Data in brief 25 (2019) 104122

Contents lists available at ScienceDirect

Data in brief

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

Data Article

Data on forecasting energy prices using machine learning

Gabriel Paes Herrera [a, b, *], Michel Constantino [b], Benjamin Miranda Tabak [c], Hemerson Pistori [b, d], Jen-Je Su [a], Athula Naranpanawa [a]

[a] Department of Accounting, Finance and Economics, Griffith University, Nathan Campus, Queensland 4111, Australia
[b] Department of Environmental Sciences and Sustainability, Dom Bosco Catholic University, Campo Grande, MS, Brazil
[c] School of Public Policy and Government, Getulio Vargas Foundation (EPPG/FGV), Brasilia, DF, Brazil
[d] Department of Computer Science, Federal University of Mato Grosso do Sul, Campo Grande, MS, Brazil

ARTICLE INFO

Article history:
Received 27 April 2019
Received in revised form 18 May 2019
Accepted 31 May 2019
Available online 12 June 2019

Keywords:
Oil
Natural gas
Coal
ANN

ABSTRACT

This article contains the data related to the research article "Long-term forecast of energy commodities price using machine learning" (Herrera et al., 2019). The datasets contain monthly prices of six main energy commodities covering a large period of nearly four decades. Four methods are applied, i.e. a hybridization of traditional econometric models, artificial neural networks, random forests, and the no-change method. Data is divided into 80-20% ratio for training and test respectively and RMSE, MAPE, and M-DM test used for performance evaluation. Other methods can be applied to the dataset and used as a benchmark.

DOI of original article: https://doi.org/10.1016/j.energy.2019.04.077.
* Corresponding author. Department of Accounting, Finance and Economics, Griffith University, Nathan Campus, Queensland 4111, Australia.
E-mail addresses: gabriel.paesherrera@griffithuni.edu.au (G.P. Herrera), michel@ucdb.br (M. Constantino), benjaminm.tabak@gmail.com (B.M. Tabak), pistori@ucdb.br (H. Pistori), j.su@griffith.edu.au (J.-J. Su), a.naranpanawa@griffith.edu.au (A. Naranpanawa).

https://doi.org/10.1016/j.dib.2019.104122

Specifications Table

Subject area	Economics
More specific subject area	Energy forecasting
Type of data	Tables, Figures and Excel file
How data was acquired	Primary data on historical prices of oil, coal and natural gas were obtained from the International Monetary Fund (IMF)
Data format	Raw, analyzed
Experimental factors	Four forecasting methods were compared using six time series with different sizes
Experimental features	Several parameters were tested for each method. The code was implemented on R software
Data source location	International Monetary Fund – IMF, 720 19th street, Washington, D.C., United States of America.
Data accessibility	The data is included in this article
Related research article	G.P. Herrea, M. Constantino, B.M. Tabak, H. Pistori, J. Su, A. Naranpanawa, Long-term forecast of energy commodities price using machine learning, Energy. 179 (2019) 214–221. https://doi.org/10.1016/j.energy.2019.04.077

Value of the data

- The data cover a large period of nearly four decades, which provides enough observations to train and test machine learning algorithms.
- Different methods can be applied to the data and compared to the ones presented here.
- The data can be used to guide policy makers, investors, companies, and others involved in the international energy market.

1. Data

The data includes monthly prices, reported in nominal U.S. dollars, period average and not seasonally adjusted of six energy commodities that were chosen according to their importance for the international energy market, i.e. Oil Brent, Oil WTI, Oil Dubai Fateh, Coal AU, Gas US, and Gas Russia. In 2017 the global primary energy sources were: oil (32%), coal (27%) and natural gas (22%) [2].

The description and summary statistics of each commodity is presented in Table 1. In addition, the data contains the log-return of each time series. Fig. 1 shows the price behavior and reveals the non-seasonality of the data in all six cases. We divided the data into two segments, the first 80% of the data for training and the remaining 20% for test as suggested by Ref. [3].

Table 1
Description and summary statistics.

Time series	Description	Period	Min.	Max.	Mean	Std. Dev.
Oil Brent	Crude Oil (petroleum). Dated Brent, light blend 38 API, fob U.K., US$/barrel.	Jan/1980–Jun/2017	9.56	133.90	41.946	30.927
Oil WTI	Crude Oil (petroleum). West Texas Intermediate 40 API, Midland Texas, US$/barrel.	Jan/1980–Jun/2017	11.31	133.93	41.309	27.720
Oil Dubai	Crude Oil (petroleum). Dubai medium Fateh 32 API, US$/barrel.	Jan/1980–Jun/2017	8.50	131.22	39.737	30.246
Coal AU	Australian thermal coal. 12,000- BTU/pound, FOB Newcastle/Port Kembla, US$/metric ton.	Jan/1980–Jun/2017	24.00	195.19	52.475	29.603
Gas US	Natural Gas spot price at the Henry Hub terminal in Louisiana, US$/Million Metric BTU.	Jan/1991–Jun/2017	1.14	13.63	3.875	2.260
Gas Russia	Russian Natural Gas border price in Germany, US$/Million Metric BTU.	Jan/1985–Jun/2017	1.44	16.02	5.097	3.510

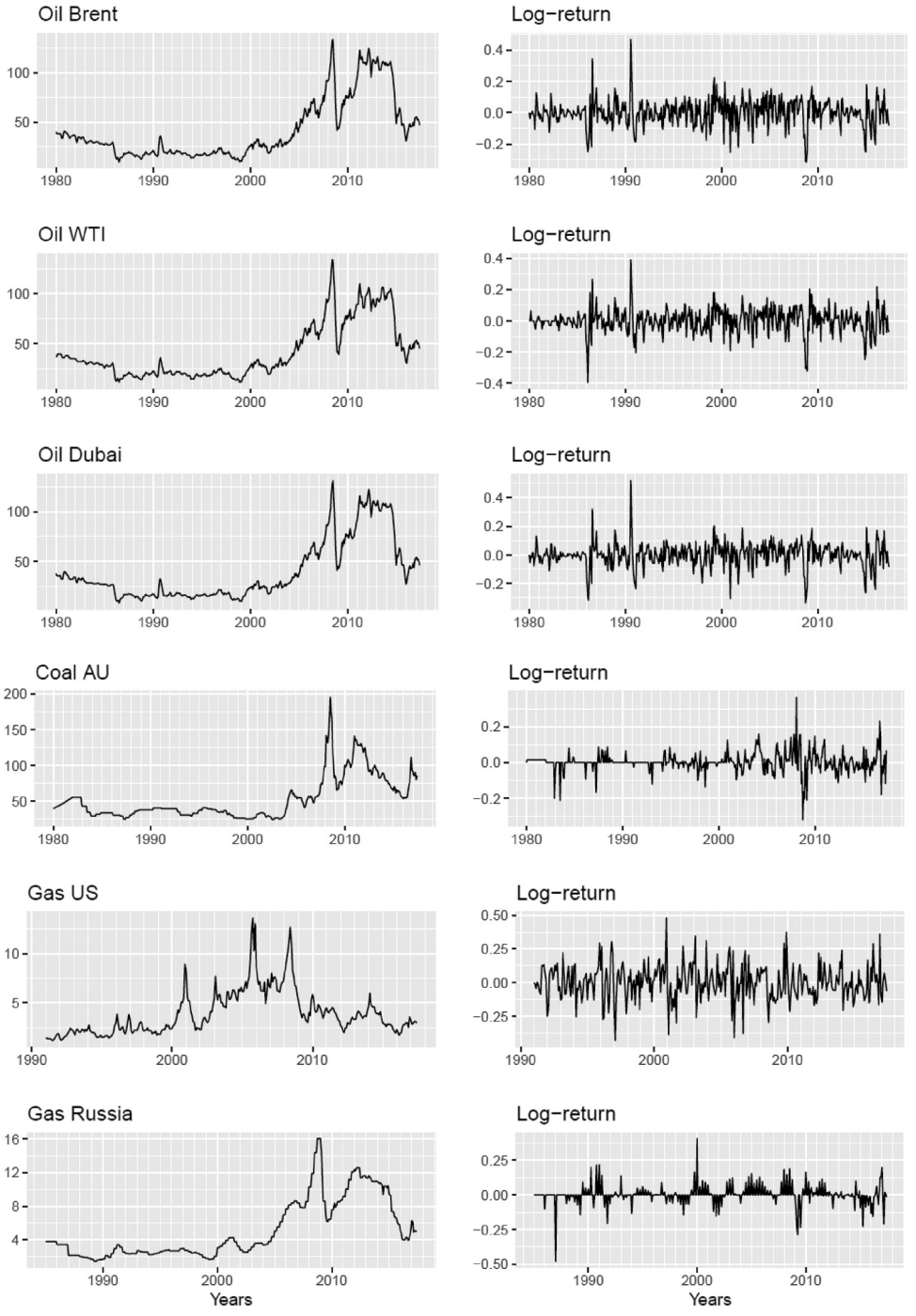

Fig. 1. Historical prices behavior.

2. Experimental design, materials, and methods

We performed and evaluated four different methods, a hybridization of traditional econometric models, artificial neural networks, random forests, and the no-change method, as described in [1]. The last one implies that changes in an observation value are unpredictable, so the best forecast is simply the current observation value. It is a convention to compare the performance of models with the no-change method as it is considered a natural benchmark [4].

The data provided within this article is the original obtained in the data source, except for the log-returns of each commodity price. However, our analyses rely on an original database since the raw information needs to go through a few steps before the actual application of the machine learning techniques. For the artificial neural network (ANN) we applied first differentiation, generated lagged values from one to twelve and then calculated the model. For the random forests (RF) we created lagged values from one to 72, which work as predictors variables. These two transformations can be easily applied using any statistical software.

The hybrid model combines autoregressive integrated moving average (ARIMA); error, trend, and seasonality (ets); seasonal and trend decomposition using Loess (stl); exponential smoothing state space model with Box-Cox transformation, ARMA errors, trend and seasonal components (tbats) and Theta model. We tested the method performance using equal weights to each individual model as well as optimal weights determined by an algorithm that uses non-rolling time series cross-validation to minimize the root mean square error (RMSE) and set the weight coefficients.

The ANN applied was a feedforward multi-layer perceptron (MLP). An iterative neural filter (INF) was used to capture the number and the period of the seasonalities that are present in the data and determine the input vector. As stated by Ref. [5], there is no methodology universally accepted to guide the architecture specification of MLPs. Therefore, different combinations of numbers of hidden layers and nodes were tested to set the optimal structure.

The random forests method operates by applying three steps: sample fractions of the data, grow a randomized tree predictor on each small piece and then aggregate these predictors by averaging. Different combinations of number of trees, lagged variables and number of variables randomly sampled for splitting at each tree node were tested to set the best architecture.

The methods were evaluated using two statistical loss functions, i.e. the mean absolute percentage error (MAPE) and the root mean square error (RMSE). Additionally, to statistically test the significant difference regarding the performance amongst the methods we used the modified Diebold-Mariano (M-DM) test as proposed by Ref. [6].

Acknowledgments

Gabriel P. Herrera would like to thank Griffith University for the Postgraduate Research Scholarship. Benjamin M. Tabak and Hemerson Pistori gratefully acknowledge financial support from CNPq Foundation. Hemerson Pistori also acknowledges financial support from FUNDECT Foundation. The authors would like to thank NVIDIA Corporation for the donation of resources utilized in this research.

Conflict of 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2019.104122.

References

[1] G.P. Herrera, M. Constantino, B.M. Tabak, H. Pistori, J. Su, A. Naranpanawa, Long-term forecast of energy commodities price using machine learning, Energy 179 (2019) 214–221. https://doi.org/10.1016/j.energy.2019.04.077.
[2] E.I.A., International Energy Outlook 2018 Technical Report, US Department of Energy, Washington, DC, 2018.
[3] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, OTexts, 2018.
[4] R. Alquist, L. Kilian, R.J. Vigfusson, Chapter 8 - forecasting the price of oil, Handb. Econ. Forecast. 2 (2013) 427–507. https://doi.org/10.1016/B978-0-444-53683-9.00008-6.
[5] S.F. Crone, N. Kourentzes, Feature selection for time series prediction - a combined filter and wrapper approach for neural networks, Neurocomputing 73 (2010) 1923–1936. https://doi.org/10.1016/j.neucom.2010.01.017.
[6] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, Int. J. Forecast. 13 (1997) 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4.