Industrial Engineering and Innovation Management (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/ieim.2024.070323
ISSN 2522-6924 Vol. 7 Num. 3

Fluctuation analysis of natural gas forecast

Mengyang Li[1,*], Yang Wang[2], Yehui Bo[2], Huiyi Zheng[2], Xuanyao Yu[3]

[1]Kunlun Digital Technology Co., Ltd., Beijing, China
[2]Petrochina Kunlun Gas Co., Ltd., Shandong Branch, Jinan, China
[3]China University of Petroleum (Beijing), Beijing, China
wumianwork@gmail.com
[*]Corresponding author

Keywords: Data volatility, Entropy value, LSTM, filtering algorithm

Abstract: This study used indicators such as dispersion coefficient, entropy value, skewness, etc. to determine the data fluctuation of users' natural gas consumption. We collected 2736 daily natural gas consumption data from 12 users in a certain region, calculated the accuracy of natural gas prediction for each user using the LSTM model, and analyzed the correlation between various volatility indicators and MAPE. The results showed that the greater the volatility, the greater the LSTM prediction error, indicating a positive correlation between the volatility value and MAPE. Adding a filtering algorithm to the original data can effectively reduce the volatility of the original data, but the MAPE values of users have not all decreased. For example, users with smaller MAPE values, such as User 5, have increased prediction errors after adding the filtering algorithm. Users with larger MAPE values, such as User 2, have reduced prediction errors after adding the filtering algorithm. The filtering algorithm reduces volatility while partially missing the original data features, making it difficult for the prediction results to be completely accurate and ultimately maintain around 5%. However, adding missing filtering algorithms can effectively reduce some user prediction errors. The specific use of this algorithm depends on the prediction results of the model itself.

1. Introduction

With people's higher requirements for environment and environmental protection, the proportion of natural gas in primary energy is increasing year by year. However, natural gas is not suitable for storage, flammable and explosive, and the daily forecast of natural gas demand can help natural gas supply enterprises to complete procurement and scheduling. The prediction of natural gas daily consumption has attracted the interest of researchers all over the world and done a lot of research work. As far as traditional models are concerned, algorithms such as linear regression model[1], random forest[2] model and SVM model[3] all have good nonlinear fitting ability and have been widely used. In the application process, people use SVR model, BP neural network and linear regression to carry out comparative experiments, and the results show that SVM model usually has more accurate results than linear regression and BP neural network models[3]. Bai and Li et al[4]. developed a structurally calibrated SVR to predict Anqing's natural gas consumption, and the MAPE of the proposed model was 2.36%. In order to further improve the accuracy of model prediction, people fuse multiple models to generate a mixed model. Compared with a single model, the mixed model combines the advantages of many methods and can produce better prediction performance, with the lowest MAPE value of 2.36%. In addition to further improvement at the model level, a large number of researchers are trying to find features that can better express the relevant characteristics of natural gas consumption. Based on MLP and RBF neural network, Tas et al. explored the influence of meteorological data such as ambient temperature, average cloud cover, relative humidity, wind speed and atmospheric pressure on daily natural gas consumption in some parts of Turkey[5]. Considering the future trend of natural gas consumption, the decisive factors such as income and natural gas price, the structural time series model is used to predict the annual natural gas demand[6].

With the rapid development of artificial neural network, neural network has achieved better results in the field of natural gas daily forecast. Among the neural network models, the recurrent neural network (RNN) model, which processes data by internal memory cycle and keeps the chain structure, is more suitable for learning the characteristics of time series data[7]. However, the deep chain structure increases the difficulty of training RNN model through back propagation. This is not the case with the long-term memory (LSTM) model[8]. Its powerful ability to analyze data sequences has been proved by successful applications in many fields, including speech recognition[9] and traffic prediction[10]. Wei et al. successively applied support vector regression (SVR)[11] and long-term and short-term memory (LSTM)[5] to predict the daily natural gas consumption based on the data of natural gas consumption in Greece. In their work, LSTM network model can achieve better performance, and its average relative deviation is less than 10%. Laib and others predicted the natural gas consumption in Algeria, and LSTM obtained more accurate results than MLP neural network and linear regression model[12].

Although the existing model has achieved good prediction results, when the model is applied, the data is processed by a series of feature engineering. In practical application, both the traditional statistical model and the deep neural network are limited in their application, and usually some obvious errors may occur. We try to explore the influence of user data quality on the prediction results of the model in order to find the most suitable model for application. At the same time, for natural gas sales enterprises, they pay more attention to the forecast results at the fluctuation point, which is also the difficulty in the forecast process. Based on this, this paper constructs a natural gas prediction model based on the LSTM model, and constructs a volatility index to evaluate users' gas consumption characteristics, analyzes the general influence of volatility index on users' prediction accuracy, and further improves the prediction of the model by using Kalman filter. On this basis, it analyzes the reasons for the large prediction error at fluctuation points.

2. Model construction

2.1. Data analysis

This study collects selected natural gas consumption data from users supplied by a gas company in Shandong, China, spanning the period from January 1, 2015, to June 30, 2022, and containing a total of 2,736 data points. In this region, there is only one supplier of this gas company, so the natural gas data is the entire consumption of each user. In order to more accurately predict natural gas consumption, we further collected local daily maximum and minimum temperature data, which are meteorological data from IBM Weather. At the same time, we consider that the gas consumption of the users is also affected by the total supply of the supplying company, so the total daily gas supply of the gas supplying company in the local area is added into consideration as a data feature.

All the users participating in the cases are industrial users, and the specifics of their data are shown in Figure 1. By observing the graph, it can be visualized that the natural gas consumption of each user presents different data characteristics. For example, users such as user 3 and user 6 show a more obvious trend and regularity, while the gas consumption data of users such as user 1 and user 9 energy show a more complex and irregular pattern of change.

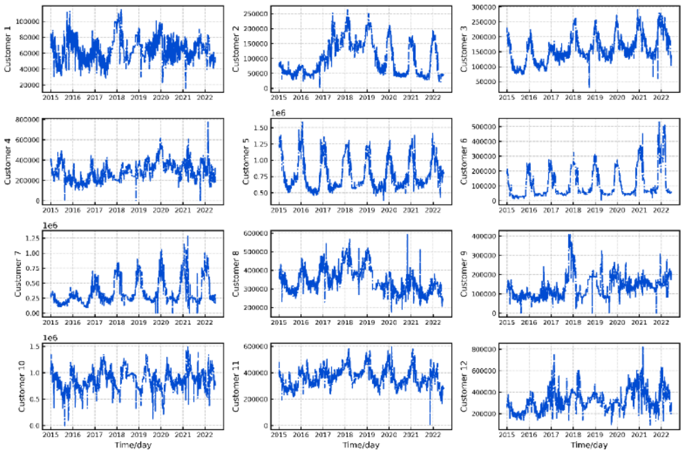

Figure 1: Trends in gas consumption by supplying customers

2.2. Data cyclicality and volatility

2.2.1. Data periodicity

The periodicity of time series data mainly includes, monthly, quarterly and annual cycles. Considering the characteristics of natural gas consumption changes, it is analyzed here only at the level of annual periodicity. Pearson correlation coefficient is used here to calculate the periodicity of natural gas consumption data. Pearson correlation coefficient is mainly used as a calculation of the correlation between two columns of data, here we segment the user natural gas consumption data according to the length of 365 (one year), calculate the correlation between the segments of data separately, and determine the final cyclicality index based on the correlation of the data in each segment. The calculation is shown in equation (1) and equation (2):

$$C\left(S_i,S_j\right)=\frac{\operatorname{cov}\left(U_i,U_j\right)}{\sqrt{D\left(U_i\right)D\left(U_j\right)}} \tag{1}$$

$$PE=\frac{1}{M\left(M-1\right)}\sum_{i=1}^{M}\sum_{j=1}^{M}C\left(S_i,S_j\right),i\neq j \tag{2}$$

2.2.2. Entropy value

Entropy, as a thermodynamic concept, can also be used to evaluate the uncertainty of sequence data by calculating the information rate. Based on the basic concept of entropy, Steven proposed the concept of approximate entropy to measure the complexity of serial data. On this basis, Richman et al. proposed a sample entropy calculation method based on approximate entropy, and the calculation method of sample entropy is shown in the equation.

$$X_i=[x(i),x(i+1),...x(i+m\text{-}1)],i=1\sim N\text{-}m+1 \tag{3}$$

$$B^m(r)=\frac{1}{N-m}\sum_{i=1}^{N-m}B_i^m(r) \tag{4}$$

$$SampEn(m,r) = \ln B^{m}\ (r) - \ln B^{m+1}(r) \tag{5}$$

A tolerance threshold *r* is set in the computation process, X_i and X_j are said to be approximated in mode *r* when the values of the corresponding points of X_j are all within the tolerance when the sequence X_i is the base. Where $B_i^m(r)$ denotes the approximation ratio based on X_i, indicating the ratio of the approximated quantity to the total quantity; $B_i^m(r)$ is the mean value of the approximation ratio; and *SampEn(m,r)* denotes the sample entropy value. Wherein, unlike the approximate entropy sample entropy calculation $B_i^m(r)$ does not include a comparison with its own data segment, the calculation error is smaller and does not depend on the length of the data. While the sample entropy calculation also involves two parameters *m* and *r*. According to the corresponding research experience *m* is taken as 2, and *r* is usually taken between 0.1 *SD (x)* and 0.25 *SD (x)* difference (*SD (x)* is the sequence standard).

2.2.3. Dispersion coefficient

Discrete coefficient is a metric for transforming continuous data into discrete data, which is usually used for feature selection and preprocessing in data mining, machine learning and other fields. It indicates the coverage of the discrete data, i.e., the number of intervals of continuous variables after discretization. The formula for the discrete coefficient is:

$$c_v = \frac{\sigma}{\mu} \tag{6}$$

Where σ is the average root of the sum of the squares of the deviations of each value from the mean for that set of data, and μ is the sum of all the values for that set of data divided by the number of data. By calculating the coefficient of discretization, continuous variables can be discretized, making the data easier to handle and analyze. In research, discrete coefficients can be used to explore the distribution and periodicity of data variables, especially variables such as natural gas prices, which have cyclical variations. Discrete coefficients can not only be used for feature selection and data preprocessing, but also play an important role in time-series forecasting, such as natural gas prices, by helping to analyze the periodicity and trend of the data variables so as to improve the accuracy of the forecasting model.

2.2.4. Skewness

Skewness is an indicator in statistics that describes the degree of asymmetry in the distribution of data, and it measures the extent to which the data set is skewed to the left or to the right of the mean. The direction of skewness of the data set is judged according to the positivity or negativity of the skewness, when the skewness is greater than 0, the data set is right skewed, i.e., the mean value is skewed to the side of the maximum value, while when the skewness is less than 0, the data set is left skewed, i.e., the mean value is skewed to the side of the minimum value. The formula for skewness is:

$$Skew(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \sum_{i=1}^{n} \frac{\frac{(x_i - \bar{x})^3}{n}}{\sigma^3} \tag{7}$$

where μ is the sample mean, σ is the sample standard deviation, *n* is the sample size, and x_i is the ith sample value.

2.3. Filtering algorithm

The Kalman filtering algorithm is suitable for predicting and filtering time series data containing noise and uncertainty. The principle is based on Bayesian probability theory, which estimates the optimal solution of the state by weighting the observed data and model. The steps of the Kalman filtering algorithm are divided into two stages: prediction and update. In the prediction stage, the previously estimated state and state transition matrix are used to calculate the predicted state and covariance matrix. In the update stage, the observed data and predicted state are used to calculate the Kalman gain and optimal state estimation. This algorithm has advantages such as efficiency, accuracy, and robustness, and is widely used in fields such as automatic control, signal processing, and machine learning. The actual formulas used include the Kalman filter equation for predicting states and the Kalman gain equation for calculating state accuracy. Equation of state:

$$x(k) = A \times x(k-1) + B \times u(k) + w(k) \tag{8}$$

Measurement equations:

$$z(k) = H \times x(k) + v(k) \tag{9}$$

Where *x(k)* denotes the state of the system at the time, *u(k)* is the external input of the system, *z(k)* denotes the measurement value, *w(k)* and *v(k)* are the system noise and the measurement noise, respectively, *A*, *B*, and *H* are the system state transfer matrix, the external input matrix, and the observation matrix, respectively, and *K(x)* is the Kalman gain.

2.4. LSTM algorithm

Recurrent Neural Network (RNN) is a type of neural network that takes sequence data as input and recurses in the direction of sequence evolution [11]. However, RNN cannot solve gradient explosion due to vanishing or long-term dependencies. In 1997, Hochrieter and Jurgen Schmidhuber proposed Long short-term memory(LSTM).LSTM is a special kind of recurrent neural network (RNN).LSTM introduces gate controller neural network whose weights can be self-updated, which makes LSTM one of the most popular deep learning networks in the world. Figure 2 shows the single neuron structure of LSTM.

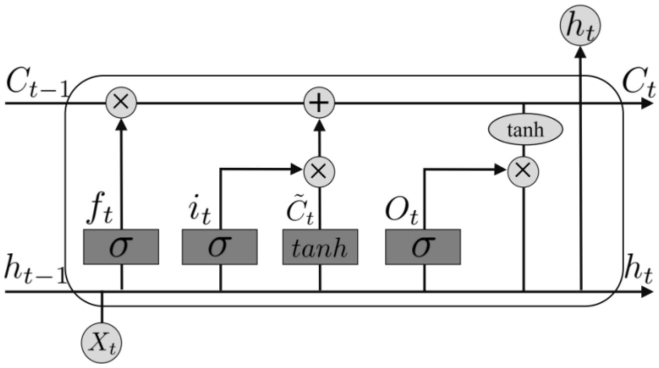

Figure 2: LSTM neurons

The LSTM model has a strong nonlinear fitting ability and can achieve better prediction results without complex data processing of the data. And it has been proven in several works to achieve more accurate and stable results than SVM, RandomForest, BP neural network and other models.

2.5. Experimental program

In order to accurately investigate the effect between data quality and change characteristics and prediction accuracy this study designed the following experimental program based on the collected data.

Case 1: Direct prediction based on raw data. The collected data are directly input into the model for prediction without any processing, and in the prediction process, only normalization, data set division and other operations are done without other processing of the data. In this way, we can observe the degree of response of the model to the original data, and at the same time, based on the calculation results of volatility and periodicity and other data indicators, we can explore the impact of the characteristics of the data itself on the prediction results.

Case 2: Predictive analysis after adding Kalman filter. Natural gas user data is processed using a filtering algorithm, and the processed data is input into the model to predict the results. Processing using the filtering algorithm can change the volatility of the data and other indicators, which can further explore the impact of data volatility on the prediction results, so as to discover the general pattern of changes in the accuracy of user predictions.

3. Results and discussion

In this section, we first predict the collected natural gas consumption of each user based on LSTM and analyze the difference of data volatility on the prediction accuracy of the model. Then based on the three cases designed to study the role of SSA decomposition to improve the prediction accuracy of LSTM models, as well as to analyze the general laws of decomposition models for time-series prediction when volatility is considered. Meanwhile each deep learning experimental process, hyperparameters such as the number of neurons, learning rate, etc. are generally determined by random search algorithms.

3.1. Direct prediction

Table 1: Results of the three calculated indicators per user

customer	dispersion coefficient	approximate entropy	sample entropy (physics)	mape
customer 1	0.23	1.31	0.93	0.0497
customer 2	0.56	0.47	0.24	0.0876
customer 3	0.28	0.75	0.57	0.0515
customer 4	0.31	0.94	0.69	0.0938
customer 5	0.31	0.6	0.33	0.0392
customer 6	0.87	0.32	0.10	0.078
customer 7	0.58	0.54	0.26	0.0719
customer 8	0.19	0.87	0.63	0.0395
customer 9	0.29	0.79	0.39	0.0554
customer 10	0.44	0.77	0.49	0.0647
customer 11	0.22	1.01	0.76	0.0557
customer 12	0.32	0.86	0.55	0.0985

Based on periodicity, volatility, and sample entropy are calculated, and the calculated metrics are shown in Table 1.

It can be seen that the periodicity of the data is relatively close, while the volatility and sample entropy are significantly different. Directly using LSTM network to predict each user, the prediction results are shown in Table 1.

From the analysis of the prediction results, the predicted MAPE values of all the users are within 10%, which indicates that the LSTM model has a strong base fitting ability and the direct use of the model to predict gives better results. Where for example users 1, 5 and 8 have MAPE values within 5%, the remaining users have MAPE values between 5% and 10%. To further reduce the prediction error, improvements are made from the model or feature engineering point of view. The correlation graph between the volatility metrics calculated from the data level and the prediction results is shown in Figure 3.

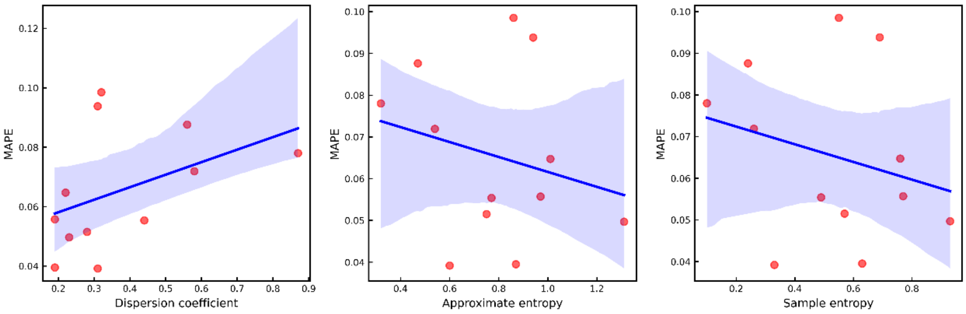

Figure 3: Location of the structure

As can be seen from Figure 3, the sample entropy and the equivalence value all present a correlation with the prediction results to a certain extent, i.e., the greater the volatility the greater the prediction error of the model. Intuitively the sample entropy value has the greatest correlation with MAPE, so from a single indicator the sample entropy is more directly affecting the prediction accuracy. However, when analyzed specifically, the correlation between volatility and MAPE is not strict, and many mutation points appear in this process, so it is inaccurate to assess the magnitude of the MAPE value directly by the correlation, and a more complex data relationship is implied between these two types of values.

3.2. Change in volatility

The filtering algorithm can perform reasonable correction on the raw data after processing, effectively eliminating the impact of noise and erroneous data on the overall data. Figure 4 shows the comparison of sample entropy values between the filtered data and the original data. The results show that for all users, the sampling entropy of the filtered data is significantly reduced, that is, the volatility is reduced. After filtering, the overall trend of the data remains unchanged, but there is a significant reduction in high-frequency data, that is, the change in mutation point data is reduced, making the original data smoother.

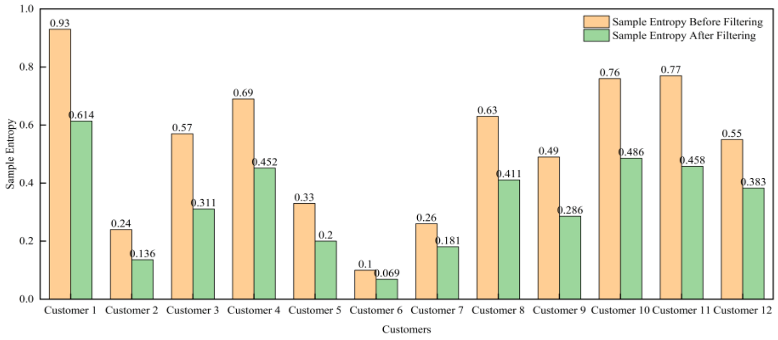

Figure 4: Comparison results of sample entropy before and after filtering for each user

The comparison of the prediction results of the processed data is shown in Figure 5. In this case for user 5 and user 8 the MAPE increased after data processing from 3.92%, 3.95% to 4.52% and 4.15%. For users such as user 2 the MAPE decreased, but the degree of decrease varied, but the overall MAPE value was around 5%. From the analysis of the users' original data, the addition of the filtering algorithm further reduces the volatility of the users and the reduced volatility shows a similar situation. However, the original volatility of user 5 and user 8 was relatively small, and although the filtering algorithm reduced the user volatility, it did not improve the prediction of the model, and the comparison of the two prediction situations of the data is shown in Figure 6.

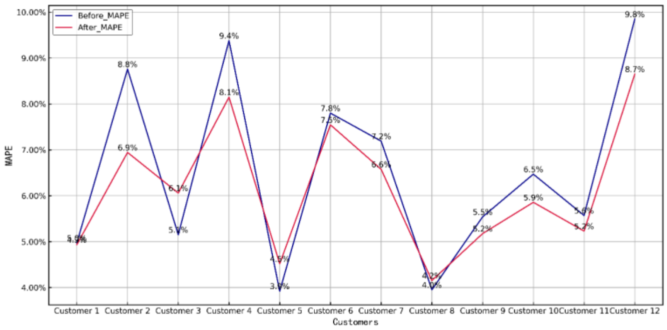

Figure 5: Plot of predicted results of processed data

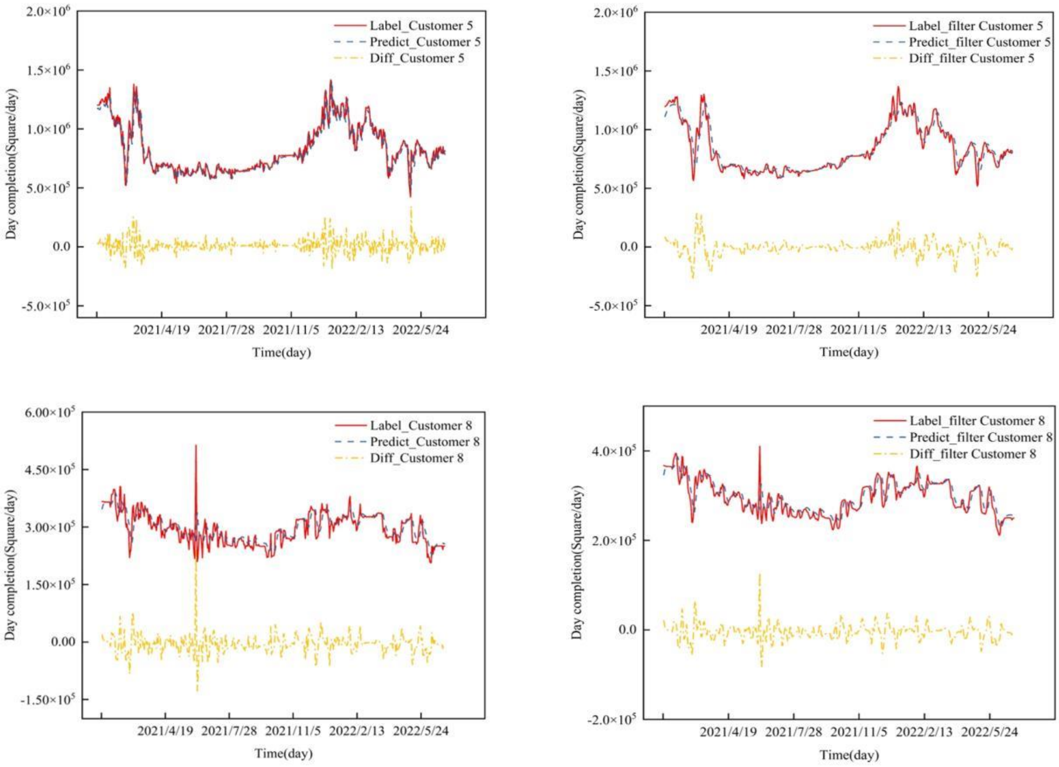

Figure 6: Calculating the test data error

Based on the graphical results, it is observed that the addition of the filtering algorithm results in a smoother trend in the data. However, due to the effect of the filtering algorithm, the prediction results at the mutation points show poorer performance, leading to an increase in the MAPE value. For users with high volatility, the smoothed prediction results can effectively avoid greater errors, but at the same time, it also reduces the model's ability to capture the changing characteristics of the data. As a result, the MAPE value obtained from the prediction is only able to reach about 5%, and it becomes more difficult to further improve the prediction accuracy.

4. Conclusions

The following conclusions can be obtained through the research in this paper:

(1) The LSTM model shows a strong fitting ability and can directly predict and obtain relatively accurate results.

(2) At the same time, the volatility of the data itself has an impact on the prediction results of the model and shows a linear relationship within a certain range, the greater the volatility, the greater the prediction error.

(3) By analyzing the sample entropy and other indicators, it is found that the sample entropy reflects the volatility of the data more effectively than the skewness and other values.

(4) Kalman filtering algorithm can make the data change more smoothly, which is directly manifested in the obvious reduction of data volatility. Although the volatility of the data is reduced after Kalman filtering, the prediction error is not completely reduced. Specifically, the MAPE value of users with large original prediction errors decreases, while the MAPE value of data with relatively accurate original predictions increases.

References

[1] Simsek Y E. Electric energy demands of Turkey in residential and industrial sectors[J]. Renewable and Sustainable Energy Reviews, 2012.

[2] Alvarez F M, Troncoso A, Riquelme J C, et al. Energy Time Series Forecasting Based on Pattern Sequence Similarity[J]. IEEE Transactions on Knowledge & Data Engineering, 2011, 23(8):1230-1243.

[3] Beyca O F, Ervural B C, Tatoglu E, et al. Using machine learning tools for forecasting natural gas consumption in the province of Istanbul[J]. Energy Economics, 2019, 80(MAY):937-949.

[4] BAI Y, LI C. Daily natural gas consumption forecasting based on a structure-calibrated support vector regression approach [J]. Energy and Buildings, 2016, 127: 571-579.

[5] Taşpinar F, Çelebi N, Tutkun N. Forecasting of daily natural gas consumption on regional basis in Turkey using various computational methods[J]. Energy and Buildings, 2013, 56: 23-31.

[6] Potočnik P, Thaler M, Govekar E, et al. Forecasting risks of natural gas consumption in Slovenia[J]. Energy Policy, 2007, 35(8): 4271-4282.

[7] RIGAMONTI M, BARALDI P, ZIO E, et al. Ensemble of optimized echo state networks for remaining useful life prediction [J]. Neurocomputing, 2018, 281: 121-138.

[8] Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey[J]. Computers and Electronics in Agriculture, 2018, 147: 70-90.

[9] Hirafuji Neiva D, Zanchettin C. Gesture recognition: A review focusing on sign language in a mobile context [J]. Expert Systems with Applications, 2018, 103: 159-183.

[10] NAGY A M, SIMON V. Survey on traffic prediction in smart cities[J]. Pervasive and Mobile Computing, 2018, 50: 148-163.

[11] ERDOGDU E. Natural gas demand in Turkey[J]. Applied Energy, 2010, 87(1): 211-219.

[12] Laib O, Khadir M T, Mihaylova L. Toward efficient energy systems based on natural gas consumption prediction with LSTM Recurrent Neural Networks[J]. Energy, 2019, 177: 530-542.