Modeling of Natural Gas Consumption Volumes in China Aided by Machine Learning Methods

I.V. Filimonova[1,2,*], V.Yu. Nemov[1,2], A.P. Samatova[1,2], N.G. Akopov[2]

[1] Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
[2] Novosibirsk State University, Novosibirsk, Russia

***Abstract* — The paper applies machine learning methods to make projections of natural gas consumption volumes in China. It is critical for Russia as a major supplier of natural gas to China to have a reasonable estimate of possible volumes of exports. This contributes to the proper allocation of available raw materials, reduces the cost of excess gas storage, and also facilitates long-term planning for future trade. These aspects are critical for sustaining Russia's economic security and developing international economic relations. It is possible to estimate possible exports based on the projected volumes of natural gas consumption in China. This study uses machine learning methods, which are considered a promising data analysis tool, to model such consumption. We used multiple models for their benchmark comparison. Ridge regression was used as a linear model, whereas random forest and gradient boosting served as nonlinear models. The simulations performed proved gradient boosting to be the best choice. The study revealed the decisive role of socio-demographic factors, such as the population and the urban area size. The most significant factors were the total population, gas reserves, urban area size, number of passenger cars, and the population in urban agglomerations with over 1 million inhabitants. The most significant factors were the total population, gas reserves, urban area size, number of passenger cars, and the population in urban agglomerations with over 1 million inhabitants.**
a modification of the discounted cash flow method. The model is tested under the assumption that carbon regulation is carried out through the introduction of a carbon border tax.

***Index Terms* — forecasting, gradient boosting, machine learning, natural gas, random forest.**

I. Introduction

It is critical for the economic security of a country to have as much information as possible about its potential exports volumes. China is currently one of the most important markets for Russian gas. Building and strengthening international economic relations depends, among other things, on the awareness of the needs of the countries with which Russia conducts trade. There is a body of published research that deals with projections of the dynamics of gas production and exports in Russia [1–3], which gives an idea of the possible exports volumes. However, without understanding the demand in other countries in the years to come, it is impossible to estimate future trade volumes.

Today's China is a country with one of the fastest growing economies in the world. Naturally, its economic boom comes at a high cost, including energy. This is

* Corresponding author.
E-mail: filimonovaiv@list.ru

http://dx.doi.org/10.25729/esr.2024.03.004

Received October 21, 2024. Revised October 23, 2024. Accepted November 11, 2024. Available online November 25, 2024.

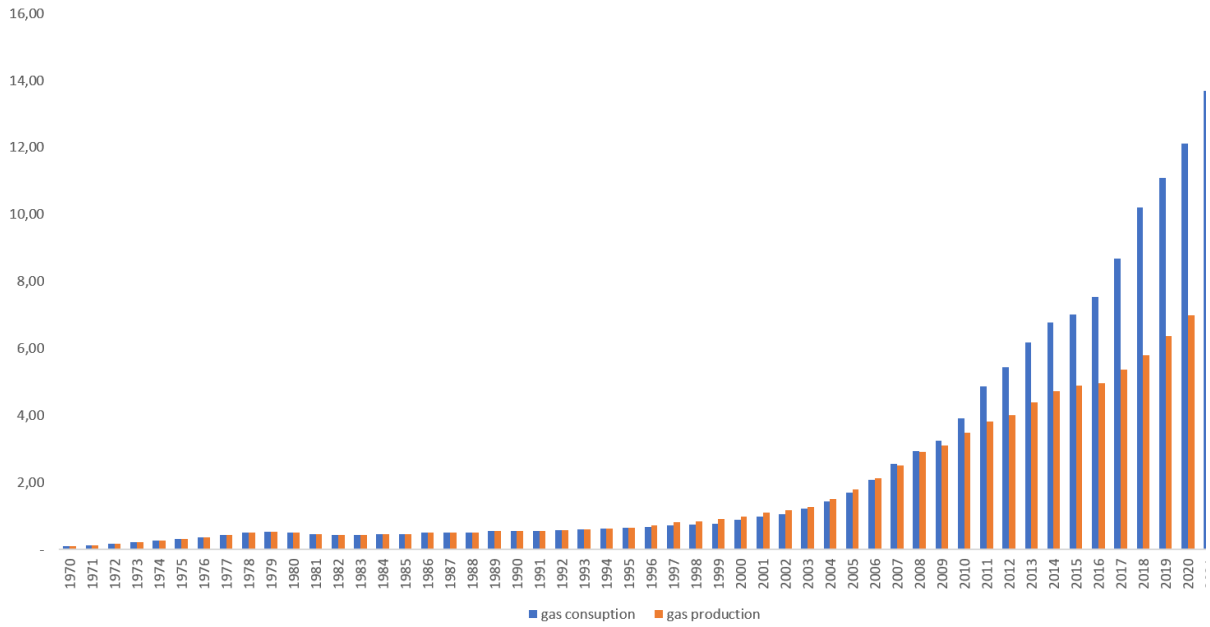

Fig. 1. Dynamics of gas consumption and production in China by year, exajoules.

attested by the fact that China has been ranked first in the world in terms of primary energy consumption since 2009 to the present day.

For a fairly long period of time, the country's natural gas production was sufficient to cover domestic consumption, but since 2007, consumption has been growing faster than production and the domestic market have been unable to cope with the growing demand anymore. Fig. 1 shows that this gap has been increasing every year.

The Chinese Government, acknowledging the importance of this challenge, is actively pursuing policies to remedy the gap. For example, the 14th Five-Year Plan targets increasing the LNG and natural gas storage capacity to approximately 2 trillion cubic feet by 2025, which amounts to a more than two-times increase since 2023. This will be achieved by additions of regasification terminals: in 2023 alone, the construction of 14 such terminals in the country was started [4].

The focus on mitigating environmental impacts also drives the demand for gas. Natural gas is considered a “cleaner” energy source compared to coal, whose share in China's energy consumption is high. Partial replacement of coal with natural gas will curb emissions of air pollutants.

We conclude from the above that it is necessary to forecast gas consumption volumes, since it becomes impossible to rely on historical consumption data: current volumes may differ too much from the past ones.

II. Literature Review

We have compiled a table to analyze the relevance of the topic of our study. Table 1 shows the number of publications on related topics and it makes it clear that the interest in natural gas demand forecasting is growing bigger every year. That being said, only a small percentage of the studies relied on machine learning models. This can be due to the fact that machine learning is still at its infancy and it is yet to be embraced by all research fields.

The above body of research employed statistical and multivariate models or time series models. The present study is based on the research contributions to the field of energy resources and the efficiency of their use made by A.G. Korzhubaev, I.V. Filimonova, A.A. Makarov, T.A. Mitrova, A.N. Dmitrievsky, V.V. Bushuev, D.M. Vasilyeva and others. It also relies on the machine learning research by Anibal Alviz-Meza, Edwin Torres-

Table 1. Number of Studies on Related Topics by Year

Years of publication	Number of studies	%
2003–2007	6	9.5
2008–2012	12	19
2013–2017	21	33.3
2017–2022	24	38.1

Salazar, Silvia Gastiaburu-Morales, Halyn Alvarez-Vasquez, Juan Orozco-Agamez, Dario Peña-Ballesteros, Wei Zhang, Jun Yang, Sunil Dhal, Dario Šebalj, Josip Mesarić, Davor Dujak.

Our literature review focused on the factors considered in the publications. Most of the works limited themselves to natural, climatic, and economic factors, whereas we assumed the key role played by socio-demographic factors. Therefore, in what follows we use both the factors that were deemed significant in other studies and the socio-demographic factors that were for the most part omitted from such studies. We also investigated the models that were used in the published research. Table 2 presents the factors and models reported in the literature.

III. METHODOLOGY

We relied on the following sources for collecting information: reports by the National Bureau of Statistics of China, BP Statistical Review, the U.S. Energy Information Administration, the research published in Russia and abroad by the experts in the field of gas production and available from ResearchGate, eLibrary, Scopus.

When deciding on a set of factors, we strived for a comprehensive analysis of their possible impact on natural gas consumption. As a result, we have identified four main groups of factors: economy, demography and urban areas, energy and climate (Table 3).

A number of studies included such factors as the GDP and GDP per unit of energy consumed [6]. The latter has been a key performance metric for the Chinese government for the last 10 years: it captures the efficiency of energy use in production and therefore it can have a significant impact on natural gas consumption. We have also included oil and coal prices as these are the main alternatives to natural gas.

Furthermore, we covered socio-demographic factors [7], since one of the uses of gas is residential heating and cooking. It was decided to consider such factors as the total population, population in urban agglomerations with over 1 million inhabitants, and urban area size. While the relationship between gas consumption and the number of people seems obvious, the relationship with the population of cities with over 1 million inhabitants warrants an explanation. The reasoning behind including this factor was as follows: the total population numbers may fail to serve as a representative metric since not all cities and settlements can be gasified; whereas the opposite is true for population numbers of major urban agglomerations because such cities are likely to already have in place district heating systems and gas available to households. It is also worth noting the fact that cities with a population of less than a million inhabitants are considered very small in China and therefore they may have less developed infrastructure.

Based on the above we can formally write down the

TABLE 2. Factors and Models Used in Related Studies

Factor	Forecasting method					
	A	B	C	D	E	F
Gas demand	8	2	2	9	4	7
Temperature	7	-	2	8	2	2
Month	3	-	-	1	1	-
Wind data	2	-	-	1	1	1
GDP	1	1	-	2	1	-
Population	1	1	-	1	1	1
Number of gas consumers	1	-	-	2	-	1
Gas price	1	-	-	1	1	1

A – neural networks, B – ANFIS, C – genetic algorithms, D – statistical models, E – hybrid models, F – time series. Source: [5]

TABLE 3. Factors Covered by the Study

Economy	Demography and urban areas	Energy	Climate
• GDP • Energy intensity • Oil price • Coal price	• Size of the urban area • Total population • Population of urban agglomerations with over 1 million inhabitants	• Coal consumption • Energy consumption from renewable sources • Passenger cars • Natural gas reserves • Availability of proven gas reserves	• Average annual temperature

problem statement. We consider a set of objects *X* and an associated set of values *Y*. It is assumed that between these two sets there exists a mapping function $y^*: X \to Y$ such that $y_t = y^*(x_t)$, and we can observe its values only on a limited subset of objects $\{x_1, \ldots, x_l\}$ from *X*. Let us call each "object–response" pair a precedent. The set of all such pairs $X^l = (x_t, y_t)_{t=1}^l$ forms a training sample. Algorithm training on these precedents is essentially to recover the y^* function based on the data of the training sample, that is, to find a decision function that will approximate the $y^*(x)$ function as accurately as possible. It is important for the proposed solution to be effective not only for the data used for training, but also for the entire set *X*, which characterizes the generalizing ability of the algorithm.

Next, we need to determine which models to use for solving the problem. To secure more accurate results, we use several models, both linear and nonlinear. Linear regression, which is trained by the least squares method, serves as the basic model. Ridge regression is used as an advanced linear model. Random forest and gradient boosting represent non-linear models.

Ridge regression

Ridge regression is a type of regularization. Regularization is a method to reduce overfitting of models. Overfitting occurs when a model "learns" the correct answers on a set of training data and therefore its generalization ability decreases. Ridge regression works by adding a penalty to the loss function used for training. Ridge regression, also known as L2 regularization, adds to the loss function a penalty equal to the square of the Euclidean norm (L2 norm) of the weight vector multiplied by the regularization parameter λ. This is expressed as:

$$L(w) = y - Xw^2 + \lambda w_2^2, \quad (1)$$

where λ – the regularization parameter of the model, w_2 – the sum of the squares of the entries of the vector *w*.

L2 regularization does not reset the coefficients of the model but makes their values closer to zero, which reduces the risk of overfitting and makes the model less sensitive to individual, possibly noisy, features.

Random forest

Random forest is a machine learning algorithm developed by Leo Breiman and Adele Cutler. The random forest uses a modified tree learning algorithm, which consists in selecting a random subset of features at each candidate split in the learning process. This is done in order to eliminate the correlation of the trees. After all, the higher the correlation of the trees in the "forest", the lower the accuracy of the forecast will be. However, the forecast accuracy gains are not only due to low correlation since there is a number of other reasons. For example, a random forest, due to the fact that trees are trained on a random subset of data, is not sensitive to various noises, outliers and distortions in the data. Moreover, the resistance to overfitting, which was referred to above, greatly improves the final quality of the model [8]. Since each tree in a random forest is built on the basis of a random subset of data and features, which is why different trees capture different aspects of the data, the chance of overfitting is significantly reduced.

Gradient boosting

Gradient boosting works by sequentially adding up predictors, usually decision trees, where each new predictor corrects for the errors of its predecessor. Unlike random forests, where trees are built independently, gradient boosting trees are built sequentially.

The algorithm starts with a very simple model, e.g., a plain average. This step is the starting point for all subsequent improvements. Next, the algorithm loops to improve the current model at each step. The goal is to close the distance between the prediction of the model and the actual data. The algorithm calculates the discrepancy produced by the current model for each instance of the data. The difference between the predicted and the actual values can be represented as an error that needs to be corrected. The algorithm then trains a new model, called a "weak learner" or "helper", which should correct for these errors. After training a new "helper", its predictions are added to the predictions of the basic model. This is done very carefully, at small increments, so as to avoid overfitting and ensure a smooth improvement of the basic model. The steps of error assessment, training of new "helpers", and their addition to the main model are repeated multiple times. Each iteration improves the model.

The coefficient of determination (R^2) and the mean absolute percentage error (MAPE) were chosen as metrics

of the quality of the models. The coefficient of determination is calculated as per the next equation:

$$R^2 = 1 - \frac{S_e^2}{S_y^2}, \quad (2)$$

where S_y^2 – variance of the response variable, S_e^2 – variance of prediction errors.

The mean absolute percentage error is calculated as per the next equation:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_i - y_i^c}{y_i}\right| \cdot 100\%, \quad (3)$$

where y_i – the observation *i* from the training sample, y_i^c – a calculated observation as predicted by the model, *N* – the number of observations.

In order to minimize the risk of overfitting, cross-validation was used. The essence of the cross-validation technique is as follows. The training sample is divided in *N* random ways into two subsamples: training and control samples. These samples must be disjoint. For each of the *N* partitions, the algorithm a_i is constructed and the error functional Q_i is calculated [9]. Then the cross-validated model performance will be defined as the averaging of all functional values on all partitions:

$$CV = \frac{1}{N}\sum_{i=1}^{N} Q\left(a_i\left(X_i^l\right)\right), \quad (4)$$

where a_i – the algorithm for the *i*-th of *N* partitions, Q_i – the functional of the *i*-th partition error.

Normalization of all factors is required for building models further. We normalize the features according to the next equation:

$$x'_{ij} = \frac{x_{ij}}{\|x_i\|}, \quad (5)$$

where x_{ij} – the *i*-th value of the *j*-th factor, $\|x_i\|$ – L2 norm of the vector x_i.

Normalization is an important step when working with linear and ridge regressions. Modeling can begin after method selection and data preparation are completed.

IV. Results and Discussion

We applied linear regression to make a forecast for benchmark comparison as was detailed above. The F-test proved the model significant but a number of factors (passenger cars, gas reserves, oil price) that are part of the model failed to be significant

The values of the metrics for the ridge regression were higher than for the linear regression. Interestingly, the temperature factor had almost zero significance for this model, although initially we assumed that it would strongly influence the response variable. The urban area size, which

TABLE 4. Comparison of the Key Metrics of Algorithm Quality and a Set of Significant Factors

	R^2	MAPE	TOP 5 factors (in descending order of importance)
Linear regression	87.13	9%	Passenger cars GDP Population in urban agglomerations with over 1 million inhabitants Coal consumption Oil price
Ridge regression	89.81	7%	Passenger cars GDP Coal consumption Oil price Natural gas reserves
Random forest	91.28	7.6%	Size of the urban area GDP Population in urban agglomerations with over 1 million inhabitants Population, total Coal consumption
Gradient boosting	92.57	6.77%	Population, total Natural gas reserves Size of the urban area Passenger cars Population in urban agglomerations with over 1 million inhabitants

is another factor whose high importance was initially assumed, proved to be of low significance as well. This is probably due to a more complex functional relationship than the algorithm could handle.

The random forest performed better than linear models. We argue that it confirmed our hypothesis of a more complex relationship at play: the urban area size topped the list of the most important factors. At the same time, the temperature factor still had the lowest significance score.

Gradient boosting yielded the best results among the considered models. We note that this model had all the factors as significant, with social and urban factors tending to prevail.

Table 4 compares the values of metrics and the most significant factors for the considered models.

The table shows that the best values of metrics are those produced by gradient boosting. Figure 2 plots the forecast generated by the model against a number of actual values.

Figure 2 shows that the values predicted by the algorithm almost always perfectly align with the demand trend dynamics. The algorithm even correctly identified the turning point in 2020, which was due to the COVID-19 pandemic, while relying on input data only, which attests to its high predictive ability.

V. Conclusion

The study revealed the factors that have the greatest impact on the dynamics of natural gas consumption in China and identified the machine learning methods that produce the most accurate forecast.

The main factors influencing the demand for natural gas inside China are socio-demographic factors such as the total population and the size of the urban area.

Consideration was given to both linear and nonlinear models (ridge regression, random forest, and gradient boosting). All these algorithms proved to be effective and are recommended for solving regression problems in machine learning. Gradient boosting performed best in forecasting natural gas consumption. Furthermore, in gradient boosting, unlike other models, there were no factors that did not contribute to the forecast. The most significant factors were overall population, gas reserves, urban area size, passenger cars, and population of urban agglomerations with over 1 million inhabitants. At the same time, all these factors were recognized as the highest-scoring factors by other models, reinforcing the reliability of the results.

The findings of the study will facilitate more accurate planning of the volumes of natural gas exports to China, which is essential for building foreign economic relations for a major supplier of natural gas such as Russia.

Acknowledgment

The research was supported by a grant from the Russian Science Foundation No. 23-78-10157, https://rscf.ru/project/23-78-10157/.

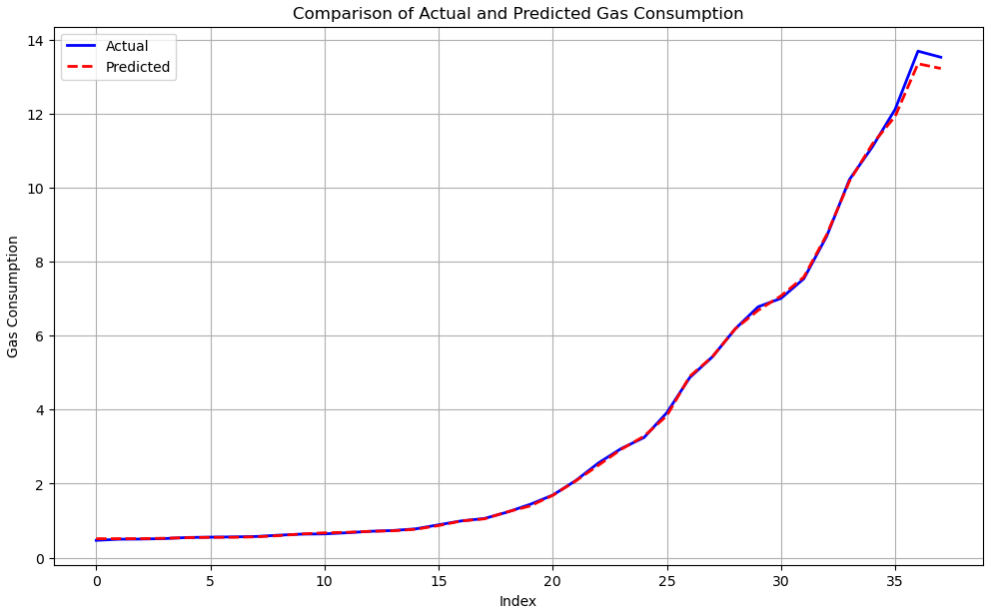

Fig. 2. Predicted and actual values of average daily consumption of natural gas in China, exajoules.

REFERENCES

[1] L. V. Eder, I. V. Filimonova, A. V. Komarova, V. Yu. Nemov, S. I. Shumilova, "Gas exports from Russia: the structure and dynamics of supplies," Gas Industry, no. 1 (179), pp. 86–92, 2019. (In Russian)

[2] I. V. Filimonova, V. Yu. Nemov, A. P. Samatova, "Forecast of the development of the gas potential of the regions of Eastern Siberia and the Russian Far East for the purposes of gasification and gas supply," in Proc. Geoeurasia-2024. Exploration technologies: science and business, vol. 12, Moscow, Russia, 2024, pp. 165–168. (In Russian)

[3] I. V. Filimonova, I. V. Provornaya, A. A. Kartashevich, A. Yu. Novikov, "Projection of gas production in the Republic of Sakha (Yakutia) as informed by the mix of proven reserves, transport availability, domestic and foreign demand," Mineral Resources of Russia. Economics and Management, no. 2 (187), pp. 46-56, 2024. (In Russian)

[4] V. B. Kashin, A. S. Pyatachkova, V. A. Smirnova, N. A. Potashev, China's energy development during the 14th Five-Year Plan period: Analytical note. Moscow, Russia: Central Committee of the Higher School of Economics, 2021, pp. 6–7. (In Russian)

[5] D. Sebalj, J. Mesaric, D. Dujak, "Analysis of methods and techniques for prediction of natural gas consumption: a literature review," Journal of information and organizational sciences, vol. 43, no. 1, pp. 99–117, 2019. DOI: 10.31341/jios.43.1.6

[6] W. Zhang, J. Yang, "Forecasting natural gas consumption in China by Bayesian model averaging," Energy Reports, vol. 1, pp. 216–220, 2015. DOI: 10.1016/j.egyr.2015.11.001

[7] M. Olgun, G. Ozdemir, E. Aydemir, "Forecasting of Turkey's natural gas demand using artificial neural networks and support vector machines," Energy Education Science and Technology Part A-Energy Science and Research, vol. 30, no. 1, pp. 15–20, 2012.

[8] O. V. Limanovskaya, T. I. Alferyeva, Fundamentals of machine learning: a textbook. Ekaterinburg, Russia: Ural, 2020, 80 p. (In Russian)

[9] A. Burkov, Machine learning in a nutshell. Saint-Petersburg, Russia: Piter Publishing House, 2020, 192 p. (In Russian)

Irina V. Filimonova, Prof. Dr. Sci. (Econ.), Head of a laboratory at the Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; Chair of the Political Economy department at Novosibirsk National Research State University, Novosibirsk, Russia. Her research interests include strategic planning of the development of the oil and gas industry in Russia and the regions of Eastern Siberia and the Russian Far East.

Vasily Yu. Nemov, Ph.D. (Econ.), Senior Researcher at the Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. Associate professor at Novosibirsk National Research State University, Novosibirsk, Russia. His research interests are built around the current state and projections of Russian and global oil and gas markets. .

Anastasia P. Samatova, a research assistant at the Trofimuk Institute of Petroleum Geology and Geophysics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; she is enrolled at the Novosibirsk National Research State University, Novosibirsk, Russia. Her research interests include energy consumption and energy efficiency of economic sectors, transportation systems in the regions of Russia and abroad

Nikita G. Akopov, is a student at Novosibirsk National Research State University, Novosibirsk, Russia. His research interests include projections of gas market dynamics in Russia and worldwide.