A Novel Electricity Sales Forecasting Method Based on Clustering, Regression and Time-series Analysis

Jiakui Zhao[1][2]

State Grid Information & Telecommunication Group, Beijing 100085, China
E-mail: `zhaojiakui@sgitg.sgcc.com.cn`

Wensheng Tang, Xuemin Fang, Jinzhi Wang
State Grid Corporation of China, Beijing 100031, China

Jian Liu, Hong Ouyang, Hongwang Fang, Yuxi Liu, Shulong Wang
State Grid Information & Telecommunication Group, Beijing 100085, China

Yaozong Lu, Jin Qiang
Xi'an Merit Data Technology Co., Ltd., Xi'an 710075, China

The forecasting of monthly electricity sales is a fundamental job for the marketing department of State Grid Corporation of China. In this paper, we propose a novel electricity sales forecasting method. Firstly, with the visual clustering algorithm, 27 provincial electric power companies of State Grid Corporation of China are clustered into different groups according to their historical monthly electricity sales curve; secondly, as to different groups, the monthly electricity sales are predicted by specific regression or time-series algorithms. Experiments show that the proposed approach has very high prediction accuracy rate in terms of forecasting the the electricity sales of the next 12 months for the power companies of State Grid Corporation of China.

---

[1]Speaker

1. Introduction

With the development of the smart grid, the forecasting of monthly electricity sales plays an important role in making monthly manufacturing and management schedule, guiding reasonable operation of the power plants and power grids, formulating the plan of using electricity in order and promoting the reasonable development of power market.

With data of six years, Taylor showed that for forecasting electricity sales, the triple-seasonal method outperformed the double-seasonal method and the univariate neural network approach [1]. A spatial Auto-Regressive and Moving Average (ARMA) model was proposed by Ohtsuka et al to consider features of the electricity demand in Japan, and thus a strategy of Markov chain Monte Carlo method was constructed to estimate parameters of the model [2]. Ringwood et al examined the application of the artificial neural network for forecasting the electricity demand [3]. With the possibility regression model, Hirano et al devoted special attention to the electricity demand and the price analysis in California [4]. A hybrid forecasting framework was proposed by Motamedi et al to combine a multi-input and multi-output forecasting engine for joint price and demand prediction with the data association mining algorithms [5]. Liu et al established Grey Model (GM) forecasting model by using the power demand load of Shandong province of China from 2004 to 2011, and forecasted the power load demand of 2012 and 2013 [6]. Kalashnikov et al proposed and analyzed numerical experiments based on the oligopolistic model of Mexican electricity market [7].

Above-mentioned papers forecasted the monthly electricity sales by using a variety of algorithms, including comparative analysis, structural analysis, regression, ARMA time-series, neural network and gray prediction, et al. These algorithms can forecast electricity sales to some extent, but the accuracy rate is unsatisfactory. Without considering the character of the historical electricity sales curves of different regions and one specific algorithm is used for all regions are the main reasons which inevitably lead to low accuracy rate.

In order to solve the problem, in this paper, we propose a global matching electricity sales forecasting method based on the combination of clustering, regression and time series analysis. The method includes two main steps: firstly, the 27 provincial electric power companies of State Grid Corporation of China (SGCC) are clustered into different groups according to the features in time and the frequency domain of their historical electricity sales curves. Then as to different groups, we choose appropriate predicting algorithms on the basis of features of the historical electricity sales curves. The algorithms include Support Vector Machine (SVM), $L_{1/2}$ sparse regression, Back Propagation (BP) neural network, and Auto-Regressive Integrated Moving Average Model (ARIMA).

The originality of our proposed electricity sales forecasting method can be summarized as follows: 1. the historical electricity sales curves are characterized by carefully selected dimensions with respect to the time domain and frequency domain; 2. the visual clustering algorithm is used to divide the historical electricity sales curves into optimized number of clusters; 3. according to different peculiarities of the regression and time-series algorithms which can be used to predict the electricity sales, we deduce a matching table between clusters and the optimized prediction algorithms.

2. Preliminary Knowledge

2.1 The Fourier Transformation

The time-domain analysis is often used to analyze the relationship between electricity sales and times according to times as abscissa and electricity sales as coordinated. The frequency-domain analysis is often used to analyze the features in frequency domain according to frequencies as abscissa and amplitudes as coordinated. Generally speaking, the time-domain analysis is visual and intuitive; however, the frequency-domain analysis is more brevity in mathematics, and hence can depict the curve characters deeply.

Let $x(n)$ whose length is N be a finite sequence, which is defined in Equation 2.1.

$$x(n) = \begin{cases} x(n) & 1 \le n \le N \\ 0 & otherwise \end{cases} \tag{2.1}$$

$$X(k) = DFS[x(n)] = \sum_{n=1}^{N} x(n) \omega_N^{(n-1)(k-1)} \tag{2.2}$$

The corresponding Fourier transformation of $x(n)$ is defined in Equation 2.2, where $1 \le k \le (\frac{2^M}{2} + 1)$, and $M = min(m)\{m \in Z^+ \mid 2^M \ge N\}$.

2.2 The Visual Clustering

The visual clustering algorithm considers a data set as an image with each light point located at a datum position [8-9]. As we blur this image, smaller light spots will gradually merge into larger ones until the whole image become one light spot at a low enough level of resolution. By identifying each light spot as a cluster, the blurring process will generate a series of clusters along the blurring hierarchy of the image.

For a data set $X = \{x_i \in R^d : i = 1, \cdots, N\}$, each data point x_i may be seen as a small light point which is commonly mathematically represented by the Dirac generalized function $\sigma(x - x_i)$. As a result, the data set X constitutes an image $p(x)$ in the *d*-dimensional space, and is mathematically represented by Equation 2.3.

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} \sigma(x - x_i) \tag{2.3}$$

According to the *visual scale theory*, the multi-scale representation of an image $p(x)$, i.e., $p(x, \sigma)$, is the convolution of $p(x)$ and the *Gaussian function* $g(x, \sigma)$, which is mathematically represented by Equation 2.4, where σ is the visual scale.

$$p(x, \sigma) = p(x) * g(x, \sigma) = (2\pi\sigma^2)^{\frac{-d}{2}} \sum_{i=1}^{N} e^{\frac{-(x - x_i)^T(x - x_i)}{2\sigma^2}} \tag{2.4}$$

Given a visual scale σ, the center of a light spot is defined as the maximum value point of $p(x, \sigma)$ with respect to x. The light spot corresponds to a center x^* is defined as the *attraction domain* of x^* with respect to the *gradient system* $\frac{dx}{dt} = \nabla_x p(x, \sigma)$, namely $B(x^*)$, which is mathematically represented by Equation 2.5, where $x(t, x_0)$ is the solution of the *initial value problem* of the gradient system defined in Equation 2.6; therefore, given a visual scale σ, by solving Equation 2.6, we can determine whether a light point x_0 belongs to a certain light spot $B(x^*)$.

$$B(x^*) = \{x_0 \in R^d : \lim_{t \to \infty} x(t, x_0) = x^*\} \tag{2.5}$$

$$\begin{cases} \frac{dx}{dt} = \nabla_x p(x, \sigma) \\ x(0, x_0) = x_0 \end{cases} \tag{2.6}$$

As the value of σ increases, the visual clustering algorithm generates a series of clusters, which form a tree structure, and hence we can obtain a series of partitions based on these clusters. Because having no any prior knowledge about the data set, we may choose the relative most important partition by using the *significance assumption*.

2.3 The SVM Regression

The SVM regression is the application of SVM classification to the functional regression field [10-11]. The sample points of SVM regression belong to only one class. The aim of SVM regression is to find an optimal hyper-plane which minimizes the total deviate which is the total

distances between all sample points and the hyper-plane. Given a data set $D : \left[x_i, y_i \right]_{i=1}^{N}$, the essence of the regression problem is to find a function $f(x)$ so that we may deduce y from any x . As Fig. 1 shows, the SVM regression problem may be converted into the optimal hard-ε zone hyper-plane finding problem, where all sample points satisfy $-\varepsilon \leq y_i - f(x_i) \leq \varepsilon$, $i = 1, \cdots, N$, and ε is minimized. The SVM regression problem is mathematically defined in Equation 2.7.

$$\min_{f(x)} \varepsilon$$
$$s.t. \ -\varepsilon \leq y_i - f(x_i) \leq \varepsilon \ , i = 1, \cdots, N \tag{2.7}$$

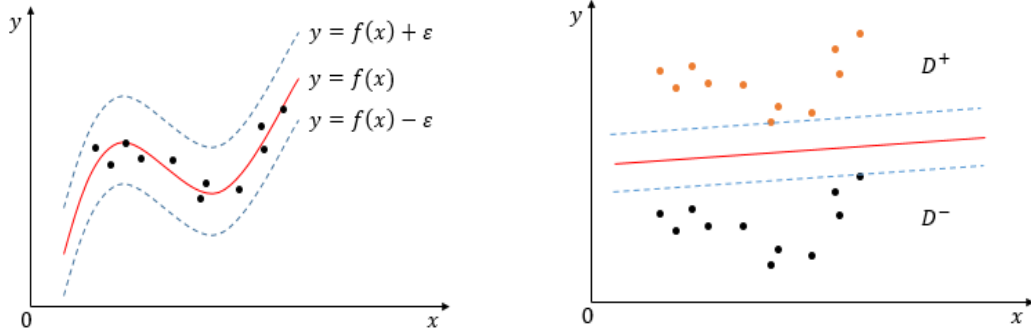**Figure 1:** Hard-ε Zone Hyper-plane   **Figure 2:** SVM Classification of D^+ & D^-

As Fig. 2 shows, the SVM regression problem may be further converted into the SVM binary-classification problem with positive class point set D^+ and negative class point set D^- , where D^+ and D^- are mathematically defined in Equation 2.8.

$$D^+ = \left\{ \left(x_i^T, y_i + \varepsilon \right)^T, i = 1, 2, \cdots, N \right\}$$
$$D^- = \left\{ \left(x_i^T, y_i - \varepsilon \right)^T, i = 1, 2, \cdots, N \right\} \tag{2.8}$$

2.4 The L$_{1/2}$ Sparse Regression

Given a data set $\left\{ \left(x_i, y_i \right)_{i=1}^{n} \right\}$, the essence of the regression problem is to find the optimal function f^* from learning machine F so that it approximates the relationship between x and y best. The popular method we're taking now is the *regularization L_p method* which may be mathematically represented by Equation 2.9 [12-13], where $l(.,.)$ is the loss function which is minimized in case of the predicted value is close to the real value, λ is the regularization parameter, and $\|f\|_p$ is normally taken as some norms of the solution, which represents the expectation, e.g., sparsity, of the solutions.

$$f^* = \arg \min_{f \in F} \left\{ \frac{1}{n} \sum_{i=1}^{n} l\left(f(x_i), y_i \right) + \lambda \|f\|_p \right\} \tag{2.9}$$

We often meet the over-fitting problem when solving regression problems. The regularization method takes this into account and reduces risks of over-fitting by adding a constraint condition $\|f\|_p$ for solutions. It had been proved that the L$_{1/2}$ sparsity method can guarantee both the solvability and the sparsity of the solution [12-13].

2.5 The BP Neural Network Regression

The BP neural network is one kind of multi-layer *Feed-Forward Network* with widest application, and is an important model of the *Artificial Neural Network* [14-17].
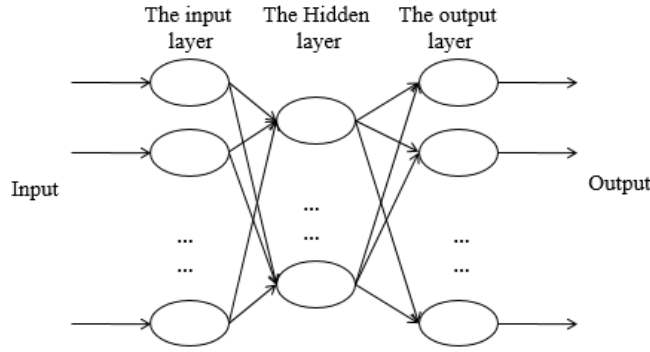
Figure 3: BP Neural Network

The structure of the BP neural network is shown in Fig. 3, which includes the input layer, the hidden layer, and the output layer. The main idea of the algorithm is that the learning process is divided into two stages. The first stage is the forward propagation stage, in which the output values are calculated from the input layer through the hidden layer and the output layer, and the states of neurons in each layer only have some impacts on the states of the neurons in the next layer. The second stage is the back propagation stage, in which if the output layer is unable of obtaining the desired outputs, it calculates the difference between the actual output and the desired output layer by layer and minimizes the error by tuning the weights of the previous layer. The forward propagation and the back propagation stages continue until desired output is obtained.

2.6 The ARIMA Time-Series Analysis

A time series is a data sequence that is chronological arranged, where the data elements are interdependent. The time-series analysis is an important area in data analysis. Traditional time-series models, e.g., AR(p), MA(q), and ARMA(p,q) are only suitable for stationary time-series analysis. As to the non-stationary time-series analysis, a differential process should be adopted to eliminate the trend and the seasonal characters of the time series so that the converted time series is a stationary series.

First Order Difference: $\nabla X_t = X_t - X_{t-1} = (1-B)X_t$

Second Order Difference: $\nabla^2 X_t = X_t - 2X_{t-1} + X_{t-2} = (1-B)^2 X_t$

…

d-th Order Difference: $\nabla^d X_t = (1-B)^d X_t$

$$\varphi(B)\nabla^d X_t = \theta(B)\varepsilon_t \qquad (2.10)$$

Let $\{X_t, t=0,\pm 1,\pm 2,\cdots\}$ be a non-stationary time series, if there exists integer d so that $\nabla^d X_t = W_t$, and $\{W_t, t=0,\pm 1,\pm 2,\cdots\}$ is an ARMA(p,q) stationary time series, then X_t is an ARIMA(p,d,q) time series [16-17], where X_t satisfies Equation 2.10, $\varphi(B)=1-\varphi_1 B-\varphi_2 B^2-\cdots-\varphi_p B^p$, and $\theta(B)=1-\theta_1 B-\theta_2 B^2-\cdots-\theta_q B^q$. The solution of the ARIMA (p,d,q) time series can be used to predict electricity sales.

3. Clustering of the Electric Power Companies

As to the 27 provincial electric power companies of SGCC, we use the visual clustering algorithm to divide them into different groups according to the historical electric sales curves, and the clustering dimensions are carefully selected from the characters of the historical electric sales curves with respect to time domain and frequency domain, so that we can get features of different groups such as long-term, periodicity, stability and disturbance. As to different groups, appropriate algorithms, e.g., the SVM regression, L$_{1/2}$ sparse regression, BP neural network

regression, and ARIMA time-series analysis are used to predict the electric sales. As to the electric power companies of the same group, the same algorithm is used. Details of the prediction idea are shown in Fig. 4. We get the historical monthly electricity sales data from January 2010 to December 2014 of the 27 electric power companies of SGCC from the marketing information system.
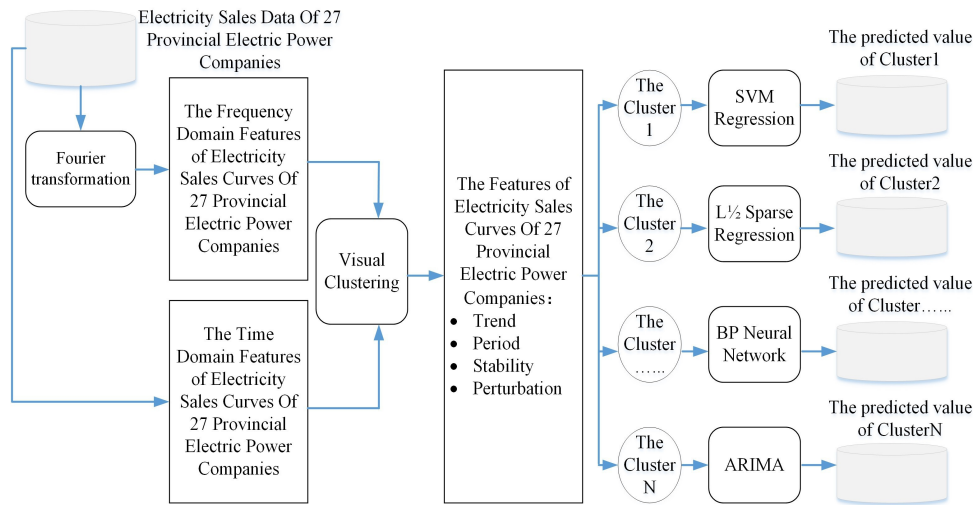

Figure 4: Prediction Idea of the Electricity Sales

For the features of electricity sales curves with respect to time domain, the 12 monthly year-on-year average values and 12 monthly month-on-month average values which demonstrate the trend of electricity sales monthly are selected as clustering dimensions. As to the features of electricity sales curves with respect to the frequency domain, As Fig. 5 shows, we obtain the Fourier coefficients of the historical electricity sales curves of 27 electric power companies of SGCC though Fourier transformation.
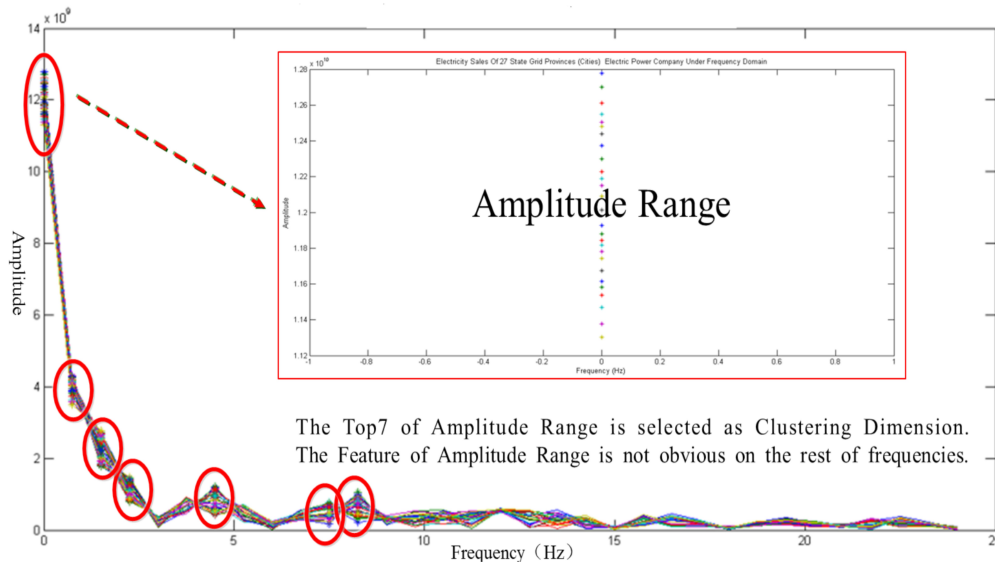

Figure 5: Electricity Sales Curves of 27 Power Companies under Frequency Domain

As shown in Fig. 5, by observation of electricity sales curves of 27 electric power companies in frequency domain, we can find that there are a total of 33 horizontal frequency nodes after the Fourier transformation, and there are seven horizontal frequency nodes which have relatively large gap of Fourier coefficients, named No.1-No.7 frequencies. In addition, this paper takes into account the Fisher Index which measures the periodicity of the electricity sales curve. Using Fisher statistics, we can detect the peak of cycle diagram to determine whether

periodicity of electricity sales sequence is present. Eventually, all the 32 clustering dimensions are shown in Table 1.

dimension category	dimension	dimension category	dimension
monthly year-on-year value (time domain)	average of January	monthly month-on-month value (time domain)	average of January
	average of February		average of February
	average of March		average of March
	average of April		average of April
	average of May		average of May
	average of June		average of June
	average of July		average of July
	average of August		average of August
	average of September		average of September
	average of October		average of October
	average of November		average of November
	average of December		average of December
No.1-No.7 Fourier coefficients (frequency domain)	NO.1 frequency	No.1-No.7 Fourier coefficients (frequency domain)	NO.5 frequency
	NO.2 frequency		NO.6 frequency
	NO.3 frequency		NO.7 frequency
	NO.4 frequency	periodicity	Fisher index

Table 1: Clustering Dimensions of the Historical Electricity Sales Curves

With the visual clustering algorithm, we obtain the clustering result of the 27 provincial electric power companies of SGCC as shown in Fig. 6. For security reasons, we omit the name of the electric power companies and use province 01 to province 27 to denote the corresponding companies. Fig. 7 and Fig. 8 shows the electricity sales curves of the clusters, due to space limitation, only two clusters, i.e., cluster 3 and cluster 9 are randomly selected for demonstration. We can see that the clustering result of the 27 electric power companies is reasonable. The specific features of the 10 clusters are shown in Table 2. Bigger month-on-month value means significant long-term, smaller year-on-year value infers significant stability, bigger Fisher index means obvious periodicity, and bigger frequency domain feature means obvious disturbance.
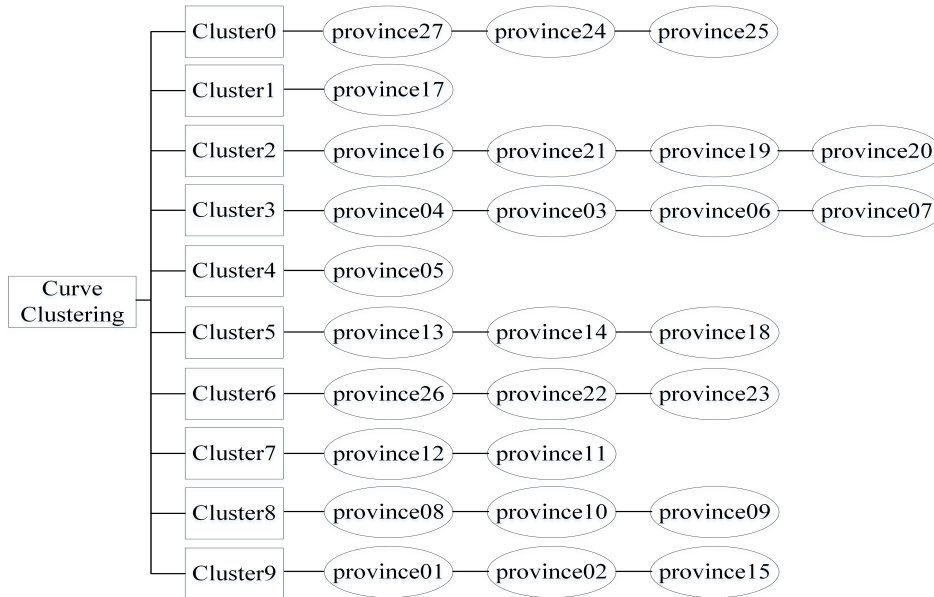

Figure 6: Clustering Result of the 27 Provincial Electric Power Companies of SGCC
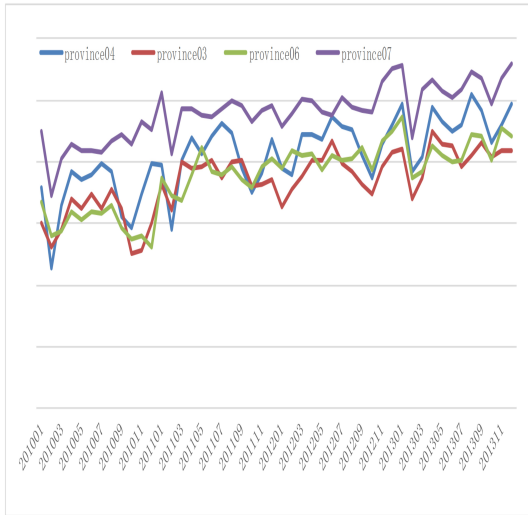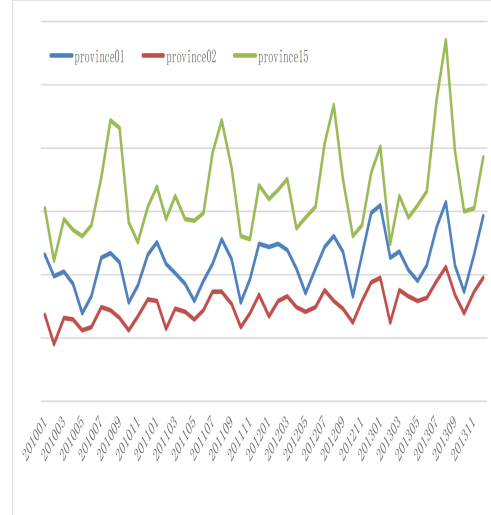
Figure 7: Electricity Sales Curves of Cluster 3

Figure 8: Electricity Sales Curves of Cluster 9

	Year-on-year value	Month-on-month value	Fisher index	Frequency feature
cluster 0	bigger	smaller	bigger	
cluster 1	bigger		bigger	bigger
cluster 2		smaller	bigger	bigger
cluster 3	bigger			bigger
cluster 4	bigger			bigger
cluster 5		smaller	bigger	bigger
cluster 6	bigger			bigger
cluster 7	bigger	smaller	bigger	
cluster 8		smaller	bigger	
cluster 9		smaller	bigger	

Table 2: Specific Features of 10 clusters

4. Electricity Sales Prediction

Combined with features such as long-term, periodicity, stability and disturbance of clusters in section 3, according to different peculiarities of prediction algorithms in Section 2.3 to 2.6, we deduce the matching table between clusters and prediction algorithms as shown in Table 3. According to the matching table, the monthly electricity sales of 27 electric power companies from January 2014 to December 2014 are predicted, where the sales data from January 2010 to December 2013 is used as training data and the data from January 2014 to December 2014 is used as testing data.

	Long-term	Stability	Periodicity	Disturbance	Selected algorithm
cluster 0	significant	significant	significant		BP neural network
cluster 1	significant		significant	significant	$L_{1/2}$ sparse regression
cluster 2		significant	significant	significant	SVM regression
cluster 3	significant			significant	$L_{1/2}$ sparse regression
cluster 4	significant			significant	$L_{1/2}$ sparse regression
cluster 5		significant	significant	significant	SVM regression
cluster 6	significant			significant	$L_{1/2}$ sparse regression
cluster 7	significant	significant	significant		BP neural network
cluster 8		significant	significant		ARIMA time-series
cluster 9		significant	significant		ARIMA time-series

Table 3: Matching Table between Clusters and Prediction Algorithms

	Maximum error (%)	Average error (%)	Minimum error (%)
province 01	5.43	3.17	0.04
province 07	3.94	1.53	0.00
province 12	9.90	4.43	1.15
province 20	9.86	3.06	0.22

Table 4: Electricy Sales Prediction Accuracy Rate

Table 4 shows the prediction accuracy rates correspond to the selected prediction algorithms with respect to Table 3. Fig. 9 to Fig. 12 show the predicted electricity sales curves in 2014. Due to space limitations, we only show the predicted results of four electric power companies, i.e., one predicted result for each prediction algorithm. For security reasons, in Fig. 9 to Fig. 12, the detailed electricity sales values are omitted.
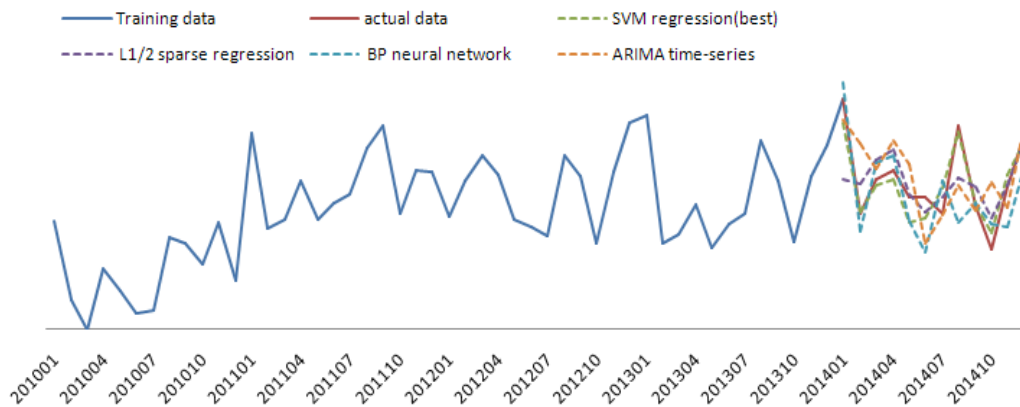

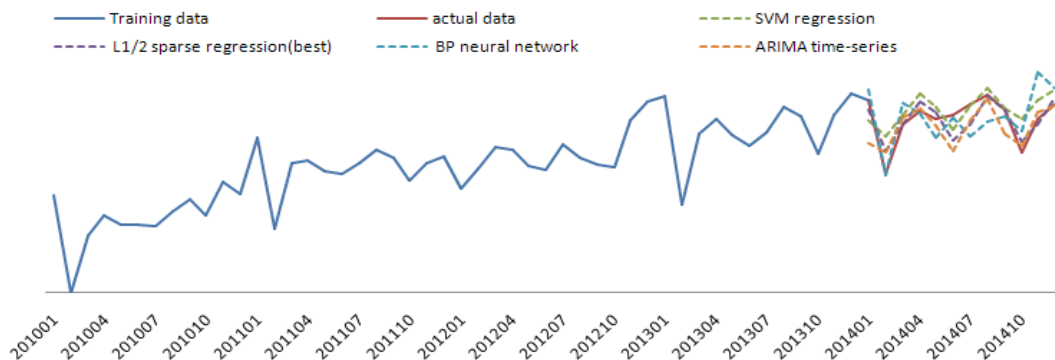**Figure 9:** Cluster 2 - Province 20 (the SVM regression is the prediction algorithm)

Figure 10: Cluster 3 – Province 07 (the L1/2 sparse regression is the prediction algorithm)
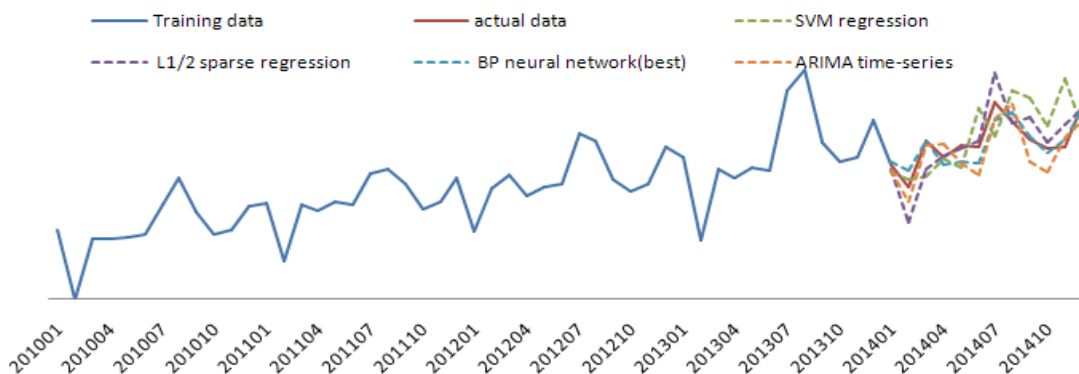

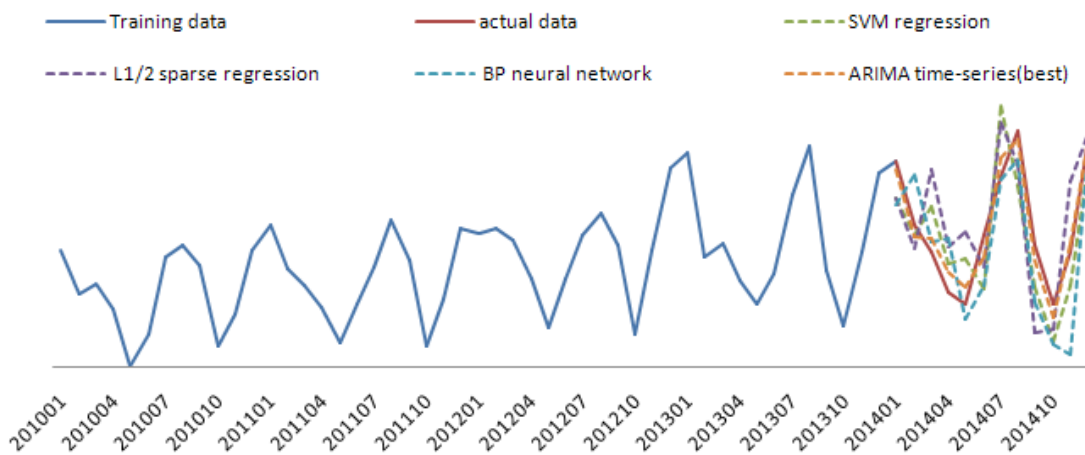**Figure 11:** Cluster 7 - Province 12 (the BP neural network is the prediction algorithm)

Figure 12: Cluster 9 - Province 01 (the ARIMA time-series is the prediction algorithm)

Above experimental results show that the selected electricity sales prediction algorithm with respect to Table 3 achieves the best accuracy rate among the four electricity sales prediction algorithms, i.e., the SVM regression algorithm, the L1/2 sparse regression algorithm, the BP neural network algorithm, and the ARIMA time-series analysis algorithm as to almost all electric power companies of SGCC.

5. Conclusion

In this paper, with the visual clustering algorithm, 27 provincial electric power companies of SGCC are clustered into 10 groups according to their historical monthly electricity sales curves. Then, as to different groups, only using historical electricity sales data, the monthly

electricity sales are predicted by using specific regression or time-series algorithms. Experiments show that the proposed approach has very high prediction accuracy rate with respect to forecasting the electricity sales of 12 months of the next year as to the 27 power companies of State Grid Corporation of China. We will further consider the influence factors for electricity sales prediction.

References

[1]   J. W. Taylor. *Triple seasonal methods for short-term electricity demand forecasting* [J]. European Journal of Operational Research. 204(1):139-152 (2010)

[2]   Y. Ohtsuka, T. Oga, K. Kakamu. *Forecasting electricity demand in Japan: a bayesian spatial autoregressive ARMA approach* [J]. Computational Statistics & Data Analysis. 54(11):2721-2735 (2010)

[3]   J. V. Ringwood, D. Bofelli, F. T. Murray. *Forecasting electricity demand on short, medium and long time scales using neural networks* [J]. Journal of Intelligent and Robotic Systems. 31(1-3):129-147 (2001)

[4]   O. Hirano, M. Kanke, K. Ozawa. *Electricity demand and price analysis in California using possibility regression model* [J]. Journal of Advanced Computational Intelligence and Intelligent Informatics. 7(2):147-152 (2003)

[5]   A. Motamedi, H. Zareipour, W. D. Rosehart. *Electricity price and demand forecasting in smart grids* [J]. IEEE Transactions on Smart Grid. 3(2):664-674 (2012)

[6]   S. Liu, L. Tian. *The study of long-term electricity load forecasting based on improved grey prediction model* [C]. Proceedings of the 2013 International Conference on Machine Learning and Cybernetics. IEEE, Tianjin, China, pp, 653-656 (2013)

[7]   V. Kalashnikov, N. Kalashnykova, F. J. C. Perez, A. C. U. Sifuentes, C. M. A. Garza. *Liberalization of Mexican electricity market: various modelling approaches with numerical simulation results* [C]. Proceedings of the 2008 International Conference on Intelligent Systems Design and Applications. IEEE, Kaohsiung, Taiwan, pp, 72-76 (2008)

[8]   C. Li, Z. Xu, T. Luo. *A heuristic hierarchical clustering based on multiple similarity measurements* [J]. Pattern Recognition Letters. 34(2):155-162 (2013)

[9]   C. Li, Z. Xu, C. Qiao, T. Luo. *Hierarchical clustering driven by cognitive features* [J]. Science China Information Sciences. 57(1):1-14 (2014) (In Chinese)

[10] V. Vapnik, O. Chapelle. *Bounds on error expectation for support vector machines* [J]. Neural Computation. 12(9):2013-2036 (2000)

[11] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee. *Choosing multiple parameters for support vector machines* [J]. Machine Learning. 46(1-3):131-159 (2002)

[12] Z. Xu, X. Chang, F. Xu, H. Zhang. $L_{1/2}$ *regularization: a thresholding representation theory and a fast solver* [J]. IEEE Transactions on Neural Network Learning System. 23(7):1013-1027 (2012)

[13] Z. Xu, H. Zhang, Y. Wang, X. Chang, Y. Liang. $L_{1/2}$ *Regularization* [J]. Science China Information Sciences. 53(6):1159-1169 (2010) (In Chinese)

[14] R. Durbin, D. E. Rumelhart. *Product units: a computationally powerful and biologically plausible extension to backpropagation networks* [J]. Neural Computation. 1(1):133-142 (1989)

[15] B. Widrow, D. E. Rumelhart, M. A. Lehr. *Neural networks: applications in industry, business and science* [J]. Communication of the ACM. 37(3):93-105 (1994)

[16] M. J. Kane, N. Price, M. Scotch, P. Rabinowitz. *Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks* [J]. BMC Bioinformatics. 15(1):276 (2014)

[17] M. Shukla, S. Jharkharia. *Applicability of ARIMA models in wholesale vegetable market: an investigation* [J]. International Journal of Information Systems and Supply Chain Management. 6(3):105-119 (2013)

PoS(CENet2015)015