CHEMICAL ENGINEERING TRANSACTIONS
VOL. 100, 2023
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Sauro Pierucci, Flavio Manenti

ISBN 979-12-81206-03-8; **ISSN** 2283-9216
DOI: 10.3303/CET23100029

Machine Learning Models for Natural Gas Consumer Decision Making: Case Study of a Colombian Company

Anibal Alviz-Meza[a,b,*], Edwin Torres-Salazar[b], Silvia Gastiaburú-Morales[a], Halyn Alvarez-Vasquez[a], Juan Orozco-Agamez[c], Darío Peña-Ballesteros[c]

[a]Grupo de investigación en Deterioro de Materiales, Transición energética y Ciencia de Datos DANT3, Universidad Señor de Sipán, Facultad de Ingeniería, Arquitectura y Urbanismo, Chiclayo, Pimentel 14001, Peru
[b]Semillero de Investigación en Corrosión de Metales, Energías Sostenibles y Análisis de Datos - COMD3S, Faculty of enginering, arquitecture and urbanism, Universidad Señor de Sipán, Chiclayo, Pimentel, 14001, Peru
[c]Grupo de Investigaciones en Corrosión, Universidad Industrial de Santander, Parque Tecnológico Guatiguará, Piedecuesta 681011, Colombia
alvizanibal@crece.uss.edu.pe

Energy systems generate a large amount of data related to their production, sale, consumption services, etc. These numbers remain deposited into large databases with valuable hidden information about trends that would help service providers and their users make better decisions. Data science, mathematics, and statistics have been advancing to respond to these needs. This revolution has brought several techniques, algorithms, and prediction models, among which machine learning emerges. This work helps to link these tools with the decisions making of companies and users related to a Colombian natural gas producer company. The public data extracted from the company were understood, cleaned, and processed before being deployed within two predictive models: time series forecasting using artificial neural networks (TS-ANN) and k-nearest neighbors (KNN). The model targets were the fixed and variable natural gas charges regarding the demographic information of the customers. The results obtained by both models presented small deviations from the test values. Both, the KNN and TS-ANN model received the data time input for predicting natural gas charges in a selected demographic sector. As a result, we found trends to identify the consumers who receive a higher service charge and the period in which these values are higher. These results are presented as an indication of what can be done today with the public information provided by companies: allowing consumers to adjust their consumption strategies.

1. Introduction

The no-return point set to the world for facing global warming is getting closer over time, boosting the energy transition processes (Kweku et al., 2018). Natural gas represents the more suitable energetic vector to lead the transition to sustainable energies, making it one of the most attractive assets in economies highly dependent on fossil fuels (Szoplik, 2015). Likewise, synthetic gases from renewable origins have become equally attractive, suitable, and interchangeable (Awasthi et al., 2020). Thus, properly managing and forecasting the consumption of these promising fuels has gained more scientific relevance in recent years (Anđelković and Bajatović, 2020). At this point, adequate measurements and interpretation of the data continuously recorded by energetic supplier companies have increased their value for customers' decision-making (Tealab, 2018). This raw material serves as the input of modern machine learning algorithms, which have proved their ability to model and forecast complex energetic services (Hashemizadeh et al., 2021).
As previously reported, being able to accurately predict natural gas costs may help save energy, manage supply contracts, and plan infrastructures (Tamba et al., 2018). One of the first approaches to understanding the natural gas demand was conducted in the middle of the last century (Hubbert, 1949), opening the path to a hundred research works seeking to predict and control this sector by means of machine learning models. A few authors discussed the prediction of natural gas prices for energy cost savings utilizing the non-linear K-Nearest Neighbor (KNN) algorithm, reaching a high accuracy between the actual and predicted variables (Wahid and Kim, 2016).

Paper Received: 24 January 2023; Revised: 6 April 2023; Accepted: 21 May 2023
Please cite this article as: Alviz-Meza A., Torres-Salazar E., Gastiaburu-Morales S., Alvarez-Vasquez H., Orozco-Agamez J., Pena-Ballesteros D., 2023, Machine Learning Models for Natural Gas Consumer Decision Making: Case Study of a Colombian Company, Chemical Engineering Transactions, 100, 169-174 DOI:10.3303/CET23100029

Meanwhile, most forecasting studies models are associated with neural networks, which are inspired by the architecture of the human brain, exhibiting features with the capacity to learn from examples and find relationships usually non-linearly correlated, as expected when modeling independent variables from energy sectors (Sen et al., 2019).
This research work represents a study case to test two well know algorithms used to forecast gas natural charges but applied to the main natural gas supplier Colombian company. Hereby, we focused on exploring two methods to predict the monthly gas fixed and variable charges. K-Nearest Neighbors and Time Series Forecasting Using Neural Networks were selected to benchmark the effectiveness and efficiency of each solution approach. This study discusses natural gas price prediction and determines its behavior in different residential sectors.

2. Materials and methods

This section aims to declare the steps followed to find the most appropriate method for forecasting gas fixed in residential and non-residential sectors from Medellin, Colombia. The forecasting was performed monthly and annually because the studied company belongs to the country's public sector. This selection allows showing changes from the gas distributor's and individual consumers' viewpoints (Yucesan et al., 2021). The machine-learning methods used, KNN and TS-ANN, were optimized and compared to obtain the best possible forecasting and generate the best possible recommendations to consumers. The KNN and TS-ANN algorithm was trained to predict natural gas charges, only receiving their corresponding time records and the demographic sector's type. This study case was developed in *python*, following the steps shown in Figure 1 and described below.

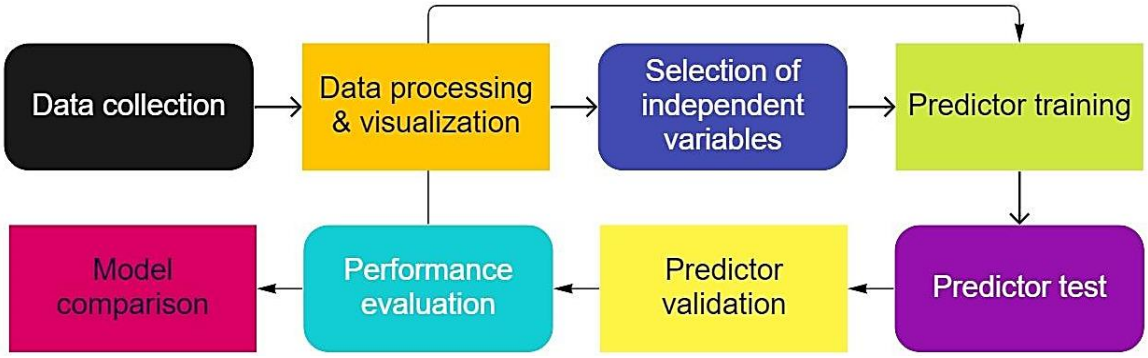


Figure 1: Proposed methodological steps for the forecasts

Phase 1: Data were collected and downloaded from the government web pages (EPM, 2022b). The downloaded file has numerical and categorical data. However, since the resolution's formula uses numerical data, this work only contemplated the categorical variables of demographic type, month, and year, avoiding reproducing the original formula used to determine gas natural charges in Colombia (EPM, 2022a). The data studied were temporarily located between January 2017 and January 2022. Otherwise, the initial data examinations allowed us to remove irregular information and identify the non-linear relationship between the target and features.

Phase 2: Data were pre-processed. As data were composed of categorical and numerical features, the numerical variables were transformed (an expression used in plots) into a scale from 0 to 1 by using *MinMaxScaler*, and to 1 and 0 using *OneHotEncoder* for the categorical ones, both strategies explored through the *Scikit-Learn* library. After the separate pre-processing of both types of features, they were merged in the same data frame.

Phase 3: The target was separated from features for further analysis.

Phase 4: The training data were aleatory selected (KNN-algorithm) and taken as 70% of the data, leaving the 30% remaining for testing the model. In the TS-ANN case, the time series required to be ordered before being trained – is essential for the *forward chaining* cross-validation.

Phase 5: The most important variables affecting the natural gas charges were determined. The KNN algorithm was optimized by modifying the appropriate number of k-neighbors to obtain the balance point between training and testing errors. Meanwhile, for TS-ANN, it was employed a multilayer perceptron time series regression

model – *MPLRegressor* from *ScikitLearn* – to predict the fixed natural gas target, adjusting the size of the hidden layer (values between 10 and 100) and the activation function (*relu*, *logistic*, and *tanh*) with the five selected folds on different k windows.

Phase 6: The accuracy of the forecasting methods was compared by using mean squared error (MSE), mean absolute error (MAE), and mean squared log error (MSLE). These metrics were applied to test the TS-ANN model, comparing the predicted values from actual and predicted data. The comparison with the KNN model was only based on the predicted values.

3. Results and discussion

3.1 Data visualization

The commercial and industrial stratum from the non-residential sector was selected for analysis and visualization. This decision was made to explore high consumption scenarios, given that residential strata typically generate consumption below 20 m^3/month (La_Republica, 2020). Figure 2 shows that the residential sector is the one that receives the higher prices -stratum 5 and stratum 6-, while strata 1, 2, and 3 perceive the lowest. Meanwhile, in the non-residential sector, commercial and industrial zones get higher charges. Moreover, the natural gas charges were found strongly related to annual periods (see Figure 3).

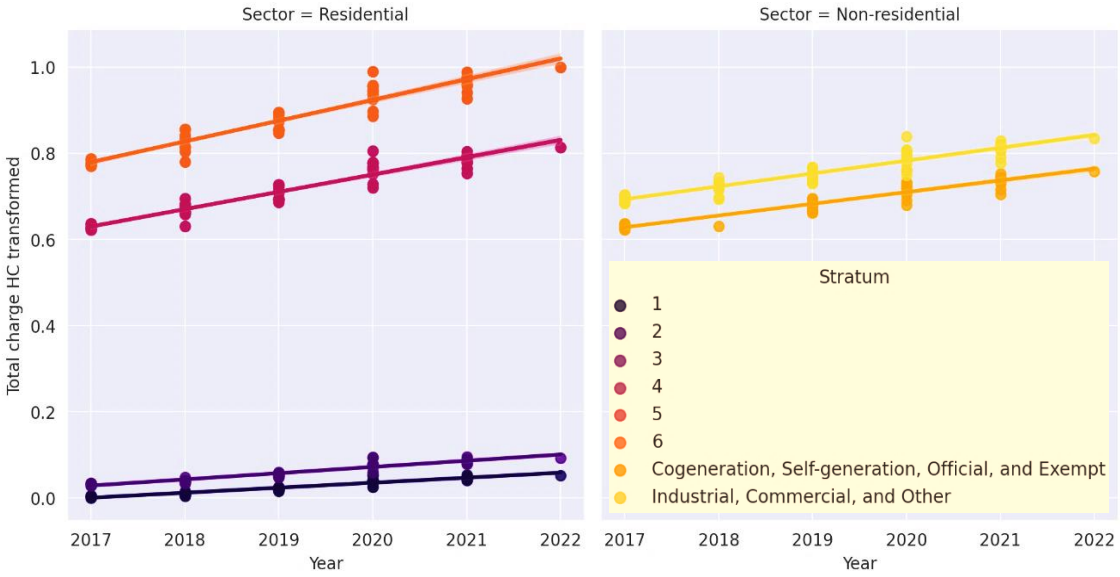


Figure 2: Natural gas charges scaled for higher consumption by sectors and strata

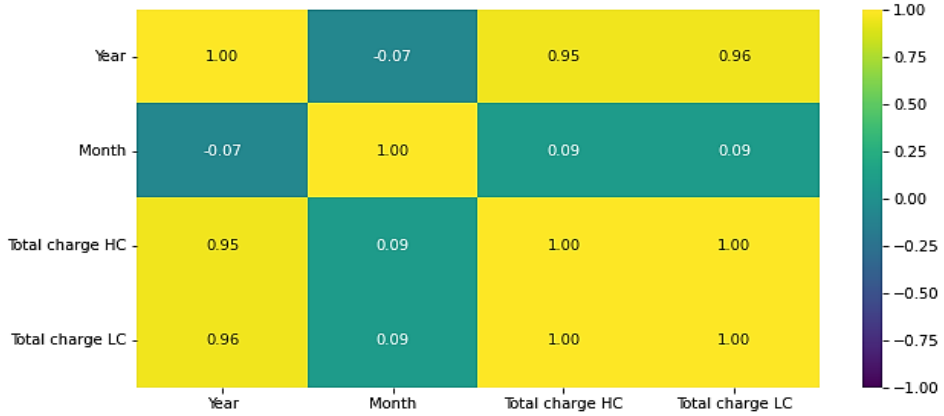


Figure 3: correlation matrix for high and low natural gas charges

Analyzing the trends of the total charges for high consumption month by month allows for obtaining the following findings (see Figure 4). The residential sector has the lowest charges in April for the first semester and October for the second semester. In the non-residential sector, the lowest charges are in February for the first semester and in October for the second. These findings may help different users from EPM to make more sustainable consumptions, even though the low quality of the open-access used data limits the reliability of these claims.

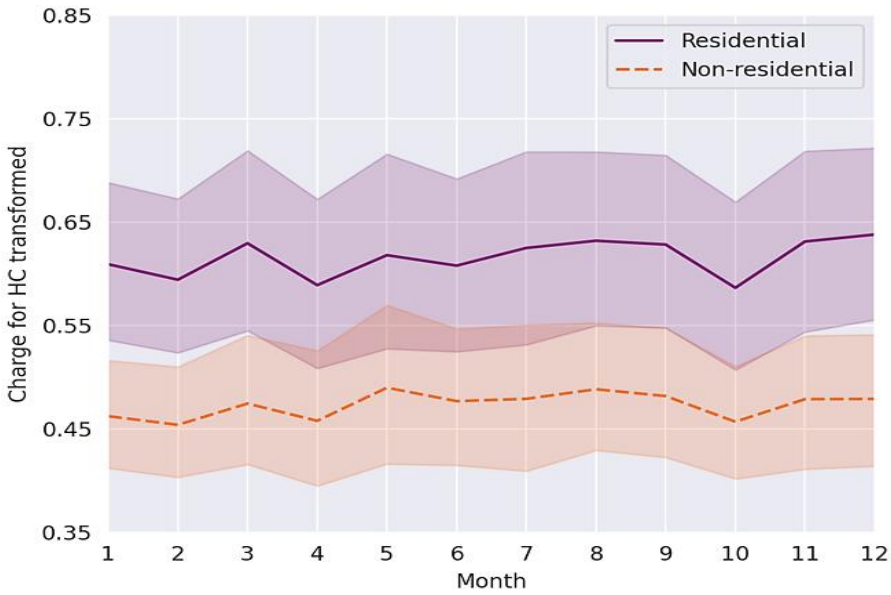


Figure 4: Monthly natural gas charges scaled for higher consumption.

3.2 KNN and TS-ANN models

Figure 5 indicates the best configuration obtained for the KNN when considering 3 nearest neighbors and 70% of data for training. The training and validation errors balance point is usually tested by a trial-and-error process to define the number of neighbors

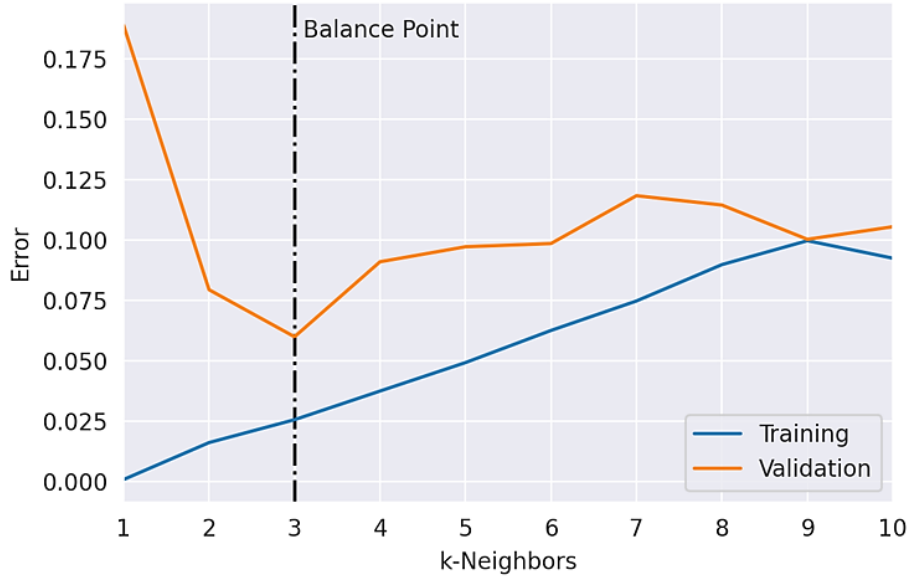
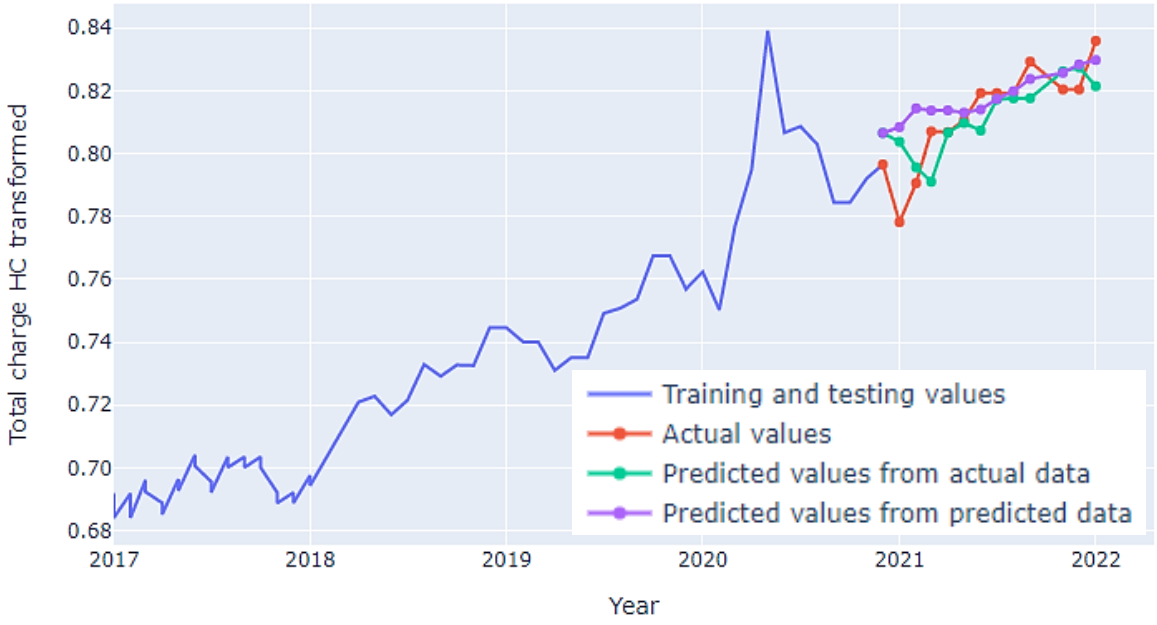


Figure 5: Fluctuation of training and validation errors with the number of k-neighbors for the commercial and industrial non-residential sector



.

Figure 6: Natural gas charges predicted for the commercial and industrial non-residential

The same metrics and data partitioning used for evaluating the KNN algorithm were replicated for the TS-ANN model. In this case, the best architecture was obtained after a cross-validation process where the best parameters found were a window size of 9, an activation function of “relu”, and a hidden layer size of 40. Figure 6 shows that the model built based on the actual data (green line) fits better than the one built based on the predicted data (purple line). This last model inherits the errors of the green line’s model and moves away from the actual values as the observation window progresses. However, the deviation of both models concerning the actual data is acceptable. These results make evident the limitations of the artificial neuronal network model to follow the fluctuation of actual values.

Table 1 resumes the metric performances of both algorithms (KNN and TS-ANN based on actual data): Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Squared Log Error (MSLE). Under said metrics, both models were found feasible predictors to be used interchangeably for natural gas charges predictions in different residential sectors from Medellin. However, it is also worth mentioning that the 2020 pandemic may have affected the natural gas charges trends in recent years, making the neural network-based model less accurate than the KNN.

Table 1: Metric performances of models from training-testing scaled data

	MSE	MAE	MSLE
KNN	0.0001	0.0079	0.0000
TS-ANN	0.0001	0.0087	0.0000

4. Conclusions

This research work focused on developing two machine-learning models to forecast natural gas charges from a Colombian company. An initial data visualization allowed us to note months in which it is more recommendable to make high natural gas consumption in residential and non-residential sectors. The KNN model fits better for 3 nearest neighbors when employing 70% of data for training. Meanwhile, the best time series artificial neuronal networks model was obtained for the same data partitioning after cross-validation. The best parameters found were: window size = 9, activation function = "relu", and hidden layer size = 40. Both algorithms presented acceptable performance metrics but the TS-ANN was less accurate, probably due to unusual charge trends in 2020 (the start of the COVID-19 pandemic).

Acknowledgments

The authors express their acknowledgment to the Ministry of Technology, Information, and Communications of Colombia. We also express our gratitude to the Universidad Señor de Sipán for supporting the presentation of this project.

References

Anđelković A. S., Bajatović, D., 2020, Integration of weather forecast and artificial intelligence for a short-term city-scale natural gas consumption prediction. *Journal of Cleaner Production*, *266*, 122096. https://doi.org/10.1016/J.JCLEPRO.2020.122096

Awasthi M. K., Sarsaiya S., Patel, A., Juneja A., Singh R. P., Yan B., Awasthi S. K., Jain A., Liu T., Duan Y., Pandey A., Zhang Z., Taherzadeh M. J., 2020, Refining biomass residues for sustainable energy and bio-products: An assessment of technology, its importance, and strategic applications in circular bio-economy. *Renewable and Sustainable Energy Reviews*, *127*(December 2019), 109876. https://doi.org/10.1016/j.rser.2020.109876

EPM., 2022a, *Tarifas del servicio de Gas Natural de EPM*. https://cu.epm.com.co/clientesyusuarios/gas/tarifas-gas

EPM., 2022b, *Tarifas Para Servicios De Gas - EPM. (Hogares - Tarifa Gas Natural)* | *Datos Abiertos Colombia*. https://www.datos.gov.co/Funci-n-p-blica/Tarifas-Para-Servicios-De-Gas-EPM-Hogares-Tarifa-G/ekup-y869

Hashemizadeh A., Maaref A., Shateri M., Larestani A., Hemmati-Sarapardeh A., 2021, Experimental measurement and modeling of water-based drilling mud density using adaptive boosting decision tree, support vector machine, and K-nearest neighbors: A case study from the South Pars gas field. *Journal of Petroleum Science and Engineering*, *207*, 109132. https://doi.org/10.1016/J.PETROL.2021.109132

Hubbert M. K., 1949, Energy from fossil fuels. *Science*, *109*(2823), 103–109. https://doi.org/10.1126/SCIENCE.109.2823.103/ASSET/26C12F7E-9E98-4A64-89CD-1260024F3BCC/ASSETS/SCIENCE.109.2823.103.FP.PNG

Kweku D., Bismark O., Maxwell A., Desmond K., Danso K., Oti-Mensah E., Quachie A., Adormaa B., 2018, Greenhouse Effect: Greenhouse Gases and Their Impact on Global Warming. *Journal of Scientific Research and Reports*, *17*(6), 1–9. https://doi.org/10.9734/jsrr/2017/39630

La_Republica, 2018, *Las empresas más grandes de 2017*. https://www.larepublica.co/especiales/las-empresas-mas-grandes-de-2017/en-el-top-100-se-concentra-mas-de-la-mitad-de-las-ventas-2728098

Sen D., Günay M. E., Tunç K. M. M., 2019, Forecasting annual natural gas consumption using socio-economic indicators for making future policies. *Energy*, *173*, 1106–1118. https://doi.org/10.1016/J.ENERGY.2019.02.130

Szoplik J., 2015, Forecasting of natural gas consumption with artificial neural networks. *Energy*, *85*, 208–220. https://doi.org/10.1016/J.ENERGY.2015.03.084

Tamba J. G., Essiane S. N., Sapnken E. F., Koffi F. D., Nsouandélé J. L., Soldo B., Njomo D., 2018, Forecasting natural gas: A literature survey. *International Journal of Energy Economics and Policy*, *8*(3), 216–249.

Tealab A., 2018, Time series forecasting using artificial neural networks methodologies: A systematic review. *Future Computing and Informatics Journal*, *3*(2), 334–340. https://doi.org/10.1016/J.FCIJ.2018.10.003

Wahid F., Kim D., 2016, A Prediction Approach for Demand Analysis of Energy Consumption Using K-Nearest Neighbor in Residential Buildings. *International Journal of Smart Home*, *10*(2), 97–108.

Yucesan M., Pekel E., Celik E., Gul M., Serin F., 2021, Forecasting daily natural gas consumption with regression, time series and machine learning based methods. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*. https://doi.org/10.1080/15567036.2021.1875082