Building Energy Consumption Prediction and Disaggregation with Basic Building Information

Sicheng Zhan (szhan)

1st author's affiliation

1st author's E-mail address

Shucheng Chao (shuchenc)

2nd author's affiliation

2nd E-mail

**ABSTRACT**

In this paper, we describe the formatting guidelines for ACM SIG Proceedings.

**Keywords**

Keywords are your own designated keywords separated by semicolons (“;”).

# INTRODUCTION

Currently, in most buildings energy meters are primarily installed to measure the total amount of energy consumption and support the agent’s billing function. However, disaggregated energy consumption information is very valuable in different phases of a building’s life cycle, and in many cases can be used to save energy.

For example, the HVAC unit capacity should be decided in the designing and construction phase. If the heating and cooling load can be accurately decided, mismatch cases in the operation phase can be avoided. Also for the retrofitting, knowing the disaggregated consumption would benefit both the auditing and the renovating process. So it’s useful to the house holders, the operators, the energy agents and the policy makers.

With the data driven research methods, the total amount of energy consumption and the basic information of the buildings, we intend to predict the energy consumption of different end use. We will compare the predicted result with the recorded data, discussing the influence of different prediction methods, and the different input features.

# METHODS OVERVIEW

With different level of data accessibility, there exist several ways to disaggregate energy consumption. In this section, the methods are divided into two main types according to sensing numbers and summarized. Then the method we intend to employ is pointed out.

## Distributed Direct Sensing

Obviously, the most reliable way to find the disaggregated energy consumption is to install meters at each devices in a building. However, in spite of the easy concept and the high accuracy, there are several deficiencies of this kind of methods: First, installing such a series of sensors will cost a lot. Additionally, the meters for most appliance are difficult to install and maintained. Also, collecting a bunch of data requires a stable and well-organized system, which is not applicable to many buildings [1]. Thus, this kind of methods are not very promising.

## Single Point Sensing

One of the widely energy consumption disaggregation methods with single point sensing is called nonintrusive load monitoring (NILM). Instead of simply directly the consumptions, NILM measure the current and the voltage going into the house. NILM measurement system is easy to use and install, but its performance is not stable and sometimes leads to misunderstanding, so the performance evaluation is necessary for each site [2]. By applying pattern recognition approach to the demand change data, energy usage of different end uses can also be found. Different appliances will come up with different patterns and be extracted from the total consumption. But this method is restricted by the characteristics of the appliances [3].

Generally, the more data is accessible, the more reliable the method is. But since the data accessibility is not sufficient for most buildings, a method with lower requirement of the data is necessary. Residential Energy Consumption Survey (RECS) and Commercial Buildings Energy Consumption Survey (CBECS) are two databases providing the basic information and the energy consumption of thousands of residential and commercial buildings in the USA. Though the data describe the buildings very thoroughly with hundreds of features, the features are mostly ready to be recorded anytime. So it would be a method applicable to all the buildings if we can find a way to predict the disaggregated energy consumption with these basic building information.

# DATABASE SELECTION

RECS and CBECS respectively record the information of residential and commercial buildings, categories including location, building type, structure, energy source and the like. With IPython and other data analysis Python packages like Pandas, both of the databases are investigated. Some details of them are discussed in this section, and finally we choose RECS, the residential database, to implement the further analysis in this project.

## CBECS

CBECS categorize the building by their functions like office building, laboratory, nursing, service and others. And record the basic structure information like the location, area, construction material, construction year and so on. Also the information about the occupants are included like the number of employees, occupants’ primary activities, time of usage and others. In total, 6720 buildings with 1119 features are recorded.

However, when we try to find the average energy consumption of buildings with different categories, the mean values of different categories do not clearly distinguish from each other except few categories like the location area. So we find that this database is not suitable for prediction because there are too many factors influencing the buildings’ energy consumption, and consequently it is difficult to find a pattern between the categories and the consumption.

## RECS

While CBECS is describing the building location by dividing the USA into 9 districts, RECS use the state for the location category. And the categories are mostly about the structure and occupants of the building, which are closely related to the energy consumption. In total, 12083 buildings with 983 features are recorded.

Using common sense, the features in RECS, like heating degree days, heating energy source, wall type and the like, are very promising in predicting the energy consumption. Thus we pick this database for the further analysis.

# DECISION TREE REGRESSION

Considering the properties of the input features, which is mostly discrete points, we decided to use decision tree regressor to predict the energy consumption. Then the question is to select the proper features to train the trees. One option is to manually pick the features based on the professional knowledge, and the other is to use a data driven way, applying evaluating models to choose the features.

## Manually picked feature

Considering the properties, 10 features are picked to train the model.

# DISCUSSION

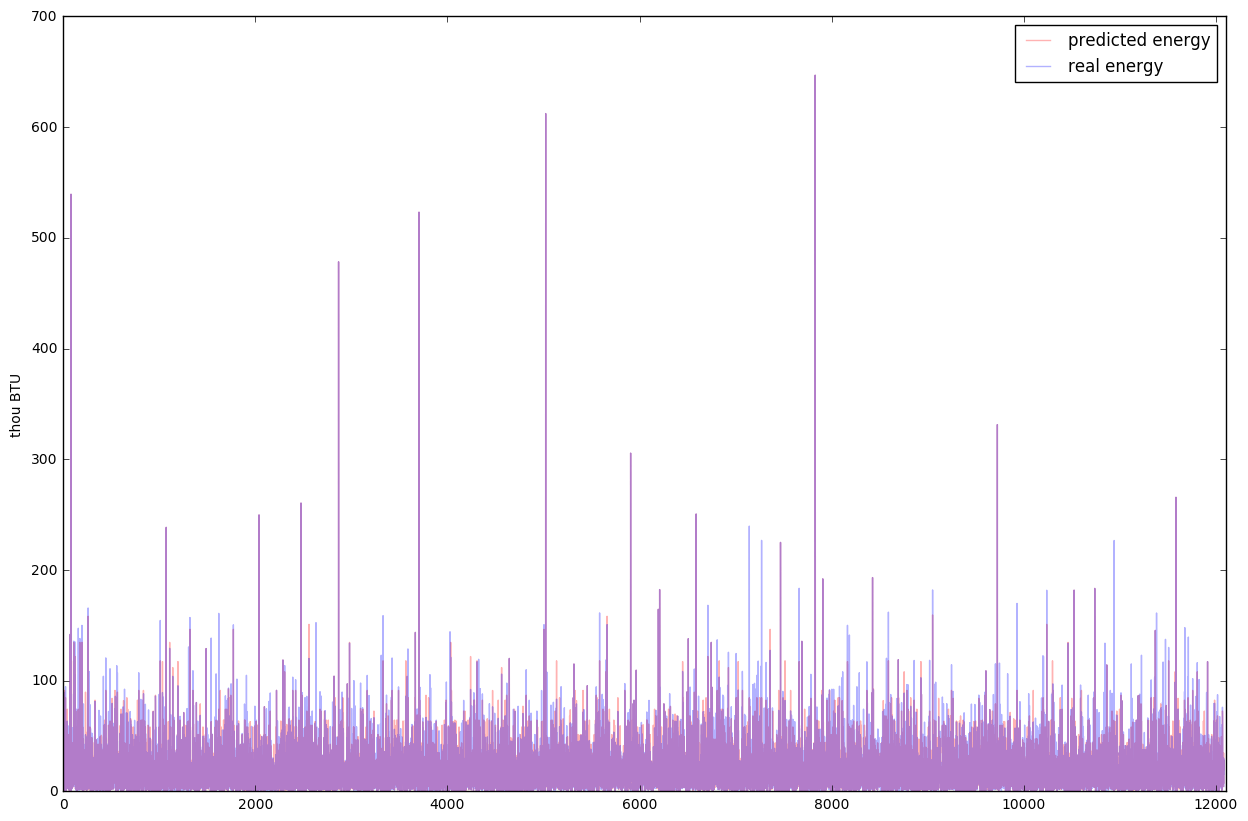


Figure 1

Figure 1 shows the prediction of the data (in red) versus the actual data value (in blue), with a correlation r-score of 0.63. The purple area indicates a pretty good overlapping between the prediction and the actual.

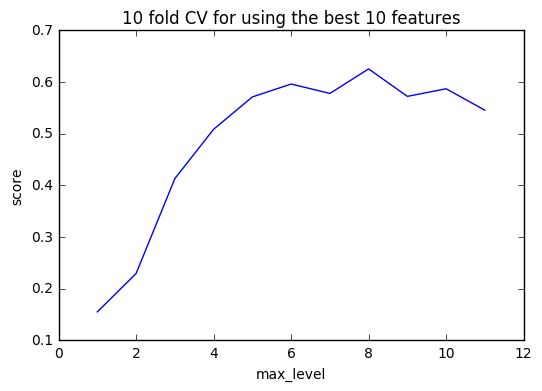


Figure 2

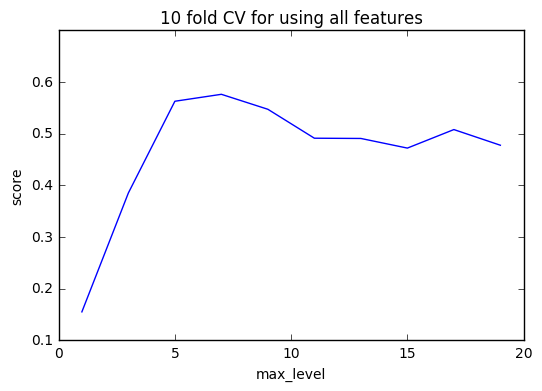


Figure 3

In order to compare the performance of using different features as training data, we need to find the best regression tree model for each of the candidate. Figure 2 and Figure 3 show the curve of r-score versus regression tree parameter “max\_level”. We can see that by using ten-fold cross validation, we were able to get a clear progress from underfitting to overfitting. The maximum of each curve illustrates the best possible r-score the model can achieve. As a result, Figure 1 shows that a regression tree trained on all features achieved its best performance of 0.58 at max\_level = 5. On the other hand, Figure 2 shows that a regression tree trained on the 10 most related features achieved a best performance of 0.63 at max\_level = 8.

As a result, by properly selecting and pruning features, we were able to increase the best performance of the model and at the mean time significantly shorten the model training time!

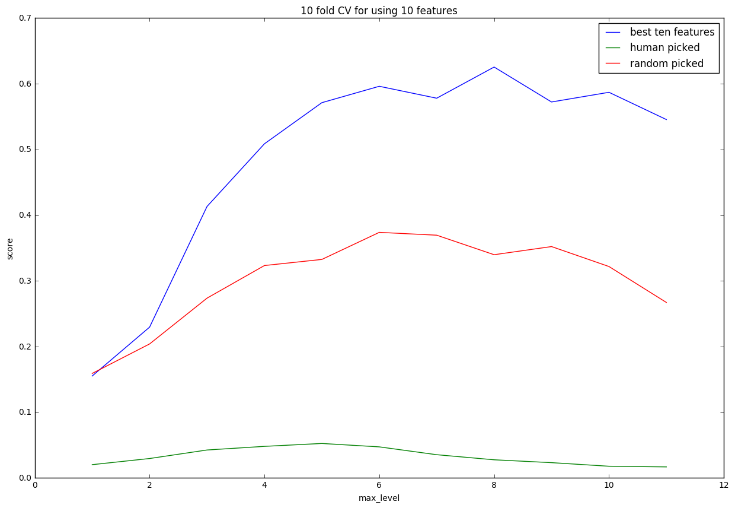


Figure 4

To further illustrate the importance of proper feature selections, Figure 3 shows a direct comparison between three ways of picking 10 training features: 10 attributes with the most relative importance fit by an Extra tree; 10 “best” attributes from human intuition; and 10 randomly chosen attributes.



Figure 1. Insert caption to place caption below figure.

.

# CONCLUSION

Our thanks to ACM SIGCHI for allowing us to modify templates they had developed.

# REFERENCES

1. Froehlich, J., Larson, E., Gupta, S., Cohn, G., Reynolds, M., & Patel, S. (2011). Disaggregated end-use energy sensing for the smart grid. IEEE Pervasive Computing, 10(1), 28-39.
2. Kolter, J. Z., & Johnson, M. J. (2011, August). REDD: A public data set for energy disaggregation research. In Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA (Vol. 25, pp. 59-62).
3. Farinaccio, L., & Zmeureanu, R. (1999). Using a pattern recognition approach to disaggregate the total electricity consumption in a house into the major end-uses. Energy and Buildings, 30(3), 245-259.