Urban Residential Load Combined Forecast Model Based on Data Mining Techniques and Panel Data Theory

Yongxiu HE†, Yuejin WANG, Tao LUO, Haiying HE, Jing WANG

College of Economic and Management, North China Electric Power University, Beijing, 102206, China

Abstract

As a result of rapid urban economic development and improvement in living standards, urban residential electricity consumption in China is increasing quickly. Although the factors which influence the urban residential load are complex, an objective analysis followed by the setting up of a logical urban residential load forecast model can offer a scientific basis for decisions regarding urban power planning and demand-side management. Firstly, based on data mining techniques, association rules mining of residential load were carried out on the relevant data for nine typical cities in China during the period 1992-2006 and the primary factors influencing the residential load were obtained, which avoids the forecast error coming from a subjective choice of factors. Next, the urban residential load was analyzed based on these factors and an urban residential load forecast model was set up based on panel data theory. The model considered not only the time effect of data but also the cross section effect, which can overcome the limitation of data deficiency. Finally, the combined forecast model is proved efficient and reliable in urban residential load forecast by comparing the forecast errors of this model with those of other models based on case studies of typical cities in China.

Keywords: Urban; Residential Load; Data Mining; Panel Data; Forecast

1. Introduction

With the development of the national economy, the proportion of urban residential power consumption to the total urban power consumption in China keeps on rising. For this reason, the forecasting of the level of residential load is of great concern to researchers. The residential load has its own inherent characteristics: it is impacted by residential living standards, habits, environment, etc. Common load forecasting methods include neural networks [1,2], the parameters optimization method of Gaussian functions based on the direction of optimal search [3,4], SVM theory [5–7], models combining features of pattern recognition and artificial neural networks [8] and so on, but the methods referred to above are usually used in short-term peak load forecast. There are also some other methods like data mining (DM), space load forecast based on the optimization of parameters, improved grey prediction model, etc. [9-11]. Although these methods have ameliorated the problems due to limitations in the integrity of the data set, they have not treated the recent changes in residential electricity structure perfectly [12]. Residential load forecasting methods focus more on multiple regression models [13], error correction models [14], dynamic analysis models [15], etc. The effect-factors considered are mainly the variation in monthly average ambient temperature [16], holiday effects [17, 18], weather conditions [19–21] and so on. However, these load forecast models chose the

† Corresponding author.

Email addresses: heyongxiu@ncepu.edu.cn (Yongxiu HE).
influencing factors subjectively, without systematic analysis; moreover, they lack association statistics data for the residential load. Differences in the residential load characteristics exist between different periods of time and different regions. How to scientifically establish a residential load forecast model needs thorough consideration.

The above points show in what ways the forecasting of urban residential load in China has to be further elaborated. DM, which collects the achievements of machine learning, pattern recognition, statistics, artificial intelligence, information management systems and so on, can be used to analyze the factors influencing the urban residential load through association rules mining. It can determine the key factors capable of decreasing the forecast error in comparison with a merely subjective choice of factors. Furthermore, because the relevant statistical data about residential load in China which is needed for the usual methods of panel data analysis is lacking and the potentially influential factors are many, the use of panel data (PD) theory should consider the section effect and time effect of every impact factor synthetically. This will have greater practical significance for modeling urban residential load in China.

This paper is organized as follows. Section 2 briefly reviews panel data theory and the technique of association rules mining. Section 3 sets up our combined forecast model of urban residential load based on DM techniques and panel data theory and make a case study of typical cities in China. Section 4 offers conclusions and proposals for further studies.

2. Association Rules Mining and PD Theory

2.1. Association Rules Mining

DM is a method of discovering previously unknown rules and relations from a large database through selection, exploring and modeling. It aims at obtaining clear and useful results for the data owner. Association rules mining is for extracting interesting relational information, i.e., association rules between item sets in the data. Over one or two decades, association rules mining have matured to become a very important component of DM techniques.

Suppose that transaction sets have s% of their transactions supporting item set X (such as influencing factors of residential load) and Y (such as residential load) simultaneously. Then s% is referred to as the support degree of the association rule \( X \Rightarrow Y \). Support degree describes the probability that the union of X and Y happened in the set of all transactions, as shown in formula (1).

\[
support(X \Rightarrow Y) = support(X \cup Y)
\] (1)

If c% of the transactions in X not only support item set X but also item set Y, c% will be called the confidence level of the association rule \( X \Rightarrow Y \). Simply put, the confidence level indicates the probability of Y’s being present in X (within the transaction set \( T \) in which the item set \( X \) appears), as shown in formula (2). The confidence level expresses the reliability of a rule.

\[
confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \times 100\%
\] (2)

If the support degree of \( X \) is not smaller than a certain minimum support degree which may be chosen by the user, then \( X \) is called a frequent item set (or a large item set), otherwise it is called a non-frequent set.
The Apriori algorithm is the classic algorithm of association rules mining, whose core ideology is the dividing of the work of finding out the association rules into two steps: first, all frequent item sets in the transaction database are found through iteration. Second, those rules from the frequent item sets which satisfy the user’s minimum trust are set up. The core of the Apriori algorithm is the mining (i.e., discovering) of all frequent item sets, namely, those which account for the largest part of the calculation. A recurrence method is used to create all the frequent item sets. The Apriori algorithm was described, simply, as follows:

1. $L_1=$\{the first frequent item set\};
2. For $(k=2; \ L_{k-1} \neq \emptyset ; k++)$ do begin
3. $C_k=$apriori-gen($L_{k-1}$); // the new candidate item set
4. For transaction set $t \in D$ do begin
5. $C_t=$subset($C_k$, $t$); //the candidate sets in the transaction set $t$
6. For every candidate set $c \in C_t$ do
7. c.count++;
8. End
9. $L_k=$\{ $c \in C_k$ | c.count$\geq$support$_{min}$\}
10. End
11. $Answer = \bigcup_{k=1}^{n} L_k$

Here, $D$ is the transaction database; $t$ is the transaction set; $L_k$ is the collection of $k$th frequent item sets; $C_k$ is the candidate $k$th item sets (a collection being considered by the algorithm for its potential to be a frequent item set), and support$_{min}$ is the minimum support degree.

The procedure of the algorithm is to start with a first frequent item set $L_1$ and then generates the second frequent item sets $L_2$, and so on. As soon as $L_r$ is the empty set, the algorithm ends. In the $k$th iteration, the process begins by generating candidate $k$-item sets $C_k$. Each such candidate item set is generated by two frequent sets belonging to $L_{k-1}$ and only contains one different item through a $(k-2)$ connection. The item sets in $C_k$ are candidate sets used to generate frequent item sets. The frequent item set $L_k$ will be a chosen subset of $C_k$. Every element in $C_k$ must be tested in the database to make the decision whether the element should join $L_k$.

The key factors influencing the increasing residential electricity consumption will be determined through this process of mining association rules.

2.2. Panel Data Theory

Residential load data and its key influencing factors are usually in the form of panel data. Panel data refers to data sets for a cohort of agents, which may be individuals or aggregated data for an entire city or region, gathered over a period of time and indexed by both the time and cohort variables. It is a multi-dimensional time series coming from the continuous observation of cross sections. The synthesis data includes information as to time, cross-section and index. There are several models for the analysis of panel data: pooled regression models, fixed effects models, random effects models, etc.

In order to establish panel data model, a test could be carried out based on EVIEW 6.0, which is the likelihood ratio test, or the Hausman test, and so on, that test should be carried out to test whether to use a
fixed or random effects model. The unit root test should be carried out before selecting a panel data model. The Fisher-ADF unit root test was used in this paper.

3. Urban Residential Load Combined Forecast Model Based on DM and PD Theory

The problem of urban residential load combined forecast is studied in this part through the analysis of typical cities in China. 20 factors related to residential load during the period 1992-2007 of nine typical cities in China were used for the association rules mining and then the load was forecast, based on the combined model. All the data are from each city’s statistics yearbook [22-30]. Some missing data-points were estimated through the moving average method. There are many factors affecting urban residential load. With a view towards the feasibility of research and the availability of the data, the factors chosen were as follows.

Nomenclature

Acronyms

| BJ | Beijing     | SH | Shanghai   | TJ | Tianjin   |
| CQ | Chongqing   | NB | Ningbo     | FZ | Fuzhou    |
| XA | Xi’an       | HEB| Harbin     | JN | Jinan     |

Symbols

| DL  | urban residential load                      |
| AWD | the average temperature                     |
| BG  | the number of freezer ownership per one hundred urban households |
| DBX | the number of refrigerator ownership per one hundred urban households |
| DCJ | the number of electrical cooker ownership per one hundred urban households |
| DN  | the number of computer ownership per one hundred urban households |
| JY  | the number of the employees in tertiary industry |
| KT  | the number of air conditioner ownership per one hundred urban households |
| M   | the number of males                          |
| PERS| housing area per capita                      |
| PYYJ| the number of exhaust fan ownership per one hundred urban households |
| SR  | disposable income per capita                 |
| TRQ | the amount of gas consumption for living per capita |
| TV  | the number of TV ownership per one hundred urban households |
| WBL | the number of microwave oven ownership per one hundred urban households |
| XCQ | the number of vacuum cleaner ownership per one hundred urban households |
| XNY | the renewable energy consumption used by the residents (giving priority to solar energy ) |
| XYJ | the number of washing machine ownership per one hundred urban households |
| ZC  | fare for water, electricity and fuel per capita |
3.1. Association Rules Mining of Urban Residential Load

The association rules mining of urban residential load was implemented in the R programming language. The core algorithm sentence of the association rules mining within the R language was:

```r
>rules<-apriori (residential, parameter=list (support=0.89, confidence=1))
```

In this situation, residential was a pretreatment data document. Because each factor is distributed in respectively different value intervals, the source data first underwent a process of normalization for convenience. 0.905 was chosen for the confidence level in the mining process. There are 63 rules with support degree up to 0.905 and confidence level up to 1 mined after the process of Apriori algorithm. These rules were sorted according to the mined frequent item sets as follows:

\{KT, M, SR, TRQ, TV, ZC, XNY\} ⇒ \{DL\} support=0.905

The result indicates that the changes of SR, M, ZC, TRQ, TV, KT, and XNY have a considerable influence on the urban residential load in 91% of the residential families. Consequently, the seven factors could be considered as the key factors influencing urban residential load.

3.2. Urban Residential Load Forecast Based on DM and PD

Based on the results of DM analysis and PD theory, the urban residential load in China’s typical cities and the key influencing factors were analyzed. Then the relationship model between the urban residential load and its key factors was set up to forecast the load.

3.2.1. Unit Root Test of the Load and the Key Influencing Factors

The ADF-Fisher Chi-square Unit root test was used to test for stationarity of the panel data of the load and the key influencing factors.

The results of the ADF panel data unit root test illustrated that all the panel data are first order stationary series. According to the definition of the co-integration, DL, SR, M, TRQ, ZC, XNY, KT and TV are all stationary series, which indicates that a co-integration relation may exist among the series.

3.2.2. Random and Fixed Effects Test

It is unsuitable to adopt pooled regression model because the differences among the typical cities are obvious. Furthermore, the likelihood test ratio was used to test whether to use a random effects or a fixed effects model for the relation between the load and the influencing factors. The results are shown in Table 1.

<table>
<thead>
<tr>
<th>Test Methods</th>
<th>Statistic Value</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groups-F statistics</td>
<td>62.6647</td>
<td>0.0000</td>
</tr>
<tr>
<td>Groups-Chi square</td>
<td>236.7419</td>
<td>0.0000</td>
</tr>
<tr>
<td>Groups-time F Statistics</td>
<td>26.2303</td>
<td>0.0000</td>
</tr>
<tr>
<td>Groups-time Chi square</td>
<td>253.6076</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
According to the results, the probabilities are approximately zero in all four test methods. This shows that the model rejects the null hypothesis and accepts the alternative hypothesis, namely, the validity of the entity fixed effects model. This illustrates that the entity fixed effects model can be used for forecasting.

3.2.3. Model Coefficient Estimation

According to the above tests, the panel data entity fixed effects model estimation was made for the urban residential load. From the results, the model’s estimated parameters’ R-squared is 0.987, which is built after taking the logarithm of each variable. The adjusted R-squared is 0.986. The Durbin-Watson statistic is 0.617. The whole exact P value is 0.00. So the model generally passes the tests. The coefficients of TRQ and XNY were negative, which illustrates that there was a negative correlation between the urban residential consumption of gas and renewable energy per capita and electricity load. More precisely, the residential load decreased by 0.224 units for every unit increase in the residential consumption of fuel gas per capita, and the residential load decreased 0.046 units in the residential consumption of new energy. It cannot be ignored that this amount is very small. The coefficients of SR, M, ZC, KT and TV were positive, which illustrates that the relationship between the factors and the urban residential load was that of a positive correlation. All these partial regression coefficients have passed with the significant level of 5%. All the chosen model variables pass the test, which illustrates that the seven influencing factors chosen do indeed exert a significant influence on the residential load.

3.2.4. Urban Residential Load Forecast and Error Analysis

In order to assess the validity of the forecasts based on the combined model (Model 1), Beijing has been taken for an example and the forecast errors of the model are compared to those of other methods. The results are shown in Fig.1. Because the database for the combined model forecast is the result of taking the logarithms of the original data, the forecasting error can’t be represented simply by exp(s). Instead, we put

\[ s = \ln x - \ln y = \ln\left(\frac{x}{y}\right) \]  
\[ \frac{x}{y} = \exp(s) \]  
\[ m = \text{abs}\left(\frac{x}{y} - 1\right) \]

Here, \( x \) and \( y \) represent the original value and forecasting value, respectively, of the load; \( s \) is the error value; and \( m \) is an index value to measure the accuracy of forecasting. At the same time, \( m \) is calculated based on the growth rate method (Model 2) and the regression forecasting method (Model 3) with per capita disposable income as independent variable. The prediction accuracy of the combined model is the smallest of that of the three models in 2007. Considering all these phenomena, the combined model is feasible.
Fig. 1 Forecast Error (m) in Beijing Based on Three Methods

The forecasting results of residential loads in typical cities in China based on the combined model are shown in Table 2.

<table>
<thead>
<tr>
<th>Year</th>
<th>BJ</th>
<th>SH</th>
<th>TJ</th>
<th>CQ</th>
<th>NB</th>
<th>FZ</th>
<th>XA</th>
<th>HEB</th>
<th>JN</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>107</td>
<td>134.28</td>
<td>45.1</td>
<td>54.15</td>
<td>23.68</td>
<td>32.04</td>
<td>28.2</td>
<td>22.73</td>
<td>21.6</td>
</tr>
<tr>
<td>2015</td>
<td>196.48</td>
<td>201.3</td>
<td>65.64</td>
<td>97.3</td>
<td>47.27</td>
<td>60.86</td>
<td>58.95</td>
<td>42.68</td>
<td>31.19</td>
</tr>
<tr>
<td>2020</td>
<td>352.5</td>
<td>381.26</td>
<td>157.52</td>
<td>178.39</td>
<td>95.32</td>
<td>160.8</td>
<td>173.14</td>
<td>95.18</td>
<td>98.21</td>
</tr>
</tbody>
</table>

4. Conclusions

This shows that the key factors influencing urban residential load could indeed be obtained by using data mining association rules technology: the Apriori algorithm. Combining the key factors determined by DM with the methods of panel data theory, a relationship model of urban residential load and its key influencing factors was set up to forecast the load. Nine typical cities in China were chosen as a case study. Association rules mining uncovered seven key factors affecting urban residential load in China.

The urban residential load combined forecast model for typical cities in China was set up based on these influencing factors and a panel data analysis. The forecast error analysis and comparison indicate that the combined application of DM technology and panel data analysis to forecasts of urban residential load is reliable and has great prospects for further research.

Acknowledgement

The work described in this paper was supported by research grants from Humanities and Social Science project of the Ministry of Education of China (Project number: 07JA790092) and the Fundamental Research Funds for the Central Universities.

References