



Contents lists available at ScienceDirect

Chinese Journal of Chemical Engineering

journal homepage: www.elsevier.com/locate/CJChE

Full Length Article

A new method to forecast multi-time scale load of natural gas based on augmentation data-machine learning model

Denglong Ma^{1,*}, Ruitao Wu¹, Zekang Li¹, Kang Cen², Jianmin Gao^{1,4}, Zaoxiao Zhang³¹ School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China² School of Civil Engineering and Architecture, Southwest Petroleum University, Chengdu 610500, China³ School of Chemical Engineering and Technology, Xi'an Jiaotong University, Xi'an 710049, China⁴ State Key Laboratory for Manufacturing System Engineering, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:

Received 27 July 2021

Received in revised form 11 November 2021

Accepted 28 November 2021

Available online 3 January 2022

Keywords:

Natural gas

Machine learning

Prediction

Neural network

Augmentation data

ABSTRACT

Gas load forecasting is important for the economic and reliable operation of the city gas transmission and distribution system. In this paper, a nonlinear autoregressive model (NARX) with exogenous inputs, support vector machine (SVM), Gaussian process regression (GPR) and ensemble tree model (ETREE) were used to predict and compare the gas load based on the gas load data in a certain region for past 3 years. The results showed that the prediction errors for most of days were higher than 10%. Further, simulation data were generated by considering the gas load variation trend, which was then combined with historical data to form the augmentation data set to train the model. The test results indicated that the prediction error of daily gas load in one year reduced to below 7% with a machine learning prediction method based on augmentation data. In addition, the model based on augmentation data set still performed better than original data in predicting the monthly gas load in last year as well as daily gas load in last month and week. Therefore, the method based on augmentation data proposed in this paper is a potentially good tool to forecast natural gas load.

© 2021 The Chemical Industry and Engineering Society of China, and Chemical Industry Press Co., Ltd. All rights reserved.

1. Introduction

In recent years, the increasing consumption of natural gas has made it important to forecast the gas load to schedule accurately. Some methods have been proposed to predict the natural gas load, including traditional regression analysis, time series methods [1–5] and intelligent prediction methods like artificial neural networks (ANNs) [6–10], their unparalleled advantages in identifying trend and prediction assistance [11].

Regression analysis and time series model are two common prediction methods. Regression analysis refers to establish a good correlation equation (regression model) for prediction based on statistical principles and amount of data.

Aras *et al.* [12] utilized typical multiple regression model and autoregressive model to forecast the local natural gas demand by considering heating period and non-heating period, which significantly reduced the forecasting error. Tratar *et al.* [13] compared the prediction performance of 16 prediction methods and pointed out that the multiple regression method was the most effective

method for short-term heat load prediction among different models while the Holt-Winters methods was fitted for long-term load prediction. However, it is difficult to find a mathematical model that accurately reflects the non-linear relationship between input and output as well as lack of self-learning ability, thus the regression analysis method still has some limitations in gas load prediction.

The time series forecasting method is a method to build the correlations of the output with time by analyzing the historical data with statistical principle. Potocnik *et al.* [14] developed a prediction model based on stepwise regression to predict short-term natural gas loads by considering statistical fluctuations in weather-related forecasts. Ediger *et al.* [15] utilized the autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA) methods to accurately predict the total primary energy demand from 2005 to 2020, and made reasonable suggestions for energy conservation and emission reduction. Jiao *et al.* [16] established an urban natural gas with auto regression (AR) time-series model for short-term load forecast with the least square method based on autocorrelation function by eliminating the trend terms. Although the time-series approaches proposed above were suitable for short-term load forecasting, they did not consider the impact of

* Corresponding author.

E-mail address: denglong.ma@xjtu.edu.cn (D. Ma).

extra sensitive factors such as weather and date characteristics on gas loads.

Intelligent prediction methods, such as artificial neural networks and optimization algorithms, were also widely applied in the field of gas load prediction. Kizilaslan and Karlik [17,18] compared eight different algorithms to forecast weekly and monthly natural gas consumption by building ANN model with the temperature as the input. The results showed that a combined strategy based on a single ANN performed better than other approaches. Kato *et al.* [19] considered the outdoor minimum temperature and used cyclic neural networks to predict dynamic heat load, which outperformed the results with triple feed-forward neural networks. Li and Wu *et al.* [20,21] utilized the traditional first-order gray-scale prediction model (GM 1.1) combining genetic algorithm to establish natural gas consumption and production models, which improved the accuracy of the traditional first-order gray-scale model in medium and long-term predictions. Barak and Sadegh [22] combined the ARIMA method with adaptive neuro-fuzzy inference system (ANFIS) method to forecast annual energy consumption. In their method, the linear part of the data was predicted by the ARIMA model, and the non-linear part of the model was predicted by ANFIS. The results indicated that the combined algorithm had a higher prediction accuracy than the single algorithm. However, aforementioned feed-forward neural network algorithm is a gradient-declining learning method, which is limited by the accuracy of the raw data, the learning time cost and the local minima.

Because the gas load prediction models based on traditional methods are constrained by the non-linearity of the problem and the inability to integrate the impact of different sensitive factors such as weather and date characteristics on the prediction model, the gas load prediction accuracy is still unsatisfactory with traditional methods like regression methods. Moreover, the machine learning methods *e.g.* artificial neural network, are constrained by the source data limitation. The historical sample cannot fully reflect the real state of present gas load, thus the original sample has natural defects. In order to overcome above problems, a machine learning method based on augmented data set will be discussed in this research. Augmentation data methods have been widely used in the fields of image processing, target detection, and deep learning [23–25]. By increasing the samples in the training set, the overfitting of the model can be effectively alleviated. But the publications have not been found for the application of augmentation data method for gas load prediction. In this research, the gas load will be first predicted by NARX method and other typical machine learning methods based on historical data. Then, the simulated data considering the variation trend will be added into the sample set to form an augmented data set for modeling and prediction. Finally, the prediction accuracy with different algorithms for multi-time scales will be compared qualitatively.

2. Forecasting Methods

In this research, the nonlinear autoregressive model with exogenous input (NARX), support vector machine model (SVM), Gaussian process regression (GPR) and ensemble tree model (ETREE) were used to predict the gas load. The daily load, weekly load, and monthly load in a certain region from Sep. 2018 to Aug. 2019 with four different methods will be predicted. All computations were operated in Matlab 2019b.

2.1. Introduction of forecasting methods

In this section, the principles of four different machine learning approaches including NARX, SVM, GPR and ETREE will be briefly described.

2.1.1. Non-linear autoregressive model with exogenous inputs

The expression of NARX model is shown with Eq. (1) and Eq. (2) [26]:

$$y(k) = f[\mathbf{u}(k)] = w_0 + \sum_{i=1}^m (w_i \varphi_i(\mathbf{u}(k)) + b_i) + e(k) \quad (1)$$

$$\mathbf{u}(k) = [\mathbf{y}(k-1), \dots, \mathbf{y}(k-n_y), \mathbf{x}(k-1), \dots, \mathbf{x}(k-n_x)]^T, \varphi_i(\cdot)(i = 1, 2, \dots, m) \quad (2)$$

This method is mainly applied for the prediction of non-linear and non-stationary time series. x is input parameters and y is the output parameter in this model. k is the time step and e is the error function. n_y is the delay time for output vector which is used as the input again to realize self-regression prediction of the data in next time step. The approximation function f is ANN in this research.

2.1.2. Support vector machines

SVM has advantages in strong generalization capabilities and excellent global seeking property, especially for the case with small sample size. For the indistinguishable character of the samples in the linear space, SVM utilizes the relaxation variables and kernel functions to find the optimal classification support surface in the high-dimensional space. Moreover, SVM improves the generalization ability through the structural risk minimization principle [27,28]. It solves the problems of small samples, nonlinearity, high dimensionality, and local minimal points. It has been applied in the fields of pattern recognition, signal processing, load prediction and other fields like this [29,30].

2.1.3. Gaussian process regression

GPR always obtains the distribution of function $f(x)$ while a general regression algorithm expects to get the value of $\mathbf{Y}$ with given input $\mathbf{X}$ [31,32].

For the data set $\mathbf{D}:(\mathbf{X}, \mathbf{Y})$, let $f(x^i) = y_i$, and $f = [f(x_1), f(x_2), \dots, f(x_n)]$. The set of x_i is defined as x^* , and the corresponding predicted value is f^* . The distribution of $f(x)$ can be obtained based on the Bayesian principle:

$$p(f^*|f) = \frac{p(f|f^*)p(f^*)}{p(f)} = \frac{p(f, f^*)}{p(f)} \quad (3)$$

The joint probability distribution $f \sim N(\mu, K)$ among the samples in the data set is GPR first calculated, where μ is the mean vector of $f(x_1), f(x_2), \dots, f(x_n)$; K is covariance matrix. Then, the posterior probability distribution of f^* is calculated according to the prior probability distributions $f^* \sim N(\mu^*, K^*)$ and $f \sim N(\mu, K)$. During the process of calculation, different kernel functions are utilized to approximate the covariance matrix expression so as to determine the covariance easily.

2.1.4. Ensemble tree model

For supervised machine learning algorithms, the goal is to train a model that is stable and performs well in all respects. But sometimes only some models with preferred performance in certain aspect can be obtained in application. Therefore, ensemble learning

Table 1
Optimal parameters for different models

Model	Parameter
NARX	Input delay = 1; Feedback delay = 1; Hidden layer size = 2;
ETREE	Method = Bag; Learn rate = 1; Minileaf = 8;
GPR	Kernel function = Matern 5/2; Kernel parameters: Kernel scale = 38; Signal standard deviation = 2×10^3 ;
SVM	Box constraint = 3.66×10^4 ; Epsilon = 3.41×10^3 ; Kernel function = Polynomial; Function order = 3; Kernel Scale = 1;

is proposed to obtain a better and more comprehensive strong supervised model by combining multiple weak supervised models. In this case, other weak classifiers can correct the error even if a weak classifier gets the wrong prediction.

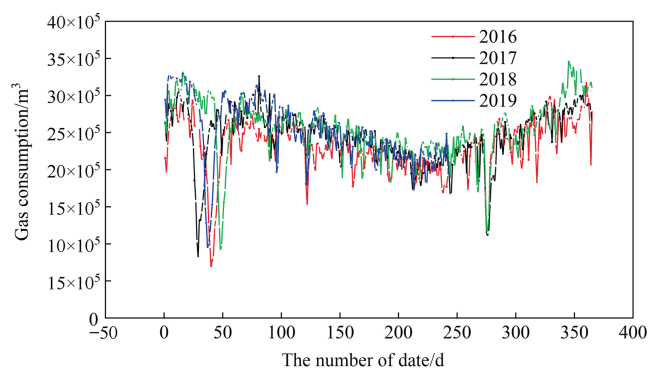
The ensemble algorithm combines several machine learning models to reduce bagging, boosting, or to improve stacking effects of the prediction. Random forest is a common ensemble algorithm [33]. In the random forest, each tree in the ensemble forest is made up of the samples drawn from the training set. In addition, only a subset of features is randomly selected instead of using all features for single tree. As a result, the estimated variance is reduced by calculating the mean for the less relevant trees although the random forest produces a slight increase in bias. Hence, the total performance of the ensemble model is better than that of single algorithm.

All parameters of machine learning models were optimized with random grid search method in our research, and main parameters of different models are listed in Table 1.

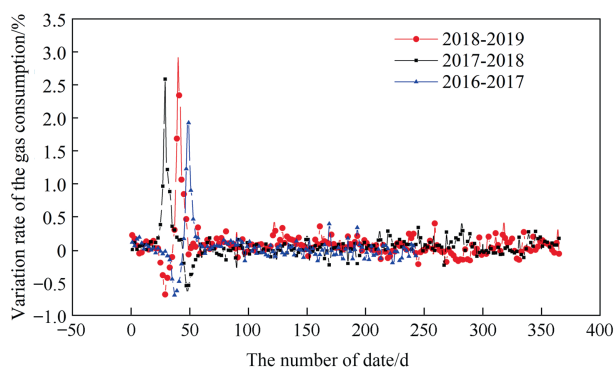
3. Results

3.1. Analysis of historical gas load data

The data collected in the article are daily gas consumption data (unit is m^3) from January 2016 to August 2019, with additional temperature characteristics (maximum, minimum, and average temperatures), weather characteristics (sunny, cloudy, rainy, and snowy days), and holiday characteristics (weekdays, weekends, and legal holidays) for that day. The gas load data were provided by the natural gas company, other data about temperature and weather characteristics were obtained from the web of public weather forecast.



(a) The variation of daily gas consumption from 2016 to 2019



(b) The daily change rate of gas consumption in adjacent years

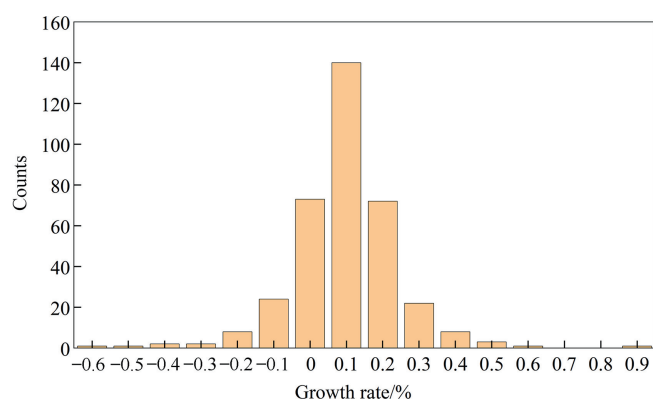
Fig. 1. The variation of gas daily consumption load from 2016 to 2019 (a) and the daily change rate of gas consumption in adjacent years (b).

3.1.1. Gas consumption changes in different years

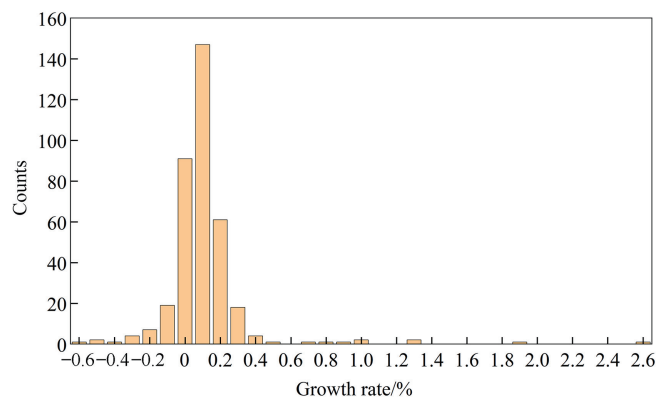
The daily gas consumption from 2016 to 2019 were analyzed and compared in Fig. 1(a). The daily change rate of gas consumption in adjacent years was also calculated, as shown in Fig. 1(b).

As seen in Fig. 1, the variation trend and law of gas consumption seemed similar for each year from 2016 to 2019. Particularly, the variation of gas consumption in the last adjacent two years was greatly similar. In February of each year, gas consumption decreased significantly, and it reached a maximum between December and January, which was consistent with the date of traditional Chinese Spring Festival holiday.

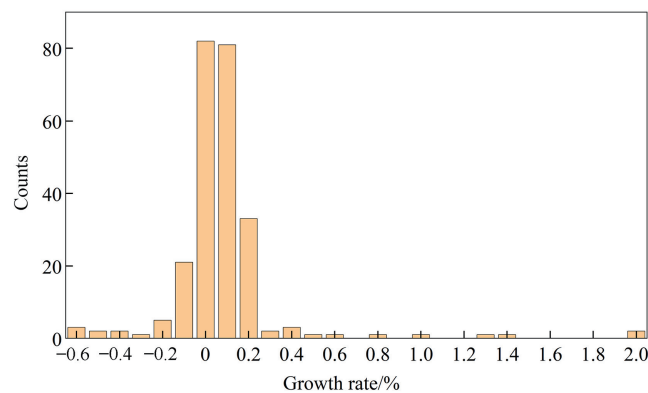
The statistic results of gas consumption variation rate in adjacent years were also obtained in Fig. 2. The horizontal coordinates of the graph are the growth rate of the latter year relative to the previous year, and the vertical coordinates are the counts of the



(a) The growth rate of gas consumption from 2016 to 2017



(b) The growth rate of gas consumption from 2017 to 2018



(c) The growth rate of gas consumption from 2018 to 2019

Fig. 2. Statistics on the rate of change in gas consumption per two years: (a) 2016–2017, (b) 2017–2018, (c) 2018–2019.

corresponding growth rates; It is noticed that the gas consumption shows an increasing trend year by year with an average growth rate between 10% and 15%.

3.1.2. Temperature changes in different years

Considering the significant impact of temperature changes on gas consumption, the temperature data and daily change rate of temperature in adjacent years were analyzed, as show in Fig. 3. It is noted that the temperature randomly fluctuates in the range of $\pm 20^\circ\text{C}$ per year. On the other hand, all gas consumption (Fig. 4 (a)) and temperature change rate (Fig. 4(b)) in three years from 2016 to 2019 were analyzed statistically. Note that the gas consumption mainly increased at a rate of 0–20% while the temperature varied within the range of $\pm 40\%$.

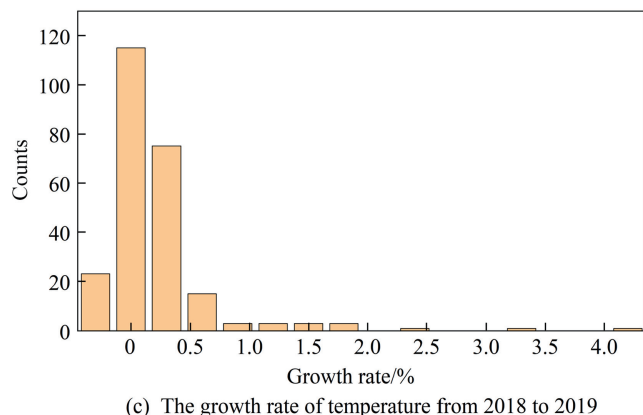
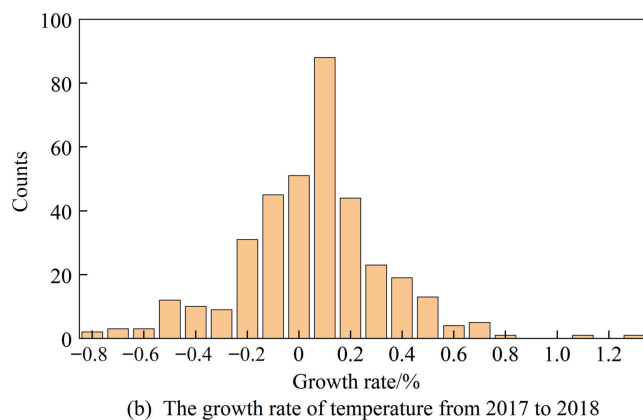
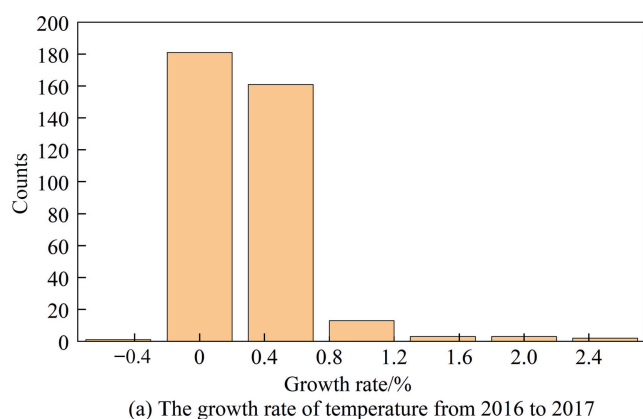


Fig. 3. Statistics results of temperature rate in the adjacent two years from 2016 to 2019: (a) 2016–2017, (b) 2017–2018, (c) 2018–2019.

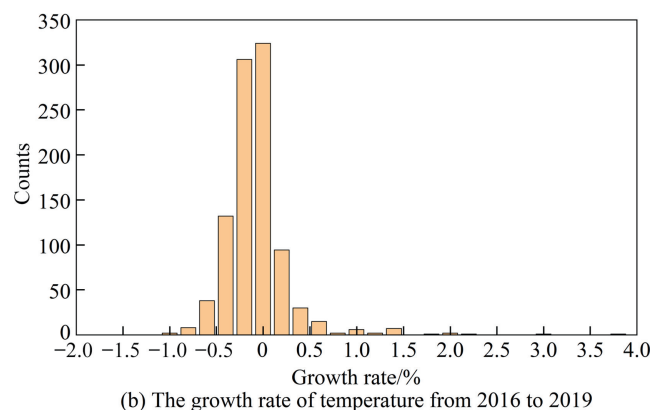
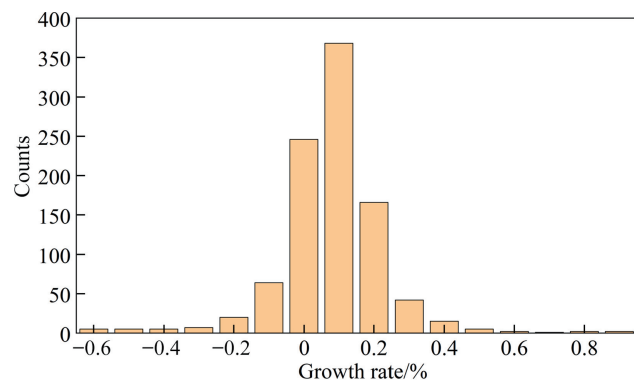


Fig. 4. Total statistic results of gas consumption (a) and temperature change rates (b) with all data.

3.2. Prediction results using NARX model

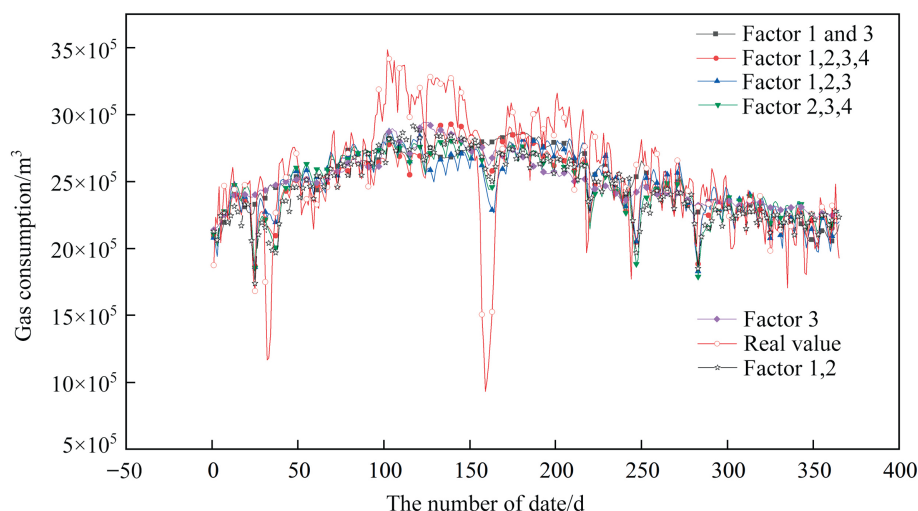
First, NARX model was used to predict the natural gas load. Following factor were considered during the prediction: Factor 1—the day on which it was in; Factor 2—daily holiday characteristics: normal, weekend, short-vacation, long holiday; Factor 3—average temperature of the day; Factor 4—weather of the day: sunny, cloudy, rainy, snowy.

The daily gas loads from Sep. 1, 2018 to Aug. 31, 2019 were forecasted with NARX model, which was trained by the other data from 2016 to 2018. Different factors were considered separately to build the model and predict the gas load, as shown in Fig. 5. The relative errors of the predictions considering different factors are compared in Fig. 5(b). Table 2 summarizes the mean absolute percentage errors (MAPE) of the gas load under the conditions with different factors.

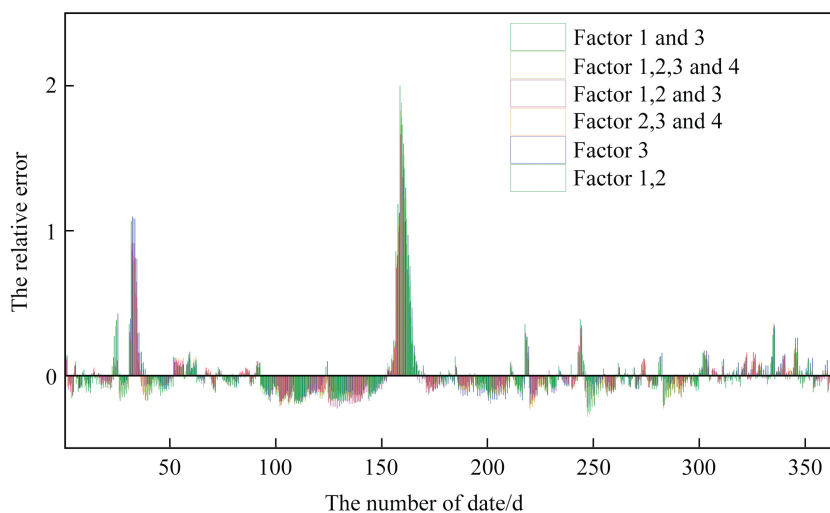
As shown in Fig. 5 and Table 2, the prediction results are slightly different with various input factors. If only temperature factor was considered, the error is larger than other cases. The prediction accuracy of prediction was improved when other more factors including days, holidays and weather features were considered. The models considering the factors related with the days (Factor 1), holidays (Factor 2) and temperature (Factor 3) had the highest prediction accuracy among different cases.

3.3. Prediction with other machine learning models

Further, different machine learning approaches were applied to predict daily gas load from Sep. 1, 2018 to Aug. 31, 2019. The prediction results based on SVM, GPR and ETREE considering input factor 1, 2 and 3, are shown in Fig. 6. The absolute errors of different models are illustrated in Fig. 6(b). It is noticed that the predic-



(a) Comparison of NARX prediction results considering different external input factors



(b) The relative error corresponding to the results in Fig. 5(a)

Fig. 5. Comparison of NARX prediction results considering different external input factors (a) and the relative errors of the predictions considering different factors (b).

tion deviations of different models in the period of March to July were quite different, where the ETREE, GPR algorithm had smaller errors than SVM and NARX.

3.4. Gas load prediction with augmentation data

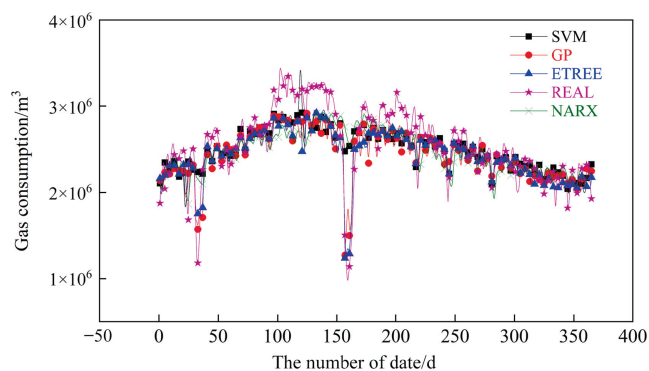
Based on the statistical analysis of the forecast results of ensemble tree model (Fig. 7), which has the lowest prediction error among four models, it is noticed that the predicted gas consumption from the model trained with the historical data is lower than actual value by 10% to 20% in most cases. The reason is that the natural gas consumption is increasing year by year, but the machine learning algorithm builds a prediction model by learning the features in historical data. The range of training data set will make the predictions lie in the range of the historical data with the highest probability, thus the predicted results will be underestimated and could not truly reflect the variation trend of the gas consumption with the year.

There may be two ways to solve this problem. One is to find a factor that correlates with the variation trend in gas consumption, such as number of users and electricity consumption. Since the

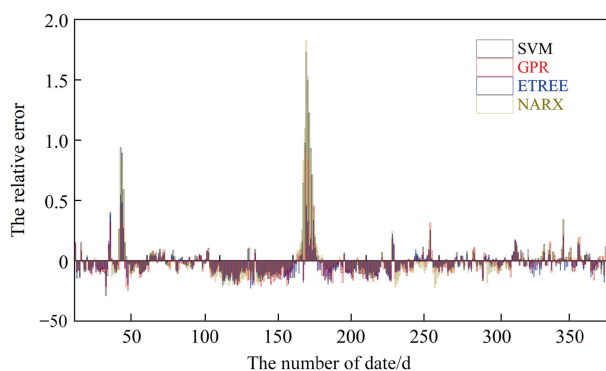
data about the number of users and electricity consumption are not yet available, they cannot be used in current research. Another idea is to expand the existing training data by using the method of data augmentation, which utilizes the model to randomly generate a large number of simulation data based on existing historical data. These simulation data include some trending information, such as gas consumption, temperature, etc., and then they are supplemented into the training set. In this method, the prediction models will be built based on the expanded training data set, which is always called augmentation data set and has been used widely in the field of image recognition.

Table 2
Comparison of prediction errors under different input factors

Factors	MAPE
1,2,3,4	0.1037
1,3	0.1047
1,2	0.1066
1,2,3	0.0979
2,3,4	0.1089
3	0.1112



(a) The predicted results of SVM, GPR and ETREE



(b) The relative error of different methods

Fig. 6. Comparison of SVM, GPR, ETREE and NARX for predicting daily load in one year period: (a) is the predicted result of SVM, GPR and ETREE, (b) is the relative error of the above four algorithms.

The augmentation data set consists of historical data and simulated data. For the problem of natural gas load prediction, the variation laws of temperature and gas consumption have been analyzed with the statistic results of historic data, as shown in Fig. 1 to Fig. 4. In this research, a batch of random data sets with the size of more than one time than original sample data were generated to compensate the information about the trend of the natural gas load into the data for training. A generation model was built based on the variation law of historical data. Thus, the simulation data of the temperature and gas consumption with the size of historical data set were randomly generated based on the statistic variation law, where the temperature varies within $\pm 40\%$ ($t_a = t_0 \pm t_0 \cdot \text{rand}(0,0.4)$) and the gas consumption increases randomly from 10% to 20% ($C_{ga} = C_{g0} + C_{g0} \cdot \text{rand}(0.1,0.2)$) compared with the same day in last year. Here t_a and C_{ga} are the augmented data for temper-

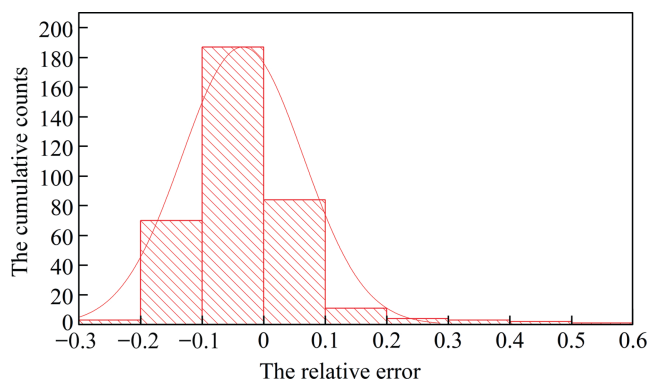
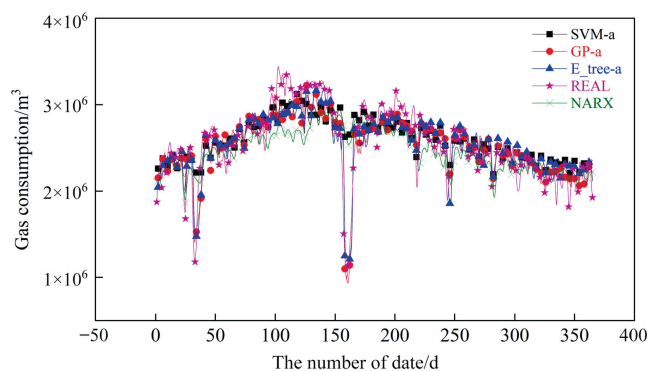
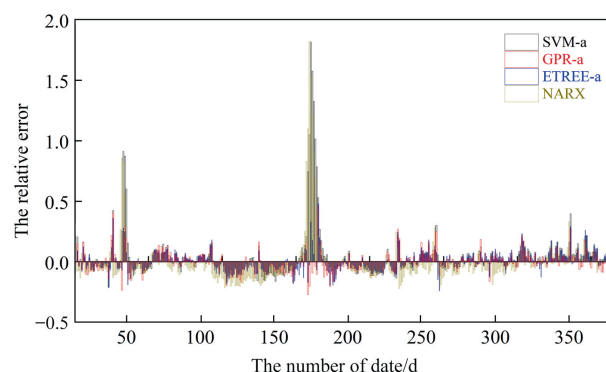


Fig. 7. The statistics results of prediction relative error with ETREE.



(a) The predicted results of SVM, GPR and ETREE



(b) The relative error of different methods

Fig. 8. Prediction results of different algorithms with the augmented data set: (a) is the predicted results of SVM, GPR and ETREE, (b) is the relative error of the above four algorithms.

ature and gas consumption while t_0 and C_{g0} are the historical data; $\text{rand}(a,b)$ is a random function from a to b . In this case, the trend information was introduced into the data set.

With the augmentation data set including both historical and simulation data, machine learning models were trained to predict the daily gas load for the period of Sep. 2018 to Aug. 2019, as shown in Fig. 8. The models trained with the augmentation data are marked as SVM-a, GPR-a, and ETREE-a, respectively.

Compared with the results in Fig. 6 and Fig. 8, the predicted results with all models based on the augmented data set (in Fig. 8) were closer to the actual values. The results with ETREE-a model performs better than other models. Fig. 9 illustrates the relative error distribution of the results with ETREE-a. It is obvious that the total prediction error of ETREE-a model is less than 50%

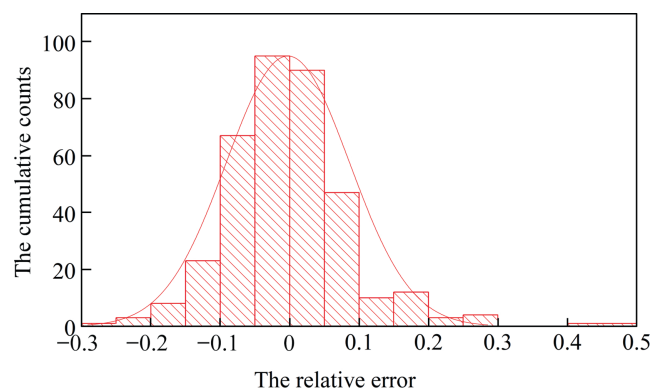
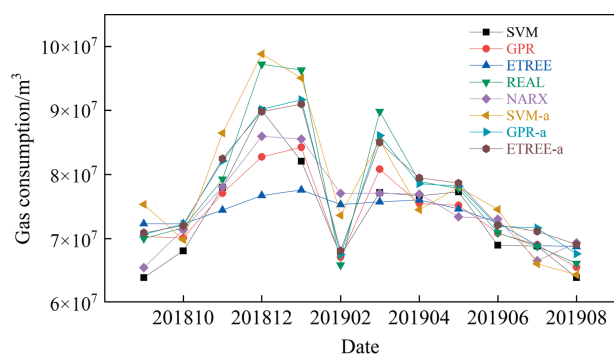


Fig. 9. Statistical results of the prediction relative bias of the Ensemble tree model under the augmented data set.

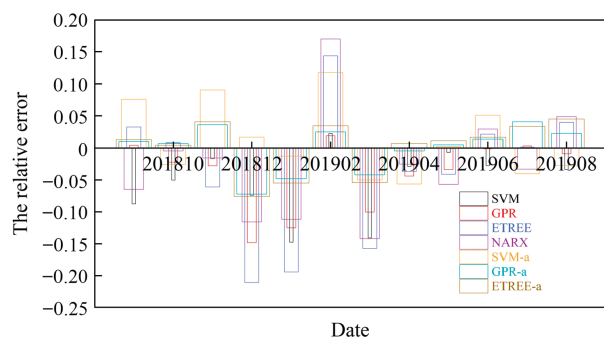
Table 3

Summary of results with different methods to forecast load from Sep. 1, 2018 to Aug. 31, 2019

Methods	SVM	GPR	ETREE	NARX	SVM-a	GPR-a	ETREE-a
MAPE	0.0979	0.0847	0.0789	0.1079	0.0952	0.0652	0.0666
COR	0.72	0.86	0.88	0.65	0.75	0.88	0.88
MAPE/COR	0.1359	0.0985	0.0897	0.1660	0.1269	0.0741	0.0756



(a) The predicted results based on historical data and augmentation data



(b) The relative error of different methods

Fig. 10. The results of monthly gas load prediction based on historical data and augmentation data from Sep. 2018 to Aug. 2019: (a) is the predicted results of different methods, (b) is the relative error of different methods.

and it is about only $\pm 5\%$ for nearly half of tested days. The number of days with a relative error less than 10% accounts for about 85%. After checking the production log data, it was found that the date with large forecast deviations were corresponded to the production records with production shutdowns and cutbacks. Hence, random cutbacks and shutdown in production could affect the prediction accuracy.

The statistic results based on original historical data and augmentation data are listed in Table 3. Here, the MAPE and the correlation coefficient (COR) are introduced to compare the accuracy of predicted results and the correlation between predicted and actual values. Further, the ratio of MAPE/COR is proposed to characterize the total performance of the methods. The smaller the value of MAPE/COR, the better the prediction performance. It is noted that the model trained with the augmentation data improves the prediction results. The absolute deviation of the prediction was

reduced below 7% with augmentation data and the correlation between predicted and actual values were also be improved. For the GPR and ETREE algorithm, the MAPE/COR was reduced by more than 20%. Compared with the SVM, GPR and ETREE algorithm, NARX algorithm has a poorer performance in predicting the daily gas load from Sep. 1, 2018 to Aug. 31, 2019. It may be due to that NARX is mainly appropriate for time series prediction, and the prediction accuracy will be higher for the series with a large time span (e.g. yearly).

4. Discussion

As discussed above, the model with augmentation data could improve the prediction accuracy of daily gas load for the time period from Sep. 2018 to Aug. 2019. In order to study the prediction performance of the methods on monthly and weekly gas load, the models based on historical data and augmented data were respectively built to predict the monthly gas load of 12 months from Sep. 2018 to Aug. 2019, as well as the daily load of 31 days from Aug. 1 to Aug. 31, 2019. Moreover, the weekly load of the last week in Aug. 2019 and the daily load of the last day in Aug. 2019 were also predicted. In this section, for the methods based on original samples, the size of training set for daily load forecast from Sep. 1, 2018 to Aug. 31, 2019 is 973 (section 3.3), and it is 32 for monthly gas load forecast from Sep. 2018 to Aug. 2019 (section 4.1). The size of training set is 942 for predicting the daily gas load in Aug. 2019 (section 4.2), and it is 135 for the last week of Aug. 2019 (section 4.3). For the methods with argumentation data set, the size of training set was tested from 5 to 100 time of the original set. Finally, it is found that the argumentation data with the size of 50 times of original data size was satisfied by considering prediction accuracy and time cost comprehensively. Therefore, the size of training set is 48,650 for daily load forecast in one year, 1600 for monthly load forecast in one year, 47,100 for daily load forecast in one month and 6750 for daily forecast in one week.

4.1. Monthly gas load forecast

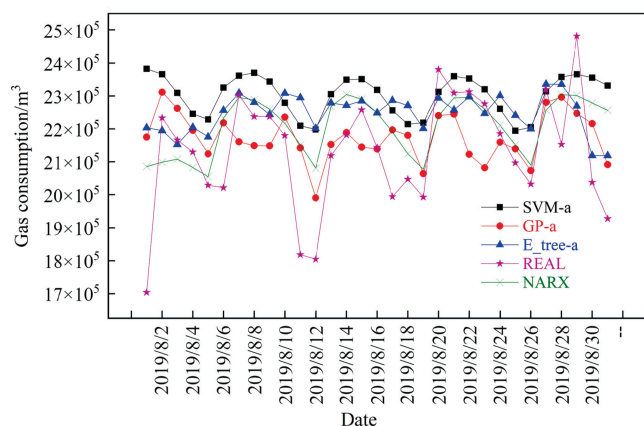
Firstly, monthly gas loads were predicted with NARX models and other machine learning models (SVM, GPR and ETREE) based on the historical data and augmentation data, as shown in Fig. 10.

It is seen from Fig. 10(b) that the maximum error of the prediction results based on the historical data set reaches 20%, especially for the month between Dec. 2018 and Mar. 2019. Meanwhile, the prediction accuracy of monthly gas load is greatly improved based on the augmentation data. The prediction accuracy of GPR and ETREE are higher than that of SVM and NARX model. For the prediction results with GPR model, it is noted that the relative error drops to less than 8% although the prediction results of the GPR

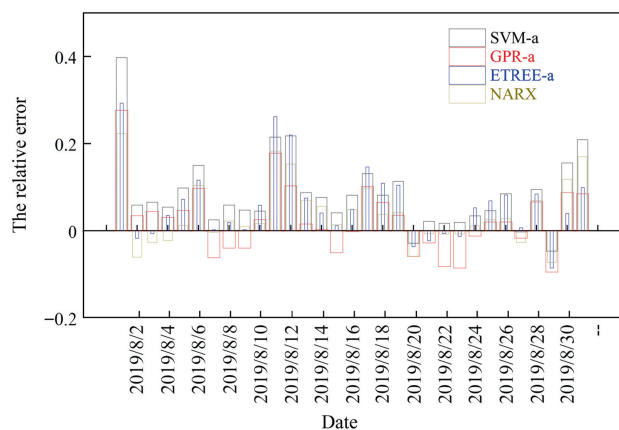
Table 4

The results of monthly load with different models predicting from Sep. 2018 to Aug. 2019

Methods	SVM	GPR	ETREE	NARX	SVM-a	GPR-a	ETREE-a
MAPE	0.0530	0.0448	0.0790	0.0741	0.0472	0.0272	0.0325
COR	0.92	0.98	0.76	0.88	0.92	0.98	0.98
MAPE/COR	0.0576	0.0457	0.1039	0.0842	0.0513	0.0277	0.0331



(a) The predicted results of SVM, GPR and ETREE



(b) The predict error of different methods

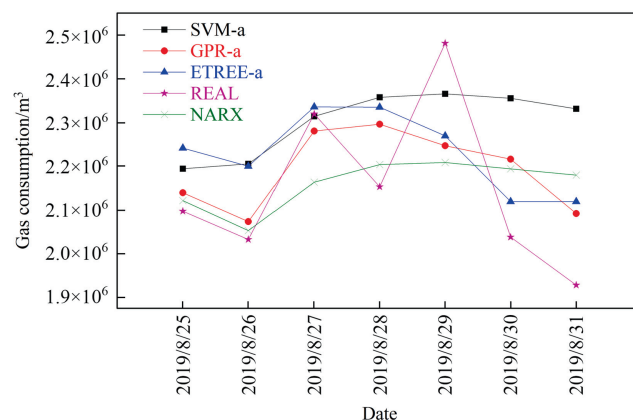
Fig. 11. Results of daily gas load in Aug. 2019 with different algorithms: (a) is the predicted results of SVM, GPR and ETREE, (b) is the relative error of different methods.

model for the months from Sep. 2018 to Aug. 2019 are still smaller than actual values. The overall deviation of the results from GPR model with augmentation data is less than 3%, as compared in Table 4. The ensemble tree model has the similar excellent performance with Gaussian process regression model.

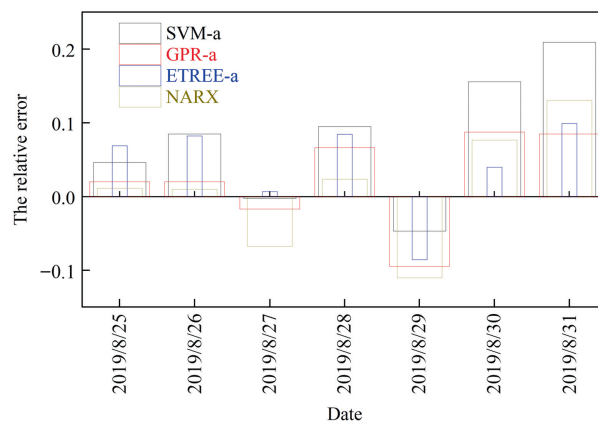
4.2. Daily gas load forecast for Aug. 2019

Further, the historical data and augmentation data sets were utilized separately to predict the daily gas load in Aug. 2019. The results with augmentation data set is shown in Fig. 11. The statistical results are listed in Table 5.

Noted from the results in Fig. 11 and Table 5 that the prediction results of daily gas load from Aug. 1 to Aug. 31 2019 are not well correlated with the actual results. It may be due to that the random fluctuations of the actual values are statistically amplified within one month of data. In contrast, the GPR model based on the augmentation data set has a better performance for daily gas load forecasting in just one month.



(a) The predicted results of SVM, GPR and ETREE



(b) The predict error of different methods

Fig. 12. Results of daily gas load prediction with different algorithms for last week of August 2019: (a) is the predicted results of SVM, GPR and ETREE, (b) is the relative error of different methods.

4.3. Daily gas load prediction for the last week of Aug. 2019

Then, the historical data and the augmentation data were utilized to predict the daily gas load for the last week of Aug. 2019 (Aug. 25 to Aug. 31 in 2019). The results of the models with augmentation data set are shown in Fig. 12, and the statistical results are shown in Table 6.

As mentioned above, the correlation between the predicted results and the actual values of the NARX improves in the daily prediction for one week in advance. The prediction performance of NARX is close to that of GPR and ETREE with the augmentation data. However, the Gaussian regression process model with the augmentation data has the best performs in forecasting daily gas load for just one week in advance.

4.4. Daily gas load prediction on the last day of Aug. 2019

Finally, the gas load of the last day of August 2019 (Aug. 31, 2019) was predicted with the models based on both historical

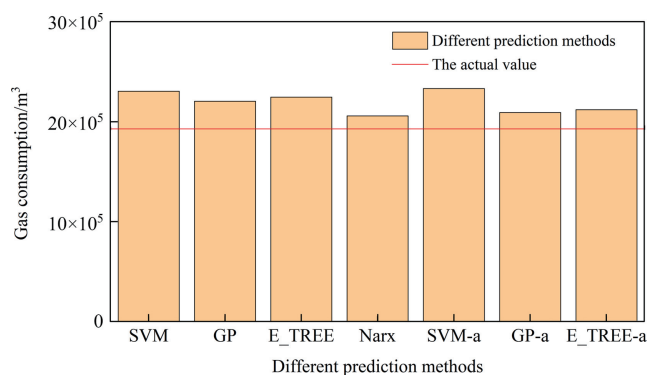
Table 5
The results of daily gas load forecast in Aug. 2019

Methods	SVM	GPR	ETREE	NARX	SVM-a	GPR-a	ETREE-a
MAPE	0.0639	0.0599	0.0637	0.0653	0.0857	0.0248	0.0724
COR	0.23	0.40	0.19	0.35	0.46	0.44	0.44
MAPE/COR	0.2778	0.2604	0.2769	0.1866	0.1863	0.0563	0.1645

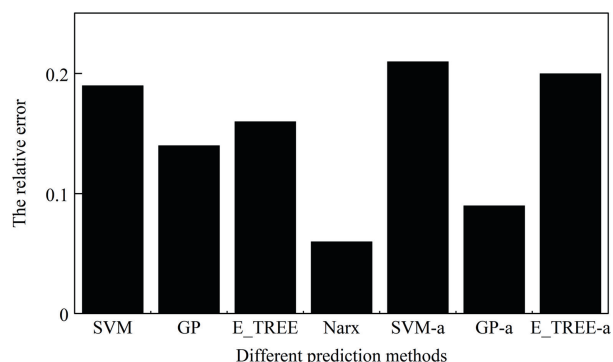
Table 6

The results of daily gas load forecast for the last week of August 2019 (Aug. 25–31, 2019)

Methods	SVM	GPR	ETREE	NARX	SVM-a	GPR-a	ETREE-a
MAPE	0.0820	0.0611	0.0748	0.0660	0.0775	0.0240	0.0678
COR	0.04	0.22	0.21	0.80	0.34	0.68	0.69
MAPE/COR	2.0500	0.2777	0.3562	0.0825	0.2279	0.0353	0.0983



(a) The predicted results of different methods



(b) The relative error of different methods

Fig. 13. Results of gas load prediction with different algorithms for the last day of August 2019. (a) is the predicted results of different methods, (b) is the relative error of different methods.

and augmentation data. The results are shown in Fig. 13 and Table 7.

As shown in Fig. 13 and Table 7, the NARX model has obvious advantages in predicting the gas load in one day advance. The result with NARX model is very close to the true value and the error is the smallest among all methods. Moreover, the results with GPR model based on the augmentation data are also close to the actual values, but the error is larger than that obtained from NARX model.

Hence, the trend information included in the augmentation data set makes the machine learning models with augmentation data set have better performance than that with original historical data for daily, weekly and monthly gas load prediction. However, for a very short-term gas load prediction, e.g. the gas load after one day, the time-series model such as NARX will perform better than other machine learning model due to that the trend information is not

necessary for a very short time scale. Therefore, the suitable model should be selected for multi-scale load forecasting for natural gas in one region.

5. Conclusions

In order to predict the natural gas load accurately, different models including nonlinear autoregressive model with external input (NARX), support vector machine (SVM), Gaussian process regression (GPR) and ensemble tree (ETREE) models based on historical data were utilized to forecast gas load in a certain region. Different factors such as vacations, temperature and weather were also considered in the models. The prediction results indicated that the models with more than one factors were better than that with common time series model based on just date. Moreover, it is demonstrated that the gas loads were underestimated 10%–20% from actual values for most test cases. It may be limited by the range of historical data, which were used to train the machine learning model.

Therefore, a load prediction method based on augmentation data was proposed. In this method, a large number of simulation data are generated based on the variation law of historical gas load data and an augmentation set including both historical and simulation data was formed to train the machine learning model. Because the information about variation trend of gas load was considered in the augmentation set, which extended the range of the training data, the prediction results of machine learning models were improved obviously. The test results for daily, weekly and monthly gas load prediction showed that the prediction accuracy was enhanced with augmentation data set, especially for Gaussian process regression and ensemble tree models. But NARX showed a better performance for just one day advance than SVM, Gaussian process and ensemble tree models.

Accordingly, the method with machine learning model based on augmentation data is a good tool to predict natural gas load accurately, although there are still some problems to be solved, e.g. the number of the users is not available and the temperature prediction is also a problem. We will improve these issues in our future work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Financial support was provided by the National Natural Science Foundation of China (21808181), China Postdoctoral Science Foun-

Table 7

Projected load for the last day of August 2019 (Aug. 31, 2019)

Methods	SVM	GPR	ETREE	NARX	SVM-a	GPR-a	ETREE-a
The actual value	192,794	192,794	192,794	192,794	192,794	192,794	192,794
Predictive value	230,499	220,367	224,518	205,686	233,150	209,182	211,923
The relative error	0.19	0.14	0.16	0.06	0.21	0.09	0.20

dation (2019M653651, 2021T140544), Basic research project of natural science in Shaanxi province (2020JM-021).

References

- [1] M. Doroodi, A. Mokhtar, Comparison of time series approaches for prediction of energy consumption focusing on greenhouse gases emission in Iran, *Int. J. Energy Sect. Manag.* 13 (3) (2019) 486–499.
- [2] M. Akpinar, N. Yumusak, Year ahead demand forecast of city natural gas using seasonal time series methods, *Energies* 9 (9) (2016) 727.
- [3] S. Ding, A novel self-adapting intelligent grey model for forecasting China's natural-gas demand, *Energy* 162 (2018) 393–407.
- [4] N. Wei, C.J. Li, C. Li, H.Y. Xie, Z.W. Du, Q.S. Zhang, F.H. Zeng, Short-term forecasting of natural gas consumption using factor selection algorithm and optimized support vector regression, *J. Energy Resour. Technol.* 141 (3) (2019) 032701.
- [5] V. Bianco, F. Scarpa, L.A. Tagliafico, Scenario analysis of nonresidential natural gas consumption in Italy, *Appl. Energy* 113 (2014) 392–403.
- [6] J. Szoplik, Forecasting of natural gas consumption with artificial neural networks, *Energy* 85 (2015) 208–220.
- [7] Z. Xiao, S.J. Ye, B. Zhong, C.X. Sun, BP neural network with rough set for short term load forecasting, *Expert. Syst. Appl.* 36 (1) (2009) 273–279.
- [8] A. Paul, S. Bhowmik, R. Panua, D. Debroy, Artificial neural network-based prediction of performances-exhaust emissions of diesel engine piloted dual fuel diesel engine under varying compressed natural gas flowrates, *J. Energy Resour. Technol.* 140 (11) (2018) 112201.
- [9] A. Azari, M. Shariaty-Niassar, M. Alborzi, Short-term and mediumterm gas demand load forecasting by neural networks, *Iran J. Chem. Chem. Eng.-Int. Engl. Ed.* 31 (4) (2012) 77–84.
- [10] D. Cinar, G. Kayakutlu, T. Daim, Development of future energy scenarios with intelligent algorithms: Case of hydro in Turkey, *Energy* 35 (4) (2010) 1724–1729.
- [11] Z.R. Jiao, P.F. Hu, H.F. Xu, Q.S. Wang, Machine learning and deep learning in chemical health and safety: A systematic review of techniques and applications, *ACS Chem. Heal. Saf.* 27 (6) (2020) 316–334.
- [12] H. Aras, N. Aras, Forecasting residential natural gas demand, *Energy Sources* 26 (5) (2004) 463–472.
- [13] L. Ferbar Tratar, E. Strmčnik, The comparison of Holt-Winters method and Multiple regression method: A case study, *Energy* 109 (2016) 266–276.
- [14] P. Potočnik, E. Govekar, Applied short-term forecasting for the Slovenian natural gas market, in: 2016 13th International Conference on the European Energy Market (EEM), Porto, Portugal, IEEE, 2016, pp. 1–5.
- [15] V.Ş. Ediger, S. Akar, ARIMA forecasting of primary energy demand by fuel in Turkey, *Energy Policy* 35 (3) (2007) 1701–1708.
- [16] W.L. Jiao, Y.L. Qin, L.B. Zhao, Research on urban gas load forecasting system, *Nat. Gas. Ind.* 25 (1) (2005) 155–157.
- [17] A. Khotanzad, H. Elragal, T.L. Lu, Combination of artificial neural-network forecasters for prediction of natural gas consumption, *IEEE Trans. Neural Networks* 11 (2) (2000) 464–473.
- [18] R. Kizilaslan, B. Karlik, Comparison neural networks models for short term forecasting of natural gas consumption in Istanbul, in: 2008 First International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), August 4–6, 2008, Ostrava, Czech Republic, IEEE, 2008, pp. 448–453.
- [19] K. Kato, M. Sakawa, K. Ishimaru, S. Ushiro, T. Shibano, Heat load prediction through recurrent neural network in district heating and cooling systems, in: 2008 IEEE International Conference on Systems, Man and Cybernetics (SMC), Singapore, Singapore, 2008, pp. 1401–1406.
- [20] Y. Xie, M. Li, Research on prediction model of natural gas consumption based on Grey modeling optimized by genetic algorithm, in: 2009 IITA International Conference on Control, Automation and Systems Engineering (CASE), Zhangjiajie, China, 2009, pp. 335–337.
- [21] H.W. Ma, Y.H. Wu, Grey predictive on natural gas consumption and production in China, in: 2009 Second Pacific-Asia Conference on Web Mining and Web-based Application, IEEE, Wuhan, China, 2009, pp. 91–94.
- [22] S. Barak, S.S. Sadegh, Forecasting energy consumption using ensemble ARIMA-ANFIS hybrid algorithm, *Int. J. Electr. Power Energy Syst.* 82 (2016) 92–104.
- [23] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, *Commun. ACM* 60 (6) (2017) 84–90.
- [24] J. Lemley, S. Bazrafkan, P. Corcoran, Smart augmentation learning an optimal data augmentation strategy, *IEEE Access* 5 (2017) 5858–5869.
- [25] S. Liang, M. Liu, Z.Y. Wang, T.X. Zhang, Infrared spectrum image inversion method for augmenting data sets, *Proc. SPIE 11428, MIPPR 2019: Multispectral Image Acquisition, Process, Analy.* 1142 (2020) 162–168.
- [26] D.L. Ma, J.M. Gao, Z.Y. Gao, H.Q. Jiang, Z.X. Zhang, J.T. Xie, Gas leakage recognition for CO₂ geological sequestration based on the time series neural network, *Chin. J. Chem. Eng.* 28 (9) (2020) 2343–2357.
- [27] R. Panja, N.R. Pal, MS-SVM: Minimally spanned support vector machine, *Appl. Soft Comput.* 64 (2018) 356–365.
- [28] P.F. Pai, W.C. Hong, Support vector machines with simulated annealing algorithms in electricity load forecasting, *Energy Convers. Manag.* 46 (17) (2005) 2669–2688.
- [29] K.X. Bi, T. Qiu, An intelligent SVM modeling process for crude oil properties prediction based on a hybrid GA-PSO method, *Chin. J. Chem. Eng.* 27 (8) (2019) 1888–1894.
- [30] H.F. Lu, M. Azimi, T. Iseley, Short-term load forecasting of urban gas using a hybrid model based on improved fruit fly optimization algorithm and support vector machine, *Energy Rep.* 5 (2019) 666–677.
- [31] C.L. Mei, Y. Su, G.H. Liu, Y.H. Ding, Z.L. Liao, Dynamic soft sensor development based on Gaussian mixture regression for fermentation processes, *Chin. J. Chem. Eng.* 25 (1) (2017) 116–122.
- [32] C. Zhang, H.K. Wei, X. Zhao, T.H. Liu, K.J. Zhang, A Gaussian process regression based hybrid approach for short-term wind speed prediction, *Energy Convers. Manag.* 126 (2016) 1084–1092.
- [33] J.L. Liu, C. Ullishney, C.E. Dumitrescu, Random forest machine learning model for predicting combustion feedback information of a natural gas spark ignition engine, *J. Energy Resour. Technol.* 143 (1) (2021) 012301.