Prediction of Time Series Analysis of Power Usage Based on Rstudio

Cuihua Tian, Mugisha Theophile, Jianhong Qian, Yiping Zhang, and Zhigang Hu

***Abstract*—Use the historical data of household energy consumption to design a valuable information model for predicting future demand. The cluster analysis of the data set of the transmission power distribution system shows the proportion of different consumption behaviors and the level of power consumption in different periods, and effectively predicts the power consumption of users. It can help power companies and users to control the load during peak hours of power demand transfer to off-peak hours.**

***Index Terms*—Big data, data analytics, RStudio language, predictive model.**

I. Introduction

At present, with the continuous development of computer technology, the energy consumption of the power system has exploded year by year.

A large amount of real-time data is stored in the database [1], which contains a large amount of time information useful to users and Company System Operators (CSO). Smart meters are part of data collection and communication in smart grid infrastructure [2]. The key element of the long-term control and detection of the future smart grid load is the smart meter. Through analysis and modeling, valuable data can be extracted from the database [3], which requires data mining technology [4].

Previous studies of cluster analysis relied on data provided by survey users [5]. According to the information provided by the smart meter, this method is not accurate [6]. This study uses unsupervised learning clustering and time series to predict energy consumption. Use smart meters to collect data and analyze receipts through R programming. Data mining technology has achieved better prediction results.

II. Design of Analysis System

The electricity consumption model based on system design analysis in this study is based on RStudio's electricity consumption cluster analysis for household electricity consumption analysis, as shown in Fig. 1. This step can be divided into five steps:

1) Data preparation. Including data cleaning and load curve normalization.
2) Reduce the dimensionality of the load distribution by compressing the group years.
3) Analyze and establish a predictive model.
4) Integrate, visualize and realize the collected data.
5) Discussion and results.

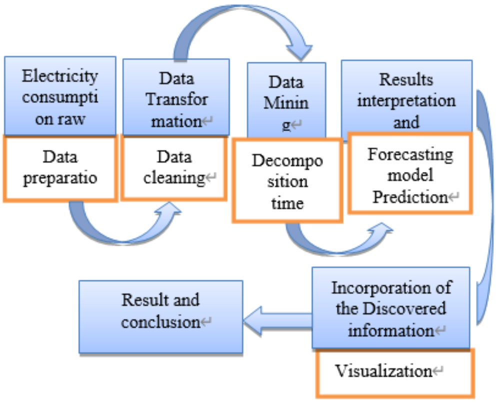
Electricity consumpti on raw
Data Transfor mation
Data Minin g
Results interpretation and
Data preparatio
Data cleaning
Decompo sition time
Forecasting model Prediction
Result and conclusion
Incorporation of the Discovered information
Visualization

Fig. 1. Analyzing power usage household using clustering analysis.

A. Date Collection

This study obtained a data set of electricity consumption of a household from December 2006 to November 2010 (47 months) from an open source database. After 4 years (1410 days) of 60-second granularity recording, the total power consumption of customer service is 2075259.By detecting the weight distribution of missing values or zeros, a poor weight distribution can be roughly identified [7].Use R and RStudio to model. Fig. 2 depicts the reduced power consumption using summarise. ALL (FUNs (sum)) function in the R script code.

1) Sub_metering_1: Kitchen (microwave, oven and dishwasher)
2) Sub_metering_2: Laundry room (washing machine, refrigerator, light and dryer)
3) Sub_metering_3: Air conditioning and water heater

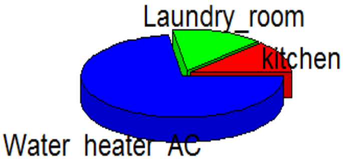

Fig. 2. Power consumptions by sub-metering.

B. Measurement Description

Preprocess the original data

Clean up data: Change the time format to a readable date format. Sort time series data and modify inconsistent data

Normalization: convert the data x={x_1, x_2, … x_N} into the range of [0,1], as shown in equation (1).To reduce the impact of abnormal data. Fig. 3 shows the electricity consumption in different months.

Manuscript received September 9, 2020; revised April 22, 2021.

Cuihua Tian is with the School of Computer and Information Engineering, Xiamen University of Technology, China (Corresponding author: Jianhong Qian; e-mail: qianjianhong2020@outlook.com, tiancuihua@xmut.edu.cn).

doi: 10.18178/ijmlc.2022.12.2.1081

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \qquad (1)$$

x_i': normalized power consumption

x_i: actual power consumption

$x_{\max}$: maximum power consumption

$x_{\min}$: minimum power consumption

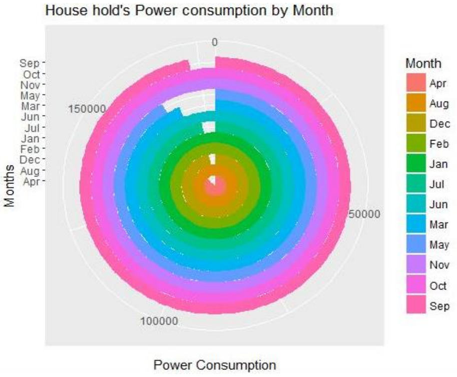

Fig. 3. Power usage per year in different months.

Data conversion: construct attributes, replace or add new class values.

C. Feature Selection and Transformation

1) Data visualization and transformation. First, read the data from the file, use the R language to convert the date format into a proportional format, and construct a data frame. As shown in Table I, analyze the data of different plots.

TABLE I: DATA FRAME READABLE SCRIPT ELECTRIC POWER CONSUMPTION

Date	Time	Global Active power	Global Reactive power	voltafge	Global_ intensity
16/12/2006	17:24:00	4.216	0.418	234.840	18.400
16/12/2006	17:25:00	5.360	0.436	233.630	23.000
16/12/2006	17:26:00	5.374	0.498	233.290	23.000
16/12/2006	17:27:00	5.388	0.502	233.740	23.000
16/12/2006	17:28:00	3.666	0.528	235.680	15.800
16/12/2006	17:29:00	3.520	0.522	235.020	15.000
16/12/2006	17:30:00	3.702	0.520	235.090	15.800
16/12/2006	17:31:00	3.700	0.520	235.220	15.800
16/12/2006	17:32:00	3.668	0.510	233.990	15.800
16/12/2006	17:33:00	3.662	0.510	233.860	15.800
16/12/2006	17:34:00	4.448	0.498	232.860	19.600
16/12/2006	17:35:00	5.412	0.470	232.780	23.200
16/12/2006	17:36:00	5.224	0.478	232.990	22.400

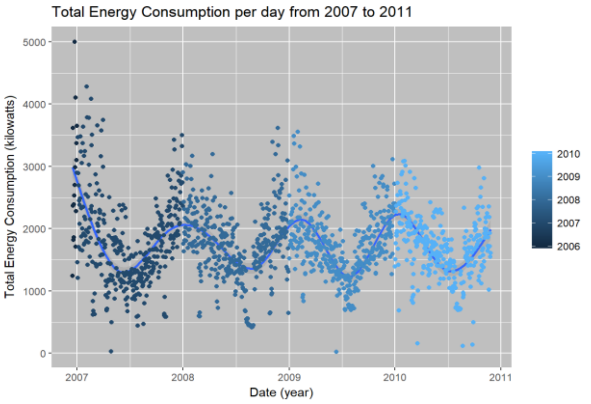

Fig. 4. Total energy consumption per day from 2007 to 2011.

2) Household electricity consumption throughout the year. Observe the energy expenditure of the household and add up the energy recorded every minute to form a Fig. 4 for the total daily energy consumption from 2007 to 2011.It can be seen from the figure that the electricity consumption trend in the past four years, the peak of electricity consumption is mainly concentrated in the year and the end of the year, and the lowest peak of electricity consumption is around the middle of the year.

D. Time Series Data

Time series is a sequence of information points, which can be divided into univariate and multivariate, usually the observation or continuous measurement of quantifiable variables in a certain time interval [8], [9]. Time series decomposition.

1) The purpose of changing the data format of time series data is to use historical data to predict the energy consumption of a household in a certain period of time. Use the aggregation script in R to format the data information and change the data every month. As shown in Table II, monthly usage (Kwh)) and maximum demand (Kwh)) for a given month are recorded during the recording period.

TABLE II: MONTHLY DATA FORMAT

	Month	Max_Demand_kW	Total_Use_kWh
1	2007-01-01	9.272	1150.1977
2	2007-02-01	9.410	941.4814
3	2007-03-01	10.670	981.0365
4	2007-04-01	8.160	586.3875
5	2007-05-01	7.672	733.4812
6	2007-06-01	7.614	594.7138
7	2007-07-01	7.240	495.0638
8	2007-08-01	8.694	568.2743
9	2007-09-01	8.110	697.8768
10	2007-10-01	9.036	821.2728
11	2007-11-01	9.326	931.9774
12	2007-12-01	9.686	1210.0695
13	2008-01-01	10.162	1086.1564

2) Analyze data exploration. According to the monthly summary in Table II, Fig. 5 is drawn, showing the pattern of high in winter and low in summer. According to the graph description this data seems that was taken in Europe.

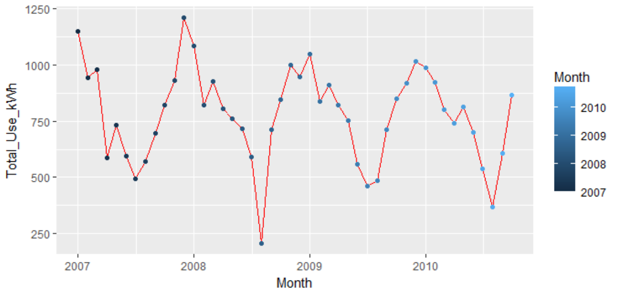

Fig. 5. Total usage plotting.

E. Construction and Time Series Forecasting Model

Use time series to build a data series model, and compare the forecast model that automatically fits the time series data with other forecasting methods.

1) Automatic model time series forecast ETS (A, N, A).Use the ts() function in the R library to construct a time series,

create an automatic model prediction function from the prediction library package, specify the monthly sequence frequency as 365, 12, and predict the total global active power in the next 25 months. Starting from December 2006, the parameters Frequency = 12 and Start = C (2006, 12) will be set once a week. The result graph and summary view of the model are shown in Fig. 6.

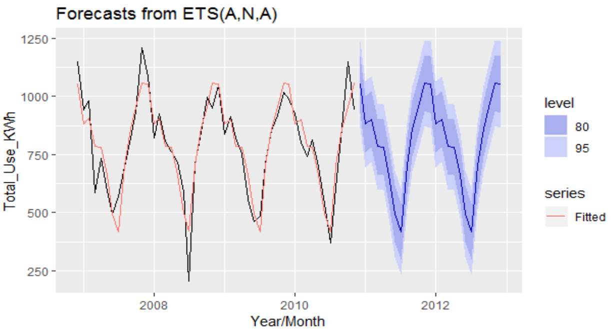

Fig. 6. Automated model time series forecasts from ETS (A, N, A).

2) Forecasting method. Simple exponential smoothing (SES).A simple exponential method is proposed to predict the global average active power. Test the construction of train data from 2006 to 2010, and predict train data values from 2011 to 2013 (25 months). Set beta and Gamma to false. Fig. 7 shows the trend of warm winter and cool summer, and the Fig. 8 shows simple exponential smoothing of seasonal component filtering. Smaller value of α would lead to fitting series values would be smoother.

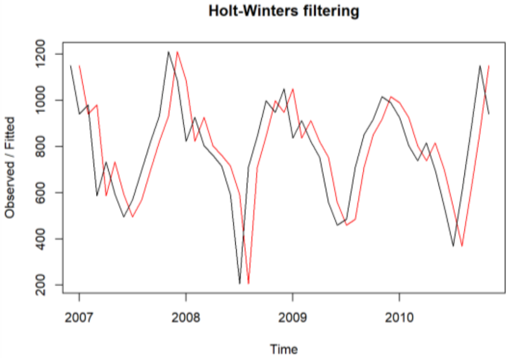

Fig. 7. Holt-winters filtering.

The forecast is further smoothed for more time points. This model is used to predict the seasonal values and trends of future time series. The trend component increases the seasonal index and estimates three parameters (smoothing). Fig. 9 shows the linear trend of alpha and beta, and gamma is set to true for seasonal adjustment.

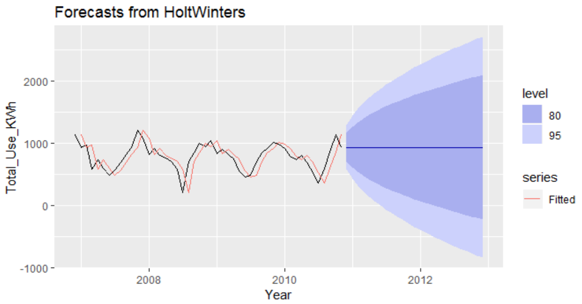

Fig. 8. Forecasts from holt-winters.

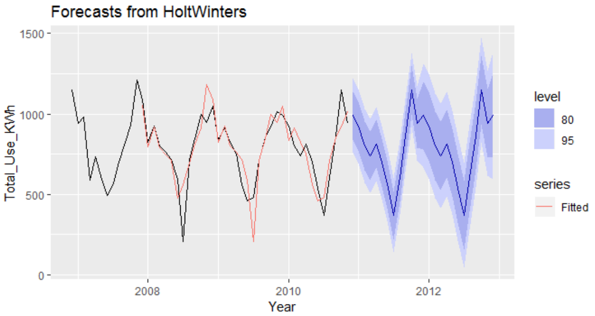

Fig. 9. Forecasts from holt-winters exponential smoothing.

3) The arima model arima is an autoregressive integrated moving average model, which is an extension of the autoregressive and moving average model (ARMA). The past error term of the model uses MA (moving average), and the lag value of the regression model [10] uses the AR(autoregressive) part. Use the auto-arima() function in the R package "Forecast" to start a new data forecast based on the forecast trend. This function calculated the best AR, difference and MA factors w.r.t the dataset used for training. Fig. 10 describes the ARIMA model.

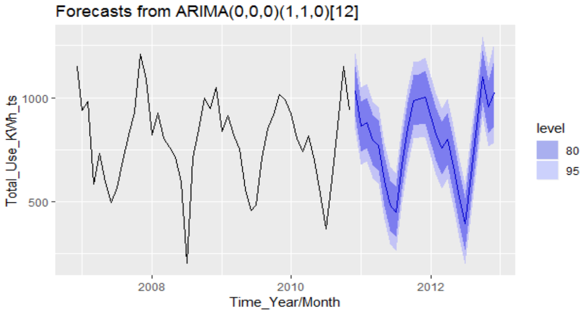

Fig. 10. Forecasts from ARIMA model.

4) Advanced forecasting methods. Due to many problems, the existing model needs to be modified. The Holt-Winter exponential smoothing model was revised by adopting the multi-season exponential smoothing method. EGRV is a multi-equation model with daily data granularity.

5) Predictive model evaluation. The evaluation of the predictive model converts the available data into training data and test data. The evaluation index of the prediction model is based on the average absolute error (MAE) and the average absolute percentage error (MAPE). The errors of different models are shown in Table III.

TABLE III: ERRORS FROM DIFFERENT MODEL ETS AND ARIMA

Criteria Model	ETS (A,N,A)	ARIMA	Holt's winter without Trend and Seasonality	Holt's winter Trend and Seasonality
ME	-5.66	-8.72	-44.4	-6
RMSE	79.28	80.36	179.26	115.53
MAE(Kw)	59.73	54.88	148.64	80.87
MPE	-2.89	-2.86	-5.47	-4.67
MAPE	9.53	8.972	22.93	15.08
MASE	0.63	0.581	1.57	0.86
ACF1	-0.11	-0.045	0	-0.17

It can be seen from the table that MAPE has the smallest error compared with other ETS models, and MAE (mean absolute error) is 54.88kW slower than other models. So the best model is ARIMA.

F. Monthly Trend and Forecasting Results

Use the key elements of monthly trend analysis to extract the time series, create an interactive HTML browser for analysis and decision-making, and use the R library package HTML Widgets for future devices. Fig. 11 predicts the monthly power consumption of html files.

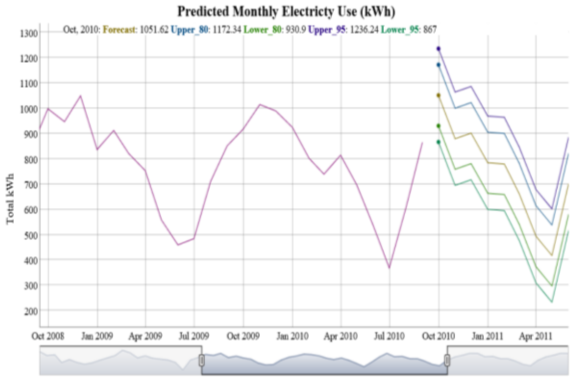

Fig. 11. Monthly prediction.

The results show that the total monthly electricity consumption is used to predict the annual cycle, with troughs in summer and peaks in winter. Electricity consumption will peak in September and October at 930-1,176 kWh (78% of the forecast range), and will steadily decrease during winter and early spring.

G. Implementation "Discovery Information"

1) Case study. Due to the large data set, in order to analyze and achieve better visualization in a short time, the 60-second time interval of the original data is summarized into a 60-minute time interval to obtain a better smooth graph.

2) Draw a power consumption diagram. The variables we check are often global active power and global average household active power per minute. In summary statistics, the energy consumption cycle is different, as shown in Fig. 12.

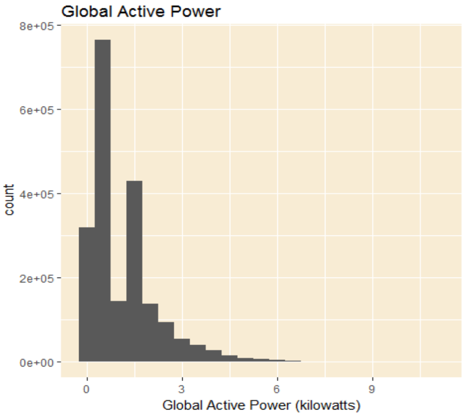

Fig. 12. Histogram of global active power over time.

3) Implemented more than one week. In order to better visualize the data set, from February 23 to March 1, 2009 (week), global active power, three-meters, voltage, and global reactive power were considered. Plot the power consumption on different days of the week to show the power consumption at different time periods. Fig. 13 shows the high power consumption from Friday night to Sunday night.

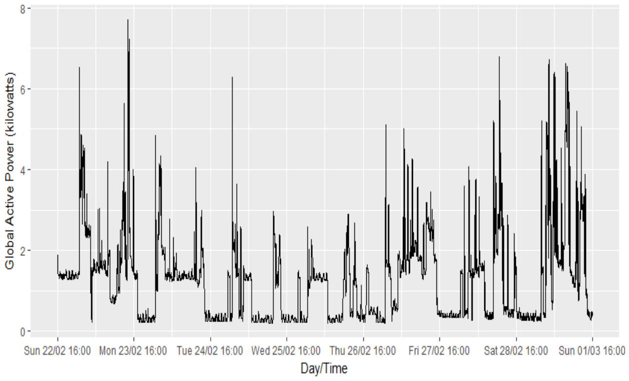

Fig. 13. Average global active power consumption over time.

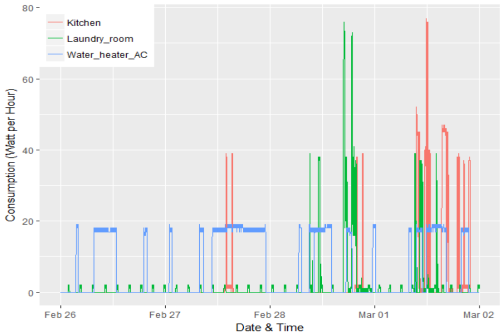

Fig. 14. Weekly power consumption of sub-metering.

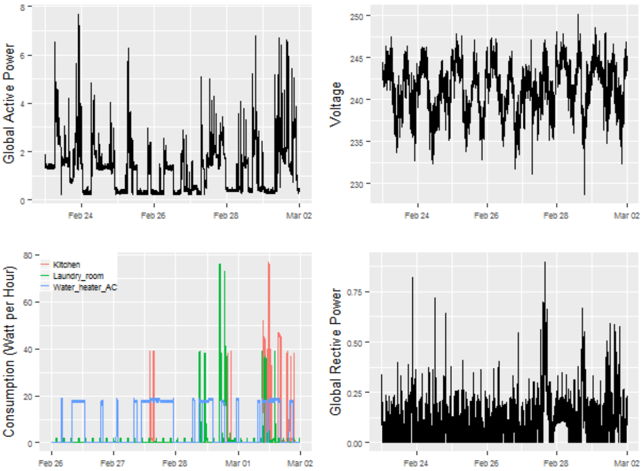

Fig. 15. Weekly power consumption for four parameters, global active power, voltage, sub-metering and global reactive power.

4) Determine the weekly energy consumption trend. According to the collected data from 26.02 to 02.03 and the weekly energy consumption trend in the following figure, it can be found that the peak power consumption is in the morning and evening on weekdays, and on Friday evening and the week with higher energy consumption and higher energy consumption. All day long. It can be seen from Fig. 14 that the electricity consumption of the kitchen is very low on Thursday, and the electricity consumption is very high on Friday, Saturday night and Sunday. Users seem to use the laundry room on Saturday morning, daytime, and Sunday noon. Air conditioners and water heaters are mainly used on Friday night, Saturday day and Sunday

day. The weekly power consumption cycle of the four parameters of spherical active power, voltage, section metering, and global reactive power is shown in Fig. 15.

III. Conclusion

The household electricity consumption data is obtained from the personal household electricity consumption [3] online library, and cluster analysis is performed based on RStudio programming to establish a qualitative prediction model for household electricity consumption data. It can better analyze and visualize the electricity consumption in a short time, let users understand the billing situation of the electric company, and help the electric company find the person who steals the electricity.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Cuihua Tian, Mugisha Theophile concuted the research; Jianhong Qian, Yiping Zhang, and Zhigang Hu analyzed the data; Cuihua Tian, Jianhong Qian wrote the paper; all authors had approved the final version.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 61976183), Science and Technology Planning Project of Xiamen (No. 3502Z20193058), Xiamen University of Technology Quality Graduate Courses and Case Library Construction Project (No. YG20190303) and China University Computer Education MOOC Alliance Construction Course: Analysis and Practice of Profit Model of Internet Thinking-E-commerce entrepreneurship.

References

[1] C. Olaru, P. Geurts, and L. Wehenkel, "Data mining tools and application in power system engineering," in *Proc. the 13th Power System Computation Conference*, 1999, pp. 324–330.

[2] R. Piacentini, "Modernizing power grids with distributed intelligence and smart grid-ready instrumentation," *Innovative Smart Grid Technologies (ISGT)*, pp. 1–6, 2012.

[3] UCI machine learning repository: Data set. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Individual+household+electric

[4] Y. N. N. Tchokonte, *Real-time Identification And Monitoring of the Voltage Stability Margin in Electric Power Transmission Systems Using Synchronized Phasor Measurements*, Kassel University Press GmbH, 2009.

[5] T. F. Sanquist, H. Orr, B. Shui, and A. C. Bittner, "Lifestyle factors in US residential electricity consumption," *Energy Policy*, vol. 42, pp. 354–364, 2012.

[6] A. Todd, P. Cappers, and C. Goldman, "Residential customer enrollment in time-based rate and enabling technology programs," *Lawrence Berkeley Natl. Lab.*, 2013.

[7] UCI machine learning repository: Individual household electric power consumption data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption

[8] H. Chen, W. L. Lee, and X. Wang, "Energy assessment of office buildings in China using China building energy codes and LEED 2.2," *Energy Build.*, vol. 86, pp. 514–524, 2015.

[9] A. Coghlan, *A Little Book of R for Time Series*, vol. 10, 2015.

[10] P. J. Brockwell, R. A. Davis, and M. V. Calder, *Introduction to Time Series and Forecasting*, vol. 2. Springer, 2002.

Cuihua Tian lives in Xiamen City, Fujian Province, China. In 1994, she obtained a bachelor of engineering degree in computer and application from Shenyang Institute of Technology. In 2002, she received a master degree in computer software and theory from Shenyang Institute of Technology. She received her doctorate degrees from the Department of Computer System Architecture, Northeastern University. From 2002 to 2010, she served as an associate professor in the School of Information Technology and Science, Shenyang University of Technology. And she has been a full-time teacher at the School of Computer and Information Engineering, Xiamen University of Technology from 2010. She have been published 2 works and about 20 papers such as "Design and analysis of algorithms," at China Metallurgical Press in August 2007, "Simulation research on Traffic Information Service of Internet of Things based on GT4," at Xiamen University in 2017, "Research on high-dimensional data reduction" at *International Journal of Database Theory and Application*, vol. 9, no. 1 in Jan 2016. She has long-term researches on the Internet of Things, cloud computing, big data mining, intelligent information processing, algorithms, and games.