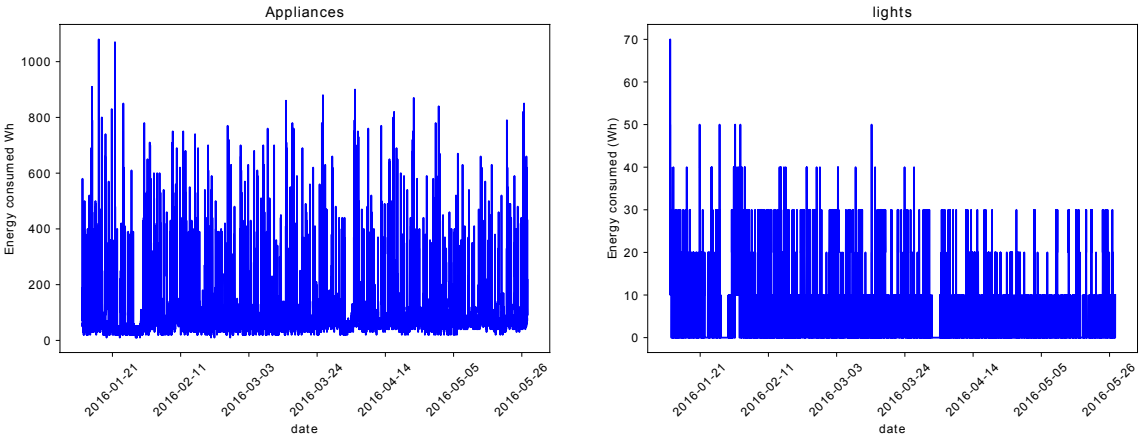


Figure 2. Energy consumed in the house by appliances (left) and lights (right).

It is clear that 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.

I calculated the total energy consumed by adding the energy used by appliances with the one consumed by lights. Figure 3 shows the KDE plot of this new variable. It can be seen that in a normal day it is consumed between 0 and 200 KWh of energy in the house.

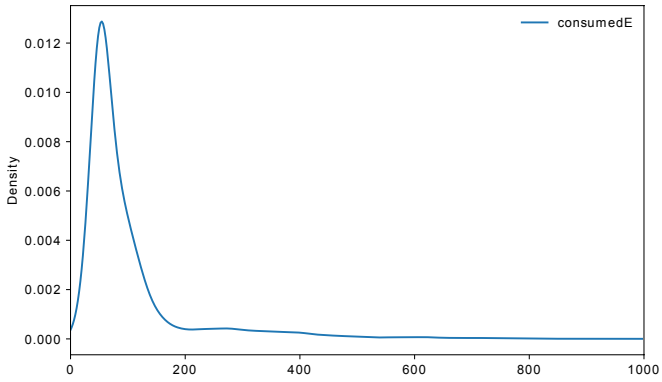


Figure 3. Energy consumption distribution. KDE plot.

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]. This method can capture any kind of statistical dependency. In the next table it can be appreciated 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.

The consumed energy is most correlated with 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. Ensemble learning methods like random forest (RF) [4,5] can also be applied to time series. In this method, multiple decision trees are built were the target variable takes into account previous values.

Before I can make the models, I have to check the stationarity of the time series, more specifically, the consumed energy.

3.1. Dickey Fuller test

Making use of the Dickey-Fuller test [6-8], it is possible to check whether or not a time series is stationary. In figure 4, it is shown the rolling mean and the rolling standard deviation of the logarithm of the consumed energy. From this plot the Dickey-Fuller test gives the next results

```
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
```

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

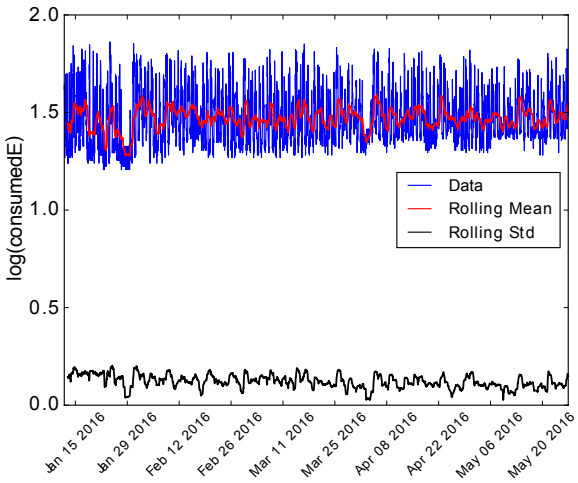


Figure 4. Rolling mean and rolling standard deviation for the consumed energy.

3.2. Model using VAR

One model was created using vector auto-regression, which is a method where each variable is a linear function of the *n* lag values for all variables in the set.

The first step is to choose the right granularity for the data. The measurements were done every ten minutes giving place to noisy behavior. I change the Δt such that the granularity now corresponds to hours.

In order to optimize the model, before the analysis, I determine the number of lag terms to use. VAR is a very good tool but this method has some limitations. For example given the length of 4 month of the data set, it is possible to predict a few steps in the future and not an entire month. I make use of cross-validation [9] to find the optimal number of lag terms.

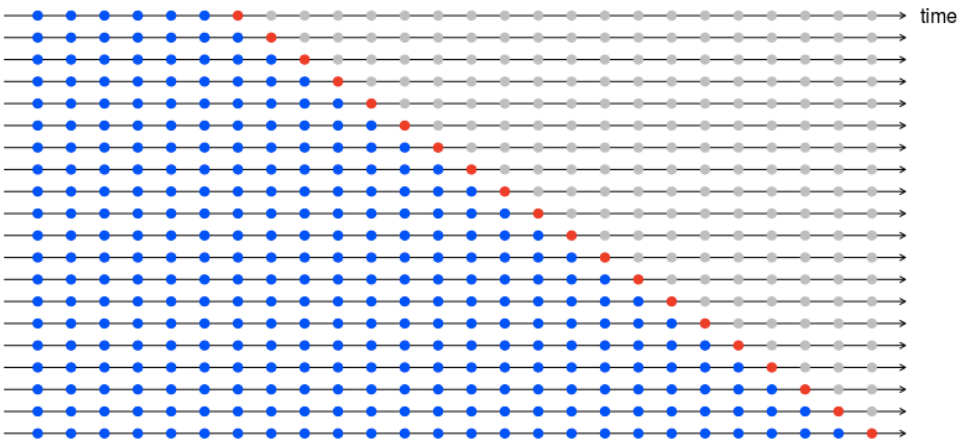


Figure 5. Diagram of the train and test sets for time series

I split the data set in training and test. I train the model and I check its performance in a test data set of 24 hours. However in time series the order is very important so that a typical K-Fold cross validation [10] cannot be performed for this kind of data. Instead of that, I used a modified version of the K-Fold cross validation [11]. In this procedure, there is a series of test sets, each consisting of a group of observations. The corresponding training set consists only of observations that occurred prior to the observation that forms the test set. Thus, no future observations can be used in constructing the forecast. Figure 5 illustrates the series of training and test sets, where the blue observations form the training sets, and the red observations form the test sets.

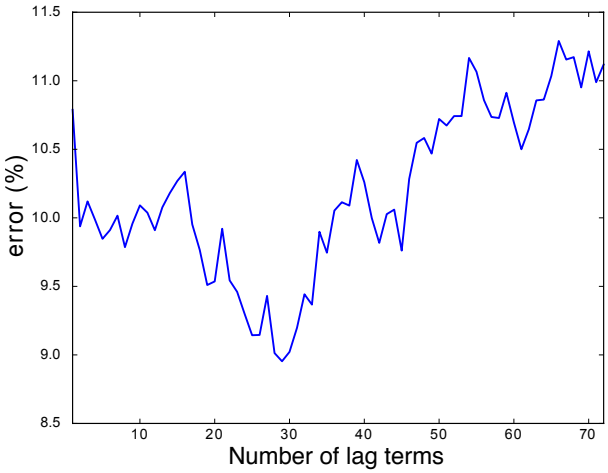


Figure 6. Mean absolute percentage error (MAPE) obtained for the forecast of the test data set.

Using the K-Fold time series cross validation, I calculated the optimal number of lag terms for the analysis. Figure 6 shows the mean absolute percentage error obtained for the test set. It can be seen that the optimal number of lag terms is 29 and shows an error close to 9%.

3.3. Model using RF

Another model was created using random forest. In this method a matrix in which every column corresponds to a variable is input in order to explain a target variable. In time series where every row in a table corresponds to a date time and the target depends on the prior observations of all the variables, I need to create a matrix that contains the previous observations as variables.

	t-7	t-6	t-5	t-4	t-3	t-2	t-1	t
1961-06-01	NaN	NaN	NaN	NaN	NaN	NaN	NaN	687.0
1961-06-02	NaN	NaN	NaN	NaN	NaN	NaN	687.0	646.0
1961-06-03	NaN	NaN	NaN	NaN	NaN	687.0	646.0	-189.0
1961-06-04	NaN	NaN	NaN	NaN	687.0	646.0	-189.0	-611.0
1961-06-05	NaN	NaN	NaN	687.0	646.0	-189.0	-611.0	1339.0
1961-06-06	NaN	NaN	687.0	646.0	-189.0	-611.0	1339.0	30.0
1961-06-07	NaN	687.0	646.0	-189.0	-611.0	1339.0	30.0	1645.0
1961-06-08	687.0	646.0	-189.0	-611.0	1339.0	30.0	1645.0	-276.0

Table 2. Example of a table modified to take into account 7 lag terms.

Table 2 shows an example of how the table is modified to take into account the previous measurements. From this table the first 7 rows are removed, playing the same roll as the lag terms in VAR.

To apply random forest to the data set, I used 29 lag terms for the consumed energy and 29 lag terms for every one of the variables in table 1.

In RF, one important parameter to fit is the depth of the trees. Making use of the time series cross validation explained in the previous subsection, I determined that value. Despite the optimal value obtained was 6 showing a MAPE with the test set equal to 5.5 %, due to hardware restrictions I used a depth equal to 4 which got a MAPE equal to 6%.

3.4. Forecast

Now that I have selected the optimal number of lag terms to use and the depth of the trees in random forest, I can use the two models to predict some days after the last day of the data set. I know that using VAR, a 24 hours forecast will show an error close to 9% and using RF it will show an error of 6%. However, the error will increase if I do the forecast for longer periods.

In the models I assume that the variables are independent between them, except for the consumed energy which is the target variable. For the random forest calculations, I perform a single variable time series calculation for every variable in table 1 at time t to forecast the value at time t+1.

In figure 7 I show the 24 hours forecast of the energy consumed using the two models (VAR and RF). It can be seen that both methods predict peaks around 9 in the morning and 5 in the afternoon whose coincides with the periods of more activity in the house. A forecast for longer periods do not capture completely the behavior, so our models predict in good agreement until one day in the future, which is a great progress given that these data can be collected daily.

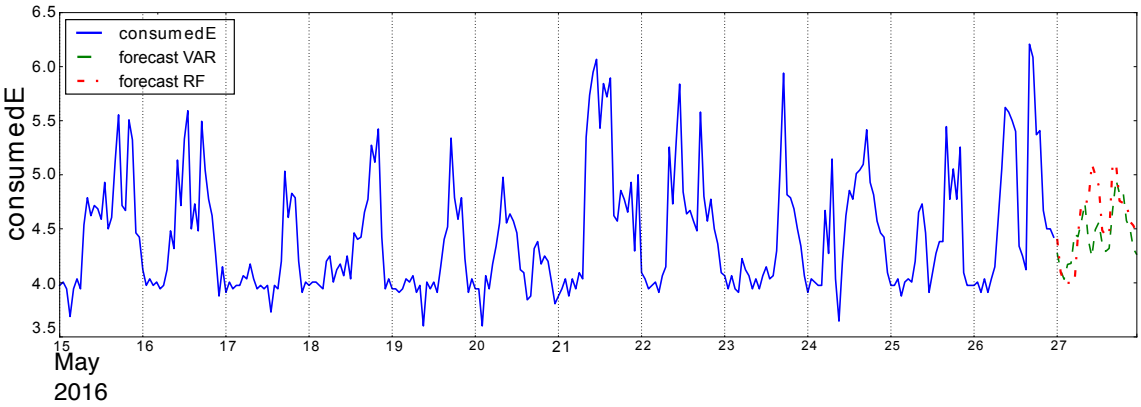


Figure 7. 4 days forecast of the consume energy in the house. The blue line shows the values in the data set while the green line show the predicted values.

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.

Two models were developed, one using vector auto regression and the other using random forest, showing errors of 9% and 6% respectively. The better performance in the random forest model is maybe because it not only assumes linear dependence between the features and the target variable.

Finally I want to encourage people to monitor their energy consumptions at home to improve their efficiency energy use and reduce 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] Tin Kam Ho, "The random subspace method for constructing decision forests," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, Aug 1998.

[5] Ho, Tin Kam (1995). Random decisión forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

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

[7] 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.

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

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

[10] McLachlan, Geoffrey J.; Do, Kim-Anh; Ambroise, Christophe (2004). Analyzing microarray gene expression data. Wiley.

[11] Christoph Bergmeir, Rob J Hyndman, Bonsoo Koo (2018) A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics and Data Analysis, 120, 70-83.

Appendix A

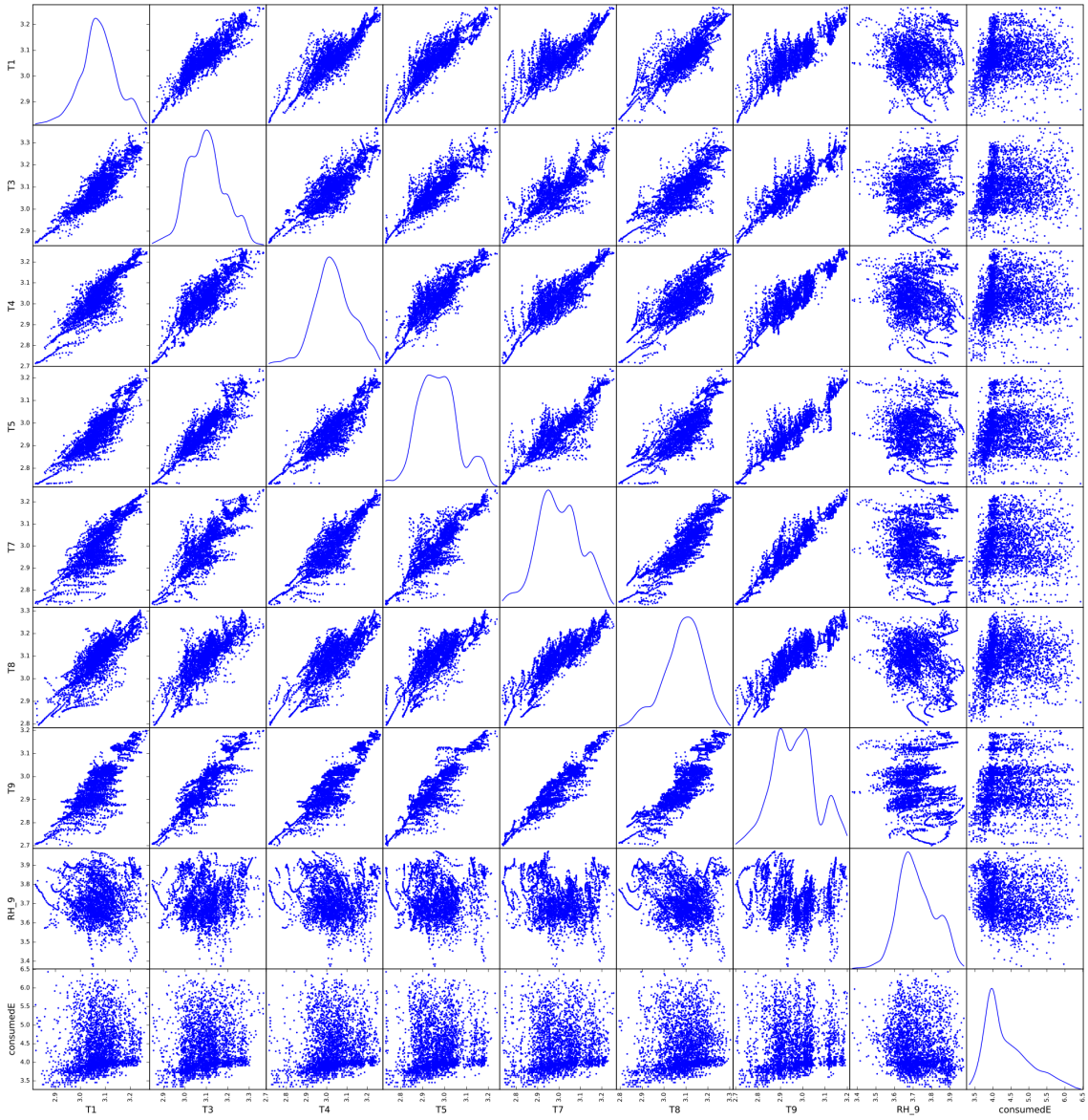


Figure A1. Pairs plot of the variables that are most correlated with the energy consumption in the house. The normalized mutual information coefficient is shown in table 1.