J Electr Eng Technol.2018; 13(1): 038-050
http://doi.org/10.5370/JEET.2018.13.1.038

ISSN(Print) 1975-0102
ISSN(Online) 2093-7423

Spatio-temporal Load Forecasting Considering Aggregation Features of Electricity Cells and Uncertainties in Input Variables

Teng Zhao†, Yan Zhang* and Haibo Chen**

Abstract – Spatio-temporal load forecasting (STLF) is a foundation for building the prediction-based power map, which could be a useful tool for the visualization and tendency assessment of urban energy application. Constructing one point-forecasting model for each electricity cell in the geographic space is possible; however, it is unadvisable and insufficient, considering the aggregation features of electricity cells and uncertainties in input variables. This paper presents a new STLF method, with a data-driven framework consisting of 3 subroutines: multi-level clustering of cells considering their aggregation features, load regression for each category of cells based on SLS-SVRNs (sparse least squares support vector regression networks), and interval forecasting of spatio-temporal load with sampled blind number. Take some area in Pudong, Shanghai as the region of study. Results of multi-level clustering show that electricity cells in the same category are clustered in geographic space to some extent, which reveals the spatial aggregation feature of cells. For cellular load regression, a comparison has been made with 3 other forecasting methods, indicating the higher accuracy of the proposed method in point-forecasting of spatio-temporal load. Furthermore, results of interval load forecasting demonstrate that the proposed prediction-interval construction method can effectively convey the uncertainties in input variables.

Keywords: Spatio-temporal load forecasting, Multi-level clustering, SLS-SVRNs, Prediction intervals, Sampled blind number

1. Introduction

1.1 Aims and Difficulties

Power map, or city map of electricity, is a useful tool for the visualization and tendency assessment of urban energy application [1-3]. The power map integrates GIS-based maps of electrical network, renewable resources, land use conditions, building types, etc., with historical & real-time data series of users' electricity consumption, weather changes, and economic fluctuations, etc. [4]. Energy audit, power system planning, customer management, high-dimensional analysis and visualization of data assets can be accomplished with a power map. In addition to the above functionalities, a prediction-based power map may also provide applications like behavioral analysis for power users, load forecasting at the level of substations, feeders, transformers, and possibly customers, and spatio-temporal load forecasting (STLF), etc. Actually, the basis of prediction-based power map is the ability of forecasting, displaying and evaluating the spatial and temporal tendency of electricity load within a utility's service area, and STLF is the foundation for building a prediction-based power map.

In STLF, the service area of a utility can be divided into a group of cells according to geographical location. In different cells, the growth pattern of cellular load takes on a situation of diversity, considering the differences in traffic conditions, land-use types, historical loads, and the access of renewable energy resources and electric vehicles. The diverse growth patterns of cellular load call for targeted modeling and forecasting method. Constructing one model for each cell is possible; however, it is not advisable, because random disturbances and indistinctive features in a single cell may lead to larger forecasting error; in addition, it is not efficient. Except for the feature of diversity, electricity cells may also present out some features of aggregation, such as spatial aggregation, load-type aggregation and load characteristic aggregation. Cells with similar properties may distribute in clusters and they share a similar load growth pattern. Thus, how to balance the diversity and similarity of electricity cells in the process of spatio-temporal load modelling, it is a problem deserves to be studied.

Under normal conditions, STLF needs a multiple scenario capability - the ability to produce a set of reasonable forecasts that cover the uncertainty of future load growth [5]. The uncertainty of STLF covers two major aspects: 1) the uncertainty relationship between cellular load and relevant factors, which arises in the process of model training; 2) the uncertainty of input variables (forecast data

† Corresponding Author: Dept. of Electrical Engineering, Shanghai Jiao Tong University, China. (zhaoteng@sjtu.edu.cn)
** Dept. of Electrical Engineering, Shanghai Jiao Tong University, China. (zhang_yan@sjtu.edu.cn)
*** State Grid Shanghai Municipal Power Company, China. (chenhaibo@sh.sgcc.com.cn)
Received: October 1, 2016; Accepted: August 9, 2017

of some relevant factors, such as GDP, population, mean annual temperature, etc.), which calls for special attention in the process of model application. Considering that the historical data for model training is factual and accurate, the forecasting model can be viewed as an objective expression of the nonlinear relationship between cellular load and relevant factors. Thus, the uncertainties in model training can be eliminated in theory. Nevertheless, in view of the uncertainties in input variables, how to obtain and evaluate the prediction intervals of spatio-temporal load is not an easy problem.

1.2 Literature review

In general, the state of the art STLF methods can be classified into three categories: land-usage simulation method, load density index method, and time series analysis method.

1) Land-usage simulation method divides the service area into cellular regions and predicts load level of different cells [6-8]. By means of analyzing the characteristics of land usage and development laws, the newly developed or redeveloped loads are allocated spatiotemporally within the service area in a top-down way [9]. In some land-usage simulation methods, fuzzy logic or fuzzy reasoning have been applied to obtain the confidence level of land-usage decision rules, in order to reduce uncertainties in the process of land-usage simulation and get more accurate point forecasts [10-12]. Nevertheless, land-usage based method is typically used in situations where urban planning is uncertain and cellular load data is insufficient, the nature of this kind of method is land-usage forecasting more than load forecasting; the nonlinear relationship between cellular load and relevant factors is less applied due to the lack of data mining. Besides, this kind of method mainly focuses on reducing the uncertainties in land-usage decision, but pays less attention to obtaining the confidence intervals of STLF results.

2) Load density index method applies to cases with a clear land-use planning. The load density of different cells can be determined by artificial experience, simple analogy, or classification [13]. However, these kinds of load density calculation methods are subjective on some level and the quantization degree of final results is not high. A better way is to divide the cells into different categories according to their load type, and find the nonlinear relationship between load density and relevant factors for each category of cells [14]; it is a way to balance the diversity and similarity of cellular load modelling. However, the existing cell classification process is mainly based on the feature of load-type aggregation, and the spatial aggregation feature of cells is ignored. In some load density index methods, the intuitionistic fuzzy theory has been utilized to describe the uncertainty that would appear in the process of load density selection, thus reducing the uncertainties in model construction and improving the accuracy of STLF [15, 16]. Nevertheless, the uncertainties in input variables of the forecasting model are not considered.

3) Time series analysis method uses historical data of cellular load to construct the forecasting model of different cells. It is a method of constructing one time-series forecasting model for each cell. Commonly used forecasting models are exponential smoothing model (ESM), grey forecasting model (GFM), and so on [17]. As a bottom-up way of STLF, time series analysis method performs better in regions with stable land-use planning and copious historical data of spatial electric load. However, indistinctive features in a single cell may lead to larger forecasting error. The aggregation features of cellular load need to be taken into further consideration.

Some strategies in short-term load forecasting can provide references for STLF. In some studies of short-term load forecasting, to solve the problem of load diversity, load series are decomposed into a set of different frequency components by wavelets or differential empirical mode decomposition [17, 18]; to balance the diversity and similarity of load patterns, different consumers are aggregated into several clusters [19, 20]. Then, each component or cluster is separately forecasted with neural networks or other regression models [21, 22]. It has been shown that careful clustering of consumers for aggregation can result in smaller forecasting errors [23]. As far as we know, there are not many researches regarding the application of similar strategies in STLF.

Generally, for cells with similar geographical conditions and load characteristics, there is a certain nonlinear relationship between cellular load and its relevant factors [14]. Based on the uncertainties in relevant factors, STLF results in the form of prediction intervals can be obtained, which can provide more information about uncertainties in input variables of STLF model than point forecasts. Even though methods have been tried to reduce the uncertainties in STLF and some results have been achieved [12, 24], the uncertainties in input variables do exist and they are inevitable to some extent; researches on how to obtain and evaluate the prediction intervals of spatio-temporal load are still relatively infrequent.

1.3 Contribution of the paper

Based on in-depth review of the available STLF methods in literature, the main contribution of this paper can be summarized as follows.

A new STLF method based on multi-level clustering and category-oriented forecasting model training is proposed, and sampled blind number is introduced to evaluate the uncertainties in input variables and build prediction intervals of spatio-temporal load.

1.4 Structure of the paper

The rest of the paper is organized as follows: Section 2

introduces the procedure, data acquisition, and subroutines of the proposed STLF methodology. Section 3 describes 5 evaluating indices. Section 4 presents results and discussions of the test case. Section 5 concludes the paper with some remarks for future study in STLF.

2. Spatio-temporal Load Forecasting Methodology

2.1 Procedure of proposed methodology

In general, the procedure of proposed STLF methodology consists of 5 steps, as shown in Fig. 1. The 5 steps are briefly introduced below.

Step 1: Different data sources for STLF of the service area, including GIS based data layers, time series data of electrical load, and socio-economic data series, are used to construct the spatio-temporal database.

Step 2: The service area is divided into a number of equal-sized cells; the intrinsic and external properties of cells can be obtained based on spatio-temporal database.

Step 3: All the cells are grouped into different categories after 3 rounds of k-means clustering, based on 3 types of vectors corresponding to 3 cellular intrinsic properties.

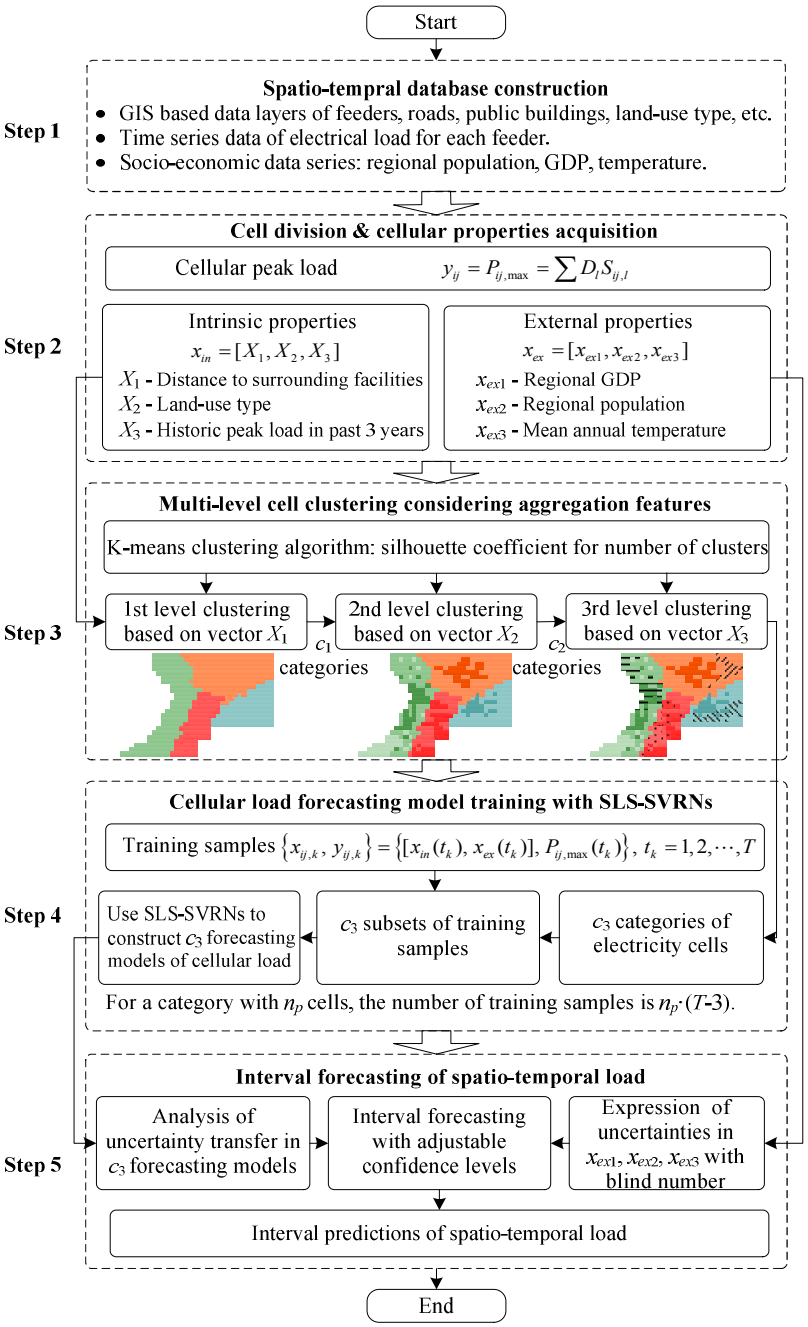

Fig. 1. Procedure of proposed STLF method

Step 4: For each category of cells, the non-linear relationship between cellular load and relevant factors (denoted by intrinsic properties and external properties) is modeled by SLS-SVRNs.

Step 5: Take cellular intrinsic properties and predicted external properties as input variables of the forecasting model, and use sampled blind number to represent uncertainties of the external properties. Considering the transmission of uncertainties in the forecasting model, prediction intervals of cellular load can be obtained with adjustable confidence levels.

2.2 Data acquisition and processing

For this study, both spatio-temporal and electrical data are used. The STLF process begins with data collection from different sources; these data include geographical landscape, land-use type, socioeconomic variations, weather conditions, power grid structure, geographic distribution of power facilities, historical peak load in different strategic points, like substations and feeders, etc. All these data are organized in a multi-dimensional spatio-temporal database that allows extracting information related to specific cells according to the needs of the STLF methodology.

Electrical load is distributed non-uniformly in the area of a utility's service zone. By dividing the service area into a group of equal-sized subzones according to geographical location, the information of spatial load distribution can be presented in the form of a grid: each subzone, known as a cell [25]. Each cell is represented as $C(i, j)$ (i=1,2,…, N_r, j=1,2,…, N_c), where N_r and N_c are the number of rows and columns in the grid, respectively. Electrical load in the area of a cell is defined as cellular load in this paper.

Historical cellular load is the basis of STLF, and it can be derived from the data of feeders [26]. Suppose that the maximum load and power supply area of feeder l in year t_k are respectively P_{l,t_k} and S_l, then load density in the power supply area of feeder l can be expressed as $D_{l,t_k} = P_{l,t_k} / S_l$. If the common area of $C(i, j)$ and feeder l is $S_{ij,l}$, then the cellular load of $C(i, j)$ in year t_k can be obtained from the following equation [27]:

$$y_{ij}(t_k) = P_{ij,\max}(t_k) = \sum D_{l,t_k} S_{ij,l} \tag{1}$$

The relevant factors of cellular load are various, including urban planning, historical load, GDP, population, etc. From an overall perspective, all the relevant factors can be divided into 2 types: the first type is intrinsic properties of cells, such as distance to surrounding facilities, land-use type and historical load, which can be used to distinguish the cell from others; the other type is external properties of cells, such as regional GDP, population, and mean annual temperature, which are all macroscopic quantities and the values are the same for different kinds of cells in the service area. The intrinsic and external properties of cells can be expressed as:

$$x_{in} = [X_1, X_2, X_3] \tag{2}$$

$$x_{ex} = [x_{ex1}, x_{ex2}, x_{ex3}] \tag{3}$$

where x_{in} and x_{ex} represent data vectors of cellular intrinsic and external properties, respectively. X_1, X_2 and X_3 denote the data vectors of distance to surrounding facilities, land-use type, and historic peak load in past 3 years, respectively. x_{ex1}, x_{ex2} and x_{ex3} are the data vectors of regional GDP, population, and mean annual temperature, respectively.

2.3 Multi-level clustering of cells considering their aggregation features

Intrinsic properties are unique for different electricity cells, and they are closely related to cellular load. In practice, electricity cells may present some aggregation features on account of their differences in intrinsic properties. For instance, if group the cells into different categories according to their distances to surrounding facilities, the cells in each category may present the feature of spatial aggregation; if the cells are classified according to their land-use type, then cells of each classification may have relevant load-types, and we can call it load-type aggregation; furthermore, if the cells are clustered based on their historical load, then the load density and load growth trend for cells in each cluster may share a high degree of similarity, and we can name it as load characteristic aggregation. For the aim of balancing diversity and similarity in the modelling of cellular load, we use multi-level clustering to group the cells into different categories, considering the aggregation features of electricity cells.

As shown in Fig. 1, in the first level of clustering, X_1= [$x_{1,1}$, $x_{1,2}$, $x_{1,3}$] works as the feature vector, where $x_{1,1}$, $x_{1,2}$, $x_{1,3}$ represents the minimum distance from a cell to surrounding main roads, transportation hubs, and public buildings (such as hospitals, schools and government offices). In the second level of clustering, X_2=[$x_{2,1}$, $x_{2,2}$,…, $x_{2,5}$] serves as the feature vector, where $x_{2,1}$, $x_{2,2}$, …, $x_{2,5}$ are the respective percentage of different land-use types, including residential land, commercial land, industrial land, municipal land, and others. In the third level of clustering, X_3=[$x_{3,1}$,$x_{3,2}$,$x_{3,3}$] is used as the feature vector, where $x_{3,1}$, $x_{3,2}$, $x_{3,3}$ represent cellular peak loads in the past 3 years, and all the cells are grouped into c_3 categories in the end.

Each level of cell clustering is implemented by k-means algorithm in this paper, and the number of clusters for k-means can be obtained by optimizing the silhouette coefficient; interested readers can refer to [28] for more details. For the initialization of cluster centers, we use the center point selection method proposed in [29]. The core idea of this center point selection method is to pick data points that are as far away from one another as possible, interested readers can refer to [29] for more details.

2.4 Construction of load forecasting model for each category of cells with SLS-SVRNs

In this section, we construct one forecasting model for each category of cells obtained in multi-level clustering. In other words, c_3 different forecasting models should be constructed, as shown in Fig. 2. Generally, for cells with similar geographical conditions and load characteristics, there is a certain nonlinear relationship between cellular load and its relevant factors [30]. Hence, we use intrinsic and external properties of the cell as the input, cellular load as the output, to train the forecasting models.

Suppose that the number of cells is n_p for category $p(p=1,2,\ldots,c_3)$, and the duration of historical data (historical cellular load, GDP, population, temperature, etc.) is $T(T>3)$ years, then cellular load $y_{ij}(t_k)$ and its relevant factors $x_{ij}(t_k)=[x_{in}(t_k), x_{ex}(t_k)]$ can form a data set for each category of cells, which includes $(T\text{-}3)\cdot n_p$ pairs of training samples. The subset of training samples for category p can be expressed as:

$$S_p = \left\{ \left[x_{ij}(t_k), y_{ij}(t_k) \right] \middle| C(i,j) \in Category\ p; t_k = 4,5,6,\cdots,T \right\} \tag{4}$$

According to Equs. (2)(3)(4), if the relevant factors are fully taken into consideration during the model construction process, input vector of the forecasting model could be of high dimension. Besides, the number of training samples for constructing the forecasting model depends on the amount of cells with similarities, which means the scale of training samples could be small or large. Thus, it is necessary to solve problems of high-dimension modeling with small samples and high-dimension modeling with large samples at the same time.

Considering the ability of neural networks in modelling unspecified nonlinear relationship between load and relevant factors [31], and the advantage of least squares support vector regression (LS-SVR) in solving problems with small samples and high dimensions, a combined algorithm, sparse least squares support vector regression networks (SLS-SVRNs), is proposed in this paper to model the growth patterns of cellular load. The structural diagram of SLS-SVRNs is shown in Fig. 3.

First, in order to improve the ability of LS-SVR in handling large samples, all the training samples are mapped to a high-dimensional feature space to obtain their maximum independent vector group; thus, a sparse least squares support vector regression machine (SLS-SVM) can be constructed. Second, use SLS-SVM to fit the nonlinear relationship between cellular load and relevant factors preliminarily, making full use of its ability in high-dimension modelling. Third, pre-trained parameters of radial basis function, obtained from SLS-SVM, are passed to radial basis function networks (RBFNs) for further optimization of the forecasting model, and gradient descent

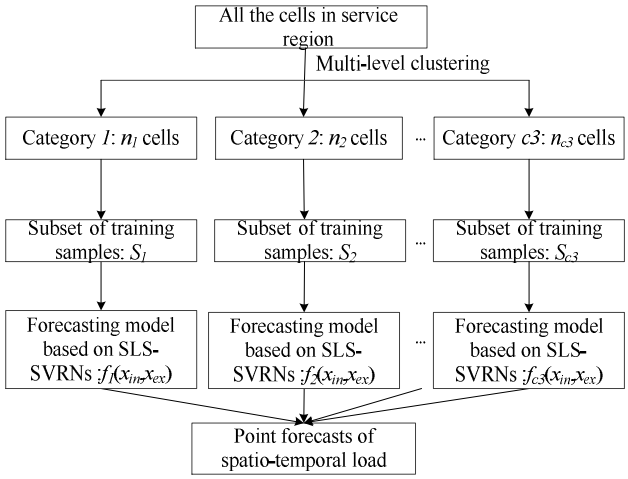

Fig. 2. Flowchart of cellular load regression

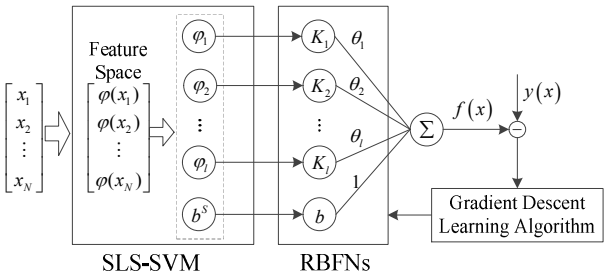

Fig. 3. Structural diagram of SLS-SVRNs algorithm

learning algorithm is used to optimize the parameters. The finally obtained SLS-SVRNs algorithm is suitable for the problem of nonlinear and high-dimensional modelling with small or large samples.

As shown in Fig. 2, with c_3 subsets of training samples, a series of forecasting models, namely $f_p(x_{in}, x_{ex})$, $p=1,2,\ldots,c_3$, can be constructed for different categories of cells, based on SLS-SVRNs. And point forecasts of spatio-temporal load can be obtained with these category-oriented forecasting models.

2.5 Interval forecasting of spatio-temporal load considering uncertainties in input variables

The input variables of the STLF model consist of intrinsic and external properties of the cells. In developed regions, the cellular intrinsic properties, including distance to surrounding facilities and land-use type, remain about the same over a period of time in the future, and cellular historical loads are exact values. Nevertheless, the forecasts for external properties, including regional GDP, population, and mean annual temperature, are the main sources of uncertainty in this research.

Historical values or forecasts of regional GDP, population, and mean annual temperature are easily accessible from local government or meteorological bureau. However, time series of these input variables are usually

not mutually independent and the forecasts of them are uncertain. In this paper, we use sampled blind number to represent the uncertainties in input variables.

Use $x_{ex,i}(i=1, 2, 3)$ to represent different external properties. For the predicted value of $x_{ex,i}$, namely $\hat{x}_{ex,i}$ with a confidence level of $(1-\alpha)\%$, divide its prediction interval into m sub-intervals, and the range of the L-th sub-interval is $[x_{ex,i}^{<L>}, x_{ex,i}^{<L+1>}]$. Suppose that the probability of $x_{ex,i} \in [x_{ex,i}^{<L>}, x_{ex,i}^{<L+1>}]$ is $P_{i,L}$, and we can define $X_{ex,i}^{L}$ as the average value of $\hat{x}_{ex,i}$ in the L-th sub-interval:

$$X_{ex,i}^{L} = \frac{1}{P_{i,L}} \int_{x_{ex,i}^{L}}^{x_{ex,i}^{L+1}} \hat{x}_{ex,i} \cdot P(\hat{x}_{ex,i}) d\hat{x}_{ex,i} \tag{5}$$

where $P(\hat{x}_{ex,i})$ is the point probability of $\hat{x}_{ex,i}$.

On this basis, we can define sampled blind number to characterize the uncertainty of $\hat{x}_{ex,i}$:

$$b(x_{ex,i}) = \begin{cases} P_{i,L} & x_{ex,i} = X_{ex,i}^{L} (L = 1, 2, \cdots, m) \\ 0 & others \end{cases} \tag{6}$$

where $b(x_{ex,i})$ is the sampled blind number of $\hat{x}_{ex,i}$; m is the order of $b(x_{ex,i})$; $(P_{i,L}, X_{i,L})$ is the L-th sample of $b(x_{ex,i})$; $X_{i,L}$ is the value of the L-th sample; $P_{i,L}$ is the confidence level of $b(x_{ex,i})$ at $X_{ex,i}^{L}$. Generally, if the confidence level of $\hat{x}_{ex,i}$ is relatively high (e.g. higher than 99.5%), then the following equation can be obtained.

$$\sum_{L=1}^{m} P_{i,L} \approx 1 \tag{7}$$

Thus, we can use sampled blind number to express the uncertainties in input variables in a discrete form.

For each of the well-trained forecasting models in 2.4, there are $p\,(p=3)$ input variables representing external properties of the cell. Use sampled blind number to characterize the uncertainties in the forecasting results of p external properties. For each external property, suppose that there are m samples and m corresponding confidence levels, and then there will be $H=m^p$ input scenarios for the forecasting model.

With H scenarios of input variables, we can obtain H point forecasts from the forecasting model (for every electricity cell at each year in the future), and all the point forecasts can also be expressed in the form of sampled blind number:

$$b(\hat{y}) = \begin{cases} P_{y,h} & \hat{y} = \hat{y}_h (h = 1, 2, \cdots, H) \\ 0 & others \end{cases} \tag{8}$$

$$\hat{y}_h = f(x_{in}, \hat{x}_{ex,1}^{L1}, \hat{x}_{ex,2}^{L2}, \cdots, \hat{x}_{ex,p}^{Lp}) \tag{9}$$

$$P_{y,h} = P_{1,L1} P_{2,L2} \cdots P_{p,Lp} \tag{10}$$

where $b(\hat{y})$ is the forecasting result in the form of sampled blind number; $\hat{y}_h$ and $P_{y,h}$ are the value and confidence level of the h-th sample; $f(\cdot)$ is the forecasting model; $\hat{x}_{ex,1}^{L1}, \hat{x}_{ex,2}^{L2}, \cdots, \hat{x}_{ex,p}^{Lp}$ are the sample values of predicted external properties; $P_{1,L1} P_{2,L2} \cdots P_{p,Lp}$ are the confidence levels of $\hat{x}_{ex,1}^{L1}, \hat{x}_{ex,2}^{L2}, \cdots, \hat{x}_{ex,p}^{Lp}$. The expected value of the forecasting result can be expressed as:

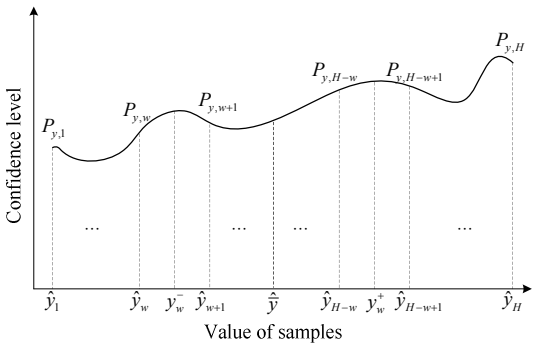

Fig. 4. Schematic diagram for confidence interval evaluation

$$\hat{\bar{y}} = E[\hat{y}_h] = \sum_{h=1}^{H} p_{y,h} \hat{y}_h \tag{11}$$

Thus, the uncertainties in input variables can be transferred to output variables via the forecasting models.

Based on sampled blind number, we can express the forecasting results of cellular load in the form of prediction intervals, with adjustable confidence levels.

As shown in Fig. 4, the entire sample values belonging to $b(\hat{y})$ and their corresponding confidence levels are listed in sequence. Use $\hat{\bar{y}}$ as the dividing point, and the sample values can be divided into 2 parts: the left part and the right part. Start from the leftmost, and remove w samples from the left part; accordingly, start from the rightmost, and remove w samples from the right part. Suppose that the gross confidence level for the removed $2w$ samples is P_{α}, and the gross confidence level for remaining samples is $P_{1-\alpha}$, then we can obtain:

$$\begin{cases} P_{\alpha} = \sum_{h=1}^{w} P_{y,h} + \sum_{h=1}^{w} P_{y,(H-h+1)} \\ P_{1-\alpha} = 1 - P_{\alpha} \end{cases} \tag{12}$$

Then, the prediction interval of spatio-temporal load, namely $[y_w^-, y_w^+]$, with a confidence level of $P_{1-\alpha}$, can be estimated as:

$$\begin{cases} (y_w^- - \hat{y}_w) P_{y,w+1} = (\hat{y}_{w+1} - y_w^-) P_{y,w} \\ (y_w^+ - \hat{y}_{H-w}) P_{y,H-w+1} = (\hat{y}_{H-w+1} - y_w^+) P_{y,H-w} \end{cases} \tag{13}$$

or

$$\begin{cases} y_w^- = \dfrac{\hat{y}_{w+1} P_{y,w} + \hat{y}_w P_{y,w+1}}{P_{y,w} + P_{y,w+1}} \\ y_w^+ = \dfrac{\hat{y}_{H-w+1} P_{y,H-w} + \hat{y}_{H-w} P_{y,H-w+1}}{P_{y,H-w} + P_{y,H-w+1}} \end{cases} \tag{14}$$

By adjusting the value of w, we can obtain the prediction intervals of cellular load for each cell with different confidence levels. Thus, interval forecasting of spatio-temporal load considering uncertainties in input variables can be accomplished.

3. Evaluating Indices

3.1 Indices for performance evaluating of point forecasting models

APE calculates the absolute percentage error between the actual and forecast values, therefore, measure the forecast accuracy of cellular load at some year with percentage values, one gets:

$$APE_{cell} = \frac{\left|x_{act} - x_{for}\right|}{x_{act}} \times 100\% \tag{15}$$

where x_{act} and x_{for} are actual and forecast values of cellular load, respectively.

MAPE calculates the mean absolute percentage error between the actual and forecast values, therefore, measure the forecast accuracy of a point-forecasting model in statistics with percentage values, we can get:

$$MAPE_{cells} \ \ or \ \ MAPE_{years} = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{\left|x_{i,act} - x_{i,for}\right|}{x_{i,act}} \times 100\% \tag{16}$$

where $MAPE_{cells}$ is the MAPE of cellular load for N_c cells in a fixed year; $MAPE_{years}$ is the MAPE of cellular load for a fixed cell in N_c years; N_c is the number of cells or years.

3.2 Indices for accuracy evaluation of interval predictions

By evaluating the accuracy of prediction intervals, both the forecasting error of STLF models and uncertainties in input variables can be taken into consideration. In this paper, we use 3 indices, including PICP (prediction interval coverage probability) [32], PINAW (prediction interval normalized average width) [33], and CDI (coverage density index), to evaluate the accuracy of prediction intervals.

PICP is a measure of the probability of target values covered by prediction intervals:

$$PICP = \frac{1}{N_c}\sum_{i=1}^{N_c}\theta_i \tag{17}$$

where N_c denotes the number of samples, θ_i is an indication of the coverage behavior of the i-th prediction interval. We denote y_i is the i-th target value, then θ_i =1 if y_i lies between the upper bound and the lower bound of the i-th prediction interval; otherwise, θ_i=0. A larger PICP means more targets are covered by the constructed prediction intervals.

PICP evaluates the coverage of prediction intervals, and PINAW is used to limit the width of prediction intervals. PINAW is defined as:

$$PINAW = \frac{1}{N_c R}\sum_{i=1}^{N_c}(U_i - L_i) \tag{18}$$

where U_i and L_i are the upper bound and lower bound of the i-th prediction interval, $U_i > L_i$; R equals to the maximum minus minimum of the target values. Obviously, PINAW is larger than 0. Normalization by the range R is to numerically compare quality of prediction intervals, which corresponds to the cellular load at different cells and different years.

From a decision-making perspective, smaller PINAW with a larger PICP are preferred. Thus, we define CDI as a combined index which simultaneously assesses the coverage probability and width of interval predictions; it can be written as:

$$CDI = PICP / PINAW \tag{19}$$

CDI measures the coverage probability for unit width of prediction intervals. It tries to compromise between the PINAW and PICP, and find a tradeoff between informativeness(PINAW) and validity (PICP) of prediction intervals. Generally, a lager CDI means higher coverage density for prediction intervals.

4. Results and Discussions

4.1 Description of the test case

In this paper, we take some region in Pudong, Shanghai as the test case. The service area of the test case includes 24 main roads, 3 transportation hubs, and 15 vital public buildings. Based on grid partitioning, the service area was divided into 515 cells, with the spatial resolution of $300\,m \times 300\,m$. As shown in Fig. 5, all 515 black cells lay scattered in the ij coordinates. Electricity raw data for the test case is obtained from local grid company with time span from 2000 to 2015; the annual sample data of external relevant factors, including GDP, population, and mean annual temperature, is obtained from Shanghai Statistics Bureau and Shanghai Metrological Bureau. In the following subsections, we use historical data of cellular load and external properties from 2000 to 2012, along with geographical condition and land-use type of each cell, to forecast the spatio-temporal load of the service area from 2013 to 2015.

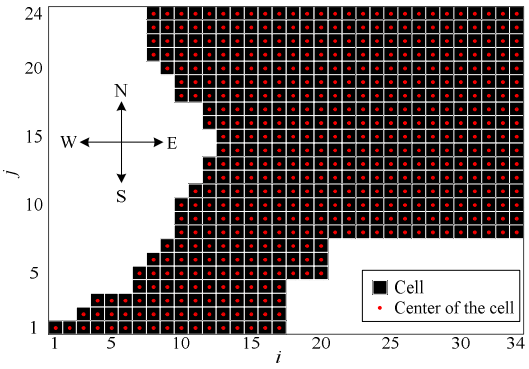

Fig. 5. Map of the service area after cellular division

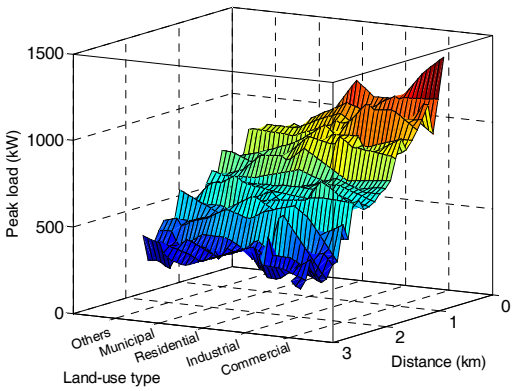

Fig. 6. Evolution of cellular peak load in terms of land-use type and distance to main roads

4.2 Multi-level cell clustering

Electrical load distribution in the service area is related to many factors. In terms of land-use type and distance to main roads, Fig. 6 shows the evolution of cellular peak load through interpolation surface fitting.

In Fig. 6, *x*-axis represents land-use type of different cells, *y*-axis represents the distance from each cell to the nearest main road, and *z*-axis represents the peak load of each cell in 2012. It is obvious that in the service area, cells with commercial and industrial land-use types entail high demand of load. Also, the distance to main roads is of great importance in determining the cellular load condition. Therefore, it is reasonable to select distance to surrounding facilities and land-use type as the feature vectors in multi-level cell clustering.

Group the cells into different categories according to X_1, X_2 and X_3, and the results of multi-level cell clustering are shown in Fig. 7. In Fig. 7(a), different colors are used to represent the results of fist-level cell clustering; all the cells are clustered into 4 categories, considering the feature of spatial aggregation. In Fig. 7(b), color depth is introduced to distinguish different categories in the second-level cell clustering; based on the results of first-level cell clustering, all the cells are further clustered into 9 categories, considering the feature of load-type aggregation. In Fig. 7

Table 1. Simulation data sets of the test case.

Data sets	Periods (Year)	Number of data samples (per category)
Training set	2003-2006, 2008-2011	128~1048
Validation set	2007, 2012	32~262
Testing set	2013-2015	48~393

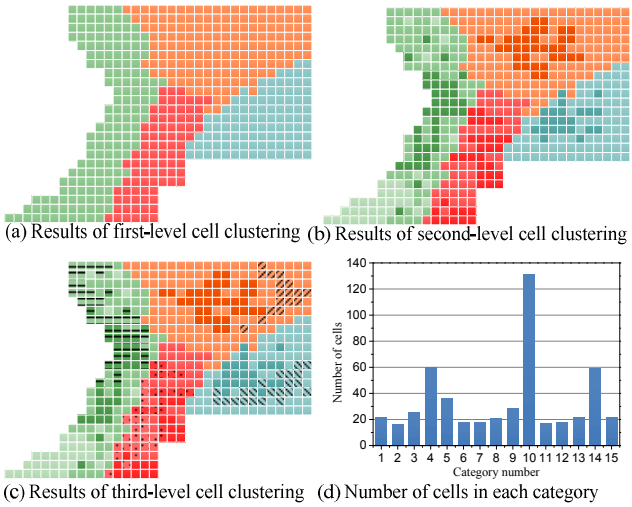

Fig. 7. Results of multi-level cell clustering

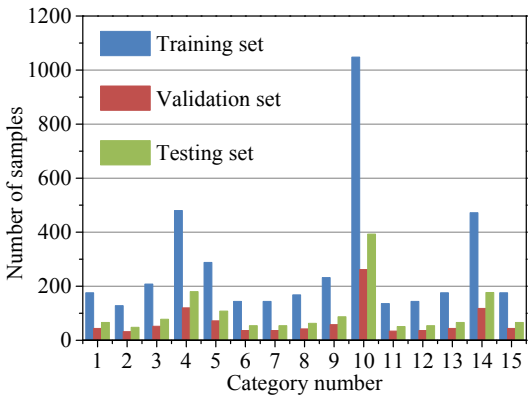

Fig. 8. Number of samples for each forecasting model

(c), different textures are added to the cells so as to differentiate diverse categories in the third-level cell clustering; finally, 15 categories are formed, considering the feature of load characteristic aggregation. The number of cells in each category is shown in Fig. 7(d).

4.3 Forecasting model training and evaluation

For cells belonging to the same category, intrinsic properties are similar and external properties are the same; in order to establish forecasting models specifically and reduce the computational complexity and random disturbances at the same time, forecasting models can be trained for each category of cells. Totally, 15 forecasting models were trained based on SLS-SVRNs.

For each forecasting model, 3 simulation data sets were constructed; they are training data set, validation data set, and testing data set. As shown in Table 1 and Fig. 8, the number of training samples for each forecasting model ranges from 128 to 1048, which covers 2 orders of magnitude. For training sets with large number of samples, the sparse process in SLS-SVRNs is helpful for reducing the computational complexity and improving the computation efficiency.

Use the 15 well-trained forecasting models to predict the cellular load of different cells, and we can obtain the forecasted spatio-temporal load of the service area in 2013, 2014, and 2015. The forecasting results and forecasting error are illustrated in Fig. 9. As shown in Fig. 9, the forecasting error (measured by APE) of cellular load ranges from 0 to approximately 7% in 2013, and 0 to almost 8% in 2014 and 2015; the forecasting error distributes uniformly in the service area.

In order to illustrate the effectiveness of the proposed method, STLF results using different methods are compared; the comparison results are shown in Table 2 and Fig. 10. The selected STLF methods for comparison include ESM (exponential smoothing method) [34], GFM (grey forecasting method) [35] and LDIM (load density index method) [14].

On the one hand, fix the year and compute the MAPE of all the cells, the MAPE at each year using different methods are shown in Table 2. It is clear that the MAPE of proposed method is less than that of ESM, GFM, and LDIM, in the year of 2013, 2014, and 2015. On the other hand, fix the cell and compute the MAPE of predicted cellular load from 2013 to 2015 for each cell; the results are shown in Fig. 10(a). As shown in Fig. 10(a), the MAPE of predicted cellular load using LDIM and GFM is mainly distributed from 4% to 6% and 6% to 8%; for ESM, the MAPE of almost 40% cells reaches the range of 8% to 15%. In contrast, the proposed method can achieve better STLF results: for 51.84% of the cells, MAPE is below 4%; comparisons between forecast results and actual spatial load data show a low spatio-temporal error.

Furtherly, the proposed method is tested with the data in [14] to prove its objectivity, and the forecasting results are illustrated in Fig. 10 (b). As with the results in Fig. 10 (a), the MAPEs of different methods in Fig. 10 (b) show that the quantity of cells with a lower forecasting error using the proposed method is larger than these using other methods, which also indicates a higher forecasting

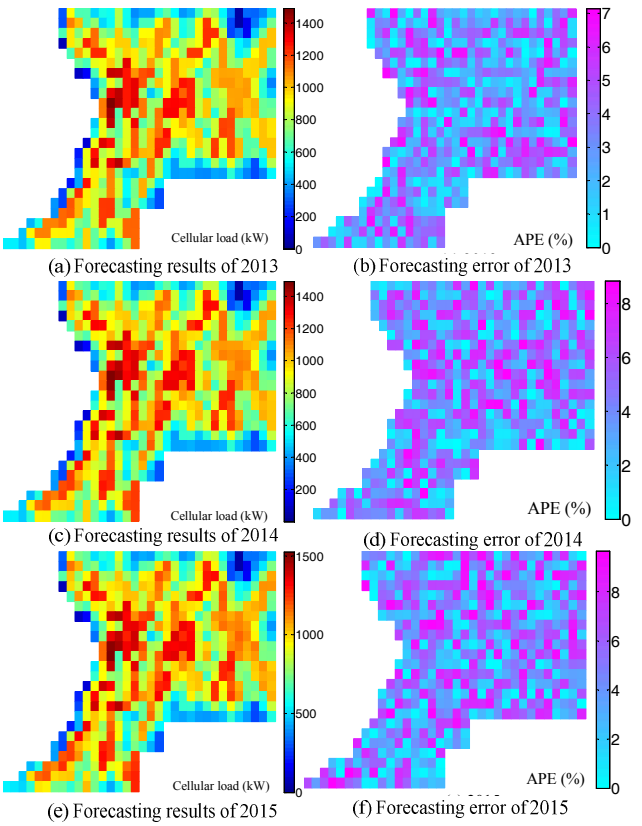

Fig. 9. Spatio-temporal load forecasting results and forecasting error of the service area in 2013, 2014 and 2015

Table 2. MAPE (year-fixed) of STLF results using different methods

Year	MAPE (%)			
	ESM	GFM	LDIM	Proposed method
2013	7.02	6.35	5.06	3.11
2014	7.88	8.26	5.45	4.16
2015	9.51	9.29	7.33	4.55
Average	8.14	7.97	5.95	3.94

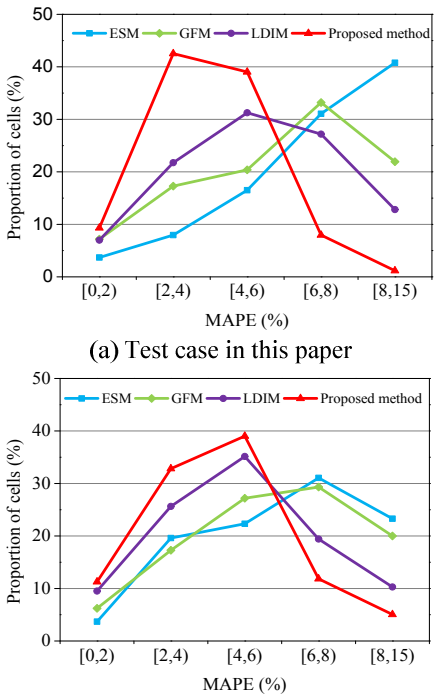

Fig. 10. MAPE (cell-fixed) of STLF results using different methods

accuracy of the proposed method.

The above results have proved that existing STLF methods, such as ESM, GFM and LDIM, may have lower accuracy in the process of STLF, without considering the aggregation features of cells. For the proposed method, by means of analyzing the relationship between cellular load and relevant factors, and constructing STLF models for cells belonging to different categories, the accuracy of STLF have been significantly improved.

4.4 Interval load forecasting

In order to deal with the problem of data uncertainty regarding external properties, interval forecasting of spatio-temporal load is introduced in this paper based on the method in subsection 2.5. Using the methods mentioned in [36-38], the interval predictions of GDP, population and mean annual temperature in 2013, 2014 and 2015 can be obtained, as shown in Table 3; the forecast error obeys Gaussian distribution, under the confidence level of 99.73% (3 sigma).

Use 5-order sampled blind number to characterize the uncertainties in interval predictions of GDP, population and mean annual temperature, the results are shown in Table 4. In this way, the interval predictions can be represented in a discrete form.

In the process of interval load forecasting, the well-trained forecasting models for each category of cells were still used. For each forecasting model, input variables denoting intrinsic properties remained the same, and input variables denoting external properties were assigned with sampled blind number. As shown in Table 4, there are 5 samples and 5 corresponding confidence levels for each external property, and then there will be $H=5^3=125$ input scenarios for each forecasting model.

With 125 scenarios of input variables, 125 point forecasts were obtained from the forecasting model for every electricity cell in each year of 2013, 2014 and 2015. Furthermore, the forecasting results of cellular load were expressed in the form of prediction intervals, with confidence levels of 95%, 96%, 97%, 98% and 99%. By comparing the forecasting results of all the cells with actual values, the indices for accuracy evaluation of interval predictions were obtained, as listed in Table 5.

As shown in Table 5, the index of PICP increases with the rising of confidence level, which indicates that the coverage of prediction intervals is improving with the growing of forecasting results reliability, or we can say that the confidence level of cellular load is consistent with PICP to a certain extent; this kind of result is what we want to see. However, if we take a look at the index of PINAW, we will find that PINAW also increases with the rising of confidence level, which demonstrates that the width of prediction intervals grows wider with the growing of forecasting results reliability.

In other words, the improving coverage of prediction intervals is to an extent influenced by the widening of prediction intervals. In this case, the index of CDI, which evaluates the coverage density of predicted intervals of cellular load, can be used to compare the forecasting results at different confidence levels. As we can see in Table 5, considering 5 confidence levels, the maximum values of CDI in the year of 2013, 2014 and 2015, are reached at confidence levels of 97%, 96% and 97%, respectively. This kind of result shows that, in the presence of uncertainties in input variables, the confidence level and accuracy of STLF results cannot be improved

Table 3. Interval forecasts of input variables (external properties)

	$\hat{x}_{GDP}$ /million dollars	$\hat{x}_{pop}$ /thousand people	$\hat{x}_{tem}$ / °C
2013	[6708.68, 6856.37]	[916.38, 935.12]	[17.32, 17.78]
2014	[7093.35, 7281.47]	[943.75, 961.07]	[17.15, 17.61]
2015	[7547.32, 7725.64]	[969.51, 986.25]	[17.05, 17.41]

Table 4. Sampled blind number of input variables (external properties)

Year	Sampled blind number	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	$x_{i,5}$
2013	$X_{GDP,1}$	6729.48	6756.26	6782.52	6808.79	6835.57
	$P_{GDP,1}$	0.0346	0.2383	0.4515	0.2383	0.0346
	$X_{pop,2}$	919.02	922.42	925.75	929.08	932.48
	$P_{pop,2}$	0.0346	0.2383	0.4515	0.2383	0.0346
	$X_{tem,3}$	17.38	17.47	17.55	17.63	17.72
	$P_{tem,3}$	0.0346	0.2383	0.4515	0.2383	0.0346
2014	$X_{GDP,1}$	7119.85	7153.96	7187.41	7220.86	7254.97
	$P_{GDP,1}$	0.0346	0.2383	0.4515	0.2383	0.0346
	$X_{pop,2}$	946.19	949.33	952.41	955.49	958.63
	$P_{pop,2}$	0.0346	0.2383	0.4515	0.2383	0.0346
	$X_{tem,3}$	17.21	17.30	17.38	17.46	17.55
	$P_{tem,3}$	0.0346	0.2383	0.4515	0.2383	0.0346
2015	$X_{GDP,1}$	7572.44	7604.77	7636.48	7668.19	7700.52
	$P_{GDP,1}$	0.0346	0.2383	0.4515	0.2383	0.0346
	$X_{pop,2}$	971.87	974.90	977.88	980.86	983.89
	$P_{pop,2}$	0.0346	0.2383	0.4515	0.2383	0.0346
	$X_{tem,3}$	17.10	17.17	17.23	17.29	17.36
	$P_{tem,3}$	0.0346	0.2383	0.4515	0.2383	0.0346

Table 5. Evaluation of interval load forecasting.

Year	Confidence level (%)	PICP (%)	PINAW (%)	CDI
2013	95	83.12	56.33	1.48
	96	87.33	58.01	1.51
	97	92.67	61.03	1.52
	98	93.01	62.11	1.50
	99	95.62	65.67	1.46
2014	95	83.54	56.50	1.48
	96	87.07	57.95	1.50
	97	91.84	61.76	1.49
	98	92.92	62.92	1.48
	99	94.09	66.39	1.42
2015	95	82.78	56.78	1.46
	96	84.39	58.77	1.44
	97	91.09	61.76	1.47
	98	90.79	63.61	1.43
	99	94.37	65.28	1.45

simultaneously in the overall process of confidence level adjustment. In other words, the optimum value of confidence level and forecasting accuracy may not be reached at the same time. Thus, when selecting the results of STLF, it is necessary to compromise between confidence level and forecasting accuracy according to demand.

5. Conclusions

In this paper, a new STLF method based on multi-level clustering and SLS-SVRNs is presented. The proposed method constructs various models to forecast the load variation and load distribution in the service area, considering aggregation features of electricity cells. In addition, sampled blind number is used to evaluate the influences of uncertainties in input variables on STLF results. One advantage of this method is that external and intrinsic properties of the cells are both considered in the process of forecasting model training.

From the results presented in this paper, we can reach the following conclusions.

(i) The spatial distribution and time variation of cellular load are closely related to intrinsic properties of electricity cells. The cellular features of spatial aggregation, load-type aggregation and load characteristic aggregation can be identified by multi-level clustering of electricity cells based on intrinsic properties.

(ii) The proposed method, in which the aggregation features of electricity cells have been taken into consideration, could achieve a higher accuracy in the point forecasting of spatio-temporal load, compared with STLF methods of ESM, GFM and LDIM.

(iii) Under a certain condition of uncertainties in input variables, the confidence level and forecasting accuracy of STLF results cannot be improved simultaneously in the overall process of confidence level adjustment. It is necessary to compromise between confidence level and forecasting accuracy according to demand.

STLF is a foundation for building the prediction-based power map. Moreover, the STLF method proposed in this paper is helpful for the visualization of cellular load aggregations and the interval prediction of spatio-temporal load, based on multi-level clustering and uncertainty transmission analysis. At present, electricity cells used in the proposed method are all defined as equal-sized squares, without considering the spatial distribution and load characteristics of power consumers, and the shape of buildings. For this deficiency, the proposed STLF method will be constantly improved and refined in future studies.

Acknowledgements

This research was conducted with the support of National High-tech R&D Program, China (Grant No. 2015AA050203). Also, this work was financed by State Grid Science & Technology Project (520900150037).

References

[1] Monika, D. Srinivasan and T. Reindl, "Real-time display of data from a smart-grid on geographical map using a GIS tool and its role in optimization of game theory," in *Smart Grid Technologies - Asia*,, pp. 1-6, 2015.

[2] X. He, Q. Ai, R. C. Qiu, J. Ni, L. Piao, Y. Xu, and X. Xu, "3D Power-map for smart grids - An integration of high-dimensional analysis and visualization," *Statistics,* 2015.

[3] A. N. Sekhar, K. S. Rajan and A. Jain, "Spatial informatics and geographical information systems: tools to transform electric power and energy systems," in *TENCON 2008-2008 IEEE Region 10 Conference*, pp. 1-5, 2008.

[4] P. B. T. Inc., "Interactive city map of electricity, "http://www.powermap.com.cn/, 2016-08-26

[5] X. Bai, Z. Chao and M. U. Gang, "Review and prospect of the spatial load forecasting methods," *Proceedings of the CSEE,* vol. 33, no. 25, pp. 78-92, 2013.

[6] X. Bai, G. Peng-wei, M. Gang, Y. Gan-gui, L. Ping, C. Hong-wei, L. Jie-fu, and B. Yang, "A spatial load forecasting method based on the theory of clustering analysis," *Physics Procedia,* vol. 24, Part A, pp. 176-183, 2012.

[7] J. D. Melo, E. M. Carreno and A. Padilha-Feltrin, "Multi-agent simulation of urban social dynamics for spatial load forecasting," *IEEE Transactions on Power Systems,* vol. 27, no. 4, pp. 1870-1878, 2012.

[8] J. D. Melo, E. M. Carreno, A. Padilha Feltrin, and C. R. Minussi, "Grid‐based simulation method for spatial electric load forecasting using power‐law distribution with fractal exponent," *International Transactions on Electrical Energy Systems,* 2015.

[9] E. M. Carreno, R. M. Rocha and A. Padilha-Feltrin, "A cellular automaton approach to spatial electric load forecasting," *Power Systems, IEEE Transactions on,* vol. 26, no. 2, pp. 532-540, 2011.

[10] J. Yu, F. Yan, W. Yang, and X. Gao, "Spatial load forecasting of distribution network based on fuzzy multi-objective multi-person decision making," *Power System Technology,* vol. 30, no. 7, pp. 69-76, 2006.

[11] S. Lei, C. Sun, Q. Zhou, and X. Zhang, "Application of fuzzy rough set theory in spatial load forecasting," *Power System Technology,* vol. 29, no. 9, pp. 26-30, 2005.

[12] X. Yang, J. Yuan, T. Zhang, and H. Mao, "Application of uncertainty reasoning based on cloud theory in spatial load forecasting," in *World Congress*

on Intelligent Control and Automation, pp. 7567 - 7571, 2006.

[13] W. B. Tao, L. Z. Zhang, P. Hong, L. I. Zhen-Yuan, and Z. Hua, "Spatial electric load forecasting based on double-level bayesian classification," *Proceedings of the CSEE,* 2007.

[14] Z. Quan, S. Wei, H. Ren, Z. Yun, C. Sun, G. Xie, and J. Deng, "Spatial load forecasting of distribution network based on least squares support vector machine and load density index system," *Power System Technology,* vol. 35, no. 1, pp. 66-71, 2011.

[15] L. J. Liu, Y. Fu, S. W. Ma, and R. Hu, "Spatial load forecasting of distribution network based on intuitionistic fuzzy entropy and fuzzy clustering," *Advanced Materials Research,* vol. 516-517, no. 100, pp. 1433-1436, 2012.

[16] Z. Sun, X. Wang, Z. Shouxiang, W. Lei, and S. Guo, "New load density forecasting method for objective network planning," in *International Conference on MEMS NANO, and Smart Systems*, pp. 114-117, 2009.

[17] M. Ghofrani, M. Ghayekhloo, A. Arabali, and A. Ghayekhloo, "A hybrid short-term load forecasting with a new input selection framework," *Energy,* vol. 81, pp. 777-786, 2015.

[18] S. Li, P. Wang and L. Goel, "Short-term load forecasting by wavelet transform and evolutionary extreme learning machine," *Electric Power Systems Research,* vol. 122, pp. 96-103, 2015.

[19] C. H. Jin, G. Pok, Y. Lee, H. Park, K. D. Kim, U. Yun, and K. H. Ryu, "A SOM clustering pattern sequence-based next symbol prediction method for day-ahead direct electricity load and price forecasting," *Energy Conversion and Management,* vol. 90, pp. 84-92, 2015.

[20] F. L. Quilumba, W. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid,* vol. 6, no. 2, pp. 911-918, 2015.

[21] C. Wang, G. Grozev and S. Seo, "Decomposition and statistical analysis for regional electricity demand forecasting," *Energy*, vol. 41, no. 1, pp. 313-325, 2012.

[22] A. S. Khwaja, M. Naeem, A. Anpalagan, A. Venetsanopoulos, and B. Venkatesh, "Improved short-term load forecasting using bagged neural networks," *Electric Power Systems Research,* vol. 125, pp. 109-115, 2015.

[23] S. Bandyopadhyay, T. Ganu, H. Khadilkar, and V. Arya, "Individual and aggregate electrical load `forecasting: one for all and all for one," in *ACM Sixth International Conference*, pp. 1653-1654, 2015.

[24] X. M. Yang, J. S. Yuan, J. F. Wang, and X. Gao, "A new spatial forecasting method for distribution network based on cloud theory," *Proceedings of the CSEE,* vol. 26, no. 6, pp. 30-36, 2006.

[25] M. Batty, *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*: The MIT press, pp. 1-565, 2007.

[26] J. D. Melo, E. M. Carreno, A. Calviño, and A. Padilha-Feltrin, "Determining spatial resolution in spatial load forecasting using a grid-based model," *Electric Power Systems Research,* vol. 111, pp. 177-184, 2014.

[27] B. Xiao, P. Nie, G. Mu, J. Wang, and L. Tian, "A spatial load forecasting method based on multilevel clustering analysis and support vector machine," *Automation of Electric Power Systems,* vol. 39, no. 12, pp. 56-61, 2015.

[28] R. Lletı´ , M. C. Ortiz, L. A. Sarabia, and M. S. Sánchez, "Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes," *Analytica Chimica Acta,* vol. 515, no. 1, pp. 87-100, 2004.

[29] A. Rajaraman, J. Leskovec and J. D. Ullman, *Mining of massive datasets*, 1 ed. Cambridge, United Kingdom: Cambridge University Press, pp. 253-254, 2011.

[30] A. Khosravi, S. Nahavandi and D. Creighton, "Load forecasting and neural networks: A prediction interval-based perspective," in *Computational intelligence in power engineering*: Springer, pp. 131-150, 2010.

[31] J. W. Taylor and R. Buizza, "Neural network load forecasting with weather ensemble predictions," *IEEE Transactions on Power Systems,* vol. 17, no. 3, pp. 626-632, 2002.

[32] A. Khosravi, S. Nahavandi and D. Creighton, "Construction of optimal prediction intervals for load forecasting problems," *IEEE Transactions on Power Systems,* vol. 25, no. 3, pp. 1496-1503, 2010.

[33] A. Khosravi, S. Nahavandi and D. Creighton, "Prediction interval construction and optimization for adaptive neurofuzzy inference systems," *IEEE Transactions on Fuzzy Systems,* vol. 19, no. 5, pp. 983-988, 2011.

[34] N. A. A. Jalil, M. H. Ahmad and N. Mohamed, "Electricity load demand forecasting using exponential smoothing methods," *World Applied Sciences Journal,* vol. 22, no. 11, pp. 1540-1543, 2013.

[35] A. W. L. Yao, S. C. Chi and J. H. Chen, "An improved Grey-based approach for electricity demand forecasting," *Electric Power Systems Research,* vol. 67, no. 3, pp. 217-224, 2003.

[36] G. I. Treyz, *Regional economic modeling: A systematic approach to economic forecasting and policy analysis*. Berlin, Germany: Springer Science & Business Media, pp. 258-262, 2013.

[37] D. B. Stephenson, C. Coelho, F. J. DOBLAS REYES, and M. Balmaseda, "Forecast assimilation: a unified framework for the combination of multi − model weather and climate predictions," *Tellus A,* vol. 57,

no. 3, pp. 253-264, 2005.
[38] A. J. Coale, P. Demeny and B. Vaughan, *Regional Model Life Tables and Stable Populations: Studies in Population*, 2 ed. New York, USA: ACADEMIC PRESS, pp. 25-36, 2013.

Teng Zhao He received bachelor's degree in electrical engineering from Shanghai Jiao Tong University, China. Currently, he is pursuing a PhD degree in power system and its automation in Shanghai Jiao Tong University. His research interests are smart grid and big data.

Yan Zhang She received the M.S. and Ph.D. degrees from China Electric Power Research Institute and Shanghai Jiao Tong University, respectively. Her research interests are power system analysis, smart grid, and big data.

Haibo Chen He received bachelor's degree in electrical engineering and master's degree in business administration from Shanghai Jiao Tong University. His research interests are power system management and electric power information technology.