A Case Study of Industrial Time Series Analysis Methods and Simulation Experiment Design - An Example of Energy Consumption Data

Weixiao Liang, Jin Hao, Ze Wang

***Abstract*— In order to prevent issues such as cost escalation and efficiency reduction in industrial production caused by energy shortage, it becomes particularly critical to forecast energy consumption. This paper intends to adopt the classic model in time series analysis methods--Seasonal Auto-regressive Integrated Moving Average Model (SARIMA), considering the tendency and seasonality traits of energy consumption data. Through statistical analysis of the data, we identify whether there exists trend and seasonality and determine the parameters of the SARIMA model based on the results of data analysis for energy consumption data modeling and forecasting. Experiments were conducted with industrial energy consumption data and the results demonstrated that this method could effectively predict energy consumption.**

Index term*—*Industrial Energy Consumption Forecasting, Time Series Analysis, SARIMA

I. INTRODUCTION

Industrial production occupies an extremely important place in the economic and social development of a country and can have a huge impact on both economic growth and environmental change. Forecasting energy consumption helps industrial companies to better plan and adjust their production schedules and to manage and conserve energy more effectively[1]. Forecasting energy consumption also helps governments and related organizations to make decisions, such as adjusting energy policies in a timely manner and rationally allocating energy resources to meet the needs of different industries and regions.

Since industrial energy consumption data are typically time-series data, time-series analysis methods are currently the main means of forecasting energy consumption data. Autoregressive Integrated Moving Average (ARIMA) model has been widely used as a typical time series model in various fields [2].

In cases where the data behave as time series, ARIMA has an advantage over other similar methods by identifying the appropriate model to best fit the respective time series [3].

Manuscript received April 12, 2024

Weixiao Liang, School of Computer Science and Technology, Tiangong University, Tianjin, China

Jin Hao, Boya Chuangzhi (Tianjin) Technology Co., LTD.Tianjin, China

Ze Wang, School of Software , Tiangong University, Tianjin, China

To recommend the nature of energy policies. Liu et al [4] used ARIMA model in wireless sensor information collection to save energy. Saab et al [5] used ARIMA model to predict monthly electricity consumption in Lebanon in one step. Contreras et al [6]used ARIMA model to predict the next day's electricity price. Pappas et al [7] proposed an ARIMA model for predicting Greek electricity consumption with an ARIMA model and compared the model with three analytical time series models, which showed that the ARIMA model is more effective than the other time series models.

On the basis of the above research, this paper starts from the trend and seasonality of the industrial energy consumption data itself, decomposes and analyzes the data, and uses the time series analysis SARIMA model to model and forecast the data, and finds that the model can be well applied to this problem.

II. PROBLEM DEFINITION

A. Time series analysis methods

This paper conducts research on industrial energy data using time series analysis methods. Time series analysis is a statistical technique used to process data generated by various observed items at successive points in time. These observational items can be stock prices, energy consumption, or any other data arranged in chronological order. This type of series is one of the most common forms of data and is characterized by the fact that each observation is recorded at a specific point in time and there is a temporal dependency between the data. The following are the general methods and steps involved in time series analysis, The process is shown in Figure 1:

Understanding and Cleaning Data: The foremost step involves understanding and preprocessing the data. The correctness and completeness of data bear great significance to the precision of the analytical outcomes. Data cleaning encompasses processes such as outlier removal and missing data imputation.

Data Exploration: In this phase, we venture to understand the fundamental characteristics and structure of the data, including its distribution, periodicity, and trends, among others. Tools such as STL((Seasonal and Trend decomposition using Loess)) decomposition and autocorrelation plots facilitate this exploratory process.

Model Construction: Subsequent to the preliminary observations and analyses, we proceed to select an

appropriate model, such as commonly employed ARMA or SARIMA models.

Model fitting: Then, utilizing the optimal parameters, we execute software computations to yield the model and fit it to our data.

Model Verification: After the model has been fitted, it's necessary to review the model for stability and validity. Common methods for this purpose include residual tests and white noise tests.

Forecasting: The final step is using the model for forecasting, provided the model is compliant with the requisite statistical assumptions.

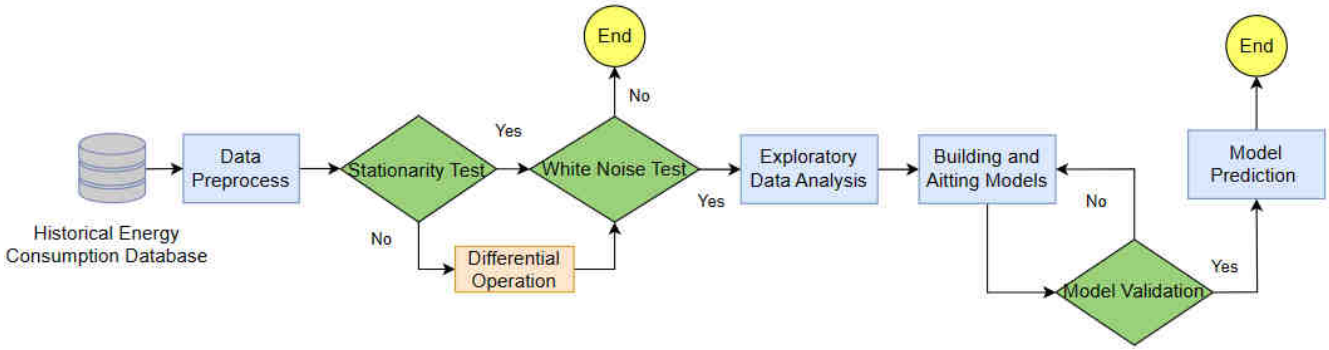

FIG 1. Time series analysis process

B. Establishment of SARIMA model

Energy consumption data possesses certain specific characteristics due to external influences. For example, energy demand in winter may be higher than in summer, leading to potential seasonal variations in energy consumption data. Furthermore, energy consumption data may exhibit long-term upward or downward trends due to policy influences. Addressing these characteristics is a key function of this model. The SARIMA model can be represented as SARIMA(p,d,q)(P,D,Q)s, where lowercase symbols represent the non-seasonal part of the model, with p denoting the autoregressive parameter, d denoting the differencing parameter, and q denoting the moving average parameter. Uppercase symbols represent the seasonal part of the model, indicating the number of periods for each season. For example, the annual cycle is represented by s = 12. The mathematical formalization of SARIMA is shown in Equation (1).

$$\varphi_p(B)\psi_P(B^S)\nabla^d\nabla_S^D y_t = \theta_q(B)\Theta_Q(B^S)\varepsilon_t \quad (1)$$

For the prediction of natural gas consumption data, where B is the lag operator, p is the order of the regular AR polynomial, q is the order of the MA polynomial, P is the order of the seasonal AR polynomial, Q is the order of the seasonal MA polynomial, d is the non-seasonal differencing order, D is the seasonal differencing order, and $\varepsilon_t$represents the prediction error. The acquisition method of the SARIMA model can be summarized into four steps: identification, estimation, diagnostic checking, and prediction [8].

III. SIMULATION AND PREDICTION

A.Data Set

To effectively forecast energy consumption sequences in the industry using the SARIMA model, this section employs the monthly data of industrial sector natural gas consumption from the U.S. Energy Information Administration (EIA) spanning January 1, 2000, to August 1, 2023. This dataset consists of 284 data points. Given that most regions worldwide sample energy consumption on a monthly, quarterly, or annual basis, the adoption of monthly data is pursued with the intent to maximize the size of the dataset and to ensure that the SARIMA model can capture an increased amount of information. Furthermore, it's considered that data prior to the year 2000 will not significantly affect the current predictions due to shifts in industrial development, economic policies, and environmental conditions. The methodology used in this study necessitates the chronological arrangement of the training set. Accordingly, the training and test sets are not arbitrarily divided but are partitioned based on their chronological sequence. The data from the final year is utilized as a test set to validate performance, while the remaining data forms the training set.

B.Exploratory data analysis

Before modeling and forecasting the energy consumption data, it is necessary to conduct an exploratory analysis of the data to understand the general situation of the data and analyze the trend and seasonality of the data, and the results of the data analysis will provide a reference for the construction of the subsequent model. The time series diagram of the natural gas consumption data is shown in Figure 2.

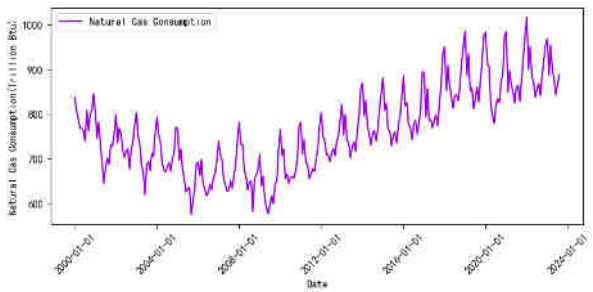

FIG 2. Natural Gas Consumption Sequence Diagram

Intuitively observing the data, it can be seen that the consumption of industrial natural gas showed a continuous downward trend before 2010, and then continued to rise until 2023. And from the time series diagram, we can roughly see that there seems to be some regular fluctuations.

Before further decomposition analysis, we need to check its stationarity and whether it is a white noise sequence.

In time series forecasting methods, conducting a stationarity test on the data is essential. This is not only because some time series models require data to be stationary, but also because stationarity testing can determine whether the data possesses stable statistical characteristics, i.e., whether the mean and variance of the data change over time. Stationary data exhibits favorable properties in time series analysis, enabling more accurate prediction and modeling. The Augmented Dickey-Fuller (ADF) test is a commonly used method for unit root testing, employed to examine whether time series data possesses a unit root (i.e., non-stationarity). The ADF test results for the two energy datasets are presented in Table1 , Where CL represents Confidence level.

Table 1 ADF test results

T-test	P-value	1% CL	5% CL	10% CL
-0.58	0.876	-3.45	-2.87	-2.57

From the results of the ADF test, it can be seen that the T-test values of the data are much greater than the 10% confidence level, indicating that they are non-stationary. This means that in the subsequent modeling and prediction process, it is necessary to use differential operations to make the data stationary. The ADF test results of the data after first-order difference operation are shown in Table 2.

Table 2
First-order differential sequence ADF test results

T-test	P-value	1% CL	5% CL	10% CL
-3.97	0.0015	-3.45	-2.87	-2.57

It can be seen that the differential sequence has already shown as a stationary sequence. In order to determine the research value of this stationary sequence, it is necessary to test whether it is a pure random sequence. The Ljung Box method is used for white noise testing, with orders set to 6th and 12th, respectively. The results are shown in Table 3.

Table 3 White Noise test results

Lags	Lb_stat	Lb_Value
6	83.75	5.98e-16
12	295.24	4.68e-56

It can be seen that the Lb_value values of both 6th and 12th orders are far less than 0.05, indicating that they are non random sequences and can be modeled using the SARIMA method for research.

In order to see the trends and seasonality in the data more clearly, we use the STL method to decompose the data into components such as trends, seasonality, and residuals, as shown in Figure 3.

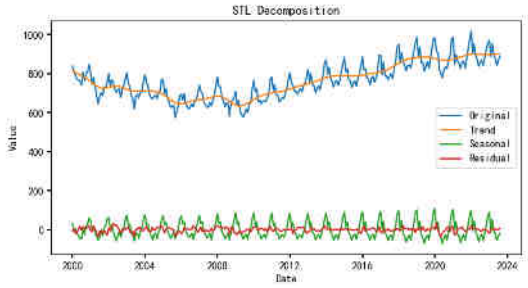

FIG 3. SLT decomposition diagram of Natural Gas consumption

Based on the results of STL analysis, natural gas consumption data shows significant periodicity in its seasonal curve, with a complete cycle occurring every 12 months. The residual plot describes the difference between the data decomposed by section and trend and the original data, representing the unexplained portion remaining after removing trends and seasonality. From the graph, it can be seen that the residual lines of both datasets show slight fluctuations around the 0 value, indicating that if a model can fully extract the trend and seasonality of the data, it should be able to capture the key features of the data.

ACF (Auto-correlation Function) and PACF (Partial Auto-correlation Function) plots are commonly used tools in auto-regressive analysis of time series, and they can be used to help us determine the best model for the time series, as well as to identify seasonality or outliers. The ACF , PACF plots for the Natural gas consumption series are shown in Figure 4.

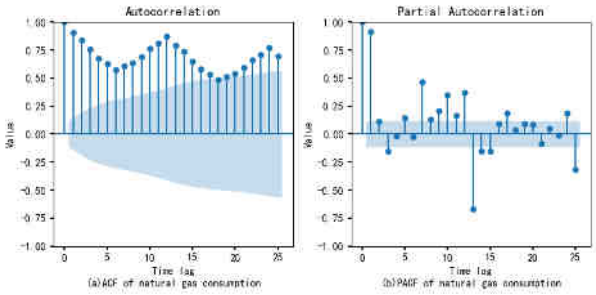

FIG 4. ACF and PACF of natural gas consumption.

Periodic peaks can be seen in the ACF plot for Natural gas consumption series, which shows the seasonal cyclical variation present in the series. The blue shaded area indicates the 95% confidence interval, and if the auto correlation coefficient lies outside the confidence interval, it suggests that the lag term may not be due to random noise. The PACF plot, on the other hand, shows the correlation between a lag and the current value after controlling for the effects of other lags. Both the ACF and PACF plots show that data have some degrees of auto-correlation, which provides a theoretical basis for subsequent SARIMA modeling.

C. Evaluation criteria

In assessing the effectiveness of models, the magnitude of the error between predicted values and actual values, i.e.,

accuracy, is undoubtedly the most important and convincing criterion. This study employs two statistical metrics, namely the Mean Absolute Percentage Error (MAPE) and the Root Mean Square Error (RMSE), to evaluate the model's performance, each defined by equations (2) and (3), respectively.

$$MAPE = \frac{1\times100\%}{N}\sum_{i=1}^{N}\left|\frac{P_{ai}-P_{fi}}{P_{ai}}\right| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_{ai}-P_{fi})^2} \quad (3)$$

Pai and Pfi represent the actual and predicted values of coal consumption at time i, and N is the number of samples involved in the prediction. MAPE is an unbiased estimator used to evaluate the predictive ability of a model, widely applied due to its intuitive interpretation of relative errors. RMSE measures the difference between the values predicted by a model or estimator and the actual observed values[9, 10].

D. SARIMA modeling and prediction

After exploratory data analysis of the data we started the prediction task. For the SARIMA model, according to the stationarity of the data, the number of differences d is set to 1, and it is determined that the domain of values of the p,q parameters should be [0,5], the domain of values of the P,Q parameters should be [0,2], D should be [0,1], and s should be 12. A grid search is used to search the parameter space to determine the AIC and BIC information criterion as a measure of the strength of the parameter combinations. The two information criteria will find a balance between the model complexity and prediction accuracy, so that the model has a high accuracy while having the smallest possible complexity, and the expressions of the two information criteria are shown in equations (4), (5).

$$AIC = 2k - 2ln(L) \quad (4)$$

$$BIC = kln(n) - 2ln(L) \quad (5)$$

where n is the number of samples, k is the number of parameters, and L is the likelihood function. The optimal parameters of SARIMA for the Natural gas consumption data are shown in Table 4.

Table 4 The optimal parameters of SARIMA

Parameters	AIC	BIC
SARIMA (1,1,1)*(1,1,1,12)	2227.908	2245.693

After finding the optimal parameters, we use the optimal model to fit the historical data and predict the next 12 time steps. The result is shown in Figure 5.

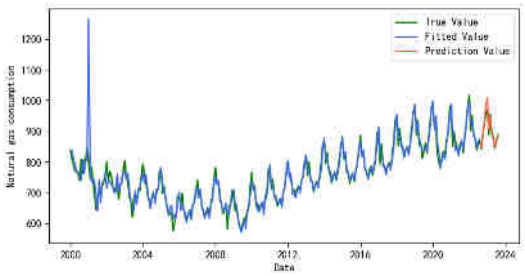

FIG 5. ACF and PACF of natural gas consumption

The graph indicates that apart from an initial few outliers, the majority of the data points are well fitted. Furthermore, the predicted values are strikingly close to the actual values, signifying the model's ability to effectively forecast the consumption of industrial natural gas. The forecast MAPE and RMSE values are 1.4% and 17.61 respectively, demonstrating the model's high level of forecast accuracy.

IV. CONCLUSION

Industrial energy consumption data often exhibit complex trends and seasonal periodicities. This paper embarks on a predictive task using the time series analysis method, taking the consumption of industrial natural gas as a case study. An in-depth statistical analysis is initially carried out on the data to understand distinct features within. Subsequently, using the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model, data modelling is performed. Parameter tuning and forecasting are accomplished via a grid search algorithm and the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) principles. Experimental results indicate that the SARIMA model is adept at fitting historical data and making precise predictions for future data points.

REFERENCES

1. Huang R, Shen Z, Yao X. How does industrial intelligence affect total-factor energy productivity? Evidence from China's manufacturing industry [J]. Computers & Industrial Engineering, 2024, 188: 109901.
2. Suganthi L, Samuel A A. Energy models for demand forecasting—A review [J]. Renewable and Sustainable Energy Reviews, 2012, 16(2): 1223-1240.
3. Sen P, Roy M, Pal P. Application of ARIMA for forecasting energy consumption and GHG emission: A case study of an Indian pig iron manufacturing organization [J]. Energy, 2016, 116: 1031-1038.
4. Chong L, Kui W, Min T. Energy efficient information collection with the ARIMA model in wireless sensor networks; proceedings of the GLOBECOM '05 IEEE Global Telecommunications Conference, 2005, F 28

Nov.-2 Dec. 2005, 2005 [C].

5. Saab S, Badr E, Nasr G. Univariate modeling and forecasting of energy consumption: the case of electricity in Lebanon [J]. Energy, 2001, 26(1): 1-14.
6. Contreras J, Espinola R, Nogales F J, et al. ARIMA models to predict next-day electricity prices [J]. IEEE Transactions on Power Systems, 2003, 18(3): 1014-1020.
7. Pappas S S, Ekonomou L, Karamousantas D C, et al. Electricity demand loads modeling using AutoRegressive Moving Average (ARMA) models [J]. Energy, 2008, 33(9): 1353-1360.
8. Box G E, Jenkins G M, Reinsel G C, et al. Time series analysis: forecasting and control [M]. John Wiley & Sons, 2015.
9. De Myttenaere A, Golden B, Le Grand B, et al. Mean absolute percentage error for regression models [J]. Neurocomputing, 2016, 192: 38-48.
10. Xu Y, Zhang J, Long Z, et al. A novel dual-scale deep belief network method for daily urban water demand forecasting [J]. Energies, 2018, 11(5): 1068.