Non-technical loss detection by multi-dimensional outlier analysis on the remote metering data

HAN Yuejun[1], LIU Fubin[2], XIN Jieqing[3*], MOU Tingting[3]
1. Shibei Electricity Supply Company, SMEPC, 200072 Shanghai, China; 2. East China Grid Company Limited, Shanghai 200120, China; 3. Department of EE, Shanghai Jiaotong University, 200240 Shanghai, China

Abstract—A multi-dimensional outlier analysis method is presented in this paper for electricity stealing suspect detection (ESSD). A preprocessing method is first presented to resolve the defects in remote metering data. After that, six features are extracted from the customers' daily consumption series, and a cluster-outlier interactive algorithm is applied for solving the ESSD problem. A case study with an estate located in the north of Shanghai, China is provided at the end of the paper. Results show that the presented method can search out the electricity stealers timely in their first months of consumption drop, and is of much higher search rate and much lower misjudgment rate relative to the one-dimensional outlier analysis method.

Index Terms—Non-technical loss analysis, remote metering, outlier detection, multi-dimensional.

I. I NTRODUCTION

ELECTRICY stealing suspect detection (ESSD) is crucial for non-technical loss management of an electricity supply company. Previously, electricity suppliers can hardly locate electricity stealing unless neighboring customers or meter readers report such activities. With the popularization of remote meter reading, electricity suppliers has less opportunity of timely catching abnormalities at the customers' side. Therefore, new methods are required for ESSD.

A series of data mining methods have been presented in the last decade for the ESSD problem, which can be classified into two kinds. One is *clustering based*, the basic idea of which is to take customers with load profiles far from medoids as suspects[1]. The other is *classification based*, which identifies suspects by a classifier previously trained based on historical load data of both fraudulent and normal customers. A classifier might be a decision tree[2], a neural network[3] or a support vector machine[4], etc. However, all such methods rely on the analysis of several months' historical data (typically monthly consumption in the past 12 months), implying that electricity stealing cannot be discovered in time.

Automatic remote metering (ARM), although reducing the chance of onsite inspection, actually provides at the same time more detailed load information that makes timely ESSD possible. A few works have done on the topic of ESSD based on the time series outlier detection technologies. In [5] for example, current or voltage outliers are detected to identify suspects in high-voltage customers. This method, however, is infeasible to low-voltage customers because only daily energy consumption are metered for such customers. Reference [6] presents a windowed Pearson coefficient analysis method to detect anomalous drops (as symptom of electricity stealing) in consumption series. However, consumption drop alone is not sufficient to claim electricity stealing, so decision tree is used in that paper as complementary technique for ESSD.

This paper focuses on ESSD in low-voltage customers among whom electricity stealing mainly happens. The multi-dimentional outlier analysis (MDOA) technique is trialed for this purpose. Unlike the existing methods, load features are extracted from the daily consumption series in a month so that suspects can be detected immediately after the first drop of consumption.

The paper is organized as follows. After describing the data preprocessing method in section II, six load features are presented in section III for describing every customer as a point in a multi-dimensional space. Section IV remarks a clustering-outlier iterative algorithm for searching multi-dimentional outliers. A case study is provided in section V towards an estate in the north of Shanghai, China, where search rate and accuracy (shown by misjudgment rate) of the presented method are examined and compared with those of the one-dimensional outlier detection.

II. DATA PREPROCESSING

Outlier detection based on load features extracted from daily consumption series enhances the timeliness of ESSD. However, the original metering data often contain detects caused by communication or device malfunction. It's therefore crucial to distinguish defects caused by malfunction and distortions caused by fraudulent activities in data preprocessing.

The ARM system collects for every low-voltage metering point the accumulated energy consumption (AEC) data frozen at 0:00 every day. For time-of-use (TOU) rate customers, metering data also include peak and off-peak AEC data. ESSD can be conducted monthly before billing, so data preprocessing is conducted towards the AEC series of the objective month by the steps as follows.

A. Missing Data Estimation

Missing data are filled in by the linear interpolation method. Let A_i^{before} be the last AEC data of the i^{th}

Project supported by State Grid Corporation of China (52091414007Y)

customer before the data missing period, A_i^{after} the customer's first AEC data larger than A_i^{before} after the data missing period and ω the number of days between them, then AEC value of the k^{th} day within the data missing period is evaluated as

$$A_{i,k} = A_i^{before} + \frac{A_i^{after} - A_i^{before}}{\omega+1} \cdot k, \quad k = 1 \cdots \omega \tag{1}$$

B. Noisy Data Smoothing

Fig.1 provides the typical AEC curves in a year of a fraudulent customer and several normal customers with data defects. It can be seen that AEC series of the electricity stealer and the normal customers are quite similar with each other, being both smooth and monotonically non-decreasing curves (here, the electricity stealer's AEC series drops at late July due to meter replacement after inspection). A sudden spurt or drop in the AEC series is therefore bound to be a data defect.

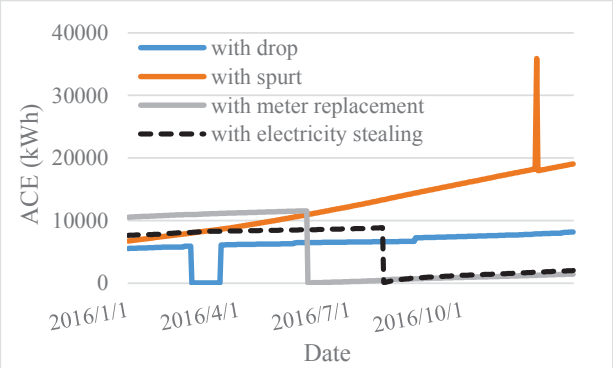

Fig.1 Typical AEC curves of an electricity stealer and several normal ones with different kinds of defect data

In this light, data smoothing is carried out as follows,

Step 1: Examine the meter replacement record. If the meter of a customer *i* is replaced in the j^{th} day of the objective month, then for every day *k* after this day, the AEC value is corrected as

$$A'_{i,k} = A_{i,j-1}^{before-rp} + A_{i,k} \tag{2}$$

Where, $A_{i,j-1}^{before-rp}$ is the AEC value of the last day before meter replacement, $A'_{i,k}$ and $A_{i,k}$ respectively the corrected and original AEC values in the k^{th} day ($k \geq j$).

Step 2: Search abnormal spurt in the AEC series. Sudden uprush of the AEC value of a customer *i* in a day *k* is viewed as data defect if and only if

$$E_{i,k} > 24 \cdot \alpha_i \cdot U_{ph,N} \cdot I_{i,k}^{max} \tag{3}$$

Where α_i is the phase number of the i^{th} customer, $U_{ph,N}$ the rated phase voltage and $I_{i,k}^{max}$ the maximum current of the meter. The AEC values during the spurt periods are corrected by (1) with A_i^{before} and A_i^{after} respectively as the AEC values of the day before and after the spurt period.

Step 3: Search drop points within the AEC series. If there's a drop in the AEC series of a customer *i*, find the AEC value of the last day before the drop (A_i^{before}) and the first one larger than A_i^{before} after the drop period, correct the AEC values of the days during the drop period by (1).

C. Data Transformation

After data smoothing, the AEC series are transformed into daily energy consumption (DEC) series, the DECs being calculated as the differences of adjacent days. Load feature selection and outlier detection are further conducted towards the DEC series.

III. FEATURE SELECTION

According to the data performance possibly caused by the fraudulent customers, six features are extracted for the outlier detection purpose. Every index is normalized into [0, 1], approaching 1 implying more possibility of being an abnormal user.

A. The Consumption Level Index (CLI)

Most electricity stealing activities lead to the drop of monthly energy consumption. So, low level of monthly consumption is an important clue for electricity stealing. The CLI is hence defined as

$$l_i = \begin{cases} \dfrac{\overline{E} - \overline{E}_i}{\overline{E} - \min\limits_k E_k}, & \overline{E}_i < \overline{E} \\ 0, & else \end{cases} \tag{4}$$

Where, $\overline{E}_i$ and $\overline{E}$ respectively denote the average DEC value of the i^{th} customer and all the low-voltage customers in the objective month. Equation implies that it's worth taking customers with DECs below the average level as suspects, and their possibilities of electricity stealing are inversely proportional to their levels of monthly consumption.

B. The Daily Temperature Relevance Index (DTR)

In months from April to October, DEC normally increases with the increase of temperature (i.e. the DEC's temperature correlation coefficient $R_i \in (0,1]$) while in the other months, DEC usually increases with the decrease of temperature (i.e. $R_i \in [-1,0]$). Contrary phenomena might relate to fraudulent activities. Hence, the DTR index is defined as

$$\tau_i = \begin{cases} (1-R_i)/2, & for\ April \sim October \\ (1+R_i)/2, & for\ other\ months \end{cases} \tag{5}$$

For months from April to October, τ_i approaches 1 when R_i approaches -1, while for other months τ_i approaches 1 when R_i approaches 1.

C. The Monthly Temperature Relevance Index(MTR)

Unlike the DTR index which reflects the abnormal degree in the correlation of a customer's daily energy consumption with the daily average temperature, the MTR index is used to show the reasonability of the change of a customer's monthly energy consumption relative to the change of the monthly average temperature. The MTR of customer *i* is calculated as

$$m_i = \begin{cases} \dfrac{E_i^{M-prev} - E_i^M}{E_i^{M-prev}}, & E_i^M < E_i^{M-prev} \ and \ \delta = 1 \\ 0, & else \end{cases} \tag{6}$$

Where, E_i^M and E_i^{M-prev} are respectively the monthly energy consumption of the i^{th} customer in the objective month and the month before. δ denotes a calculation

condition, which equals 1 if the average temperature of the objective month is higher (for April~October) or lower (for other months) than that of the month before. Equation (6) implies that drop of consumption in hotter summer months or in colder winter months probably relates to fraudulent activities, and the probability is proportional to the rate of decrease of the customer's monthly energy consumption.

D. The Daily Consumption Variation Index (DCV)

Consumption variation is measured by the standard deviation of the DEC data. Sharply drop of DEC constructs a high variation but after that, DEC series usually manifests low variation. In the light of this, the DCV index is defined as

$$v_i = \left|\sigma_i - \overline{\sigma}\right| \Big/ \max_k \left|\sigma_k - \overline{\sigma}\right| \tag{7}$$

Where, σ_i is the DEC standard deviation of the i^{th} customer, while $\overline{\sigma}$ the average DEC standard deviation of all the low-voltage customers in the objective month. Equation means that the more the DEC standard deviation deviates from its average value, the more likely the corresponding customer is to be an electricity stealer.

E. The Chain Growth Rate Index (CGR)

The CGR index shows the relative reasonability of a customer's chain growth rate of monthly energy consumption compared to the others. For customer *i*, the CGR index is calculated as

$$f_i = \begin{cases} 0 \quad , E_i^{M-prev} < 5 \text{ or } \eta_i > 0 \text{ or } \overline{\eta_-} < \eta_i < 0 \\ \dfrac{\overline{\eta_-} - \eta_i}{\overline{\eta_-} - \min\limits_k \eta_k} \quad , \text{else} \end{cases} \tag{8}$$

Where, η_i is the chain growth rate of monthly energy consumption of customer *i* while $\overline{\eta_-}$ the average value of all the low-voltage customers' chain growth rates of monthly energy consumption. The first line of implies that there's little possibility for a customer of little consumption (<5kWh) in the previous month or for a customer whose consumption doesn't change apparently to start electricity stealing in the objective month. For a customer with considerable consumption decrement however, as shown in the second line of, his possibility of electricity stealing is proportional to the drop rate of consumption.

F. The Valley Consumption Ratio Index (VCR)

This index is considered only for TOU rate customers, since such customers might evade electricity charges by tampering with their peak/ off-peak consumption. VCR of a customer *i* is defined as

$$g_i = \begin{cases} \dfrac{E_i^{M-G}}{E_i^{M-F} + E_i^{M-G}} \quad , E_i^M \neq 0 \\ 0 \qquad\qquad\qquad , else \end{cases} \tag{9}$$

Where, E_i^{M-F} and E_i^{M-G} are respectively the customer's energy consumption in the peak and off-peak periods. The higher the proportion of valley consumption he has, the more possible he is to be a suspect.

IV. THE OUTLIER DETECTION ALGORITHEM

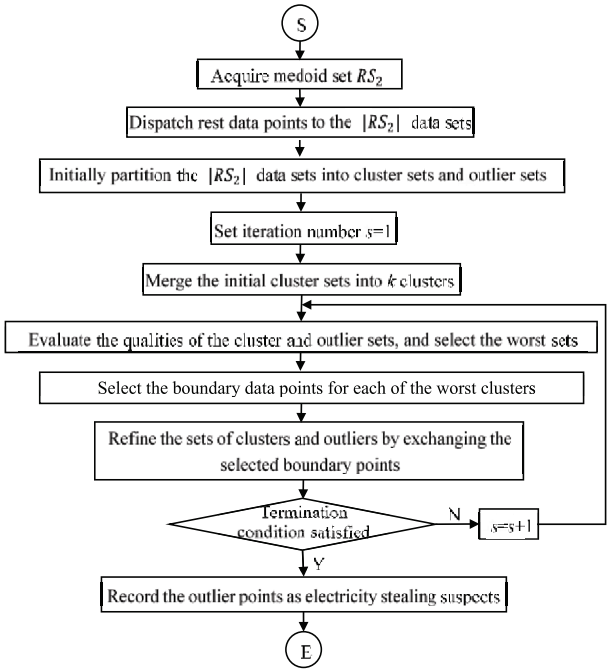

Fig. 2 Flowchart of the COID algorithm for multi-dimensional outlier detection

The cluster–outlier iterative detection (COID) approach in [7] is used for solving the MDOA problem. The flowchart of the algorithm is shown in Fig.2. A brief description is as follows,

Step 1: *Acquiring original medoid set RS_2*. To form RS_2 as a well-scattered data set, a random set RS_1 is first chosen from the original data set. RS_2 is then constructed by selecting for each time a data point in RS_1, which is farthest to its closest data point in RS_2. Here, both RS_1 and RS_2 should be multipliers of the objective number of clusters *k*.

Step 2: Dispatching the data points left in RS_1 to the original medoids. The medoid a point belonged to is the nearest one in RS_2. The Manhattan distance shown below is used to measure the distance between two points (x_1 , x_2) in the *n*-dimensional space:

$$d\left(\vec{x}_1, \vec{x}_2\right) = \sum_{i=1}^{n} \left(\left(x_{1i} - x_{2i}\right)^{0.1} \right)^{10} \tag{10}$$

Step 3: Initially partitioning the data sets into cluster and outlier sets. After the previous step, there're $|RS_2|$ numbers of data sets, E_j ($j = 1 \cdots |RS_2|$). To partition these sets into clusters and outliers, calculate

$$T_j = \left|E_j\right| \Big/ d_t\left(E_j\right), \quad j = 1 \cdots |RS_2| \tag{11}$$

Where, $\left|E_j\right|$ is the size of E_j, $d_t\left(E_j\right)$ is the diversity between E_j and its nearest set E_s. Diversity shows the distance between two clusters, whose detailed definition can be found in [7]. The smaller the size of E_j is and the farther it is from other data sets, the smaller T_j is and the more possible the cluster E_j is to be an outlier set. So, sort E_j ($j = 1 \cdots |RS_2|$) by T_j in ascending order and select the sets whose T_j are below the first sharp point in the ascending T_j series as outlier sets, leaving others as normal cluster sets.

Step 4: Merging the initial clusters into k clusters. Merge for each time two nearest clusters until the total number of clusters equals to the objective value *k*.

Step 5: Evaluating the clusters and outliers. The quality of a cluster and an outlier set is defined respectively as

$$Q_c(C)=\left[\frac{\sum_{C'\neq C} D_2(C,C')}{k_c-1}+\frac{\sum_{O} D_1(C,O)}{k_o}\right]\Big/ CPT(C) \quad (14)$$

$$Q_O(O)=\left[\frac{\sum_{O'\neq O} D_3(O,O')}{k_O-1}+\frac{\sum_{C} D_1(C,O)}{k_C}\right] \quad (15)$$

Where, k_O and k_C are respectively the number of cluster sets and outlier sets. D_1 denotes the diversity between a cluster and an outlier while D_3 the diversity between two outliers. Detailed definition of D_1 and D_3 can be found in [7]. Obviously, the larger the diversity of cluster *C* is from the outliers and other clusters, the higher quality it has. Similarly, the larger the diversity of an outlier set *O* is from the other outliers and the clusters, the higher quality it has.

Step 6: Selecting the boundary data points for the worst cluster. Boundary data points of a cluster are those with the far distance to the medoids and having small number of neighboring points. Hence, define

$$b_j = d_{ij}/\varsigma(j) \quad (16)$$

Where d_{ij} is the distance between a data point $\bar{x}_j$ and the medoid it belongs to, while $\varsigma(j)$ the number of data point within a certain radius of x_j. Sort the data points within the worst cluster by b_j in descending order, and choose the points whose b_j are larger than the first sharp downward point in the descending b_j series as boundary data points.

Step 7: Refine the cluster sets and outlier sets. Assign the worst outlier set to its nearest cluster, and change each of the boundary data points of the worst cluster as a new outlier set.

Step 8: Examine the termination condition. Repeat step 5~7 until either of the following termination condition is satisfied: the data points within each clusters and outliers do not change dramatically anymore, or the uplimit of the iteration number is reached. The final data points left in the outlier sets are viewed as electricity stealing suspects.

V. CASE STUDY

Experiment is conducted with a housing estate located in the north of Shanghai, China. There're 881 residential customers in the estate, to whom a carpet inspection without prior notice was conducted in July 2nd, 2015. Totally 16 customers were found stealing electricity. By observing the DECs of these customers in the previous months, it's realized that they started electricity stealing from different months (as shown in TABLE 1).

TABLE 1 Observed start months of electricity stealing of the 16 customers

Customer ID	Month	Customer ID	Month
1116282784	Feb.	2002983682	Apr.
2002965382	Feb.	2002983684	Apr.
2002967930	Mar.	2002659556	Apr.
2002659433	Mar.	1116276635	Apr.
2002965605	Mar.	2002976475	Apr.
2002965609	Apr.	2002976496	May
1116283917	Apr.	2002965383	May
2002965323	Apr.	2002967908	Jun.

A. ESSD result

For each month from February to June, 2015, six indices of every customer are evaluated based on the DEC series. After that, MDOA is conducted month by month. All the 16 fraudulent customers are searched out as suspects in their start months of electricity stealing. Apart from them however, there're also other customers judged to be suspicious of stealing (as listed in TABLE 2). TABLE 3 gives the statistical data by comparing TABLE 1 and 2, where *N* is the total number of customers being detected, N_{ss} and N_{sn} the numbers of the customers within the 16 stealers while judged respectively as suspects and non-suspects by MDOA, N_{ns} and N_{nn} numbers of the customers outside the stealer list while judged respectively as suspects and non-suspects by MDOA. Taking the five months' detection results as a whole, the search rate reaches 100% and the misjudgment rate is only 1.93%.

TABLE 2 Detected suspects of the five months

Month	Customer ID of the suspects
Feb.	1116283057,1107201543, 1116283704
Mar.	1116283141, 1116279242, 1116284833
Apr.	1142181902, 1116276525, 1116283205, 1116283057, 1116280253, 1116283064,
May	1116283536, 1116283769, 1116283141
Jun.	1116283536, 1142186170

TABLE 3 Statistics of the ESSD results

Detection result / Inspection result	Suspect	Non-suspect
Stealer	N_{ss}=16	N_{sn}=0
Non-stealer	N_{ns}=17	N_{nn}=848
Search rate=N_{ss} /(N_{ss} + N_{sn})		100%
Misjudgment rate=(N_{ns} + N_{sn})/N		1.93%

It's worth notice that carpet inspection hasn't been conducted from April to June in that year, whereas a customer might not steal electricity continuously. Taking the two customers in Fig.3 for example, the DECs of a misjudged suspect (1142181902) might be very similar with that of a verified stealer (2002965609). So, customers outside TABLE 1 should not be definitely excluded from suspicion of stealing. In this sense, accuracy of the MDOA method might be better than what it looks like in TABLE 2.

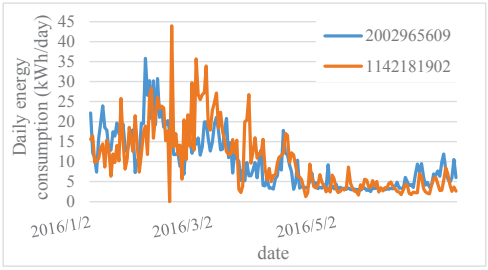

Fig. 3 Daily energy consumption curves of a misjudged suspect and a verified stealer

Another phenomenon is that electricity stealers are searched out only in their first month of stealing. This is because two features, MTR and CGR, relate to the extent of decrease of the monthly energy consumption. Since the monthly consumption of a stealer rarely drops sharply again after the first drop due to electricity stealing, there's little chance for a stealer to be discovered repeatedly in the subsequent months.

B. Comparison with one-dimensional detection

Under one-dimensional detection, the consumption level index is the only feature for outlier analysis. In this case, outlier number is related with the threshold of CLI. Fig. 4 shows the misjudgement rates of one-dimensional outlier analysis under various thresholds of CLI.

It can be seen that both search rate and misjudgment rate rise steadily with the increase of the threshold of CLI. However, search rate only reaches 18.8% even under the threshold of 70% (i.e. customers with CLIs higher than 70% of the customers are viewed as suspects), being much lower than that of MDOA. Similarly, the misjudgment rate is much higher than that of MDOA, being 15.3% even under the threshold of 90%. This is because low level of monthly consumption alone is neither a sufficient nor necessary evidence for electricity stealing. Vacant houses also features in low consumption while a customer stealing electricity intermittently might still use a considerable amount of energy every month.

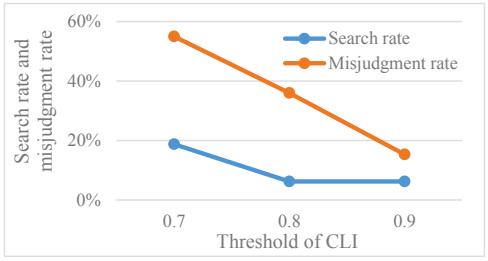

Fig.4 Search rate and misjudgment rate of one-dimensional outlier detection under various thresholds of CLI

IV. CONCLUSIONS

A multi-dimensional outlier analysis (MDOA) method is presented for electricity stealing suspect detection purpose. Six features are extracted from customers' daily consumption series collected by the ARM system. A cluster-outlier iterative detection algorithm is used to solve the problem. A case study towards an estate in Shanghai shows that the presented method can precisely search out the stealers with a misjudgment rate much lower than that of the one-dimensional outlier analysis. The MDOA method also features in searching out the suspects in their first month of consumption drop, which makes timely onsite inspection possible and avoids financial loss before billing.

REFERENCES

[1] A. H. Nizar, Z. Y. Dong, M. Jalaluddin and M. J. Raffles. Load profiling method in detecting non-technical loss activities in a power utility[C]// Proceedings of the first International Power and Energy Conference PECon 2006, November 28–29, 2006, Putrajaya, Malaysia: 82–87.

[2] J. R. Filho, E. M. Gontijo, E. Mazina, J. E. Cabral, J. O. P. Pinto and A. C. Delaíba. Fraud identification in electricity company customers using decision tree[C]// Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics, October 10-13, 2004, Hague, Netherlands: 3730–3734.

[3] M. M. Badjian, J. Nagi, S. K. Tiong, K. S. Yap, S. P. Koh and F. Nagi. Comparison of supervised learning techniques for non-technical loss detection in power utility[J]. International Review on Computers and Software, 2012, 7(2): 626-636.

[4] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed and M. Mohamad. Nontechnical loss detection for metered customers in power utility using support vector machines[J]. IEEE Transactions on Power Systems, 2010, 25(2): 1162–1171.

[5] C. Cheng, H. J. Zhang, Z. M. Jing, M. Chen, L. Jiao and L. X. Yang. Study on the anti-electricity stealing based on outlier algorithm and the electricity information acquisition system[J]. Power System Protection and Control, 2015, 43(17): 69-74.

[6] I. Monedero, F. Biscarri, C. León, et al. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees[J]. International Journal of Electrical Power & Energy Systems, 2012, 34(1): 90-98.

[7] S. Yong and L. Zhang. COID: A cluster–outlier iterative detection approach to multi-dimensional data analysis[J]. Knowledge and Information Systems, 2011, 28(2011): 709-733.

HAN Yuejun (1981-), male. He got master degree in 2011 and is currently a senior engineer of Shibei Electricity Supply Company, SMEPC. His research interest includes management and data analysis of remote metering. Mail address: No. 2511, Gonghe Xin Road, Shanghai 200070, P. R. China. Email: hawking_1981 @hotmail.com; Tel: 0086-13816075742.

LIU Fubin (1975-), male, PhD. He got PhD in 2003 and is presently a senior engineer of East China Power Grid Company Co., Ltd. His research interest includes power market and economy in power system operation. Mail address: No.882, Pudong South Road, Shanghai 200120, P.R.China. Email:34984004@qq.com; Tel: 0086-13482704511.

XIN Jieqing (1973-), female, PhD, corresponding author. She got PhD in 2003 and is presently an associate professor in the Department of EE, Shanghai Jiaotong University. Her main research interest includes demand side management

and power market. Mail address: No. 800, Dongchuan Road, Shanghai 200240, P.R.China. Email: jqxin@sjtu.edu.cn; Tel: 0086-13817871308.

MOU Tingting (1992-), female. She is presently a master student in Shanghai Jiaotong University. Her research interest include data mining analysis on the remote metering data. Mail address: No. 800, Dongchuan Road, Shanghai 200240, P.R.China. Email: 1534849507@qq.com; Tel: 0086-13262602628.