DOI: 10.20472/EFC.2023.019.009

KRZYSZTOF DRACHAL

Faculty of Economic Sciences, University of Warsaw, Poland

CHOOSING PARAMETERS FOR BAYESIAN SYMBOLIC REGRESSION: AN APPLICATION TO MODELLING COMMODITIES PRICES

Abstract:

This study examines the application of Bayesian Symbolic Regression (BSR) for in-sample modelling of various commodities spot prices. The studied method is a novel one, and it shows promising potential as a forecasting tool. Additionally, BSR offers capabilities for handling variable selection (feature selection) challenges in econometric modeling. The focus of the presented research is to analyze the suitable selection of the initial parameters for BSR in the context of modelling commodities spot prices. Generally, it is a challenge for (conventional) symbolic regression to properly specify the set of operators (functions). Here, the analysis is primarily focused on specific time-series, making the presented considerations especially tailored to time-series representing commodities markets. The analysis is done with an aim to assess the ability of BSR to fit the observed data effectively. The out-of-sample forecasting performance analysis is deferred for investigations elsewhere. Herein, the main objective is to analyze how the selection of initial parameters impacts the accuracy of the BSR model. Indeed, the already known simulations were based on synthetic data. Therefore, herein real-word data from commodities markets are used. The outcomes can be useful for researchers and practitioners further interested in econometric and financial applications of BSR. (Research funded by the grant of the National Science Centre, Poland, under the contract number DEC-2018/31/B/HS4/02021.)

Keywords:

Bayesian symbolic regression, Commodities, Genetic algorithms, Modelling, Symbolic regression, Time-series

JEL Classification: C32, C53, Q02

1 Introduction

Forecasting commodities spot prices is a complex task, addressing two main challenges. Firstly, over the years, researchers have identified numerous potential drivers of commodities prices, including supply and demand factors, exchange rates, financial market interactions, speculative pressures, various uncertainty indices, etc. Consequently, selecting the most relevant variables (feature selection) when constructing an econometric model, such as multilinear regression, has become a challenging task (Chen et al., 2010; Gargano and Timmermann, 2014; Salisu et al., 2019; Steermer, 2018).

The second problem is that the impact of specific drivers on a commodity price can vary across different time periods, leading to a time-varying nature of the model's parameters and structure (Huang et al., 2021). Moreover, the functional form specification of the model needs to account for non-linear effects and the complicated market structure (Caginalp and DeSantis, 2011).

Symbolic regression offers a solution to these challenges. It begins with creating a set of operators (functions) and then employs evolutionary processes like crossover, mutation, and selection to discover a suitable functional form for the model (Koza, 1998). Symbolic regression is a form of regression analysis that aims to discover an explicit mathematical expression or equation that best fits a given dataset. Unlike traditional regression methods that rely on pre-defined functional forms (e.g., linear, quadratic), symbolic regression is more flexible and allows the model to find the most suitable mathematical structure and coefficients directly from the data. In symbolic regression, the algorithm searches through a space of mathematical expressions using techniques inspired by evolutionary algorithms, such as genetic programming. It starts with a population of randomly generated mathematical expressions, represented as trees of mathematical operators and variables. These expressions are then evaluated based on how well they fit the training data using a specified fitness function. The evolutionary process involves iteratively selecting the best-performing expressions from the population, applying genetic operations like crossover and mutation to create new variations, and repeating the evaluation and selection steps until a stopping criterion is met. Over successive generations, the algorithm tends to converge towards a mathematical expression that provides the best fit to the data while also being interpretable (Koza, 1998).

Conventionally, symbolic regression has used genetic algorithms for this purpose. However, a more recent approach, i.e, Bayesian Symbolic Regression (BSR), was proposed by Jin et al. (2019). It replaces genetic algorithms with Bayesian symbolic trees. This method replaces evolutionary processes with Bayesian prior-posterior inference and is claimed to result in better predictions and to be computationally more efficient.

Herein, this novel method is employed to in-sample fit spot prices of 56 commodities and various specifications of this method are studied.

2 Methodology and Data

Bayesian symbolic regression (BSR) proposed by Jin et al. (2019) represents a novel approach to symbolic regression designed to address challenges faced by traditional methods, such as difficulties in integrating prior knowledge into genetic programming, complexities arising from outcome expression, and reduced interpretability. These issues have been well-documented in existing approaches to symbolic regression (Korns, 2011). BSR offers a promising solution to these problems, aiming to improve the efficiency and interpretability of symbolic regression models.

Herein, the in-sample analysis is performed. The reason for that is primarily to focus on analysing how the specification of certain parameters can impact the data fitting. Indeed, the researcher must specify certain initial parameters to BSR, but even in case of the genetic algorithm based symbolic regression this is not an easy task (Nicolau and Agapitos, 2021). As a result, this study focuses on comparing different BSR models applied to 56 different commodities spot prices. This makes it particularly tailored to specific time-series. As a result, further studies can use these real-market simulations outcomes for further elaborations. The evaluation of out-of-sample forecasting with BSR is reserved for other studies. For example, an application to forecasting crude oil spot price can be found in the paper by Drachal (2023).

Another crucial aspect of BSR is its emphasis on improving the interpretability of the derived expressions. To achieve this goal, the method aims to capture concise yet informative signals with a linear and additive structure. Prior distributions are employed to control the complexity of the symbolic trees, a representation commonly used in similar problems (Weiss, 2014). The core of BSR lies in Markov Chain Monte Carlo (MCMC) sampling, which generates symbolic trees from the posterior distribution. Although computationally intensive, Jin et al. (2019) demonstrated that this method can even enhance memory usage in computer computations compared to standard genetic programming approaches for symbolic regression. However, simulations conducted by Jin et al. (2019) were based on synthetic data, therefore this paper tries to fill the literature gap and focuses on very specific real-market, i.e, commodities one. The obtained outcomes can be helpful for researchers willing to apply BSR further in forecasting and analysing commodities prices time-series, as this time-series possess quite specific features (Kent Baker et al., 2018).

A quick sketch of BSR is presented below. Details can be found in the original paper by Jin et al. (2019). Let y_t be the forecasted commodity spot price (possibly transformed, for example, into logarithmic differences). In particular, Brent, Dubai and WTI crude oil, Australian and South African coal, U.S. and European natural gas and Japan liquefied natural gas, cocoa, Arabica and Robusta coffee, Colombo, Kolkata and Mombasa tea, coconut oil, groundnuts, fish meal, palm oil, soybeans, soybean oil, soybean meal, maize, Thai 5% broken rice, U.S. soft red winter and hard red winter wheat, U.S. bananas, orange, beef, chicken meat, Mexican shrimps, European, U.S. and world sugar, U.S. import tobacco, Cameroon and Malaysian logs, Malaysian sawnwood, plywood, cotton (A index), Singapore traded rubber, phosphate rock, diammonium phosphate, triple superphosphate, urea, potassium chloride, aluminium, iron ore, copper, lead, tin, nickel, zinc, gold, platinum and silver spot prices were taken (The World Bank, 2022).

Let $x_{1,t}$, …, $x_{n,t}$ be the explanatory variables. Following, Gargano and Timmermann (2014) and Drachal (2018) dividend to price ratio (Schiller, 2022), U.S. 3-month treasury bills secondary market rate representing short-term rate and U.S. long-term government 10-year bond yields representing long-term rate, term spread (i.e., the difference between the long-term rate of U.S. bonds and U.S. treasury bill rate), default return spread (i.e., the difference between U.S. long-term corporate bonds yield and U.S. treasury bill rate, where long-term corporate bond yield was taken as the index based on bonds with maturities 20-years and above), U.S. Consumer Price Index (transformed into logarithmic differences), U.S. industrial production (transformed into logarithmic differences), U.S. M1 money stock (transformed into logarithmic differences), Kilian global economic activity index, U.S. unemployment rate, Australian dollar to U.S. dollar exchange rate (transformed into logarithmic differences), Indian rupee to U.S. dollar exchange rate (transformed into logarithmic differences), S&P GSCI Commodity Total Return Index (transformed into logarithmic differences) and U.S. dollar open interest (transformed into logarithmic differences) were taken (Bloomberg, 2022; Commodity Futures Trading Commission, 2022; FRED, 2022).

Then, the following equation is considered $y_t = \beta_0 + \beta_1 * f_1(x_{1,1,t-1}, \ldots, x_{1,i,t-1}) + \ldots + \beta_K * f_k(x_{K,1,t-1}, \ldots, x_{K,i,t-1})$, with $x_{i,j,t}$ standing for some of the explanatory variables out of some n available ones (herein, n = 14) which are present in the i-th component expression, i.e., f_i, with j = { 1, …, n } and i = { 1, …, K }. The number of components, K, is fixed and must be set up at the initial stage, whereas coefficients β_i are estimated by the Ordinary Least Squares linear regression method. Jin et al. (2019) claimed that higher values of K result in higher forecast accuracy, but that this increment diminishes with the further growth of the parameter K. Herein, K = {1, 2, …, 10 } were tested.

Component expressions f_i are represented by symbolic trees (Weiss, 2014) constructed from some set of operators, such as, for example +, *, 1 / x, etc. Nicolau and Agapitos (2018) and Keijzer (2004) noticed that even for the genetic algorithm based symbolic regression the optimal selection of this set is not an easy task. Therefore, herein, 6 different sets were studied. F = 1 represents the set consisting of unary $neg(x_{i,t}) = - x_{i,t}$ and binary $add(x_{i,t},x_{j,t}) = x_{i,t} + x_{j,t}$ operators. F = 2 represents as for F = 1, but expanded by unary $square(x_{i,t}) = (x_{i,t})^2$. F = 3 represents as for F = 1, but expanded by unary 12-periods back moving average, i.e., $ma12(x_{i,t}) = (x_{i,t} + x_{i,t-11}) / 12$, and unary $lag(x_{i,t}) = x_{i,t-1}$. F = 4 represents as for F = 2, but expanded by binary $mul(x_{i,t},x_{j,t}) = x_{i,t} * x_{j,t}$. F = 5 represents as for F = 4, but expended by unary $inv(x_{i,t}) = 1 / x_{i,t}$, unary $cubic(x_{i,t}) = (x_{i,t})^3$, unary $sqrt(x_{i,t}) = \sqrt{}x_{i,t}$, unary $log(x_{i,t}) = \ln(|x_{i,t}|)$, unary ma12 and unary lag. F = 6 represents as for F = 1, but expanded by the unary operator $lt(x_{i,t}) = a * x_{i,t} + b$, with a and b being some real numbers. Following Nicolau and Agapitos (2018) and Keijzer (2004) this operator can improve the set of obtained expressions. The selection of various F was based on potential usefulness for modelling economic time-series and to have some economic motivation. Also the aim was to consider small and big set of operators.

The Bayesian inference takes over the symbolic trees representing the expressions. A symbolic tree is represented by $g(\cdot ; T, M, \Theta)$ with g representing a function as above, i.e., $g = f_1 + \ldots + f_k$ and T – the set of nodes, M – nodes' features, and Θ – parameters. Uniform priors are taken at the initial stage. A node feature represents whether the given node is a terminal one, or extends to one child node, or splits into two child nodes. The probabilities of these transformations were taken as in the paper by Jin et al. (2019). Priors for a and b for operators lt were Gaussian and centred around the identity function, also following Jin et al. (2019). The prior-posterior inferences of the entire model were performed with Metropolis-Hastings algorithm (Jin et al., 2019). 50 iterations were performed as advised by Jin et al. (2019).

432 observations were used, as the data set spans from January 1986 to January 2022. (Monthly frequency was used.) Additionally to the previously stated transformations, time-series were further standardized before inserting into the BSR. Explanatory variables were also lagged 1 period back. Computations were done in Python and R (Jin, 2021; R Core Team, 2018; Van Rossum and Drake, 1995).

Evaluation was done with respect to the raw time-series of commodities prices, i.e., BSRs forecasts were scaled back from standardization and logarithmic transformation before the evaluation. In particular, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean

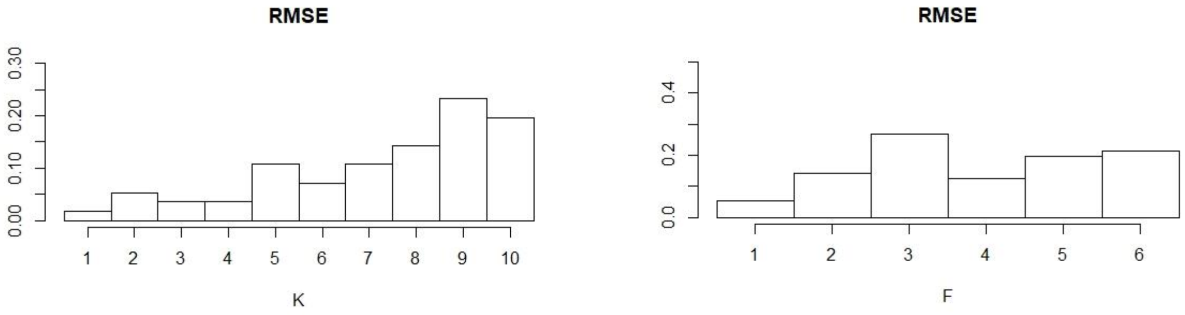
Absolute Scaled Error (MASE) were considered (Hyndman and Koehler, 2006). Model Confidence Set (MCS) was also employed (Hansen et al., 2011).

3 Results

For each commodity, the BSR model minimizing RMSE was chosen. Figure 1 presents frequencies of parameters K and F of the models selected in such a way. It can be seen that, findings by Jin et al. (2019) can be confirmed. In other words, errors tend to smaller with higher values of K. However, this cannot be taken as a general rule. Secondly, the simplest set of operators is rarely selected. Sets with more operators, especially the one consisting of neg, add, ma12 and lag operators are preferred. These are typical operators used in financial time-series analyses. But also the sets connected with some scaling transformations and possible non-linear effects are often selected. Despite the fact that the data were already transformed before inserting into the BSR models, the specific features of commodities markets time-series are still detected.

Moreover, there is no much discrepancy between the outcomes based on the analysis of RMSE, MAE and MASE. Therefore, to keep the paper concise, mostly the outcomes based on RMSE are reported. Indeed, in case of 50% of commodities the selected values of K and F were exactly the same for RMSE, MAE and MASE.

Figure 1 Frequencies of selected BSR parameters for models minimizing RMSE

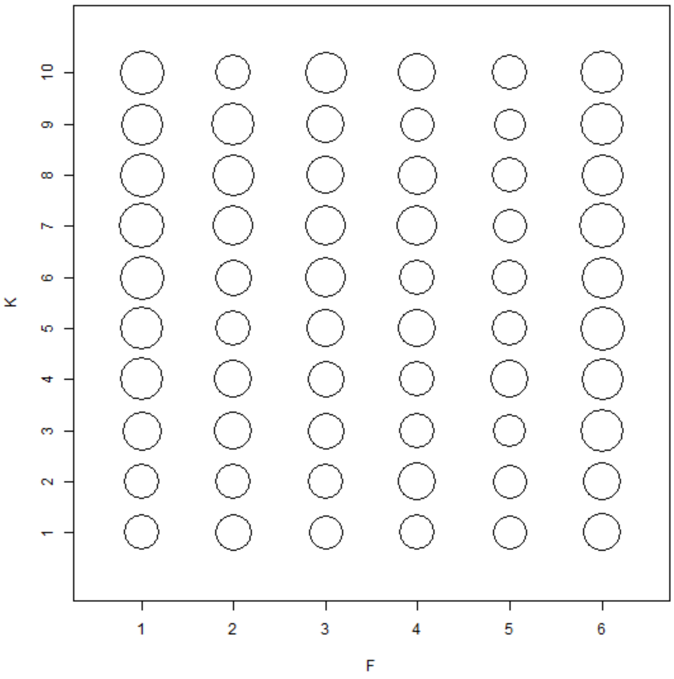


Source Own estimation

Figure 2 presents inverses of MASEs averaged over all analyzed commodities for fixed K and fixed F. (Therefore, bigger circle is preferred.) Unfortunately, there is no clear pattern visible. However, K = 7 and F = 6 results with smallest MASE on average.

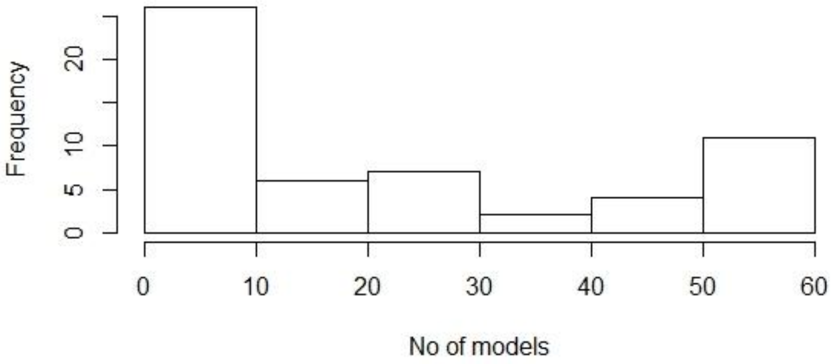
Next, for every commodity, MCS with 90% confidence level, 1000 bootstrapped samples and “TR” statistic (Hansen et al., 2011) was performed for forecasts for all values of K and F (i.e., for 10 * 6 = 60 models). First, it was checked how many models survived MCS procedure for each commodity. From Figure 3 it can be seen that in many cases relatively large number of models remained. However, also for a reasonable number of models, MCS procedure was quite harsh and excluded some models. This means that, indeed, in certain cases the proper selection of K and F is significant, but in some cases it actually does not matter much which values of parameters are taken, as the differences in the obtained forecasts are statistically not significant. Also none of the specification is clearly superior over others, neither any one can be definitely excluded. The least frequently remaining model still remained for 16 commodities, and the one most often surviving remained for 33 commodities. Indeed, Figure 4 presents how often models with a given K (or a given F) survived MCS procedure. The pictures are rather flat showing no clear tendency towards any preferred value of K or F, being rather uniformly distributed. Figure 5 presents the overall picture.

Figure 2 Inverses of MASEs averaged over analyzed commodities



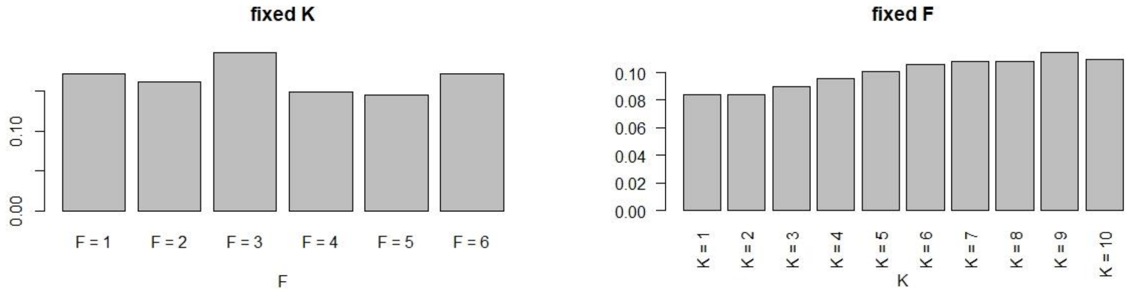
Source Own estimation

Figure 3 Number of commodities for which a given number of BSR models with different K and F parameters survived MCS procedure



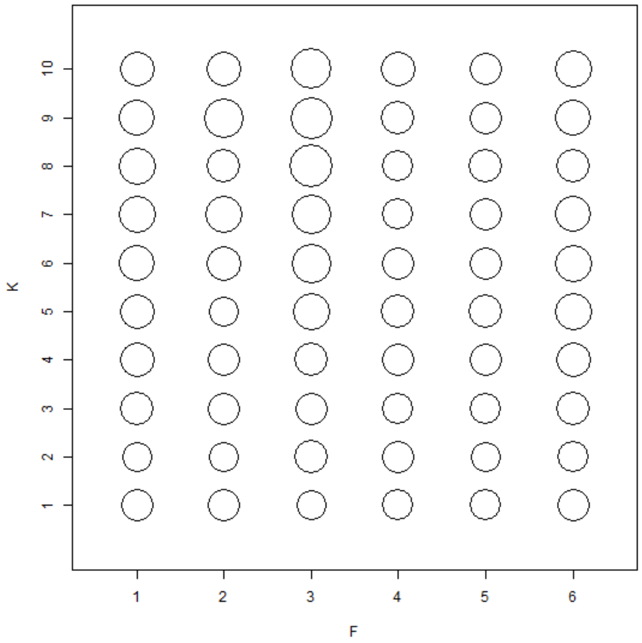
Source Own estimation

Figure 4 Frequencies of survival MCS procedure by various BSR models with different K and F parameters (one parameter is fixed)



Source Own estimation

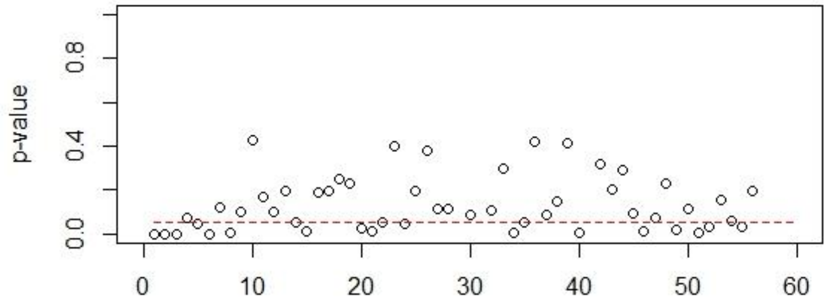
Figure 5 Frequencies of survival MCS procedure by various BSR models with different K and F parameters (bigger circles are preferred)



Source Own estimation

Finally, seeing from Figure 4 that the BSR specification with K = 9 and F = 3 is most commonly left by MCS procedure, it can be asked if forecasts from BSR with this specification, indeed, significantly differ from those from the models with the specification which minimized RMSE for a given commodity. This was checked with Diebold-Mariano test (Diebold and Mariano, 1995). The null hypothesis of the test is that forecasts from both models have the same accuracy. The alternative – that the one from the model minimizing RMSE has greater accuracy. At 5% significance level, the null hypothesis was rejected only for 18 commodities. Figure 6 presents p-values for all commodities.

Figure 6 P-values from Diebold-Mariano test (dotted line corresponds to 5% significance level)



Source Own estimation

Table 1 P-values from Diebold-Mariano test (K fixed)

	F = 2	F = 3	F = 4	F = 5	F = 6
K = 1	0.5515	0.5178	0.5617	0.5501	0.4544
K = 2	0.6422	0.4991	0.5482	0.6229	0.4869
K = 3	0.5382	0.6192	0.6340	0.7056	0.5326
K = 4	0.5693	0.5425	0.6818	0.6513	0.4860
K = 5	0.6568	0.5262	0.7066	0.6840	0.4768
K = 6	0.6566	0.5501	0.7179	0.6818	0.5139
K = 7	0.6307	0.4884	0.6818	0.6950	0.5233
K = 8	0.6234	0.4966	0.6915	0.6836	0.5492
K = 9	0.5620	0.4895	0.6539	0.6979	0.4455
K = 10	0.6894	0.5060	0.6512	0.7189	0.4922

Source Own estimation

Table 1 presents p-values from Diebold-Mariano test averaged over commodities, performed in the following way. For every commodity and for every parameter K, for every F = { 2, …, 6 } the forecast from a BSR model with the given K and F was tested against the one from the model with the same K, but with F = 1 (i.e., from the model with the simplest set of operators). The alternative hypothesis was that the forecast from the model with F = 1 is less accurate. High p-values suggest that there is little evidence to treat the simplest set of operators as leading to statistically significantly less accurate forecasts under the same specification of the parameter K. In other words, if the parameter K is already set, then choosing a set of operators has small impact on the accuracy of the obtained forecast. The more deep analysis showed that the null hypothesis was rejected only in approximately 7% of individual cases, if 5% significance level was assumed; and in approximately 12% cases, if 10% was assumed.

Finally, for robustness check, it can be interesting to consider if, instead of 432 observations, as it was taken, the repetition of the above analysis with, for example, the only first 100 observations, would lead to the similar conclusions. In short, in such a case RMSE was not minimized by the models with exactly the same K and F parameters for every commodity. Although, for a reasonable number of them, it was. However, in case of the parameter K, for the majority of commodities this parameter in both cases took quite close values. Rarely, the model based on the first 100 observations from the sample taken herein, and the model based on all 432 observations, minimizing RMSE, would have very different than each other parameter K (i.e., differing by more or less than 1 or 2 units).

4 Conclusions

When applying the novel Bayesian Symbolic Regression (BSR) to modelling or forecasting commodities spot prices, there is no one, particular set of initial parameters and operators set that can be roughly suggested to employ. However, in general, the higher value of the admissible number of components in linear regression expression can lower errors of the model. Secondly, inclusion of moving average and lagging operators seem to be slightly more useful, than some operators dealing with non-linear effects. Surprisingly, for quite many commodities, under the fixed specification of the number of linear regression components, even taking only some very simple operators does not lead to statistically significantly less accurate forecasts, than if more rich set of operators would be taken.

Depending on computational resources, a researcher should either consider quite a rich set of operators (e.g., dealing with moving averages, lagging, non-linearity, etc.), or narrow to just some small simple set of operators. However, in case of the number of admissible linear regression components, this number should not be taken too small. For practical applications, it would be a good advice to perform some initial training period simulations on some first observations from the analyzed sample, to select these parameters for estimations and forecasting over the whole sample. The issues with suitable selection of the set of operators seem to be more challenging than the selection of the number of admissible linear regression components.

5 Acknowledgement

Research funded by the grant of the National Science Centre, Poland, under the contract number DEC-2018/31/B/HS4/02021.

6 References

BLOOMBERG (2022). S&P GSCI Total Return CME. Online. Available from: https://www.bloomberg.com

CAGINALP, G. and DESANTIS, M. (2011). Nonlinearity in The Dynamics of Financial Markets. Nonlinear Analysis: Real World Applications, Vol. 12, pp. 1140-1151.

CHEN, Y. C.; ROGOFF, K. S. and ROSSI, B. (2010). Can Exchange Rates Forecast Commodity Prices?. The Quarterly Journal of Economics. Vol. 125, pp. 1145-1194.

COMMODITY FUTURES TRADING COMMISSION (2022). Historical Compressed. Online. Available from: https://www.cftc.gov/MarketReports/CommitmentsofTraders/HistoricalCompressed/index.htm

DIEBOLD, F. X. and MARIANO, R. S. (1995). Comparing Predictive Accuracy. Journal of Business and Economic Statistics. Vol. 13, pp. 253-263.

DRACHAL, K. (2018). Some Novel Bayesian Model Combination Schemes: An Application to Commodities Prices. Sustainability, Vol. 10, pp. 2801. https://doi.org/10.3390/su10082801

DRACHAL K. (2023). Forecasting The Crude Oil Spot Price with Bayesian Symbolic Regression. Energies, Vol. 16, pp. 4. https://doi.org/10.3390/en16010004

FRED (2022). Economic Data. Online. Available from: https://fred.stlouisfed.org

GARGANO, A. and TIMMERMANN, A. (2014). Forecasting Commodity Price Indexes Using Macroeconomic and Financial Predictors. International Journal of Forecasting, Vol. 30, pp. 825-843.

HANSEN, P. R.; LUNDE, A. and NASON, J. M. (2011). The Model Confidence Set. Econometrica, Vol. 79, pp. 453-497.

HUANG, J.; LI, Y.; ZHANG, H. and CHEN, J. (2021). The Effects of Uncertainty Measures on Commodity Prices From A Time-varying Perspective. International Review of Economics and Finance, Vol. 71, pp. 100-114.

HYNDMAN, R. J. and KOEHLER, A. B. (2006). Another Look at Measures of Forecast Accuracy. International Journal of Forecasting, Vol. 22, pp. 679-688.

JIN, Y. (2021). A Bayesian MCMC Based Symbolic Regression Algorithm. Online. Available from: https://github.com/ying531/MCMC-SymReg

JIN, Y.; FU, W.; KANG, J.; GUO, J. and GUO J. (2019). Bayesian Symbolic Regression. Online. Available from: https://arxiv.org/abs/1910.08892

KEIJZER, M. (2004). Scaled Symbolic Regression. Genetic Programming and Evolvable Machines, Vol. 5, pp. 259-269.

(Eds.) KENT BAKER, H.; FILBECK, G. and HARRIS, J. H. (2018). Commodities: Markets, Performance, and Strategies. Oxford: Oxford University Press.

KORNS M. F. (2011). Accuracy in Symbolic Regression. In: (Eds.) Riolo, R.; Vladislavleva, E. and Moore, J.; Genetic Programming Theory and Practice IX. New York, NY: Springer.

KOZA, J. (1998). Genetic Programming. Cambridge, MA: MIT Press.

NICOLAU, M. and AGAPITOS, A. (2021). Choosing Function Sets with Better Generalisation Performance for Symbolic Regression Models. Genetic Programming and Evolvable Machines, Vol. 22, pp. 73-100.

R CORE TEAM (2018). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Online. Available from: https://www.R-project.org

SALISU, A. A.; ISAH, K. O. and RAHEEM, I. D. (2019). Testing The Predictability of Commodity Prices in Stock Returns of G7 Countries: Evidence from A New Approach. Resources Policy, Vol. 64, pp. 101520. https://doi.org/10.1016/j.resourpol.2019.101520

SCHILLER, R. (2022) Online Data. Online. Available from: http://www.econ.yale.edu/~shiller/data.htm

STEERMER, M. (2018). 150 Years of Boom and Bust: What Drives Mineral Commodity Prices?. Macroeconomic Dynamics, Vol. 22, pp. 702-717.

THE WORLD BANK (2022). Commodities Markets. Online. Available from: https://www.worldbank.org/en/research/commodity-markets

VAN ROSSUM, G. and DRAKE, JR F. L. (1995). Python Reference Manual. Amsterdam: Centrum voor Wiskunde en Informatica.

WEISS, M. A. (2014). Data Structures and Algorithm Analysis in C++. Upper Saddle River, NJ: Pearson Education, Inc.