RANDOM FOREST AS A MODEL FOR CZECH FORECASTING*

Katerina Gawthorpe[a]

Abstract
Random forest models have recently gained popularity for economic forecasting. Earlier studies demonstrated their potential to provide early warnings of recession and serve as a competitive method to older prediction models. This study offers the first evaluation of the random forest forecast for the Czech economy. The one-step-ahead forecasting results show high accuracy on the Czech data and are proven to outperform forecasts from the Czech Ministry of Finance and the Czech National Bank. The following multi-step random forest forecast, estimated for the next four quarters, shows results similar to those from the central institutions. The main difference stems from the household and industrial confidence variables, which significantly impact on the random forest forecast. The variable-importance analysis further emphasizes the soft variables as valuable determinants for Czech forecasting. Overall, the findings motivate other forecasters to exercise this method.

Keywords: Random forest, Czech Republic, forecast, regression tree
JEL Classification: C63, E37, C32

1. Introduction

Random forest models represent a powerful novel method of forecasting. Despite its popularity, only a few articles apply it to forecasting, and this is the first study to evaluate its prediction potential for the Czech economy. The model assessment depends on accuracy measures from over a hundred prediction estimates for eleven target variables and further comparison to the quarterly projections of the two leading Czech forecasting institutions, the Ministry of Finance (MF) and the Czech National Bank (CNB).

The MF and the CNB econometric and agent-based forecasting models differ from the random forest approach. While these older methods rely on economic theory,

* This work was supported by the Internal Grant Agency of the Faculty of Business Administration, University of Economics, Prague under grant No. IP300040.

a Prague University of Economy and Business, Prague, Czech Republic
Email: xzimk04@gmail.com

the random forest is a crushing data technique. The absence of a theoretical background avoids formulating equations necessary for the macroeconomic agent-based models (see the Ministry of Finance model from Aliyev *et al.*, 2014, extended by Gawthorpe, 2020, and the CNB g3 model, www.cnb.cz).

The model assumptions are only as permanent as the economic theory that defines them. However, economic theories tend to change. For example, we now doubt the relevance of concepts vastly accepted in the 1990s, such as long-run purchasing power parity or the Phillips curve (Engel, 2000; Atkeson and Ohanian, 2001; Blanchard, 2016), and struggle to model new phenomena, such as the zero bound on interest rates (Wolman, 2003; Kucharčuková *et al.*, 2013). Agent-based models require reformulation to adopt these changes. Random forest, in contrast, learns these changes from data and advances with new observations. Learning from data appears very attractive in the presence of such a complex and continuously changing economic environment.

Although unknown in economics until recently (Boulesteix *et al*., 2012; Biau and D'elia, 2010), this flexible and powerful method is proven to outperform classification methods from 17 families, such as Bayesian models, generalized linear models, decision trees or principal component regression (Fernández-Delgado *et al*., 2014), but also logistic regression, Gaussian discriminant analysis, quadratic discriminant analysis and support vector machines in time-series forecasting (Khaidem *et al*., 2016), the ANN and ARMA approaches in forecasting real-time prices on the NY electricity market (Mei *et al*., 2014), neural networks and support vector machines in forecasting Malaysian exchange rate (Ramakrishnan *et al*., 2017), econometric methods in forecasting primary energy commodities and anticipating their turning points (Herrera *et al*., 2019) and neural networks, discriminant analysis and logit models in forecasting stock index movements (Kumar and Thenmozhi, 2006). Baybuza (2018) finds the random forest method to be a useful forecasting tool for Russian inflation as autoregression. Biau and D'elia (2010) prove random forest as a valuable technique in forecasting with large datasets. Besides dealing with large datasets, Woloszko (2020) finds it useful for capturing nonlinearities in the complex and changing economic reality. At the same time, Woloszko (2020) criticizes the unrealistic assumptions of econometric models concerning stable relationships and stable data distribution across history. Random forests also show an exciting potential to serve for early recession warnings (Alessi and Detken, 2011); Nyman and Ormerod (2017) capture the economic downturn in 2009, and Coulombe (2020) the unemployment growth in 2008.

However, despite the evidence on the forecasting accuracy of the method, the random forest suffers from several drawbacks. The more complex the machine-learning algorithm is, the better it captures complex economic reality, but this growing accuracy comes at the price of lower interpretability, the Occam dilemma (Woloszko, 2020, p. 7).

The random forest calculation process is also slower and harder to interpret than a single regression tree forecast (Baybuza, 2018). The interpretation is facilitated with the so-called variable importance analysis, which measures impact of variables on the prediction result (Li *et al*., 2019). The variable importance analysis also complements the outcome estimation in this study.

Other issues relate to time series forecasting. Woloszko (2020) discusses difficulties in time series forecasting with machine-learning models, such as the process of classification of data into sub-samples, the if-then concept and abstracts from simultaneous interactions among forecast variables. In contrast, agent-based macroeconomic models understand interdependent economic relationships (Gawthorpe, 2019). This study introduces mutual dependence among variables with sequential forecasting, where a prediction estimate for one target variable becomes a feature value for forecasting another target variable. Machine-learning models also have no awareness of time and struggle to predict trends (Pavlyshenko, 2019). The present paper confronts this problem in the pre-processing data stage.

This study builds on previous research and evaluates a random forest model for forecasting Czech macroeconomic variables. The model features eleven variables collected from the third quarter of 2003 to the third quarter of 2019. The forecasting process consists of two steps. In the first step, the random forest estimates one-step-ahead forecasts for every variable. In the second step, multivariate multi-stage forecasts for these variables provide new prediction values for the entire year 2020 (not controlling for the recent pandemic). The final evaluation depends on comparing the model outcome and its accuracy to forecasts by the Czech central institutions, the Ministry of Finance, and the Czech National Bank. The variable importance approach closes the analysis with information about the variables most significant for forecasting the Czech GDP.

The structure of the present study is as follows. Methodology describes the random forest model and lays out the estimation process. The Data section summarizes observations for the training and the testing set. The Results section starts with a single regression tree presentation and continues with a random forest analysis and ends with an evaluation of the model accuracy. The Conclusions section summarizes the findings at the end of the paper.

2. Methodology

To be valuable for forecasting the Czech economy, the random forest must outperform forecasts from the two leading Czech forecasting institutions or bring new information for Czech forecasting. This study tests both the model accuracy and the significance of variables unusual in the Czech forecasting models.

The random forest, being an ensemble of regression trees, starts by estimating a regression tree on a randomly selected feature and data samples. These data samples are divided into training and testing sets, where the tree practices on the training samples to deliver estimates comparable to the testing samples (Breiman, 2001). The tree estimation process is a continuous splitting of a node into two child nodes and each child node into another two child nodes. A left-hand child node consists of samples that meet a feature argument (for example, x < 5), while a right-hand child node consists of those that violate it (Woloszko, 2020). The node splitting stops when the model reaches a maximum tree depth or meets a restriction on the number of training samples per leaf. The trained model then delivers a prediction estimate for a target variable. These prediction estimates from number of regression trees f_{RT} enter the final random forest f_{RF} prediction outcome $\tilde{y}$ (Wager and Athey, 2018):

$$f_{RF} = \frac{1}{N}\sum_{n=1}^{N} f_{RT}\left(\tilde{y}\right). \tag{1}$$

The algorithm in this study estimates a separate random forest model for every target variable. The tested variables include, besides hard data commonly used in forecasting models (see Gawthorpe, 2020; Aliyev *et al*., 2014; CNB g3 model, www.cnb.cz), also soft data as recommended by Woloszko (2020). The variables consist of gross domestic product, consumption, gross domestic product of the Eurozone, export, import, interest rate, household confidence, PMI manufacturing, construction confidence, industrial confidence, and wage. The model links these 11 variables K in time t = 1, 2, 3, …, T, T = 66 in the following way:

$$\begin{bmatrix} y_{1,t} \\ . \\ . \\ . \\ y_{K,t} \end{bmatrix} = f\left[y_{1,t-1}, \quad y_{2,t-1} \quad \cdots, \quad y_{K,t-1} \right], \tag{2}$$

where the model relates a target variable $y_{K,t}$ to the same set of lagged features and the previous state of the variable. We need to estimate forecasts only on the lagged values if we want to compare the forecasting accuracy to previous forecasts from the central institutions, which also had only past values available when making predictions. Nyman and Ormerod (2017) also use lagged features to predict GDP growth and note that only past values would have been available to a forecaster. The author's previous study (Gawthorpe, 2020) recommends applying only one lag while forecasting growth rates for the Czech macroeconomic variables.

This study evaluates the random forest method with one-step-ahead and multi-step forecasts, where error measures reflect the one-step-ahead forecast performance. Ten one-step-ahead forecasts for every target variable result in over 1600 prediction estimates. This rich dataset provides representative error measures for evaluating the model accuracy.

The multi-step prediction takes the form of a multivariate forecast; multivariate because the forecast includes the lagged character of both the target and the explanatory variables and multi-step because it predicts many steps into the future. While the random forest is a one-step-ahead forecasting method, the multi-step approach enables this model to predict values further into the future. The algorithm predicts the new values in a step-by-step manner. In every step, the code estimates a one-step-ahead forecast for every target variable and gradually expands the original data set with these new values. The prediction in the current step then depends on the values predicted in the past step.

This iterative forecasting procedure assimilates the method of Nyman and Ormerod (2017). In contrast to their article, the present prediction benefits from a multivariate forecast character where all explanatory variables are subject to multi-step forecasting. Unfortunately, the multi-step forecasting process suffers from error accumulation: errors from the past propagate into future predictions (Cheng *et al*., 2006). This issue means growing bias and variance of the model with increasing the prediction window. The present study, therefore, concentrates on forecasting only four quarters.

The forecasting outcome is further analysed using the variable importance method. The variable importance analysis measures importance of features for the prediction result, simplifying outcome interpretation, and excluding insignificant variables. Similar to principal component analysis, the method enables decreasing the number of regressors to reduce model overfitting.

The variable importance method operates with out-of-bag data, which is the fraction of data left out during the random sample selection process. The importance of including a feature in the model depends on the comparison between the mean squared error (MSE) for the model estimated on the out-of-bag data to the MSE for the model estimated on the original dataset (Nguyen *et al.*, 2015).

This study applies the mean decrease impurity method as the variable importance method. Impurity reflects how well observations fit a model and is measured as a residual sum of squares in a node. The mean decrease impurity approach selects those features that decrease the impurity most across all trees. However, this technique can display much smaller significance for less but still significant, although strongly correlated features (Li *et al*., 2019). This study verifies the mean decrease impurity outcome using separate simulations of a target variable as a function of only one other feature. In other words, the method tests the sensitivity of the random forest forecast to the inclusion of a particular variable.

3. Data

The dataset consists of eleven time-series variables collected quarterly from the third quarter of 2003 to the third quarter of 2019. The dataset consists of both soft and hard data. The OECD study of Woloszko (2020) recommends soft data for machine-learning forecasting of GDP. In this study, the soft data consist of household confidence (HHs. conf.), Purchasing Managers' Index for manufacturing (PMI), construction confidence (Const. conf.)and industrial confidence (Ind. conf.), and the hard data include Czech gross domestic product (GDP), consumption, gross domestic product of the Eurozone (GDP EA), export, import, 3-month PRIBOR (IR), and wage.

The Czech National Bank publishes the variables PMI and IR (www.cnb.cz); Eurostat the GDP EA (ec.europa.eu) and the Czech Statistical Office the remaining series (www.czso.cz). The GDP of the Eurozone and the interest rate enter the model as exogenous variables. The exogenous character of the Eurozone GDP originates in the limited impact of the Czech economy on the Eurozone. The exogenous character of the interest rate stems from the CNB's control over the interest rate. The model adopts the Eurozone GDP prediction from the Ministry of Finance (www.mfcr.cz) and the three-month PRIBOR from the Czech National Bank (www.cnb.cz).

Forecasting with random forests requires a data pre-processing step. The series enter the model as seasonally adjusted year-on-year (YoY) growth rates, as the Czech Statistical Office, the Czech National Bank, and the Eurostat supply data in a seasonally adjusted form. The transformation resembles the dataset used for forecasting dynamic stochastic general equilibrium (DSGE) models, a forecasting instrument used by the Ministry of Finance as well as the Czech National Bank (see the Ministry of Finance model, Aliyev *et al*., 2014; or see Pfeifer, 2020). Applying a dataset similar to that used for other models simplifies outcome comparison across different approaches. This data pre-processing is also convenient as machine-learning algorithms struggle to forecast trends (Pavlyshenko, 2019). The growth rate transformation can be formalized as $\log(y_t) - \log(y_{t-1})$. While differencing helps stabilize the mean, log transformation helps stabilize the variance. The data are randomly split into two datasets: 25% as the testing set and 75% as the training set. This testing-training ratio is the default in the scikit-learn algorithm (scikit-learn.org).

Table 1 provides summary statistics for the selected variables.

The random forest algorithm also requires parameter selection. The so-called hyper--parameter tuning method can test an appropriate parameter mix. This method runs the initial random forest numerous times to decide over the best parameter combination, which provides the most accurate prediction (scikit-learn.org).

Table 1: Summary statistics

Variable	Mean	Mode	Variance	St. dev.	Min	Max	Obs.
Output	2,86	2,50	8,81	2,99	–5,60	7,30	66
Consumption	2,43	2,90	3,45	1,87	–1,70	5,40	66
GDP EA	2,64	2,70	4,51	2,14	–4,50	6,20	66
Export	8,33	12,50	86,85	9,39	–16,90	37,40	66
Import	7,37	14,80	75,19	8,74	–17,00	35,00	66
IR	1,57	0,30	1,18	1,09	0,30	4,20	66
HHs. conf.	0,21	1,90	64,84	8,11	–28,90	18,40	66
PMI	0,74	3,50	16,90	4,14	–10,80	8,00	66
Const. conf.	0,63	2,90	141,28	11,98	–38,60	35,50	66
Ind. conf.	1,64	–4,60	255,27	16,10	–34,80	52,20	66
Wage	4,40	2,30	7,53	2,76	–1,60	10,20	66

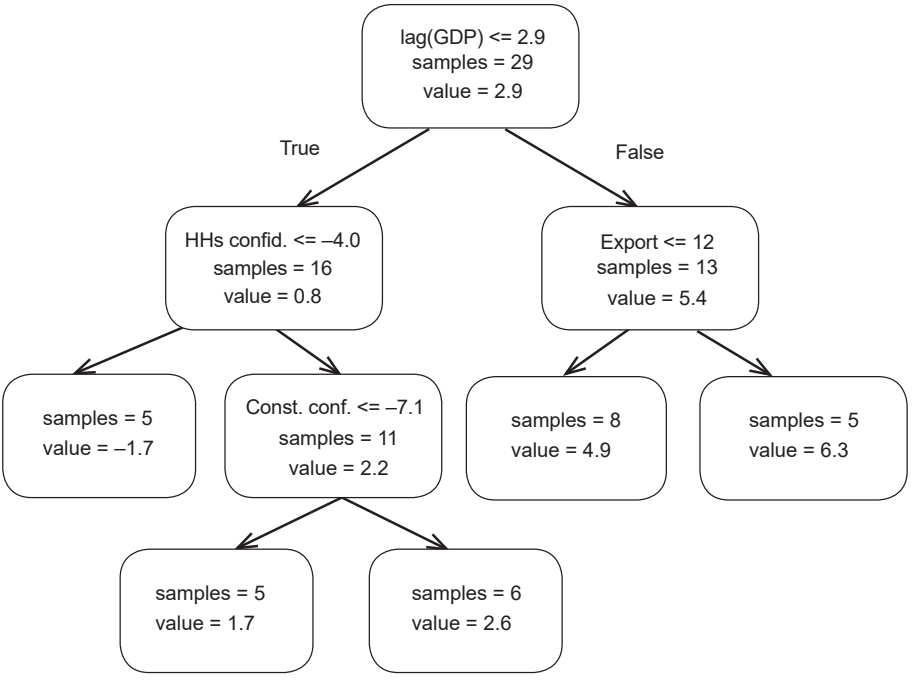
Source: Author's calculations

The hyper-parameter tuning tests the following parameter combinations: bootstrapping as the sample selection method; a number of trees varying by ten between 200 and 2000; 2, 5, or 10 as the minimum number of samples required to split a node; 1, 2, 4, or 5 as the minimum number of samples required at each leaf node; all features or subset of features allowed for tree consideration; and no limit on the maximum tree depth (see Koehrsen, 2018, for parameter selection). The parameter tuning is proven to be the most accurate random forest model estimated with 1000 trees, a minimum of two samples required to split a node, a minimum of one sample to be at a leaf node, and which randomly selects 42 features. This selected parameter set enters the random forest estimation in this study.

4. Results

This section first presents a regression tree, a random forest component, to forecast the Czech GDP.

Figure 1: Regression tree



Source: Author's calculations. All these variables are first lags of their YoY growth rates. The lag(GDP) label emphasizes the GDP variable being lagged as opposed to the non-lagged GDP target variable.

Figure 1 illustrates one regression tree out of the 1000 simulated trees in the random forest. In the regression tree, each node presents a feature used for splitting, a number of samples in a node, and an average forecast value. Every parent node is split into two child nodes: a left-hand child node with data smaller or equal to the splitting feature value, and a right-hand child node with data higher than the value. The last leaf nodes show the final conditional forecast for the Czech GDP. The weighted average across the leaf nodes is 2.98, representing a one-step-ahead forecast for one regression tree.

The regression tree predicts a lower GDP growth rate with smaller household or construction confidence; or smaller export values. This tree splitting and the selected feature set seem intuitive, which is a good sign for the random forest forecast being a combination of the tree predictions. The small open character of the Czech economy explains the export feature; the lagged GDP feature reflects the autoregressive character of the GDP series; and the household and construction confidence features signal the significance of soft data for GDP forecasting, as suggested by Woloszko (2020).

The final random forest forecast is visible in Figure 6 in the Appendix. The red dots are one-step-ahead forecasts estimated on a randomly selected testing sample, and the red line is a multi-step forecast from the last quarter of 2019 to the last quarter of 2020. The forecast is estimated on the dataset until the third quarter of 2019, thus neglecting the recent pandemic. The random forest method for predicting the Czech GDP shows high accuracy with R^2 equal to 0.9.

The tables 2 and 3 summarize the 2020 forecasts for different target variables:

Table 2: Yearly forecast

		2018	2019	2020
GDP	*mld.,c.p.*	4,735	4,866	4,971
	y/y,%	2.9	2.8	2.2
Consumption	*mld.,c.p.*	2,272	2,339	2,407
	y/y,%	3.4	2.9	2.9
Export	*mld.,c.p.*	3,996	4,057	4,213
	y/y,%	4.4	1.5	3.8
Import	*mld.,c.p.*	3,706	3,767	3,881
	y/y,%	5.9	1.6	3.0
GDP EA	*mld.,c.p.*	11,549	11,764	11,882
	y/y,%	3.2	1,9	1.0
3M IR	%	2.1	2.3	2.1

Source: Author's calculations, where c.p. stands for constant prices of the year 2010.

The random forest model enables the so-called variable importance analysis, which simplifies the result interpretation. The Methodology section explains the drawback for the mean decrease impurity if individual features are correlated. The section subsequently suggests verifying the method outcome by analysing the importance of every feature separately. The mean decrease impurity results are visible in Figure 1, and the second approach outcome, with only the GDP lag and one other variable, is illustrated in Figure 2.

The soft indicators, household confidence and industrial confidence, prove their importance for the GDP prediction in both Figure 1 and Figure 2. Similarly, Woloszko (2020) demonstrates a significant role of soft data, namely business survey and consumer

confidence, for French GDP dynamics. Lagged GDP is the third most crucial feature as evaluated by the mean decrease impurity method in Figure 1. In contrast, the least essential features are the variables PMI manufacturing, wage and interest rate.

Table 3: Quarterly forecast

		2019 Q1	2019 Q2	2019 Q3	2019 Q4	2020 Q1	2020 Q2	2020 Q3	2020 Q4
GDP	*mld.,c.p.*	1,138.0	1,222.0	1,242.0	1,264.1	1,161.6	1,251.6	1,266.7	1,290.8
	y/y,%	2.7	2.8	2.5	2.5	2.1	2.4	2.0	2.1
Consumption	*mld.,c.p.*	558.0	585.0	593.0	603.0	573.0	600.5	611.2	622.0
	y/y,%	3.2	3.0	3.3	2.4	2.7	2.7	3.1	3.1
Export	*mld.,c.p.*	1,006.0	1,031.0	983.0	1,037.4	1,027.0	1,075.9	1,033.8	1,076.7
	y/y,%	1.4	1.8	3.8	1.8	2.1	4.4	5.2	3.8
Import	*mld.,c.p.*	917.0	935.0	925.0	989.6	928.6	967.7	965.8	1,019.1
	y/y,%	1.9	1.2	2.6	1.0	1.3	3.5	4.4	3.0
GDP EA	*mld.,c.p.*	2,939.5	2,959.2	2,921.0	2,944.8	2,963.1	2,986.2	2,949.5	2,983.1
	y/y,%	2.8	2.8	1.0	0.9	0.8	0.9	1.0	1.3
3M IR	*%*	2.0	2.1	2.2	2.2	2.6	2.5	2.2	2.1

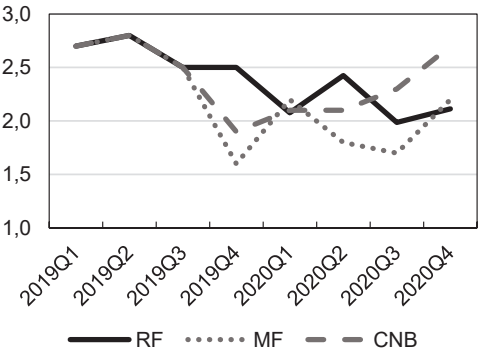
Source: Author's calculations, where c.p. stands for constant prices of the year 2010.

The variables import and export turn out to be very significant for the 2020 forecast in Figure 2, but less critical for the forecast in Figure 1. The lower importance of foreign trade and GDP EA in the mean decrease impurity method stem from the high correlation between these variables and the soft data. For example, the correlation coefficient between import and household confidence is over 70 percent. As we already discussed, the correlation between features can undermine the importance of the relatively less essential variables when applying the mean decrease impurity method. Nevertheless, the second approach illustrated in Figure 2 proves the unambiguous dependence of the Czech economic situation on foreign trade. Overall, the future Czech GDP forecast significantly depends on the soft indicators and the Czech foreign trade; this outcome also supports the regression tree findings.

Finally, Table 4 compares the prediction findings from the random forest model (RF), the Ministry of Finance prediction (MF) and the Czech National Bank (CNB), the two latter ones published in January 2020 (see www.mfcr.cz, www.cnb.cz).

Table 4: Comparison of GDP forecasts

	RF	MF	CNB
2019 Q1	2.70	2.70	2.70
2019 Q2	2.80	2.80	2.80
2019 Q3	2.50	2.50	2.50
2019 Q4	2.50	1.60	1.90
2020 Q1	2.08	2.20	2.10
2020 Q2	2.43	1.80	2.10
2020 Q3	1.99	1.70	2.30
2020 Q4	2.11	2.20	2.70



Source: Author's calculations, Ministry of Finance of the Czech Republic (Macroeconomic Forecast, January 2020), Czech National Bank (Inflation Report, I/2020).

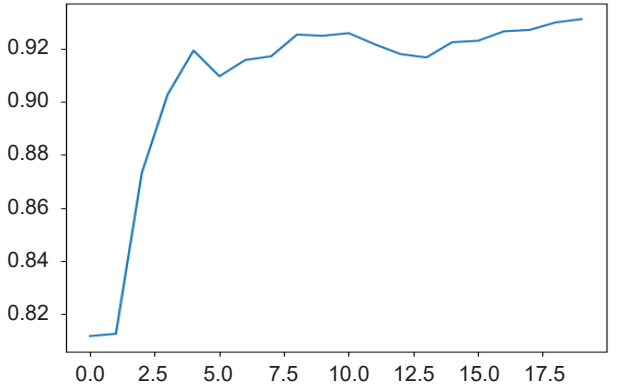
Table 4 displays the similarity of the results from the supervised machine-learning algorithm and the two central institutions. However, there is an apparent delay in the Czech GDP drop for the random forest forecast. The absence of lead variables in the random forest model could explain this difference (see Methodology section).

The MF prediction assumes the Eurozone GDP to drop in the first quarter of 2020 (www.mfcr.cz). Both the central institutions and the random forest model understand the impact of foreign trade on the small open Czech economy. However, what differs is the transmission of foreign shock into the domestic demand. While the random forest operates with lagged features, the DSGE models used by the Ministry of Finance and the Czech National Bank utilize lead variables based on rational agents' assumptions. The rational agents in these models are assumed to incorporate their future expectations into their present behaviour. The random forest lacks such economic intuition. Therefore, the expected Eurozone GDP drop translates into GDP slowdown in the same quarter as predicted by the random forest but one-quarter ahead as predicted by the CNB and the MF.

4.1. Forecasting accuracy

The random forest model, presented in the previous section, explains 90 percent of data variation. Figure 2 illustrates how the R-squared statistics vary with increasing number of regression trees in the random forest model.

Figure 2: Prediction accuracy with increasing number of trees



Source: Author's calculations

Table 5: Error measures for forecasting

	R^2	MAE	MSE	RMSE
Output	0.90	0.69	0.88	0.94
Consumption	0.78	0.62	0.80	0.89
GDP EA	0.87	0.52	0.67	0.82
Export	0.62	3.67	5.73	2.39
Import	0.63	3.49	4.97	2.23
IR	0.89	0.19	0.29	0.54
HHs. conf.	0.60	3.19	4.85	2.20
PMI	0.77	1.38	1.92	1.39
Const. conf.	0.67	7.03	9.65	3.11
Ind. conf.	0.60	4.60	6.99	2.64
Wage	0.69	1.18	1.50	1.22

Source: Author's calculations

Nevertheless, the 25 percent training/testing set split together with 66 observations results in only sixteen testing values for a target variable. We can increase the testing sample if we repeat the random forest estimation process. Every round, we estimate

a random forest; the model selects random data points for the training set and the testing set. In other words, a new random forest delivers predictions for a new set of dates, with replacement. Ten random forests predict over a hundred and sixty values in total for every target variable; see Figure 7 in the Appendix. The multitude of prediction points close to the actual data indicates that the random forest is a useful prediction tool.

Performance of machine-learning models is usually evaluated with error measures (see Hou *et al.*, 2015). Every random forest yields its own set of accuracy statistics. Table 5 averages these statistics across the ten different random forests for every forecast variable.

The R^2 statistics, a goodness-of-fit measure, is the highest for the Czech GDP and the lowest for the confidence variables. The household, construction and industrial confidence indices reflect subjective feelings, which depend on the media and future expectations and respond only partially to past economic performance. Nevertheless, the 90 percent R^2 statistics in Table 5 proves high forecasting accuracy for the Czech GDP.

This part compares the GDP growth in the last three years to the values forecast from the random forest model and the Czech central institutions.

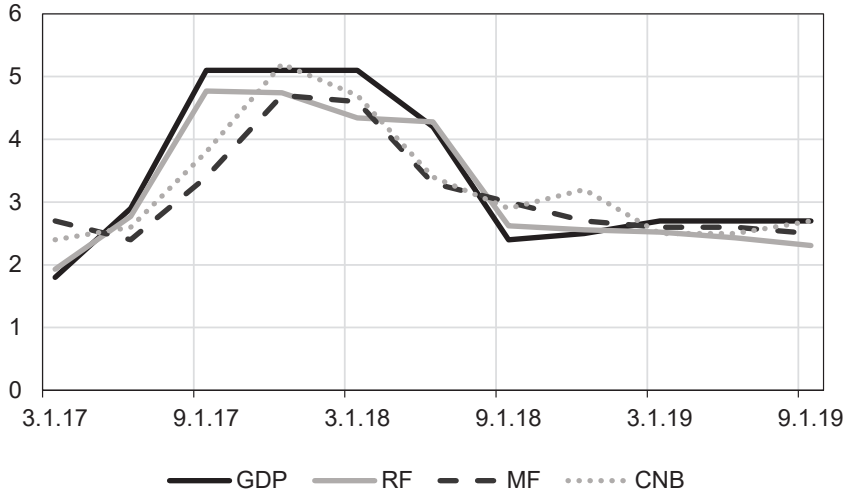
Table 6: Comparison of different GDP forecasts

Date	GDP	RF	MF_F	CNB_F	MF_N	CNB_N
17-Mar	1.80	1.93	2.70	2.40	2.30	2.50
17-Jun	2.90	2.77	2.40	2.60	3.10	3.50
17-Sep	5.10	4.77	3.40	3.80	4.80	5.00
17-Dec	5.10	4.74	4.70	5.20	5.30	5.40
18-Mar	5.10	4.34	4.60	4.70	5.00	4.90
18-Jun	4.20	4.28	3.30	3.40	2.60	2.60
18-Sep	2.40	2.62	3.00	2.90	2.60	2.70
18-Dec	2.50	2.56	2.70	3.20	2.50	2.30
19-Mar	2.70	2.52	2.60	2.50	2.70	2.60
19-Jun	2.70	2.43	2.60	2.50	2.70	2.70
19-Sep	2.70	2.31	2.50	2.70	2.50	2.70
RMSE		0.33	0.72	0.59	0.53	0.58

Source: Author's calculations. The GDP variable is expressed as a year-on-year growth rate.

Table 6 provides the actual GDP values (GDP), the random forest forecast (RF), the Ministry of Finance forecast (MF) and the Czech National Bank forecast (CNB). The suffix _F labels whether the institution predicts the current quarter and _N whether it nowcasts the last three months. When nowcasting, the central institutions already dispose of high-frequency data, such as monthly sectoral indices or monthly confidence time series. Table 6 shows the most accurate predictions for the random forest model, based on the root means square error (RMSE), which is the lowest even compared to the Ministry and the National Bank's nowcasts.

Figure 3: Different GDP forecasts



Source: Author's calculations. RF labels the random forest forecast, MF the Ministry of Finance forecast and CNB the Czech National Bank forecast.

Figure 3 illustrates the GDP prediction estimates from Table 6. The lines MF and CNB stand for the institutions' forecasts of the current quarter. The random forest model again visibly outperforms forecasts from the latter institutions.

5. Conclusions

The random forest is proven valuable for Czech forecasting; the model outperforms forecasts by the Czech Ministry of Finance (MF) and the Czech National Bank (CNB) and proves the significance of soft data for Czech forecasting. This study assesses forecasting

accuracy with one-step-ahead and multi-step forecasts. Estimation of ten random models for every target variable provides over a hundred one-step-ahead prediction estimates. The subsequent error measures show high forecasting accuracy, explaining around 90 percent of the Czech GDP. Furthermore, the model provides more accurate one-step-ahead predictions relative to forecasts and nowcasts from the two Czech forecasting institutions, tested on the data sample for the last three years.

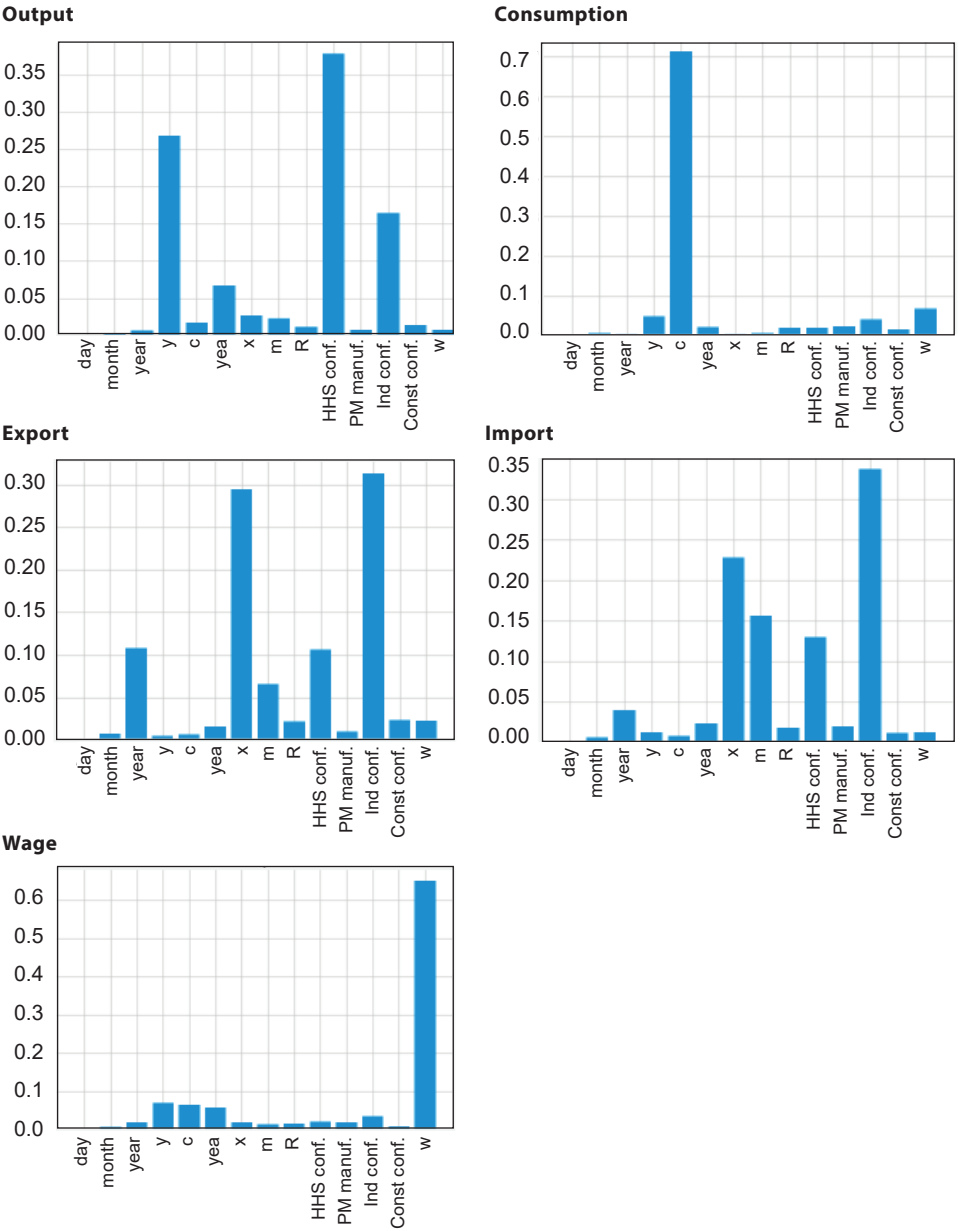
The multivariate multistep forecast for the Czech GDP assimilates the forecasting results from the MF and the CNB; however, the random forest forecast acts with a delay relative to the central institutions' forecasts. The MF and the CNB models assume rational agents and thus introduce lead variables for forecasting. The absence of the lead variables in the random forest model makes the outcome more past-dependent.

The random forest model also differs from the institutions' agent-based models by incorporating soft variables. The selected regression tree and the variable importance analysis stress the importance of these soft variables, especially the household and industrial confidence indices. The variable importance analysis thus reveals new information about variables crucial for prediction. Forecasters could use this approach to spot significant variables from large data sets before forecasting with other models. While the variable importance analysis shows the most appropriate variables, the hyper-parameter tuning identifies appropriate parameters to maximize prediction accuracy. The random forest outcome sensitivity to selected parameters and variables necessitates this data pre-processing step. Overall, the random forest withstands as a useful and surprisingly accurate forecasting alternative.

Further research could use the random forest method to complement other forecasting approaches; the variable importance analysis could help reduce the number of variables in econometric models or serve as a counterpart to the historical decomposition in dynamic stochastic general equilibrium models. Researchers could also apply this method to verify assumed Czech economic linkages. The presented random forest results could also serve for comparison with outcomes from other machine-learning models. Finally, lengthening the dataset with new observations could further improve the model forecasting accuracy.

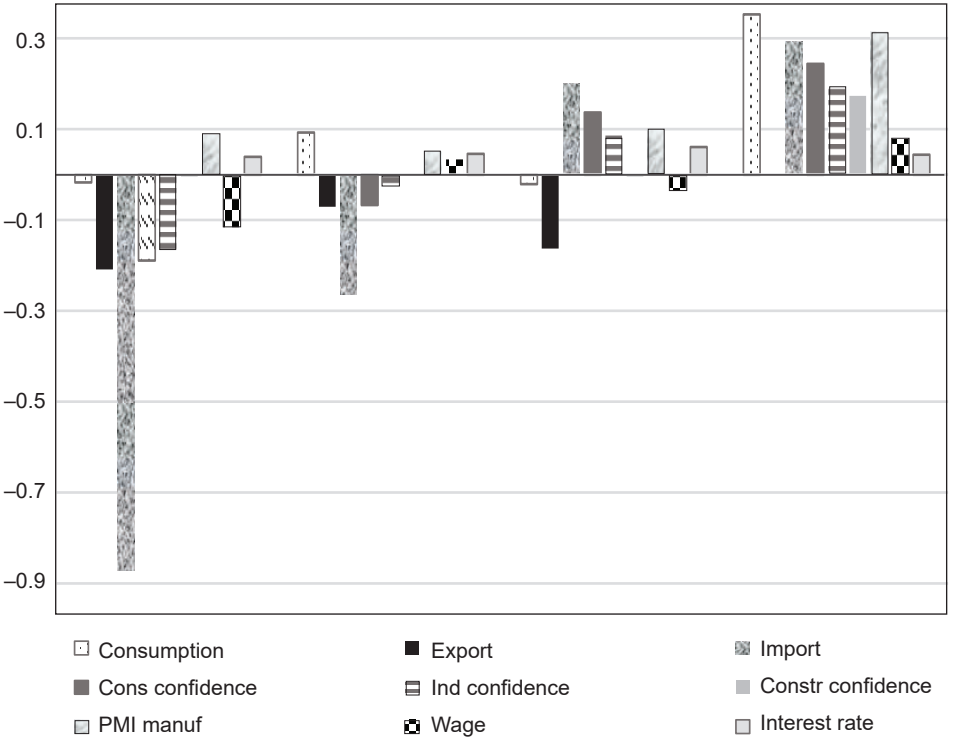
Appendix

Figure 4: Variable importance



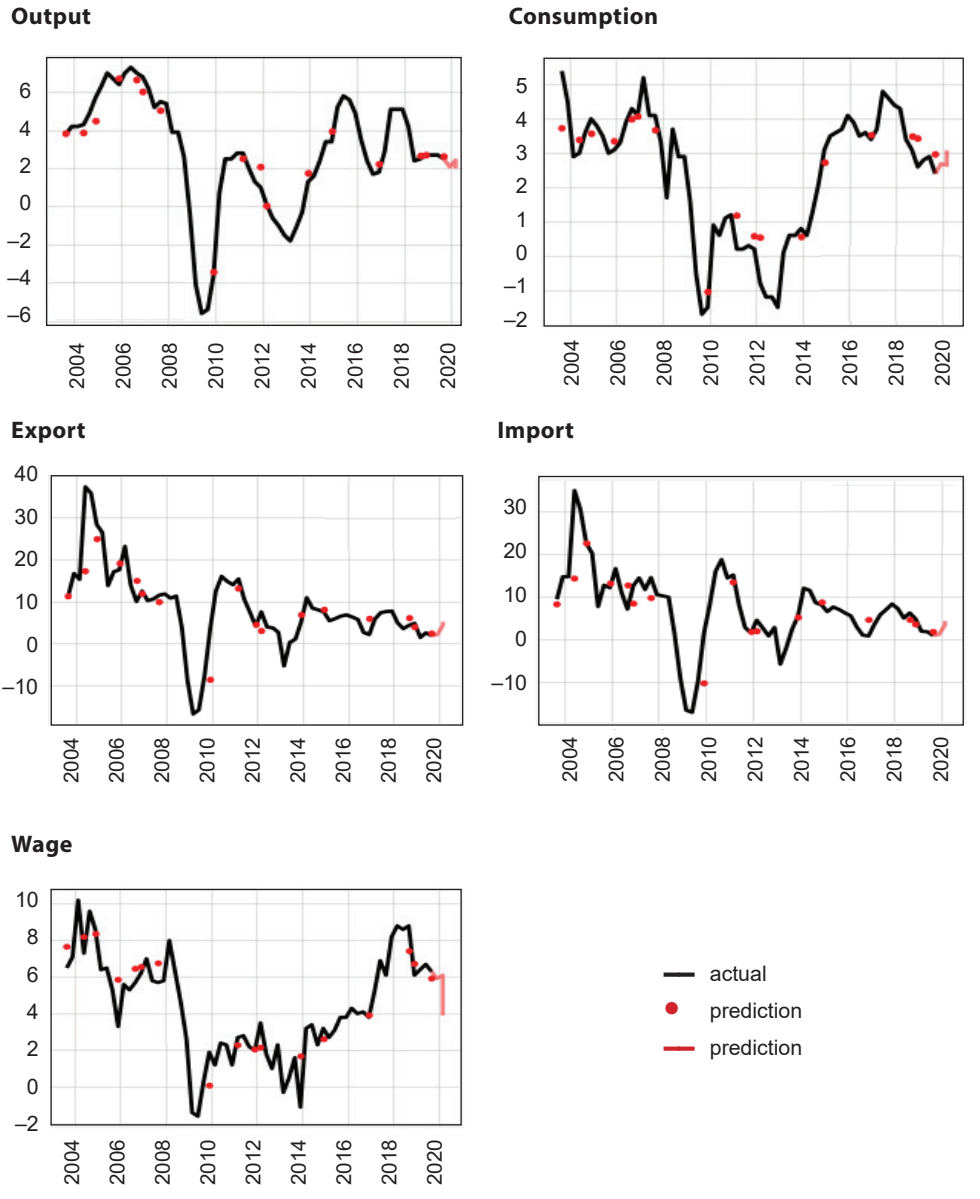
Source: Author's calculations. All the features on the x-axis are first lags of their YoY growth rates.

Figure 5: Impact of variables on prediction



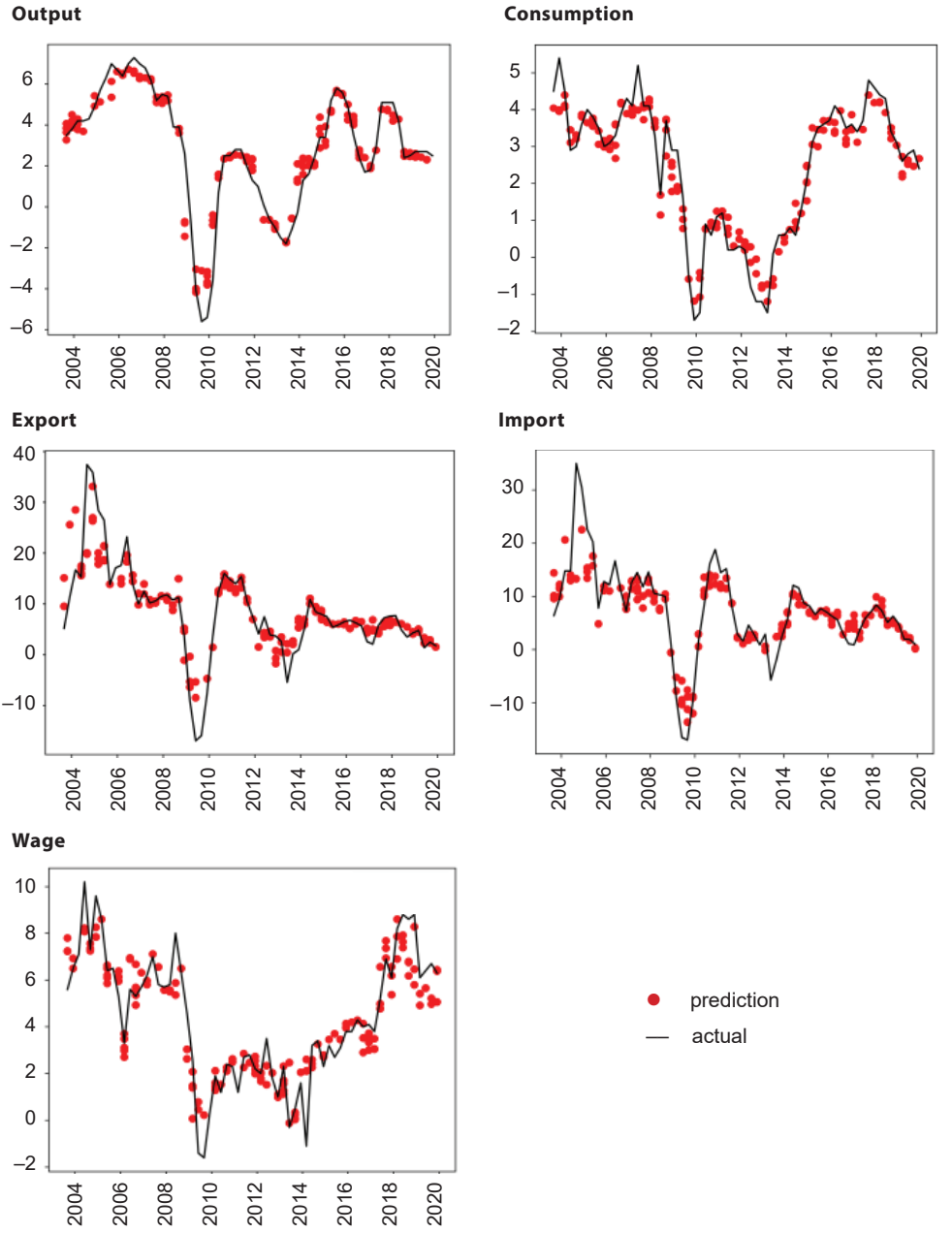
Source: Author's calculations

Figure 6: Actual and predicted values



Source: Author's calculations. All the variables are expressed as YoY growth rates.

Figure 7: Forecasting every target variable with ten random forests



Source: Author's calculations. All the variables are expressed as YoY growth rates.

References

Alessi, L., Detken, C. (2011). Quasi Real Time Early Warning Indicators for Costly Asset Price Boom/Bust Cycles: A Role for Global Liquidity. *European Journal of Political Economy*, 27(3), 520–533, https://doi.org/10.1016/j.ejpoleco.2011.01.003

Aliyev, I., Bobková, B., Štork, Z. (2014). *Rozšířený DSGE model české ekonomiky*. Praha: Ministerstvo financí České republiky. Available at: https://www.mfcr.cz/cs/o-ministerstvu/ odborne-studie-a-vyzkumy/2014/rozsireny-dsge-model-ceske-ekonomiky-17282

Atkeson, A., Ohanian, L. E. (2001). Are Phillips Curves Useful for Forecasting Inflation? *Quarterly Review*, 25(1), 1–11, https://doi.org/10.21034/qr.2511

Baybuza, I. (2018). Inflation Forecasting Using Machine Learning Methods. *Russian Journal of Money and Finance*, 77(4), 42–59, https://doi.org/10.31477/rjmf.201804.42

Biau, O., D'elia, A. (2010). *Euro Area GDP Forecast Using Large Survey Dataset—A Random Forest Approach*. EcoMod2010 No. 259600029. Available at: https://ideas.repec.org/p/ekd/002596/259600029.html

Blanchard, O. (2016). The Phillips Curve: Back to the '60s? *American Economic Review*, 106(5), 31–34, https://doi.org/10.1257/aer.p20161003

Boulesteix, A.-L., Janitza, S., Kruppa, J., et al. (2012). Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics: Random Forests in Bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6), 493–507, https://doi.org/10.1002/widm.1072

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32, https://doi.org/10.1023/A:1010933404324

Česká národní banka (2020). *Nový strukturální model „g3"*. Praha: Česká národní banka. Available at: https://www.cnb.cz/cs/menova-politika/zpravy-o-inflaci/tematicke-prilohy-a-boxy/Novy-strukturalni-model-g3

Česká národní banka (2020). *Zprávy o inflaci.* Praha: Česká národní banka. Available at: https://www.cnb.cz/cs/menova-politika/zpravy-o-inflaci/

Český statistický úřad (2020). Český statistický úřad. Available at: https://www.czso.cz/csu/czso/domov

Cheng, H., Tan, P.-N., Gao, J., et al. (2006). Multistep-Ahead Time Series Prediction. *Proceedings of the 10th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 765–774, https://doi.org/10.1007/11731139_89

Database—Eurostat (2020). Available at: https://ec.europa.eu/eurostat/data/database

Engel, C. (2000). Long-run PPP May Not Hold After All. *Journal of International Economics*, 51(2), 243–273, https://doi.org/10.1016/S0022-1996(99)00011-2

Fernández Delgado, M., Cernadas García, E., Barro Ameneiro, S., et al. (2014). *Do We Need Hundreds of Classifiers to Solve Real World Classification Problems*? 11(1), https://doi.org/10.1117/1.JRS.11.015020

Gawthorpe, K. (2019). Input-Output DSGE Model for the Czech Republic. *Prague Economic Papers*, 28(5), 612–630, https://doi.org/10.18267/j.pep.724

Gawthorpe, K. (2020). *Forecasting VAR Analysis for a DSGE–VAR Model*. Ministerstvo financí České republiky. Prague Working Paper No. 1/2014.

Goulet Coulombe, P. (2020). *The Macroeconomy as a Random Forest*. Social Science Research Network, https://doi.org/10.2139/ssrn.3633110

Herrera, G. P., Constantino, M., Tabak, B. M., et al.(2019). Long-term Forecast of Energy Commodities Price Using Machine Learning. *Energy*, 179, 214–221, https://doi.org/10.1016/j.energy.2019.04.077

Hou, Y., Edara, P., Sun, C. (2015). Traffic Flow Forecasting for Urban Work Zones. *IEEE Transactions on Intelligent Transportation Systems*, 4(16), 1761–1770, https://doi.org/10.1109/TITS.2014.2371993

Khaidem, L., Saha, S., Dey, S. R. (2016). Predicting the Direction of Stock Market Prices Using Random Forest. *ArXiv:1605.00003 [Cs]*, 1–20. Available at: http://arxiv.org/abs/1605.00003

Kucharčuková, O. B., Franta, M., Hájková, D., et al. (2013). *What We Know About Monetary Policy Transmission in the Czech Republic: Collection of Empirical Results*. Czech National Bank. Prague CNB Research and Policy Notes No. 1/2013.

Kumar, M., Thenmozhi, M. (2006). Forecasting Stock Index Movement: A Comparison of Support Vector Machines and Random Forest. *Indian Institute of Capital Markets 9th Capital Markets Conference Paper*, pp. 1–16, https://doi.org/10.2139/ssrn.876544

Li, X., Wang, Y., Basu, S., et al. (2019). A Debiased MDI Feature Importance Measure for Random Forests. *Advances in Neural Information Processing Systems*, 32(2019), 8049–8059. Available at: http://papers.nips.cc/paper/9017-a-debiased-mdi-feature-importance-measure-for-random-forests.pdf

Mei, J., He, D., Harley, R., et al. (2014). A Random Forest Method for Real-time price Forecasting in New York Electricity Market. 2014 *IEEE PES General Meeting | Conference Exposition*, pp. 1–5, https://doi.org/10.1109/PESGM.2014.6939932

Nguyen, T.-T., Huang, J., Nguyen, T. (2015). Unbiased Feature Selection in Learning Random Forests for High Dimensional Data. *The Scientific World Journal*, 2015, 1–18, https://doi.org/10.1155/2015/471371

Nyman, R., Ormerod, P. (2017). Predicting Economic Recessions Using Machine Learning Algorithms. *ArXiv:1701.01428 [q-Fin]*, 1–14. Available at: http://arxiv.org/abs/1701.01428

Ministerstvo financí České republiky (2020). Available at: https://www.mfcr.cz/cs

Pavlyshenko, B. M. (2019). Machine-Learning Models for Sales Time Series Forecasting. *Data*, 4(1), 15, https://doi.org/10.3390/data4010015

Pfeifer, J. (2020). *A Guide to Specifying Observation Equations for the Estimation of DSGE Models*. 81. Available at: https://bbcb79fe-a-62cb3a1a-s-sites.googlegroups.com/site/pfeiferecon/Pfeifer_2013_Observation_Equations.pdf?attachauth=ANoY7cp7k-tSpM7wwJb6nm7QIIBSe6wZV2PlWyAJYT84Tjjh_mdrAXqDPNgA_5R7bPGKEBMMyyQa6aEY7I7d8czTAV-m9H-XgTGWHSkXDJM1gxZktgDQxQ26nVWEY6cf0kst26-L_uKNvmzCEmTGKtTTlNPadMlM-QpA7MVnSrd9fEQ-qZ4wcQ4XIwUekxIMq_8KvEXufe1M9wWLS5w8njupRJKKE6z_Bkwk9qkbmOz7TuQAmySkgIXN_wkQhFinqEjE5IcdUMqh&attredirects=2

Ramakrishnan, S., Butt, S., Chohan, M. A., et al. (2017). Forecasting Malaysian Exchange Rate Using Machine Learning Techniques Based on Commodities Prices. 2017 *International Conference on Research and Innovation in Information Systems (ICRIIS)*, pp. 1–5, https://doi.org/10.1109/ICRIIS.2017.8002544

Wager, S., Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228–1242, https://doi.org/10.1080/01621459.2017.1319839

Wolman, A. L. (2003). *Real Implications of the Zero Bound on Nominal Interest Rates*. Federal Reserve Bank of Richmond. Richmond Working Paper No. 03-15, https://doi.org/10.2139/ssrn.2184489

Woloszko, N. (2020). *Adaptive Trees: A New Approach to Economic Forecasting*. OECD Economics Department. Paris Working Papers No. 1593, https://doi.org/10.1787/5569a0aa-en *3.2. Tuning the Hyper-parameters of an Estimator—Scikit-Learn 0.22.2 Documentation*. (2020). scikit-learn developers. Available at: https://scikit-learn.org/stable/modules/grid_search.html