Energy Reports 10 (2023) 2352–2360

Contents lists available at ScienceDirect

Energy Reports

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

Short-term forecasting of natural gas consumption by determining the statistical distribution of consumption data

Ivan Smajla *, Domagoj Vulin, Daria Karasalihović Sedlar

University of Zagreb, Faculty of Mining, Geology and Petroleum Engineering, Pierottijeva 6, 10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 21 February 2023
Received in revised form 1 August 2023
Accepted 4 September 2023
Available online 15 September 2023

Keywords:
Forecasting methods
Natural gas consumption
Statistical distribution
Smart metering
Remote reading

ABSTRACT

The development of gas smart meters has enabled the collection of data on daily natural gas consumption which can be used to develop and improve methods and models for natural gas consumption forecasting. This paper presents the development of a model for the short-term forecasting of total natural gas consumption, which is applicable in different distribution areas where smart meters are installed in large numbers. The advantages of this model are the use of only two input parameters (daily natural gas consumption and average daily temperature), forecasting the total consumption in the determined area by analyzing the consumption data of less than 10% of the total consumers as well as robustness to consumer types. Daily natural gas consumption data collected from the more than 3300 gas smart meters over a period of six months was used for the determination of correlations between lognormal distribution variables and temperature. The defined correlations between distribution variables and temperature were used for upscaling consumption to a specific number of final consumers, i.e., to obtain the total consumption of natural gas in the observed area. Best results were achieved using the "two-day rolling average temperature" in the consumption scenario up to 250 m^3 per day (MAPE was 7.26%). When compared to using "average temperature" as an input parameter, "two-day rolling average temperature" and "shaving peaks temperature" produced better results due to the mitigated impact of sudden temperature changes that significantly affected the simulated consumption in the model while the actual consumption is a little more inert. Also, consumption scenarios up to 250 m^3 can be considered the most representative for forecasting total natural gas consumption since it achieved the best results.

1. Introduction

In the last ten years, the natural gas market in Europe has developed significantly and has become very dynamic. More and more natural gas suppliers are moving away from long-term contracts and turning to trading natural gas on a short-term basis, i.e., the spot market. In order to achieve quality business results, it is important for suppliers to have an accurate forecast of the natural gas consumption of their end users. More than 20 methods for forecasting natural gas consumption developed or tested between 2015 and 2021 have been described in the scientific work of Smajla et al. (2021). Since the last literature overview, several other authors dealing with this topic have achieved the following results.

Li et al. (2021) combined the decomposition-fusion technique with a replacement data function, feature selection, and a diversified Stacking ensemble learning model for short-term natural gas load forecasting for cities in China. Data from 1460 samples were used where 25 different features were considered and rearranged according to their importance. Empirical results of using this model showed that base learners' capabilities and discrepancies significantly affect the model's performance.

Liu et al. (2021) gave a historical overview of the development of methods for forecasting natural gas consumption. They concluded that the development of computer science and artificial intelligence has significantly influenced the improvement of forecasting methods, mostly those for the short-term. They also concluded that the long-term consumption forecast is mostly influenced by production, population, and economic factors, while the medium-term forecast of natural gas consumption is most significantly influenced by economic and weather variables.

Peng et al. (2021) suggested the use of a combination of long short-term memory, local mean decomposition, and wavelet threshold denoising algorithm for daily natural gas load forecasting in the city of London. They used from 335 to 355 input data to predict consumption for a period of 10, 20 and 30 days and compared the results obtained with the proposed model with four other models. Their model proved to be the best in the

* Corresponding author.
E-mail addresses: ivan.smajla@rgn.hr (I. Smajla), domagoj.vulin@rgn.hr (D. Vulin), daria.karasalihovic-sedlar@rgn.hr (D.K. Sedlar).

https://doi.org/10.1016/j.egyr.2023.09.051

case of forecasting consumption for a period of 20 days where it achieved MAPE of 7.09%.

Sharma et al. (2021) used real world natural gas consumption and weather data, both of which were recorded at 6-h intervals. Data were used to forecast natural gas consumption using four individual machine learning models. The results obtained using individual models were compared with the results obtained by the combined use of individual models and Sharma et al. concluded that by combining individual models 15% better MAPE can be achieved.

Wei et al. (2021) proposed the use of a hybrid white box model that combines the PCA algorithm, weighted parallel model architecture (WPMA), and MLR. The model was used to forecast daily natural gas consumption in four major cities located in three different climate zones. For the city of Melbourne, the proposed model achieved the highest accuracy with a MAPE of 7.79% compared to the remaining six machine learning models (MAPE ranging from 9.31% to 35.88%).

Xiong et al. (2021) used a new fractional-order accumulation-based incomplete gamma gray forecasting model to forecast annual natural gas consumption from 2008 to 2018 in the Asia-Pacific region. The achieved results showed that the MAPE of this model is lower compared to other models, which means that this model can be considered favorable for forecasting the annual consumption of natural gas and other energy prediction problems.

Zhang et al. (2021) proposed the use of a novel polynomial discrete gray prediction model with fractional order accumulation to forecast annual natural gas consumption in different regions of China over the next 5 years. The results obtained using this model (overall MAPE of 6.33%) proved to be superior compared to the four remaining models.

Due to the impact of data complexity on natural gas consumption forecasting Wei et al. (2022) proposed using a hybrid method of complexity measure, named CMLS. The results showed various forecasting performances at different complexity levels. In the case of very hard level, daily natural gas consumption is very hard to forecast (R^2 of forecasts are all negative).

Ma et al. (2022) used a nonlinear autoregressive model (NARX) with exogenous inputs, support vector machine (SVM), Gaussian process regression (GPR), and ensemble tree model (ETREE) to forecast daily gas load based on previous 3-year gas load data. With achieving forecast error below 7% by using an augmentation data set to train the model they concluded that this was a potentially good tool to forecast natural gas load.

Lao and Sun (2022) developed a novel discrete fractional nonlinear grey Bernoulli model with power term (DFNGBM(1,1,)) in order to achieve quality prediction of natural gas consumption and production in China. This model showed advantages when compared to the existing grey prediction models, such as fractional-order accumulation operation and time power term.

Song et al. (2022) developed a novel hybrid model that was proposed to predict the daily natural gas consumption in the district heating systems based on the seasonal decomposition and temporal convolution network (SDTCN) under the principle of divide and conquer strategy and deep learning algorithm. The model managed to achieve a prediction accuracy of 94.4% which proved to be a superior result when compared to several other models.

Safiyari et al. (2022) used multi-layer perceptron and support vector machine as two neural network models and vector autoregression and multivariate generalized autoregressive conditional heteroskedastic as two econometric models to forecast monthly natural gas demand. Analyzed data from March 2003 to February 2021 was used in order to forecast consumption from the beginning of March 2019 to the end of February 2021. The used model achieved the lowest measure of error compared to three other models.

Barbiero and Grillenzoni (2022) provided statistical methodology to forecast gas demand at the municipal level in the absence of historical data on gas consumption, by using spatial autoregressive (SAR) models with exogenous (X) variables. Novel natural gas infrastructure network in the region of Sardagena was designed and its economic sustainability was evaluated. The analysis showed a break-even point within 15 years with a market penetration rate being under 70%.

Panek and Włodek (2022) used machine learning algorithms, neural networks, and two regression algorithms, to forecast several variants of natural gas demand with different lengths of the forecast horizon. Obtained results showed that the Random Forrest algorithm achieved the best results for the tested input data but the differences between algorithms were not significant.

Ma et al. (2023) proposed using wavelet kernel-based machine learning and grey system modeling which takes advantage of the features of nonlinearity and periodicity of the wavelet kernel. The proposed model was used in three case studies based on the real-world data sets of urban natural gas consumption and it outperformed other 15 time series forecasting models.

By integrating domain knowledge into association graph construction and capturing temporal-spatial features via a hybrid deep learning network, Du et al. (2023) proposed a novel deep learning prediction method (KE-GB-TSN) for predicting natural gas consumption. Results demonstrate that the proposed model can outperform more sophisticated models like decision trees and gated recurrent units in terms of predicted results that are both more accurate and efficient. The proposed model's Mean Absolute Relative Errors and Root Mean Squared Relative Errors are all less than 0.11 and 0.14, respectively, indicating an improvement over earlier research.

Bilici et al. (2023) compared the performances of 4 different metaheuristic algorithms for estimating Turkey's natural gas demand. Normalized meteorological data (temperature, pressure, humidity, wind, and precipitation) over a period of 96 months were used as training input parameters, while the demand was tested over a period of 36 months. The PSO-Quadratic model showed the most successful forecasting results for observed natural gas consumption.

On the basis of the traditional nonlinear grey Bernoulli model, Tong et al. (2023) expanded the development coefficient a and grey action quantity b, proposed a new self-adaptive time-varying grey Bernoulli prediction model, deduced the time response formula of the model, and explored the relationship between model parameters and model accuracy. The results demonstrated that, in comparison to other traditional grey prediction models, the new model has a better simulation effect, higher prediction accuracy, and stronger applicability. The effectiveness of the model was examined by forecasting the natural gas consumption of China, the United States, and Russia.

Wang and Zhang (2023) added a novel fractional reverse accumulation method to the traditional grey model to construct a novel grey prediction novel. Natural gas consumption in the Commonwealth of Independent States was forecasted for the period from 2022 to 2025 and the new accumulation method performed well and enabled the model to capture the system's most recent trends, achieving accurate forecasts.

Ding et al. (2023) proposed Dual Convolution with Seasonal Decomposition Network method for forecasting natural gas consumption in different areas, from residential quarters to whole countries in different time spams. Simulations demonstrated that the proposed approach outperforms state-of-the-art approaches on city-level forecasting in terms of overall prediction accuracy and variation sensitivity regardless of different time intervals.

Ou and Chen (2023) proposed a novel dynamic parameter discrete grey model after studying the dynamics of parameters based on the parameter sequence obtained from the discrete grey model. The results demonstrated that this model performs better than other models and that its forecasts are more accurate and trustworthy. In order to provide useful information to the energy sector, the proposed model was then used to forecast China's natural gas production and domestic consumption from 2021 to 2024.

This paper additionally reviewed the literature that was published after the Smajla et al. (2021) paper, which thoroughly reviewed more than 20 forecasting methods for natural gas consumption. After reviewing the literature from both papers, it can be concluded in general that a variety of forecasting methods and models, ranging from mathematically straightforward ones to extremely sophisticated machine learning techniques, can be used to predict natural gas consumption. Some methods and models (usually the more complex ones) have shown to be somewhat more accurate than others in terms of forecasting, but this is highly subjective because the same methods and models produced high-quality results as well as low-quality ones depending on the paper in which they were used and analyzed. Additionally, forecasting can be done for time periods ranging from short-term (hourly or daily) to long-term (multi-year consumption forecast) consumption.

In conclusion, it can be said that there is no method or model that is universally better or worse for predicting the consumption of natural gas, but for the most accurate forecasting of natural gas consumption, it is necessary to carefully tune the method or model chosen, taking into account the available input and required output data.

A literature overview showed that the determination of a statistical distribution of consumption in relation to temperature has not been used so far for the purpose of forecasting the total consumption of natural gas in a particular distribution area. As a novelty, this paper shows that with the availability of high resolution consumption data (smart metering), using the determination of the statistical distribution of consumption (mathematically relatively simple method), it is possible to achieve high-quality forecasting results. This model uses historical data on daily natural gas consumption in order to determine the statistical distribution of consumption in correlation with temperature, i.e., to determine the total natural gas consumption in the observed distribution area. The advantage of this model is its robustness, i.e., independence from data on types of consumers, independence from all weather forecast parameters except the temperature forecast for the next day, and high-quality forecasting results for different consumption ranges. Apart from the above, the biggest advantage is the forecasting of the total consumption by analyzing the consumption data of less than 10% of the total consumers.

2. Smart metering data

Smart meters are increasingly being used around the world, mostly in terms of smart meters for measuring electricity consumption, but also smart meters for measuring gas and water consumption (Smajla et al., 2022a). Given that the use of smart meters provides very high-resolution consumption data, for example, electricity consumption every 15 min (Liu and Nielsen, 2016), there will be a large amount of data that needs to be properly processed (Wilcox et al., 2019). Data processing can be done using different computer software, algorithms, and/or machine learning, where it is important to choose the right platform so that the processing can be done as fast as possible (Liu et al., 2016). Also, Haben et al. (2016) and Mohajeri et al. (2020) propose data clustering in order to reduce computational complexity, high stochasticity, and irregularity of household-level demand.

High-quality processed data collected by smart meters can be used to forecast energy consumption, create energy efficiency certificates (Chambers and Oreszczyn, 2019), better understand user characteristics and optimize the services of utility companies (Wang et al., 2019a), create official statistics (Carroll et al., 2018) and define household appliance consumption patterns (Gajowniczek and Zabkowski, 2015; Weiss et al., 2012). Also, smart meters enabled the recognition of fraud attempts, i.e., theft of electricity (Sahoo et al., 2015), recognition of "energy" poverty (Hurst et al., 2020a,b), and the reduction of energy consumption with the goal of financial savings, because consumption data is available to end users in real-time (Hurst et al., 2020a,b).

Many analysis possibilities of collected smart meter data have caused different public opinions about the privacy of smart meter users and the misuse of data (Efthymiou and Kalogridis, 2010). Gough et al. (2022) proposed an innovative algorithm compliant with differential privacy (DP) to ensure the protection of data from consumers' smart meters, while Asghar et al. (2017) provided a comprehensive overview, shortcomings, and research recommendations for security solutions that are needed for privacy-preserving meter data delivery and management.

2.1. Data collection and database design

For the purpose of this research, data on daily natural gas consumption was collected by using the modules for remote reading of natural gas consumption. The installation of a module for remote consumption reading actually represents the implementation of smart meters, because by installing them, the classic membrane meter acquires the characteristics of a smart gas meter. This module collects consumption data every 6 h and sends the sum of measured values to the central module every 24 h. This will provide high-resolution data that needs to be processed in a quality and organized manner, so it does not adversely affect the efficiency and cost-effectiveness of business decisions (Lu et al., 2012; Wang et al., 2019b; Zakariazadeh, 2022). Also, Wei et al. (2022) state that the complexity of data often has a negative impact on the accuracy of forecasting daily natural gas consumption. Considering that and the literature overview given by Smajla et al. (2021) in previous research, the model developed in this paper observes only natural gas consumption and temperature as input parameters.

This model uses data on the daily natural gas consumption of around 3300 end users from the area of two closely located small cities with 124 593 inhabitants in the east of the Republic of Croatia. Consumption data refers to different types of users from households, which are certainly the most common to small business facilities, public buildings, small industrial facilities, etc. The mentioned consumption data was collected for a period of six months, from October 1st, 2021, to March 31st, 2022 (around 450 000 measurements).

For the same time period, temperature data was collected from a meteorological station near the observed cities. Temperature data contain temperature measurements in periods of every 3 h, i.e., 8 measurements in a period of 24 h. Temperatures were measured in the following hours during the day: 0 h, 3 h, 6 h, 9 h, 12 h, 15 h, 18 h and 21 h. Since most remote reading modules (more than 85 percent) send consumption information to the central module in the morning hours (between 4 and 8 a.m.), an average of 8 measurements starting from 6 a.m. was taken to calculate the average daily temperature. This means that, for example, to calculate the average daily temperature for the date 5th of January, 6 measurements (at 6 h, 9 h, 12 h, 15 h, 18 h, and 21 h) will be taken from that day as well as measurements at 0 am and 3 a.m. from the next day. The reason for this is to match the timeframe between the temperature data and natural gas consumption, which mostly corresponds to the consumption in one gas day (from 6 a.m. to 6 a.m. the next day). The gas day as a timeframe is important because it serves as the foundation timeframe for the nominations made for the transfer of natural gas from the transportation system to the distribution network the following day.

Since the data on natural gas consumption (in which the consumptions for areas and time periods that are not relevant for this research were also recorded) was transferred to SQLite database, the obtained data was filtered to different temporary tables (table views). Also, as Tureczek and Nielsen (2017) and Tureczek et al. (2019) concluded, filtering and pre-processing of input data can greatly speed up the run time of simulations. The data was first filtered only for the area of the observed cities and then for 5 different daily consumption ranges (consumption up to 40, 50, 60, 250, and 500 m^3 per day). The reason for observing the maximum daily consumption in 5 different cases is to observe the influence of larger consumers (daily consumption > 60 m^3) on the parameters that define the statistical distribution of consumers according to daily consumption. Also, in the collected consumption data, slightly more than 10% of the data was declared invalid due to incorrectly recorded consumption (examples: −1, NULL, double recorded consumption, abnormally high consumption, etc.). It was very important to separate such improperly recorded data because if that data was included in the simulation model, forecasting errors would be significantly higher.

3. Methodology

In terms of methodology, the first step was to determine the distributions of consumption for each day in the observed period, similar to the research published by Smajla et al. (2021). This was done in a way that all gas consumption data on each observed day was distributed in steps of 1 m^3, that is, for each step of 1 m^3, the number of users who had such consumption on the observed day was determined. After processing all the data, the consumption distribution curve was obtained for each day in the observed period from October 1st, 2021, to March 31st, 2022.

During the determination of consumption distributions, the minimum analyzed consumption was set to 0.001 m^3 in order to avoid the negative effect of consumption that was recorded as 0. It was necessary to separate the consumption that was recorded as 0, especially in the winter months, because it represents users who do not currently reside in the location where the smart meter is located, users who once used natural gas and have not terminated the contract with the supplier even though they are not using natural gas anymore, and users who have switched to alternative heating sources. It is very important to single out consumption 0 because analyzing it in the model would lead to a significant shift of the distribution curve to the left (the number of simulated very low consumptions would increase significantly), that is, the simulated total consumption would be significantly less than the real one. The maximum analyzed daily consumption for all cases was equal to the previously determined 5 different consumption ranges (consumption up to 40, 50, 60, 250, and 500 m^3 per day). Most of the analysis in terms of selecting the observed period (October 1st, 2021, to March 31st, 2022) and minimum and maximum consumption was done by SQL query in Python programming language (Pandas, 2022).

Using the conclusions presented in the work of Smajla et al. (2021) it was assumed that consumption distributions for each day best fit a lognormal statistical distribution. The same was proven to be correct by the analysis of distributions for the observed area but with a smaller number of users and a consumption range of up to 50 m^3 (Fig. 1) (Smajla et al., 2022b).

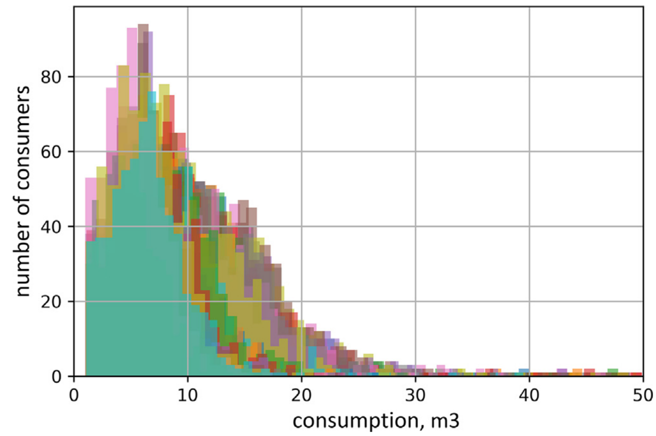

Fig. 1. Consumption distributions for the observed area (Smajla et al., 2022b).

The next step was to determine the variables (shape, loc and scale) that define the shape of statistical lognormal distributions for each day of the observed period by using scipy.stats library (Jones et al., 2001). The obtained variables (shape, loc, and scale) were then correlated with average daily temperatures by matching the dates (Fig. 2) and the parameters of the linear equation were calculated automatically in the code for each variable.

In order to forecast total natural gas consumption in the area the developed model needs a temperature forecast for the next day as well as the total number of users in the analyzed and tested area. Temperature data is then used in the correlations for calculating shape, loc, and scale parameters in order to determine the shape of lognormal statistical distribution for each day. The defined distributions are then used for upscaling consumption to a specific number of final consumers, i.e., to obtain the total consumption of natural gas in the observed area. Considering that about 15% of consumption data was reported outside the morning hours, unknowns about large temperature-independent consumers, and neglecting consumptions that amount to 0, a "flat deviation" appears in the model, which is canceled after the simulation by fine-tuning the model with median consumption of large consumers in every case.

In addition to using the average temperature for each day, this model also uses the "two-day rolling average temperature" and the "shaving peaks temperature". "Two-day rolling average temperature" is obtained as an average of the temperature for the observed day and the temperature from the previous day, while "shaving peaks temperature" is obtained by reducing/increasing the average temperature on the observed day by 50% of the difference compared to the temperature on the previous day.

4. Discussion and results

To obtain results, i.e., validate the accuracy of this model, total natural gas consumption was simulated for the same area and period as observed in data collection (two small cities, from October 1st, 2021, to March 31st, 2022) but for a significantly larger number of users.

Since the temperature forecast for the next day can be realized with very high precision, in the validation of this model the historical temperatures represented the temperature forecast for the next day. The total number of end users in the observed area was also well known (around 46 870), and it changed insignificantly over a longer period (+−1%). With the use of the data mentioned above, this natural gas consumption forecasting model managed to upscale the consumption data of around 3300 users with gas

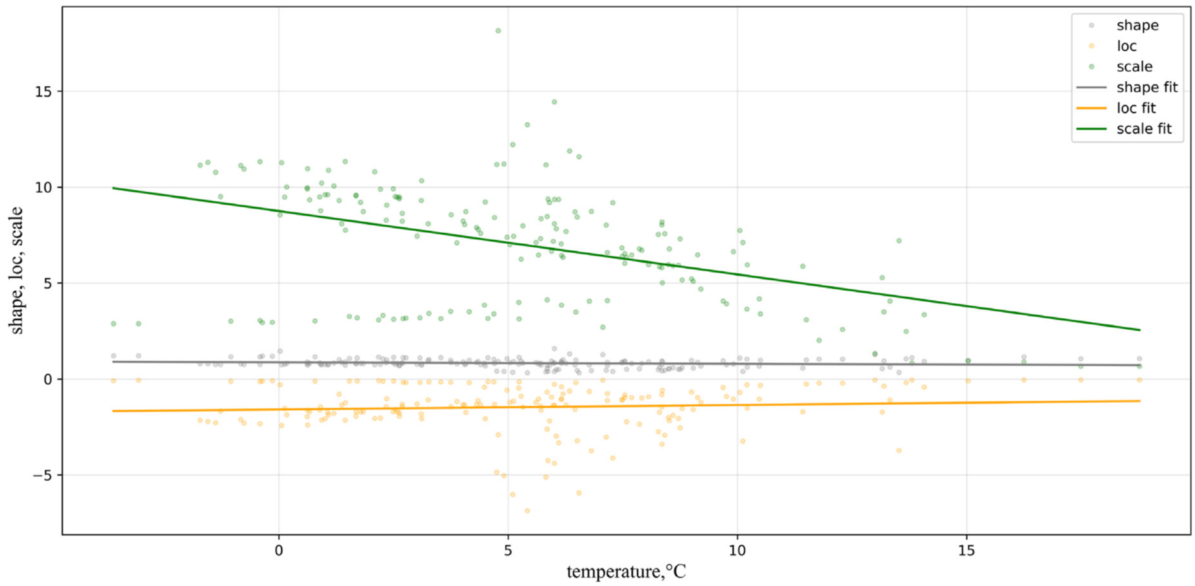

Fig. 2. Correlation between distribution variables shape, loc and scale and temperature for observed period.

Table 1
MAPE for all simulated consumption ranges and scenarios.

	Percent of small users				
	99%	98%	97%	96%	95%
40avg_temp	8.80	8.77	8.80	8.77	8.85
40rolling_2_days	7.78	7.81	7.88	7.79	7.91
40shavingpeaks_2_50	7.91	7.90	7.91	7.88	7.96
50avg_temp	8.84	8.81	8.80	8.80	8.76
50rolling_2_days	7.70	7.77	7.70	7.63	7.67
50shavingpeaks_2_50	7.89	7.91	7.83	7.79	7.78
60avg_temp	8.79	8.74	8.68	8.72	8.66
60rolling_2_days	7.69	7.64	7.62	7.58	7.51
60shavingpeaks_2_50	7.83	7.68	7.70	7.70	7.64
250avg_temp	8.93	8.80	8.75	8.65	8.75
250rolling_2_days	7.40	7.41	7.31	7.27	7.26
250shavingpeaks_2_50	7.55	7.50	7.40	7.43	7.35
500avg_temp	12.15	11.95	11.79	11.66	11.44
500rolling_2_days	11.00	10.53	10.51	10.31	10.16
500shavingpeaks_2_50	11.09	10.85	10.62	10.47	10.34

smart meters in order to forecast the total consumption of natural gas in the entire observed area for around 46 870 end users.

From the data collected by using gas smart meters (consumption scenario up to 250 m^3), it is evident that the share of consumers who have never had a consumption greater than 250 m^3 is more than 99%. Given that consumption data is available for around 3300 users out of a total of 46 870, it is not possible to determine the exact number of small (consumption less than 250 m^3) and large (consumption equal to or more than 250 m^3) consumers. For this reason, five different scenarios were prepared for each consumption range, where the share of small consumers ranges from 95% to 99%. Mean Absolute Percentage Errors (MAPE) for all consumption ranges and scenarios are shown in Table 1.

The results obtained by using the "two-day rolling average temperature" achieve the best forecasting results, where the MAPE is from 0.94% to 1.52% better than using the daily average temperature. The reason for this is the mitigation of the effect of sudden temperature changes that significantly affect the simulated consumption in the model, while in reality, the consumption is a little more inert (Figs. 3 and 4). For example, in the case of a warm day after several cold days, gas consumption in the majority of households will not decrease proportionally to the rise in temperatures (customer habits). Also, it can be concluded that the consumption scenario up to 250 m^3 achieves the best results, which means that it can be considered the most representative sample for forecasting total natural gas consumption in the observed area. Real and normalized values of reported total consumption, simulated total consumption, and temperatures in the observed period for the cases "250avg_temp" and "250rolling_2_days" are shown in Figs. 3 and 4. It should be noted that consumption was simulated for a total of 175 out of a possible 182 because there is no temperature data for 7 days (repairs at the meteorological station).

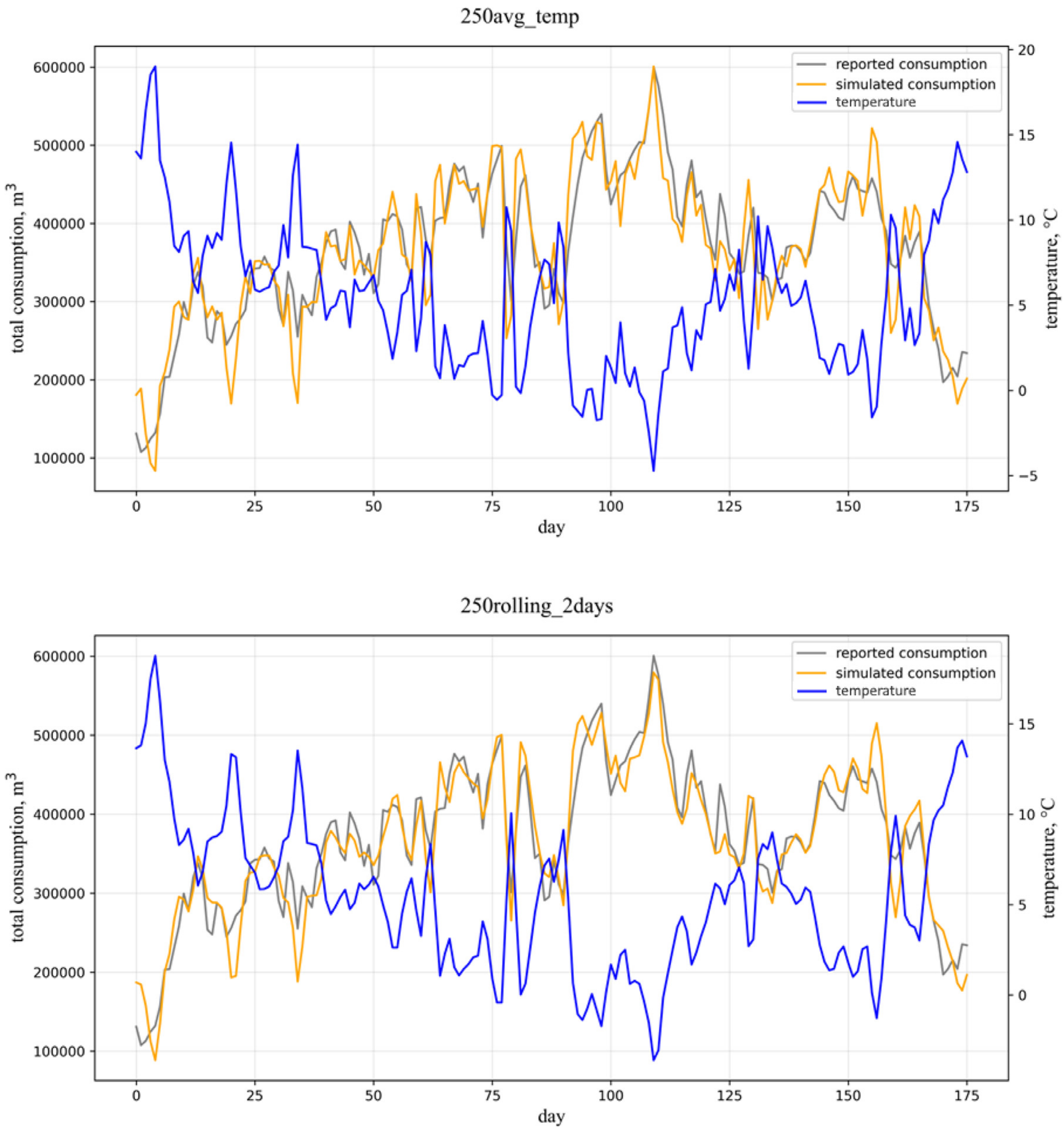

Fig. 3. Reported total consumption, simulated total consumption, and temperatures for cases "250avg_temp" and "250rolling_2_days".

5. Conclusion

This research has shown the use of statistical distributions as one of the possible methods for forecasting natural gas consumption which has not been encountered in the literature so far. The advantages of this model are the use of only two input parameters (daily natural gas consumption and average daily temperature), forecasting the total consumption by analyzing the consumption data of less than 10% of the total consumers as well as robustness to consumer types (households, business facilities, public facilities). This model achieved Mean Absolute Percentage Errors from 7.26% to 12.15% depending on the observed consumption range, used temperature datasets, and chosen scenario for the number of small users. The best results were achieved using the "two-day rolling average temperature" for the observed consumption range from 0.001 m^3 to 250 m^3. As far as the temperature datasets are concerned, in all cases the best results were achieved using the "two-day rolling average temperature", then the "shaving peaks temperature" dataset, while the worst results were achieved using the average daily temperature.

The reason behind this is the partial cancellation of the effect of sudden temperature changes that significantly affect the simulated consumption in the model, while in reality, the consumption is a little more inert. The model also showed relatively high-quality results for all other consumption ranges, except for the largest one where the consumption of up to 500 m^3 was observed. It can also be concluded that the selection of a high-quality temperature dataset has a much greater impact on accurate forecasting than the selection of a scenario for the share of small consumers.

Future research will be based on the formation of a better database of input data with the aim of preventing the recording of unwanted and incorrect data and further automation of the entire forecasting process. In order to improve the results, it is necessary to ensure that all smart meters report in the morning hours and that all errors during reporting are removed. In this way, the number of useful data will increase, which will ultimately lead to a more accurate consumption forecast. Also, with the further installation of gas smart meters in the observed distribution system, the percentage of total installed smart meters will increase,

Fig. 4. Normalized reported total consumption, simulated total consumption, and temperatures for cases "250avg_temp" and "250rolling_2_days".

and thus the number of input data. As this research has shown so far, a significant amount of input data is of utmost importance and has the greatest impact on increasing the accuracy of natural gas consumption forecasting. Furthermore, this research should be expanded in the future in terms of integration of the proposed model with the nomination systems of different suppliers in more distribution areas. After the integration, the development of a mobile application for end consumers with smart meters is also proposed, so that they too have insight into their own consumption in real time.

Nomenclature

CMLS — novel hybrid method of complexity measure
DFNGBM (1,1,) — novel Discrete Fractional Nonlinear Grey Bernoulli Model with power term
ETREE — Ensemble Tree Model
GPR — Gaussian Process Regression
KE-GB-TSN — Knowledge Enhanced Graph Based Temporal Spatial Network
MAPE — Mean Absolute Percentage Error
MLR — Multiple Linear Regression
NARX — Nonlinear Autoregressive Model
PCA — Principal Component Analysis
PSO — Particle Swarm Optimization
SDTCN — Seasonal Decomposition and Temporal Convolution Network ()
SVM — Support Vector Machine
WPMA — Weighted Parallel Model Architecture

CRediT authorship contribution statement

Ivan Smajla: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Domagoj Vulin:** Conceptualization, Software, Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Visualization. **Daria Karasalihović Sedlar:** Conceptualization, Formal analysis, Data curation, Methodology, Writing – original draft, Visualization, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ivan Smajla reports administrative support was provided by HEP Plin d.o.o.

Data availability

The data that has been used is confidential.

References

Asghar, M.R., Dán, G., Miorandi, D., Chlamtac, I., 2017. Smart meter data privacy: A survey. IEEE Commun. Surv. Tutor. 19 (4), 2820–2835. http://dx.doi.org/10.1109/COMST.2017.2720195.

Barbiero, T., Grillenzoni, C., 2022. Planning a novel regional methane network: Demand forecasting and economic evaluation. Energy Convers. Manage.: X 16, http://dx.doi.org/10.1016/j.ecmx.2022.100294.

Bilici, Z., Özdemir, D., Temurtaş, H., 2023. Comparative analysis of metaheuristic algorithms for natural gas demand forecasting based on meteorological indicators. J. Eng. Res. 100127, http://dx.doi.org/10.1016/J.JER.2023.100127.

Carroll, P., Murphy, T., Hanley, M., Dempsey, D., Dunne, J., 2018. Household classification using smart meter data. J. Off. Stat. 34 (1), 1–25. http://dx.doi.org/10.1515/jos-2018-0001.

Chambers, J.D., Oreszczyn, T., 2019. Deconstruct: A scalable method of as-built heat power loss coefficient inference for UK dwellings using smart meter data. Energy Build. 183, 443–453. http://dx.doi.org/10.1016/j.enbuild.2018.11.016.

Ding, J., Zhao, Y., Jin, J., 2023. Forecasting natural gas consumption with multiple seasonal patterns. Appl. Energy 337, 120911. http://dx.doi.org/10.1016/J.APENERGY.2023.120911.

Du, J., Zheng, J., Liang, Y., Wang, B., Klemeš, J.J., Lu, X., Tu, R., Liao, Q., Xu, N., Xia, Y., 2023. A knowledge-enhanced graph-based temporal-spatial network for natural gas consumption prediction. Energy 263, 125976. http://dx.doi.org/10.1016/J.ENERGY.2022.125976.

Efthymiou, C., Kalogridis, G., 2010. Smart grid privacy via anonymization of smart metering data, pp. 238–243, http://dx.doi.org/10.1109/smartgrid.2010.5622050.

Gajowniczek, K., Zabkowski, T., 2015. Data mining techniques for detecting household characteristics based on smart meter data. Energies 8 (7), 7407–7427. http://dx.doi.org/10.3390/en8077407.

Gough, M.B., Santos, S.F., Alskaif, T., Javadi, M.S., Castro, R., Catalao, J.P.S., 2022. Preserving privacy of smart meter data in a smart grid environment. IEEE Trans. Ind. Inform. 18 (1), 707–718. http://dx.doi.org/10.1109/TII.2021.3074915.

Haben, S., Singleton, C., Grindrod, P., 2016. Analysis and clustering of residential customers energy behavioral demand using smart meter data. IEEE Trans. Smart Grid 7 (1), 136–144. http://dx.doi.org/10.1109/TSG.2015.2409786.

Hurst, W., Montañez, C.A.C., Shone, N., 2020a. Time-pattern profiling from smart meter data to detect outliers in energy consumption. IoT 1 (1), 92–108. http://dx.doi.org/10.3390/iot1010006.

Hurst, W., Montanez, C.A.C., Shone, N., 2020b. Towards an approach for fuel poverty detection from gas smart meter data using decision tree learning. In: ACM International Conference Proceeding Series. pp. 23–28. http://dx.doi.org/10.1145/3416028.3416034.

Jones, E., Oliphant, T., Peterson, P., 2001. SciPy: Open source scientific tools for python. http://www.scipy.org.

Lao, T., Sun, Y., 2022. Predicting the production and consumption of natural gas in China by using a new grey forecasting method. Math. Comput. Simulation 202, 295–315. http://dx.doi.org/10.1016/J.MATCOM.2022.05.023.

Li, F., Zheng, H., Li, X., Yang, F., 2021. Day-ahead city natural gas load forecasting based on decomposition-fusion technique and diversified ensemble learning model. Appl. Energy 303, http://dx.doi.org/10.1016/j.apenergy.2021.117623.

Liu, X., Golab, L., Golab, W., Ilyas, I.F., Jin, S., 2016. Smart meter data analytics: Systems, algorithms, and benchmarking. ACM Trans. Database Syst. 42 (1), http://dx.doi.org/10.1145/3004295.

Liu, X., Nielsen, P.S., 2016. A hybrid ICT-solution for smart meter data analytics. Energy 115, 1710–1722. http://dx.doi.org/10.1016/j.energy.2016.05.068.

Liu, J., Wang, S., Wei, N., Chen, X., Xie, H., Wang, J., 2021. Natural gas consumption forecasting: A discussion on forecasting history and future challenges. J. Nat. Gas Sci. Eng. 90, 103930. http://dx.doi.org/10.1016/J.JNGSE.2021.103930.

Lu, N., Du, P., Guo, X., Greitzer, F.L., 2012. Smart meter data analysis. In: Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference. http://dx.doi.org/10.1109/TDC.2012.6281612.

Ma, X., Lu, H., Ma, M., Wu, L., Cai, Y., 2023. Urban natural gas consumption forecasting by novel wavelet-kernelized grey system model. Eng. Appl. Artif. Intell. 119, 105773. http://dx.doi.org/10.1016/J.ENGAPPAI.2022.105773.

Ma, D., Wu, R., Li, Z., Cen, K., Gao, J., Zhang, Z., 2022. A new method to forecast multi-time scale load of natural gas based on augmentation data-machine learning model. Chin. J. Chem. Eng. 48, 166–175. http://dx.doi.org/10.1016/J.CJCHE.2021.11.023.

Mohajeri, M., Ghassemi, A., Gulliver, T.A., 2020. Fast big data analytics for smart meter data. IEEE Open J. Commun. Soc. 1, 1864–1871. http://dx.doi.org/10.1109/OJCOMS.2020.3038590.

Ou, J., Chen, Y., 2023. A novel dynamic parameter discrete grey model and its application. Energy Rep. 9, 4941–4950. http://dx.doi.org/10.1016/J.EGYR.2023.04.009.

Pandas, 2022. https://pandas.pydata.org/pandas-docs/version/1.1.1/index.html.

Panek, W., Włodek, T., 2022. Natural gas consumption forecasting based on the variability of external meteorological factors using machine learning algorithms. Energies 15 (1), http://dx.doi.org/10.3390/en15010348.

Peng, S., Chen, R., Yu, B., Xiang, M., Lin, X., Liu, E., 2021. Daily natural gas load forecasting based on the combination of long short term memory, local mean decomposition, and wavelet threshold denoising algorithm. J. Nat. Gas Sci. Eng. 95, 104175. http://dx.doi.org/10.1016/J.JNGSE.2021.104175.

Safiyari, M.H., Shavvalpour, S., Tarighi, S., 2022. From traditional to modern methods: Comparing and introducing the most powerful model for forecasting the residential natural gas demand. Energy Rep. 8, 14699–14715. http://dx.doi.org/10.1016/J.EGYR.2022.10.397.

Sahoo, S., Nikovski, D., Muso, T., Tsuru, K., 2015. Electricity theft detection using smart meter data. In: 2015 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference, ISGT 2015. http://dx.doi.org/10.1109/ISGT.2015.7131776.

Sharma, V., Cali, Ü., Sardana, B., Kuzlu, M., Banga, D., Pipattanasomporn, M., 2021. Data-driven short-term natural gas demand forecasting with machine learning techniques. J. Petrol. Sci. Eng. 206, 108979. http://dx.doi.org/10.1016/j.petrol.2021.108979.

Smajla, I., Karasalihović Sedlar, D., Jukić, L., Vištica, N., 2022a. Cost-effectiveness of installing modules for remote reading of natural gas consumption based on a pilot project. Energy Rep. 8, 5631–5639. http://dx.doi.org/10.1016/J.EGYR.2022.04.019.

Smajla, I., Sedlar, D.K., Vulin, D., Jukić, L., 2021. Influence of smart meters on the accuracy of methods for forecasting natural gas consumption. Energy Rep. 7, 8287–8297. http://dx.doi.org/10.1016/J.EGYR.2021.06.014.

Smajla, I., Vulin, D., Karasalihović Sedlar, D., Jukić, L., 2022b. Short-term natural gas consumption forecasting by using smart meters to determine the statistical distribution of consumers. In: Ban, M., Duić, N., Vučinić, A. Anić, Baleta, J., Fan, Y. Van, Gjoka, K., Kabashi, S., Klemeš, J.J., Pula, G., Varbanov, P.S., Vujanović, M. (Eds.), Digital Proceedings of the 5th SEE Conference on Sustainable Development of Energy, Water and Environment Systems. University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture.

Song, J., Zhang, L., Jiang, Q., Ma, Y., Zhang, X., Xue, G., Shen, X., Wu, X., 2022. Estimate the daily consumption of natural gas in district heating system based on a hybrid seasonal decomposition and temporal convolutional network model. Appl. Energy 309, 118444. http://dx.doi.org/10.1016/J.APENERGY.2021.118444.

Tong, M., Qin, F., Dong, J., 2023. Natural gas consumption forecasting using an optimized grey Bernoulli model: The case of the world's top three natural gas consumers. Eng. Appl. Artif. Intell. 122, 106005. http://dx.doi.org/10.1016/J.ENGAPPAI.2023.106005.

Tureczek, A.M., Nielsen, P.S., 2017. Structured literature review of electricity consumption classification using smart meter data. Energies 10 (5), http://dx.doi.org/10.3390/en10050584.

Tureczek, A.M., Nielsen, P.S., Madsen, H., Brun, A., 2019. Clustering district heat exchange stations using smart meter consumption data. Energy Build. 182, 144–158. http://dx.doi.org/10.1016/j.enbuild.2018.10.009.

Wang, Y., Chen, Q., Gan, D., Yang, J., Kirschen, D.S., Kang, C., 2019a. Deep learning-based socio-demographic information identification from smart meter data. IEEE Trans. Smart Grid 10 (3), 2593–2602. http://dx.doi.org/10.1109/TSG.2018.2805723.

Wang, Y., Chen, Q., Hong, T., Kang, C., 2019b. Review of smart meter data analytics: Applications, methodologies, and challenges. IEEE Trans. Smart Grid 10 (3), 3125–3148. http://dx.doi.org/10.1109/TSG.2018.2818167.

Wang, H., Zhang, Z., 2023. A novel grey model with fractional reverse accumulation for forecasting natural gas consumption. Comput. Ind. Eng. 179, 109189. http://dx.doi.org/10.1016/J.CIE.2023.109189.

Wei, N., Yin, L., Li, C., Li, C., Chan, C., Zeng, F., 2021. Forecasting the daily natural gas consumption with an accurate white-box model. Energy 232, 121036. http://dx.doi.org/10.1016/J.ENERGY.2021.121036.

Wei, N., Yin, L., Li, C., Liu, J., Li, C., Huang, Y., Zeng, F., 2022. Data complexity of daily natural gas consumption: Measurement and impact on forecasting performance. Energy 238, http://dx.doi.org/10.1016/j.energy.2021.122090.

Weiss, M., Helfenstein, A., Mattern, F., Staake, T., 2012. Leveraging smart meter data to recognize home appliances. In: 2012 IEEE International Conference on Pervasive Computing and Communications, PerCom 2012. pp. 190–197. http://dx.doi.org/10.1109/PerCom.2012.6199866.

Wilcox, T., Jin, N., Flach, P., Thumim, J., 2019. A big data platform for smart meter data analytics. Comput. Ind. 105, 250–259. http://dx.doi.org/10.1016/j.compind.2018.12.010.

Xiong, P., Li, K., Shu, H., Wang, J., 2021. Forecast of natural gas consumption in the Asia-Pacific region using a fractional-order incomplete gamma grey model. Energy 237, 121533. http://dx.doi.org/10.1016/J.ENERGY.2021.121533.

Zakariazadeh, A., 2022. Smart meter data classification using optimized random forest algorithm. ISA Trans. 126, 361–369. http://dx.doi.org/10.1016/j.isatra.2021.07.051.

Zhang, J., Qin, Y., Duo, H., 2021. The development trend of China's natural gas consumption: A forecasting viewpoint based on grey forecasting model. Energy Rep. 7, 4308–4324. http://dx.doi.org/10.1016/j.egyr.2021.07.003.