

Urban Electricity Consumption Forecasting Based on SARIMA and Random Forest Modeling

Sijian Zhao*

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Wuhua District, Kunming City, Yunnan Province, 650221, China

Zhao_Si-Jian@hotmail.com

**Corresponding author*

Keywords: Power Consumption Forecasting, SARIMA, Random Forest, Time Series Analysis, Machine Learning

Abstract: This project proposes to use a combination of machine learning and time series analysis to provide in-depth analysis and forecasting of electricity consumption in a city in North Africa. The dataset used in this study contains a range of information including date, temperature, humidity, wind speed, total flow, and electricity consumption. The project proposes to reveal patterns and patterns of electricity consumption behavior through data preprocessing, normalization, and seasonal decomposition. The project proposes to use two models: Seasonal Autoregressive Integrated Sliding Average (SARIMA) and Random Forest based on feature engineering. The SARIMA method is used to analyze the seasonality and trend of the time series data, and the Random Forest method is used to study the nonlinear relationship between electricity consumption and environmental factors. On this basis, we add more information such as rolling rolling standard deviation, minimum large value, and time-delayed features to the random forest. This method greatly improves the prediction accuracy of power consumption. The experimental results show that compared with the single SARIMA model, the random forest model using j combined with the feature engineering method can better predict the load changes of the power system. The results show that the Random Forest model can capture the complexity of power consumption more effectively, especially after adding detailed feature items. At the same time, the good interpretability and flexibility of Random Forest makes the model able to better understand and predict the urban power demand, which can effectively help the power grid enterprises to realize the optimal allocation of resources and reduce energy consumption.

1. Introduction

1.1 Research Background

At a time of growing global energy thirst and accelerating urbanization, the issue of power consumption in cities is particularly critical and has become a central issue in energy layout and scheduling. This explosive growth in demand is not only a product of economic expansion,

technological innovation, and government policy, but also highlights the urgency of its development [1]. Predictive power is also a mainstream direction at present, because people's lives have been inseparable from the supply of electricity, predictive power can be a great degree of resource allocation optimization, the significance of which is beneficial to the country and the people. In addition, power load forecasting is the key to ensure the safe and stable operation of the power grid, as well as to realize the rational allocation of energy. On this basis, a new new technology is proposed based on power density. It is necessary to reduce the loss of electric energy caused by improper configuration, and also to guide the power supply and demand to achieve the optimal balance through the forecast value [2]. Load forecasting is an important part of the power system, especially important in real-time grid scheduling, energy market transactions. Accurate load forecasting is of great significance in reducing the uncertainty of energy use, lowering energy costs, and improving the safe and reliable operation of the power grid [3]. On this basis, the project proposes to combine the SARIMA model with the machine learning-based random forest model to predict the power consumption of power plants, so as to better serve the energy conservation and emission reduction.

1.2 Research purpose and significance

Aiming at this problem, this project intends to compare and evaluate SARIMA and Random Forest, to explore the application effect of the two in the prediction of electric energy data, and to improve the two from the perspective of feature engineering. The research results of this project will optimize energy management and planning, promote model innovation, assist decision-making, improve decision-making level, promote scientific decision-making, promote the combination of theory and practice, and provide new research ideas for solving complex data problems.

2. Literature Review

2.1 Development status of the electricity market

In recent years, some scholars have used traditional statistical methods to predict electricity consumption, while Niyozov NN et al. used the PCA method to construct an electricity consumption prediction model for the electricity market and compared it with the actual electricity consumption data, and concluded that the difference between the two is very small [4]. Bilgili M et al. proposed a machine learning model based on long and short-term memory (LSTM) to predict electricity consumption [5]. In view of this, this project proposes to build on the previous work by proposing a methodology that incorporates time series and machine learning to improve the accuracy of electricity consumption prediction.

3. Modeling and Algorithm Design

3.1 Modeling

The SARIMA model is as follows [6]:

$$(1 - \sum_{i=1}^p \phi_i L^i)(1 - L)^d (1 - \sum_{i=1}^P \Phi_i L^{i \times s})(1 - L^s)^D X_t = (1 + \sum_{i=1}^q \theta_i L^i)(1 + \sum_{i=1}^Q \Theta_i L^{i \times s}) \varepsilon_t \quad (1)$$

L is the lag operator, ϕ_i is the autoregressive parameter, Φ_i is the seasonal autoregressive parameter, θ_i is the moving average model parameter, Θ_i is the seasonal moving average parameter,

X_t is the original time series at time point t , and ε_t is the random error term.

Model evaluation was measured using the MAE (Mean Absolute Error) metric. The mathematical model is as follows [7]:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

n is the number of observation points, y_i is the i -th actual observation, and \hat{y}_i is the i -th predicted value.

3.2 Algorithm construction

(1) Data Preprocessing and Feature Engineering

The data is first loaded and cleaned, where DateTime columns are converted to datetime format and set as indexes. The data was aggregated by day to calculate the average power consumption, followed by normalization using MinMaxScaler to ensure consistency in model training. In addition, multiple features were constructed, including rolling statistics, lagged features (consumption on the previous day and the previous seven days), and date-based features to provide rich input variables for subsequent model training.

(2) Time Series Analysis and Model Selection

Seasonal decomposition is executed on the normalized time series to clarify its trend, seasonality and stochastic components. The auto_arima automatic selection algorithm is utilized to determine the appropriate SARIMA model parameters for this data to ensure that the model adequately captures the seasonality and other time series characteristics of the data.

(3) Model Training and Forecasting

The SARIMA model was trained based on the automatically selected parameters and used to forecast electricity consumption for the next 30 days. At the same time, the random forest model is adapted and trained by grid search with time series segmentation methods in order to find the best prediction method among different data features and model dynamics.

(4) Result Evaluation and Visualization

Finally, the prediction results of SARIMA and Random Forest are combined to optimize the prediction accuracy by weighted averaging, and the performance is evaluated by calculating the Mean Absolute Error (MAE) of the model on the test set. The model's prediction effect is visualized by plotting the actual values against the predicted values, which can clearly show the model performance and prediction accuracy.

4. Case Studies and Analysis

The dataset used in this paper is derived from the UCI database of electricity consumption data for the city of Tetouan, Morocco over a certain period of time [8].

In this project, Zone 1 power consumption in the dataset is selected as the independent variable, and the blue bars in Figure 1 indicate the power consumption at different times, with each bar representing a point in time. We can observe that the consumption fluctuates within a certain range, with several distinct peaks and troughs. Peaks usually represent an increase in electricity consumption during a specific time period.

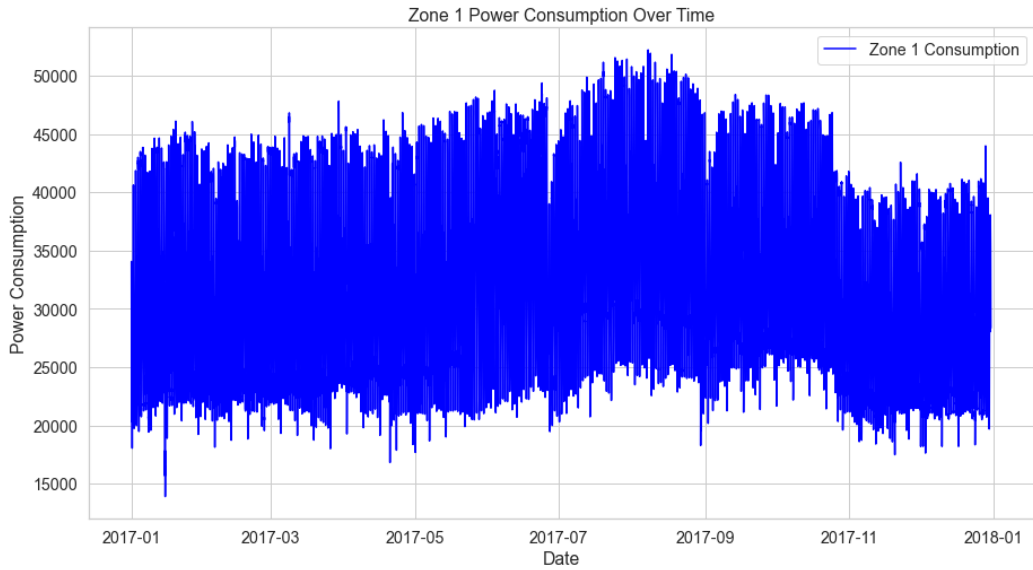


Figure 1: Zone 1 power consumption over time

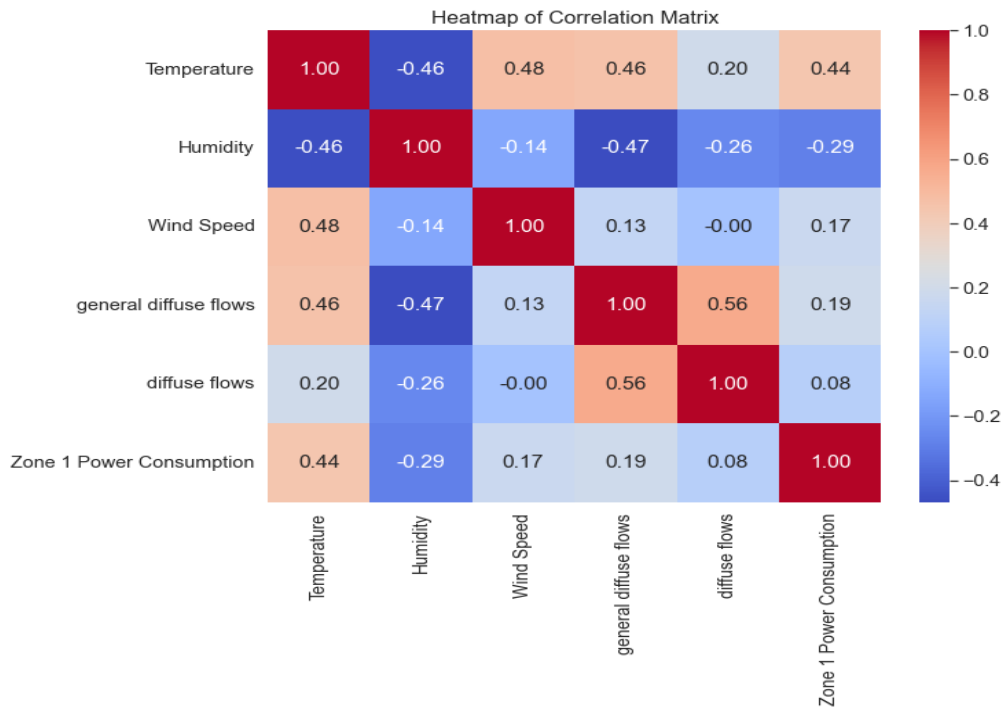


Figure 2: Heatmap of correlation matrix

The Figure 2 is a heat map of a correlation matrix that shows the correlation coefficients between different variables. The correlation coefficients range from -1 to +1, where +1 means perfectly positive correlation, 0 means no correlation, and -1 means perfectly negative correlation. Colors closer to red indicate a positive correlation and colors closer to blue indicate a negative correlation. Lighter colors indicate weaker correlations and darker colors indicate stronger correlations.

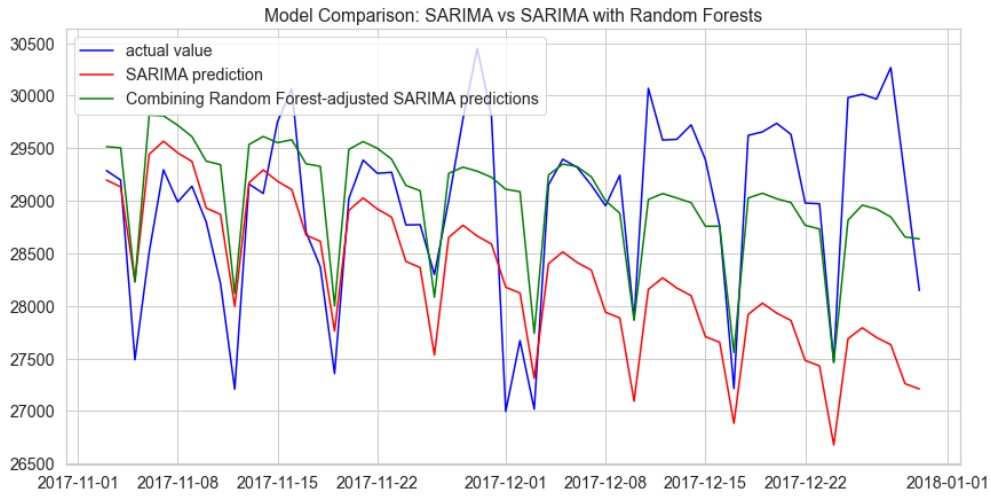


Figure 3: Model Comparison: SARIMA vs SARIMA with Random Forests

As can be seen from the Figure 3, the predictions of the SARIMA model (red line) deviate significantly from the actual values (blue line) at some points, whereas the SARIMA predictions adjusted by incorporating random forests (green line) are closer to the actual values at many time periods. This indicates that the accuracy of the predictions is improved by introducing random forests to adjust the SARIMA model. Overall, the SARIMA model combined with random forests provides more accurate forecasts over this time horizon, demonstrating the potential of machine learning techniques used in conjunction with traditional statistical models in time series analysis.

Table 1: MAE of the models.

MAE	VALUE
SARIMA Model	944.903
Adjusted SARIMA Model	562.767

As shown in Table 1, the forecasting accuracy of the SARIMA model corrected with random forest is significantly improved over the original SARIMA model, with the error reduced by about 40%. The results of the study show that the use of machine learning (here Random Forest) to optimize the classical time series data can effectively improve the forecasting effect of the model.

5. Conclusion and Recommendations

5.1 Conclusion

This project intends to carry out an in-depth study on urban power consumption based on the previous work, and on this basis, a random forest machine learning model based on feature engineering optimization is used to realize the prediction of urban power consumption based on the SARIMA statistical model. Compared with the traditional SARIMA model, the mean absolute error (MAE) of the random forest model constructed using the feature engineering method is reduced from 944.903 to 562.767, which is about 40%. The research results of this project not only show that combining machine learning and traditional statistical methods organically has good application prospects, but also has high computational efficiency.

The research results of this project will provide new ideas and technical support for the study of electricity load forecasting, and also provide scientific data support for the development of power grid enterprises. On this basis, the enterprises can make accurate prediction and management of electricity demand, realize the optimal allocation of resources and reduce energy consumption; and

promote the organic combination of urban and rural planning and sustainable development on a larger scale. This is of great practical significance for the development of China's electric power industry itself and how to cope with the energy crisis and realize green development.

5.2 Recommendations

The research results of this paper show that the urban power consumption has been effectively predicted by using SARIMA, Random Forest and other methods, and combining with feature engineering techniques. On this basis, for future research and practice, this project intends to discuss the following recommendations:

1) Expansion of data sources: Future research should include more diversified data sources, such as economic indicators, population density and urban development, to enrich the data sources and discover more laws that have an impact on power consumption.

2) Real-time data adjustment: By correcting the real-time data to make it more dynamic and instantaneous, it facilitates the short-term load balancing of the power system and contingency planning for emergencies.

3) Interaction test with multiple cities: The method is empirically tested by utilizing a variety of climatic, socio-economic and other factors to enhance the practicality and credibility of the method.

4) Explore hybrid models: This project needs to further explore the integration of other hybrid models to overcome the shortcomings of a single model by integrating traditional statistics and cutting-edge machine learning algorithms.

References

- [1] Jiang, P., Li, R., Lu, H. et al. *Modeling of electricity demand forecast for power system. Neural Comput & Applic* 32, 6857–6875 (2020).
- [2] Liu, Y. Q., & Chen, X. F. (2023). *Research on Power Consumption Forecasting Based on CNN+LSTM Hybrid Model. Computer and Telecommunications*, (07), 0065-0070.
- [3] Daneshi, Hossein, and Ali Daneshi. "Real time load forecast in power system." *2008 Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies. IEEE, 2008.*
- [4] Niyozov N N, Rikhsitillaev B K, Jalilova D A, et al. *Forecasting of electricity consumption by industrial enterprises with a continuous nature of production based on the principal component method (PCA)[C]//E3S Web of Conferences. EDP Sciences, 2023, 384: 01031.*
- [5] Bilgili M, Pinar E. *Gross electricity consumption forecasting using LSTM and SARIMA approaches: A case study of Türkiye [J]. Energy*, 2023, 284: 128575.
- [6] Zhang, Jingxin. *Mathematical Modeling: Algorithms and Programming Implementations. Beijing: China Machine Press, 2022.*
- [7] G éron, Aur élien. *Hands-On Machine Learning with Scikit-Learn and TensorFlow. Beijing: Mechanical Industry Press, 2018.*
- [8] <https://archive.ics.uci.edu/dataset/849/power+consumption+of+tetouan+city>