Energy Disaggregation via Clustered Regression Models: A Case Study in the Convenience Store

Hsiao-Hui Chen[1], Ping-Feng Wang[2], Ching-Tien Sung[2], Yi-Ren Yeh[3] and Yuh-Jye Lee[1]
[1]*Dept. of Computer Science & Info. Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan*
[2]*Smart Network System Institute, Institute for Information Industry, Taipei, Taiwan*
[3]*Department of Applied Mathematics, Chinese Culture University, Taipei, Taiwan*
h20117h@hotmail.com, pfwang@iii.org.tw, today@iii.org.tw, yyr2@ulive.pccu.edu.tw, yuh-jye@csie.ntust.edu.tw

***Abstract*—Global warming and the depletion of natural resources are two of the most difficult problems we have ever faced. To address this problem, people have begun paying more attention to carbon emission reduction and energy saving. For the residential electricity use, many studies have demonstrated that feedbacks, such as energy consumption of each appliance in the home, can help consumers reduce electricity consumption usage. In this article, we propose a novel framework for the disaggregation of energy consumption, which is looking forward to reaching reducing the number of smart meters installed and providing usage statistics as a feedback for consumers to decrease their energy cost. In our proposed framework, we have a chief meter which measures total energy consumption, and install smart meters at few key appliances. Based the energy consumption from these meters, we proposed a clustered regression models for energy disaggregation. More specifically, we first cluster appliances by the correlation between the using behavior of appliances, and select one of them as the key appliance in each cluster. By using the appliance with installed meter, we apply regression model to estimate the energy consumption for other appliances within each cluster. Our experimental results confirmed our proposed framework can achieve high accuracy for energy disaggregation while reducing the number of smart meters.**

***Keywords*-energy disaggregation, clustering, support vector regression;**

I. Introduction

Due to global warming and the depletion of natural resources, people have begun paying more attention to carbon emission reduction and energy saving. Many studies have demonstrated that feedbacks, such as how energy is used in the home, can help consumers reduce electricity consumption [2, 3, 5, 7, 8, 9, 15]. Thus, power load disaggregation for appliances could be regarded as feedback information to consumers.

An intuitional idea for disaggregation is direct sensing [3]. That is, one needs to install smart meters at all appliances to collect the energy consumption. The concept of direct sensing is straightforward, and can achieve high accuracy for energy disaggregation. However, it requires expensive cost to install smart meter at each appliance. To overcome problem, we propose a novel framework for the disaggregation of energy consumption, which is looking forward to reaching reducing the number of smart meters installed and providing usage statistics as a feedback for consumers to decrease their energy cost. In our proposed framework, we have a chief meter which measures total energy consumption, and install smart meters at few key appliances. Based the energy consumption from these meters, we proposed a clustered regression models for energy disaggregation. More specifically, we first cluster appliances by the correlation between the using behavior of appliances, and select one of them as the key appliance in each cluster. By using the appliance with installed meter, we apply regression model to estimate the energy consumption for other appliances within each cluster. In addition, we also used various strategies to segment data for training the regression models, such as including different levels of sensing electricity consumption and different time intervals. Our experiment confirmed these setting can achieve better performance based on our proposed framework.

II. Related Work

In recent years, with the development of sensor became matures, more and more researchers focused on non-intrusive load monitoring (NILM) which uses total energy signal aggregated from a whole power monitor to estimate individual loads of appliances. There are many surveys [3, 17] focusing on disaggregation. Current studies for disaggregation can be divided into two categories according to sensor reading frequency. One is based on high sampling rate [4, 6, 10, 11, 13], using current waveforms and voltage noise as features to do pattern recognition. It belongs to immediate probe. The other is based on low sampling rate [1, 7, 8, 9, 15, 16], using aggregate power consumption to decompose. Generally, sampling rate more than 1 Hz is high-frequency sampling rate [1].

One of methods requiring high-frequency sampling rate is using harmonics of current to recognize. Lee et al. [11] complete the task by three main parts: collecting current waveforms of various appliances; building a database of load signatures for identifying individual loads; and then classifying and recognizing individual loads utilizing multi-features such as magnitude of current and length of durations

978-1-4799-2529-2/13 $31.00 © 2013 IEEE
DOI 10.1109/TAAI.2013.21

immediately. Leeb et al. [13] add a feature, starting pulses. Any two appliances with similar current waveforms exhibit different startup features so that it makes them easy differentiated. However, it is still unable to distinguish appliances from the same model. Ex: the same model air conditioner in different rooms. Gupta et al. [4] propose a new solution, ElectriSense, to solve this strait. Changes of electric power and current generate different levels of electromagnetic interference (EMI). ElectriSense relies on detecting the voltage noise signatures to recognize appliances. Moreover, EMI has a characteristic that other methods are absent: the signal is weakened when transmitted from the source of noise to the point of sensing. Thus, two identical appliances at two positions will generate identical EMI but look not the same because of different distances from the detection sensor, different degrees of recession. Approaches described above, all of them need huge signature databases and precision instruments to do disaggregation. These methods require high cost for the hardware, ElectriSense especially. Therefore, there are many studies focusing on low sampling rate disaggregation.

The popular strategy for disaggregation by low sampling rate instruments is based on hidden Markov model (HMM) [1, 7, 9, 15] that consider the probability of on and off of appliances. Many variants of HMM have been proposed. Kim et al. [7] propose four extended methods changed from HMM. The authors add factorial, conditional, and semi-Markov ideas based on HMM. Finally, the authors merge various models into conditional factorial hidden semi-Markov model (CFHSMM). Another method used in low sampling rate is disaggregation via discriminative sparse coding proposed by Kolter et al. [8]. The authors attempt to find best basis functions of each appliance to build dictionary. Then the authors let coding be sparse and arrange the combination to minimize disaggregation error. However, if similar basis functions exist in each appliance, sparse coding will be easy to get improper encoding coefficients.

III. Framework and Methodology

In this section, we will introduce our framework and methodology for energy disaggregation. In our framework, we first cluster appliances by the correlation between the using behavior of appliances, and select one of them as the key appliance in each cluster. By using the appliance with installed meter, we apply regression model to estimate the energy consumption for other appliances within each cluster. After the prediction of regression, we re-adjust the power consumption for each appliance to make their sum equal to the power consumption collected from the chief meter. The whole procedure is illustraed in Figure 1.

For better explanation, we first give a brief introduction to our data where a full explanation of dataset is described in Section 4.A. We have 13 kinds of appliances (A1-A13) and a chief meter (A_Total). Their electricity consumption

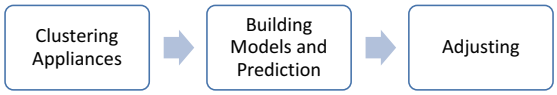


Figure 1. The flow chart of our framework.

is collected from April to December, a total of 275 days. Power data is returned by every minute. Our goal is to use the information from April to June to provide the usage statistics from July to December.

A. Clustering Appliances

To achieve energy disaggregation, an intuitional idea is direct sensing, which is installing smart meters at all appliances anywhere to measure consumption. Actually, it is not easy to be realized, and needs to pay expensive cost. Another way is only installing a smart meter in chief meter and disaggregating aggregating consumption collected from the chief meter. However, the information is insufficient for having high accuracy in disaggregation. This is a trade-off problem between reducing the installing number of smart meters and obtaining more information. Therefore, we propose clustered regression models by grouping high correlated appliances together. In our proposed method, we use key appliances to estimate other appliances in the same group.

1) Phase 1: Mean and Standard Deviation: Our clustering work starts with observing mean and standard deviation. Table I shows the mean and standard deviation of A1-A4. For instance, A1 uses 8.12 kWh every day and standard deviation of A1 is 0.07 in April. According to mean and standard deviation, the appliances can be divided into two categories, stable and intermittent appliances. Stable

Table I
The mean and standard deviation of each appliance from April to June.

Mean (kWh)	A1	A2	A3	A4
APR	8.12	8.19	58.24	1.44
MAY	8.11	7.57	58.12	1.48
JUN	8.08	7.52	57.57	1.55
Std. (kWh)	A1	A2	A3	A4
APR	0.07	0.25	0.20	0.29
MAY	0.06	0.06	0.43	0.21
JUN	0.05	0.05	0.49	0.34

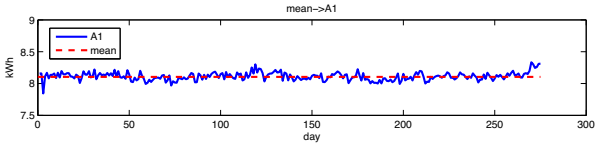


Figure 2. Appliance A1 is predicted by mean.

appliances including A1-A4, their electricity consumption

is steady and standard deviation is slight. This kind of appliances is not affected by seasonal changes so we can estimate them by past mean (see Figure 2).

2) Phase 2: Correlation Coefficient Matrix: After filtering in the first phase, we can divide appliances into stable and intermittent. Intermittent appliances are not as simple as stable appliances. They will be affected by day and night, seasonal changes and occasional discounts. Thus, it is not suitable to use the mean of past observations to estimate future measurements.

Here, we want to use regression through the relation between A and B, and let A estimate B or be estimated by B. Therefore, we need to find some key appliances as independent variables from these intermittent appliances. In this work, we adopt daily electricity consumption to produce correlation coefficient matrix and group high correlation appliances together. The correlation coefficient matrix is displayed in Figure 3.

$$C_{ij} = \frac{\sum_{k=1}^{m}(A_{ik}-\bar{A}_i)(A_{jk}-\bar{A}_j)}{\sqrt{[\sum_{k=1}^{m}(A_{ik}-\bar{A}_i)^2][\sum_{k=1}^{m}(A_{jk}-\bar{A}_j)^2]}}. \quad (1)$$

	A5	A6	A7	A8	A9	A10	A11	A12	A13	A_Total
A5	1.00									
A6	1.00	1.00								
A7	-0.12	-0.15	1.00							
A8	-0.06	-0.09	0.67	1.00						
A9	-0.07	-0.09	0.59	0.71	1.00					
A10	0.08	0.11	-0.40	-0.02	-0.14	1.00				
A11	0.16	0.20	-0.50	-0.22	-0.32	0.83	1.00			
A12	0.08	0.13	-0.65	-0.32	-0.36	0.73	0.74	1.00		
A13	-0.06	-0.01	-0.48	-0.33	-0.24	0.47	0.51	0.66	1.00	
A_Total	0.18	0.22	-0.62	-0.27	-0.32	0.80	0.82	0.98	0.65	1.00

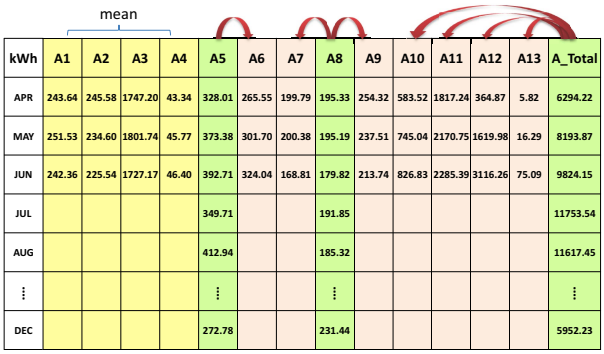
Figure 3. The correlation coefficient matrix.

According to the matrix, we can get three clusters, A5-A6, A7-A9 and A10-A13 with A_Total. We select an appliance which relations with each appliance are higher than others as independent variable from every cluster. From cluster 1 to 3, independent variables are A5, A8 and A_Total.

Over clustering, we will start our prediction work. Our goals are using the relation of appliances from April to June to predict energy consumption from July to December and producing the energy pie charts for consumers. As above, we estimate stable appliances by past mean and build regression models for intermittent appliances (see Figure 4). In order to avoid overfitting, we adopt ε-insensitive support vector regression (ε-SVR) rather than conventional regression. The ε-SVR will be illustrated clearly in next.

B. ε-insensitive Support Vector Regression (ε-SVR)

The ε-SVR has the fault-tolerant mechanism that tolerating small errors in fitting the given dataset linearly and nonlinearly. The loss function is presented in $(|A_i\mathbf{w}+b-y_i|_\varepsilon) = max\{0, |A_i\mathbf{w}+b-y_i|-\varepsilon\}$, that means if it is not over the critical range, the error will be



kWh	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A_Total
APR	243.64	245.58	1747.20	43.34	328.01	265.55	199.79	195.33	254.32	583.52	1817.24	364.87	5.82	6294.22
MAY	251.53	234.60	1801.74	45.77	373.38	301.70	200.38	195.19	237.51	745.04	2170.75	1619.98	16.29	8193.87
JUN	242.36	225.54	1727.17	46.40	392.71	324.04	168.81	179.82	213.74	826.83	2285.39	3116.26	75.09	9824.15
JUL					349.71			191.85						11753.54
AUG					412.94			185.32						11617.45
⋮					⋮			⋮						⋮
DEC					272.78			231.44						5952.23

Figure 4. The diagram of our prediction work.

zero. Figure 5 diagrams the ε-insensitive linear regression. This problem can be formulated as an unconstrained

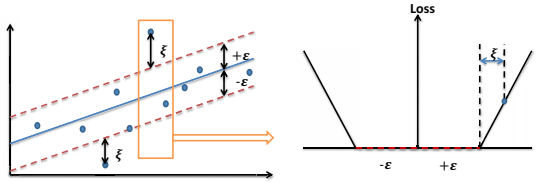


Figure 5. ε-insensitive linear regression.

minimization problem with squares of 2-norm ε-insensitive loss function given in Equation (2):

$$\min_{(\mathbf{w},b)\in R^{n+1}} \quad \tfrac{1}{2}\mathbf{w}^\top\mathbf{w} + \tfrac{C}{2}\textstyle\sum_{i=1}^{m}(|A_i\mathbf{w}+b-y_i|_\varepsilon)^2. \quad (2)$$

By utilizing the fast Newton-Armijo method, Lee et al. [12] proposed smooth ε-SVR (SSVR), which can be formulated as follows:

$$\min_{(\mathbf{w},b)\in R^{n+1}} \quad \tfrac{1}{2}(\mathbf{w}^\top\mathbf{w}+b^2) + \tfrac{C}{2}\textstyle\sum_{i=1}^{m} p_\varepsilon^2(A_i\mathbf{w}+b-y_i,\alpha) \quad (3)$$

where $p_\varepsilon^2(A_i\mathbf{w}+b-y_i,\alpha) = (p((A_i\mathbf{w}+b-y_i)-\varepsilon,\alpha))^2 + (p(-(A_i\mathbf{w}+b-y_i)-\varepsilon,\alpha))^2$. After finding $(\mathbf{w},b)$, the decision function can be expressed as follows:

$$\mathbf{y} \approx A\mathbf{w} + \mathbf{1}b. \quad (4)$$

In order to extend from linear cases to nonlinear cases, we also apply nonlinear SSVR in our experiments. As shown in [12], the nonlinear SSVR is formulated as follows:

$$\min_{(\mathbf{u},b)\in R^{m+1}} \quad \frac{1}{2}(\mathbf{u}^\top\mathbf{u}+b^2) + \frac{C}{2}\sum_{i=1}^{m} p_\varepsilon^2(K(A_i,A^\top)\mathbf{u}+b-y_i,\alpha), \quad (5)$$

where $K(\mathbf{x},\mathbf{z}) =< \Phi(\mathbf{x})\cdot\Phi(\mathbf{z}) >$ is the kernel function. Note that we use Gaussian (Radial Basis) kernel, which is

expressed in the following, in all our experiments.

$$K(A, A^\top)_{ij} = e^{-\mu\|A_i - A_j\|_2^2}, i, j = 1, 2, \ldots, m. \quad (6)$$

Similar to linear case, the decision function of nonlinear SSVR can be expressed as follows: (7):

$$\mathbf{y} \approx K(A, A^\top)\mathbf{u} + \mathbf{1}b. \quad (7)$$

C. Adjusting

After predicting by our clustered regression models, we obtain energy consumption of every appliance. It is worth noting that the sum of estimated energy consumption for A1-A13 will be not exactly the same with A_Total. The discrepancy is inevitable. To address this problem, we need to perform adjustments to make the sum of the estimated electricity consumption satisfy real total energy consumption. Thus, we can make a pie chart which correctly summed to 100% at the same time.

While total energy consumption and the sum of the estimated electricity consumption exist the discrepancy, we need to perform the adjustments. In our adjustments, we take the ratios of energy consumption of each appliance to the total consumption from April to June as weights which are used to redistribute the remaining energy. For example, A1 occupies 3% of total energy consumption from April to June. Suppose that there are 10 kWh exceeded, the estimated energy of A1 needs to be reduced 0.3 kWh. On the contrary, if there are 5 kWh not enough, A1 will be added 0.15 kWh. When complete adjusting, we finish all the processes of our framework.

IV. Experiments

A. Dataset and Experimental Setting

In our experiments, we evaluate our method by the data collected from a convenience store in Taiwan. All the data are collected from April 1^{st}, 2012 to December 31^{st}, 2012 (275 days in total). The smart meters will return consumption information per minute. The detailed information for appliances is listed in Table II. To evaluate our proposed method, we use the first 91 days as the training set for computing the correlation coefficient matrix and regression models, and use the remainder of data (184 days) as the testing set for the evaluation. Note that we set ε = 0.01 for SSVR since the instrument's precision is the second decimal place. For the nonlinear SSVR, we use Gaussian kernel as the kernel function where the width parameter is tuned by the cross-validation.

B. Different Segmentations for the Training Data

In the following, we will introduce three different ways to segment data. The three different settings will be applied to linear and nonlinear regression models, respectively.

Table II
The list of appliances.

Appliance Code	Appliance Name
A1	open refrigerator A
A2	open refrigerator B
A3	indoor lighting
A4	warehouse lighting
A5	corridor lighting
A6	signboard lighting
A7	coffee maker A
A8	coffee maker B
A9	microwave×3 and water dispenser
A10	combined refrigerator
A11	refrigerator's server
A12	air conditioner
A13	air conditioner in the warehouse
A_Total	the chief meter

*The codes described in this article can be contrasted with this list.

1) Setting A: As mentioned above, the smart meters return reading per minute. However, it can not reflect the relation between appliances when data granularity is too small. Thus, we use the power consumption, which is summarized by a day for our settings. In the setting A, we do not segment the training data. That is, we have 91 instances for each regression model (see Figure 6(a)).

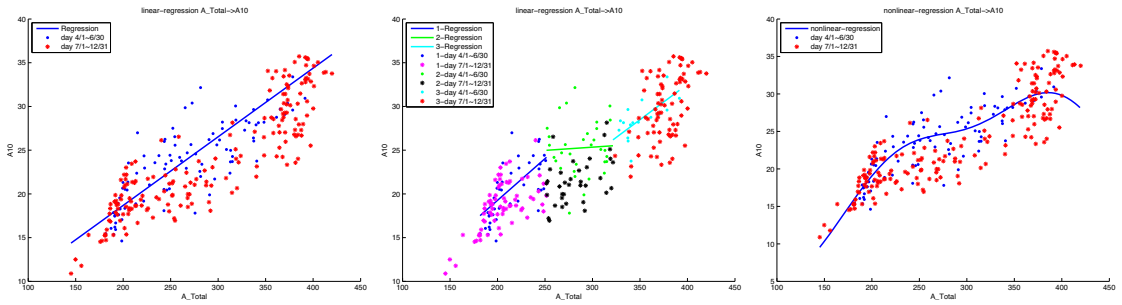
2) Setting B: We use piecewise regression [14] for the estimation in the setting B. The piecewise regression model splits the sample space into two or more sub-spaces to produce models for each sub-space. In this setting, we equally slice sensing electricity consumption into three sections and learn independent models at different levels (see Figure 6(b)). The effect of piecewise regression is similar to nonlinear regression (see Figure 6(c)), so that we also apply nonlinear models to each setting.

3) Setting C: One important factor will affect the usage behavior of appliances is time so we try to separate energy consumption by time-slices. In our experiment, we slice a day into 6 parts in order to close to our lifestyle: 02:00∼05:59, 06:00∼09:59, 10:00∼13:59, 14:00∼17:59, 18:00∼21:59, 22:00∼01:59.

C. Experimental Results

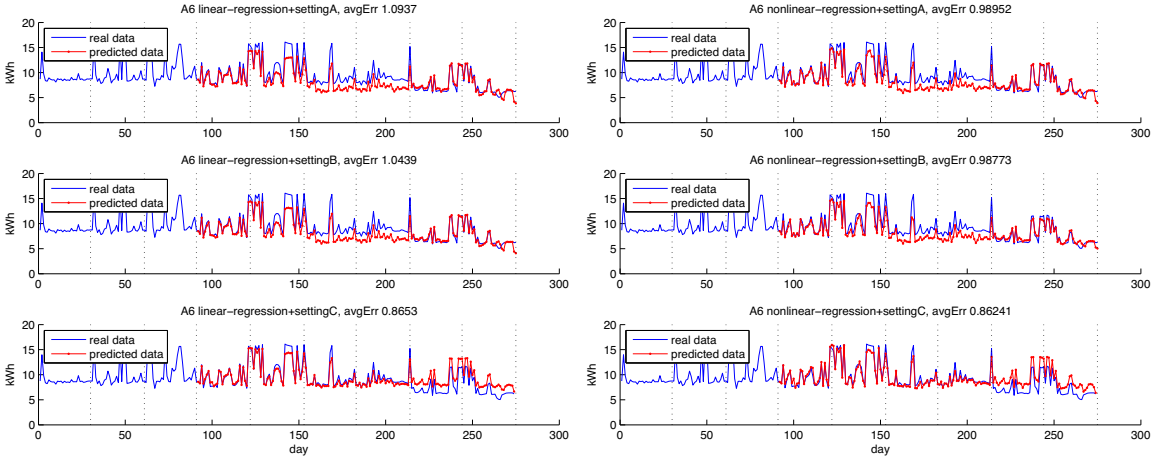
The disaggregation results are shown in Figure 7. Note that we only present the results for A6 due to the space limitation. Based on our experiments, the setting C can achieve better disaggregation results for most appliances. This indicates that the behavior of energy consumption for appliances is different during different time periods. On the other hand, we observe that nonlinear models have better results in estimating the energy consumption for larger power appliances, which contain complicated behavior of energy consumption.

As mentioned above, the setting C is the better solution so we only show the results of setting C in the following paragraphs. To provide different views for the evaluation, we



(a) (b) (c)

Figure 6. The diagram of (a) linear regression in setting A, (b) piecewise regression in setting B and (c) nonlinear regression in setting A.



(a) linear (b) nonlinear

Figure 7. Estimated result of signboard lighting for setting A, setting B, and setting C with linear and nonlinear regression models.

comparing the ranking of energy consumption for appliances between the actual ranking and the estimated one. The ranking value (RV) is defined as follows:

$$\mathrm{RV} = \frac{\sum \text{electricity consumption ordered correctly}}{\text{total electricity consumption}}. \quad (8)$$

The results are shown in Table III. It shows that the proposed method can produce a correct ranking ,which is able to occupies 80% of total energy consumption at least.

Table III
THE SORTING ORDER OF REAL DATA AND PREDICTED RESULT.

		Ranking Value			
JUL	AUG	SEP	OCT	NOV	DEC
97%	92%	95%	89%	80%	80%

*The linear and nonlinear regression have the same ranking result.

Besides the evaluation of ranking, we also compare our clustered regression models with an baseline mehtod, which directly uses past average to estimate energy consumption (see Figure 8). As shown in the figure, the baseline will produce much worse results by comparing with ours.

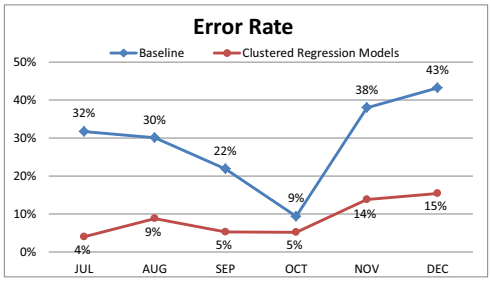


Figure 8. The error rate of baseline and our method.

For better visualization of the results, we also show the pie chart of energy consumption for a month (see Figure 9). Note that we also only show the result for July due to the space limitation. As presented in Figure 9, our models can produce a similar pie chart by comparing with the actual one. This is sufficient to make customers know their energy

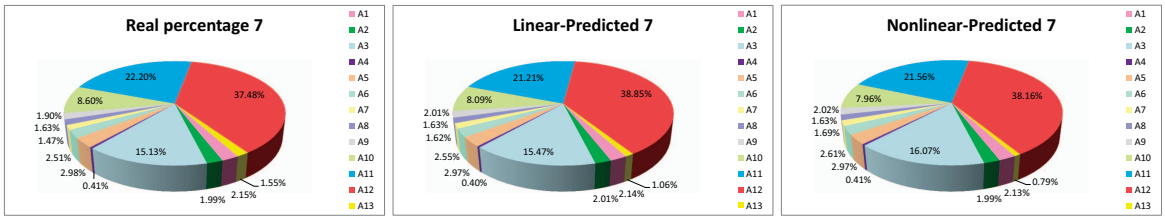


Figure 9. Energy consumption pie chart in July.

consumption distribution roughly.

V. Conclusion

In this article, we propose a novel framework for disaggregation with clustered regression models. First, we cluster high correlated appliances together and select key appliances, which are most correlated to other appliances, to estimate other appliances in the same group. Then, we build these clustered models via ε-insensitive SSVR and do post-adjustments to make the sum of the estimated electricity consumption satisfy real total energy consumption. In our results, maximum estimated error occurs in A3. We think it is because A3 occur twice unknown declined in November and December. If abnormal data are excluded, we will have 89% ranking value. Moreover, our framework can significantly reduce demand number of smart meters from 14 to 3 and without sacrificing too much accuracy. This also evidences that our proposed framework can be applied to practical usage.

Acknowledgment

This study is conducted under the Advanced Sensing Platform and Green Energy Application Technology Project of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

References

[1] F. Chen, J. Dai, B. Wang, S. Sahu, M. Naphade, and C.T. Lu. Activity Analysis based on Low Sample Rate Smart Meters. *ACM KDD*, 2011.

[2] J. Froehlich. Promoting Energy Efficient Behaviors in the Home through Feedback: The Role of Human-Computer Interaction. *HCIC Workshop*, 2009.

[3] J. Froehlich, E. Larson, S. Gupta, G. Cohn, M. Reynolds, and S. Patel. Disaggregated End-Use Energy Sensing for the Smart Grid. *Pervasive Computing*, 10(1):28–39, 2011.

[4] S. Gupta, M.S. Reynolds, and S.N. Patel. ElectriSense: Single-Point Sensing using EMI for Electrical Event Detection and Classification in the Home. *ACM ICUC*, 2010.

[5] H.A. He and S. Greenberg. Motivating Sustainable Energy Consumption in the Home. *Defining the Role of HCI in the Challenges of Sustainability Workshop*, 2009.

[6] D.A. Kelly. Disaggregating Smart Meter Readings using Device Signatures. *Thesis, Computing Science of Imperial College London*, 2011.

[7] H. Kim, M. Marwah, M. Arlitt, G. Lyon, and J. Han. Unsupervised disaggregation of low frequency power measurements. *SDM*, 2011.

[8] J.Z. Kolter, S. Batra, and A.Y. Ng. Energy Disaggregation via Discriminative Sparse Coding. In *NIPS*, 2010.

[9] J.Z. Kolter and M.J. Johnson. REDD: A Public Data Set for Energy Disaggregation Research. *SustKDD Workshop on Data Mining Applications in Sustainability*, 2011.

[10] C. Laughman, K. Lee, R. Cox, S. Shaw, S. Leeb, L. Norford, and P. Armstrong. Power Signature Analysis. *IEEE Power and Energy Magazine, IEEE*, 1(2):56–63, 2003.

[11] W.K. Lee, G.S.K. Fung, H.Y. Lam, F.H.Y. Chan, and M. Lucente. Exploration on Load Signatures. *ICEE*, 2004.

[12] Y.J. Lee, W.F. Hsieh, and C.M. Huang. ε-SSVR: A Smooth Support Vector Machine for ε-Insensitive Regression. *IEEE TKDE*, 17(5): 678–685, 2005.

[13] S.B. Leeb, S.R. Shaw, and J.L. Kirtley. Transient Event Detection in Spectral Envelope Estimates for Nonintrusive Load Monitoring. *IEEE Transactions on Power Delivery*, 10(3):1200–1210, 1995.

[14] V.E. McZgee and W.T. Carleton. Piecewise Regression. *Journal of the American Statistical Association*, 65(331):1109–1124, 1970.

[15] B. Wang, H. Dong, A. Boedihardjo, F. Chen, and C.T. Lu. A Hierarchical Probabilistic Model for Low Sample Rate Home-Use Energy Disaggregation. In *SIAM Data Mining Conference*, 2013.

[16] M. Zeifman. Disaggregation of Home Energy Display Data using Probabilistic Approach. *IEEE Transactions on Consumer Electronics*, 58(1):23–31, 2012.

[17] M. Zeifman and K. Roth. Nonintrusive Appliance Load Monitoring: Review and Outlook. *IEEE Transactions on Consumer Electronics*, 57(1):76–84, 2011.