Residential Energy Consumption Prediction and Disaggregation with Basic Building Information

Sicheng Zhan
Carnegie Mellon University
szhan@andrew.cmu.edu

Shucheng Chao
Carnegie Mellon University
shuchenc@andrew.cmu.edu

ABSTRACT

Knowing the disaggregated energy consumption of buildings is very helpful for constructing, managing and retrofitting. But collecting data is usually consuming and even difficult. In this project, we intend to use data driven methods to predict the consumption with basic building information, which can be collected anytime during the building's life cycle. By comparison, extracting the most related features by evaluating the relevancy and then using them to train the decision tree regressor is the most reliable way. Heating and cooling consumption can be predicted respectively with the 10-fold cross validation score of 0.646 and 0.756. Consumption of other usage is turn out to be hard to predict, but it can be derived by subtracting the HVAC part from the total consumption. Judging from the high availability of the data and the high score of the prediction validation, this method is promising in predicting the disaggregated consumption of buildings.

Keywords

Energy consumption disaggregation, extra-trees regressor, feature selection, regression decision tree, cross validation.

1. 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.

2. 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.

2.1 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.

2.2 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 basic building information.

3. 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.

3.1 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.

3.2 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.

4. 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.

4.1 Manually picked features

Considering the properties, 10 features are picked to train the model for heating consumption as shown in table 4.1.1. Among them, the first fours are describing the weather condition of where the building is. Other features involve the insulation condition of the building, the building configuration, the building history and the quality of the heating source.

Feature ID	Feature name
REPORTABLE_DOMAIN	Reportable states and groups of states
HDD30YR	Heating degree days, 30-year average 1981-2010, base 65F
CLIMATE_REGION_PUB	Building America Climate Region (collapsed for public file)
AIA_ZONE	AIA Climate Zone, based on average temperatures from 1981 - 2010
TYPEHUQ	Type of housing unit
YEARMADE	Year housing unit was built
WALLTYPE	Major outside wall material
ROOFTYPE	Major roofing material
TOTSQFT	Total square footage (includes all attached garages, all basements, and finished/heated/cooled attics)
HEATOTH	Main space heating equipment heats other homes, business, or farm

Table 4.1.1: features selected for heating prediction

And the 10 features selected for cooling prediction are listed in table 4.1.2. Most of them are similar to those for heating prediction. And the occupants' behavior are also considered.

Table 4.1.2: features selected for heating prediction

Feature ID	Feature name
REPORTABLE_DOMAIN	Reportable states and groups of states
HDD30YR	Heating degree days, 30-year average 1981-2010, base 65F
CDD30YR	Cooling degree days, 30-year average 1981-2010, base 65F
AIA_ZONE	AIA Climate Zone, based on average temperatures from 1981 - 2010
TYPEHUQ	Type of housing unit
UR	Urban or rural
WALLTYPE	Major outside wall material
ROOFTYPE	Major roofing material
TOTSQFT	Total square footage (includes all attached garages, all basements, and finished/heated/cooled attics)
KOWNRENT	Housing unit is owned, rented, or occupied without payment of rent

4.2 Numerically picked features

In order to determine the "best" features, we decided to use an extra-tree regressor from scikit-learn. An extra-tree regressor implements a meta estimator that fits several randomized decision trees on various sub-samples of the dataset. While creating the forest of trees, the algorithm uses mean decrease accuracy method to permute the values of each feature and measure how much the permutation decreases the accuracy of the model. The output object thereby contains performance scores for each feature from the RECS data.

Therefore, by fitting an extra-tree model onto the RECS dataset using the target as total space heating power consumption and total cooling power consumption. The top ten most important features are listed in Table 4.2.1 and Table 4.2.2.

Table 4.2.1: features selected for heating prediction

Feature ID	Score
TOTSQFT_EN	0.077509
TOTSQFT	0.073985
NCOMBATH	0.06679
AIA_Zone	0.066381
HEATOTH	0.061899
FUELHEAT	0.047722
UGWARM	0.04287
REPORTABLE_DOMAIN	0.040212
FOWARM	0.028245
PGASHEAT	0.023065

Table 4.2.2: features selected for cooling prediction

Feature ID	Score
CDD30YR	0.239172

AIA_Zone	0.174042
CDD65	0.149317
USEWWAC	0.043049
AIRCOND	0.033918
TEMPHOMEAC	0.024297
TEMPGONEAC	0.017537
TOTUCSQFT	0.016991
TEMPNITEAC	0.0162
USECENAC	0.013327

5. RESULTS

5.1 Heating

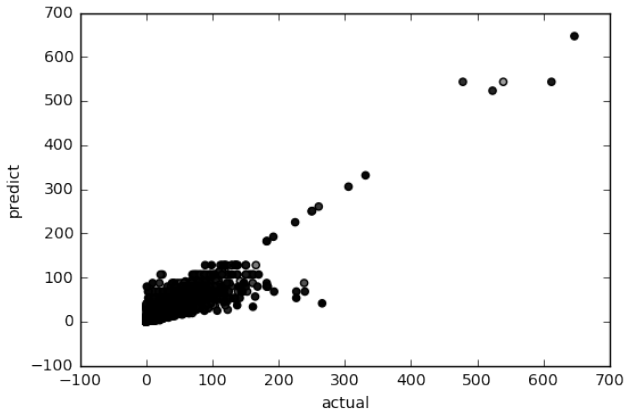

Figure 3: Prediction result for heating power consumption

Figure 1 shows the predicted heating power consumption versus the actual heating consumption. The best regression tree managed to provide a prediction r-score of 0.646. We can see that all the (actual, predicted) tuples fall around the diagonal line which represents the idea prediction result, predict = actual.

5.2 Cooling

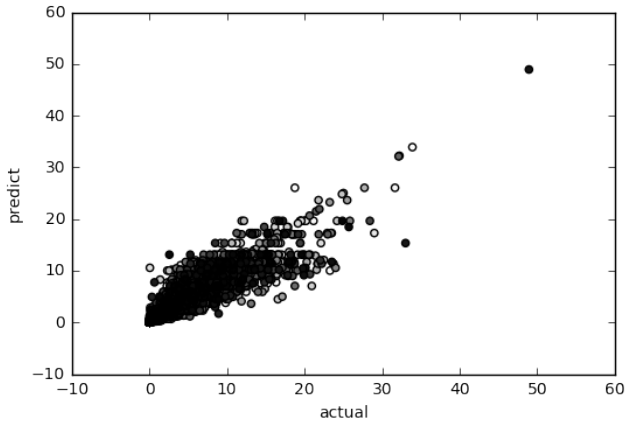

Figure 4: Prediction result of cooling power consumption

Figure 2 shows the predicted cooling power consumption versus the actual cooling power consumption. For the cooling power consumption, the best regression tree managed to reach an r-score of 0.756.

6. Discussions

6.1 Feature selection

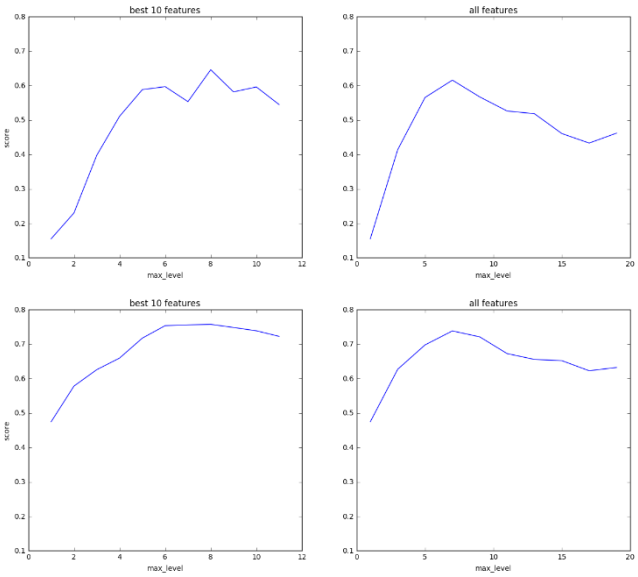

Figure 5: Performance of regression tree using best 10 features vs using all features (above: heating, below: cooling)

In this paper, we are interested in correctly predicting heating and cooling power consumption while using only several basic features of the building. However, instead of using all features, it would take much less time to train the model if we can reduce the feature space.

Figure 3 shows the curves of r-score versus regression tree parameter "max_level" for the ten-fold cross validation process. The upper two graphs show the results of heating while the bottom two graphs show the results of cooling. The left two show the results from best 10 features while the right two show the results from using all features.

The first thing we noticed is that by using ten-fold cross validation, we were able to get a clear indication from the state of under-fitting to the state of over-fitting. The curves of regression tree trained on ten features reached maxima at about max_level = 8. On the other hand, even though the curves trained on all features reached maxima at max_level = 7, since the latter curves trained on a total 838 features, the curves actually reached their best performance at a much shallower stage than the curves trained on only 10 good features.

In terms of actual numbers, the maximum of each curve illustrates the best possible r-score the model can achieve. Therefore, for heating data, the regression tree trained on all features achieved its best performance of 0.616. On the other hand, regression tree trained on the best 10 features achieved a best performance of 0. Therefore, reducing the feature space from 838 features to only 10 features not only didn't harm the overall performance score but instead increased the model's best possible performance by about 3% in both heating and cooling cases. Of course it should be mentioned that since the cross validation process requires random selection of test/train dataset, the final performance score as the mean from all 10 folds would fluctuate depending on the randomness. However, we can still confidently claim that using only 10 most related features to train the regression tree, if not significantly better, performs as well as the model trained on all 838 features.

As a conclusion, by properly selecting and pruning features, we were able to increase the best possible performance of the regression tree model and at the mean time significantly shorten the training time.

6.2 Human expert versus extra tree

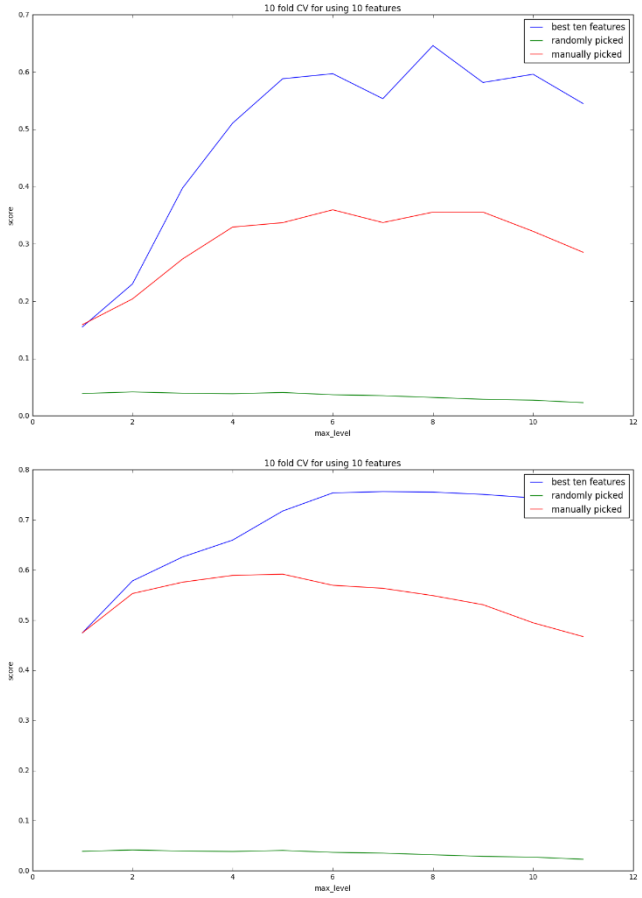

Figure 6: Performance comparison between different feature selection method (above: heating, below: cooling)

To further validate that the 10 features are actually good picks, we performed 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 as a control group, shown in Figure 4. We can see that in both heating and cooling cases, it is clear that algorithm picked > manually picked > randomly picked. For regression tree trained on expert's picks, the model reached a best performance of r-score = 0.360 for heating and 0.592 for cooling.

Interestingly, models trained on manually picked features have a very similar performance as the models trained on algorithm picked features with max_level limited to 1. However, the gap between the performances quickly increased afterwards. This makes sense since human intuition and knowledge can tell correlation between two attributes easily but have a much harder time searching for combination of multiple predictors.

7. CONCLUSION

7.1 Summary

The prediction result turns out to be pretty good in this project with the provided database. While manually picked data successes in providing relatively good prediction, using data driven ways brings even better results. Since the heating and cooling consumption can be well predicted, though prediction on the other consumptions shows lower performance, it can be derived by subtracting the HVAC part from the total consumption, which is very easy to collect. So it is promising to predict the disaggregated energy consumption with basic building information and implement corresponding decisions.

7.2 Future Work

Due to time limit and the computer's limited calculation capability. There are still series of problem to investigate in this project. For example, we directly picked ten features to train the model instead of looking for the optimal input dimension. And we directly use the decision tree regressor to predict the consumption, but other models can bring better result. Also we find that the model performs better for the high energy consumptions then the lowers, where most data sample lie in. The model can be better if the few points with high consumption are removed before training. So it still has potential to improve the performance of prediction.

8. 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.