CPB Netherlands Bureau for
Economic Policy Analysis

Forecasting World Trade Using Big Data and Machine Learning Techniques

We compare machine learning techniques to a large Bayesian VAR for forecasting world merchandise trade. We focus on how the predictive performance of the models changes when they have access to a big dataset with 11,017 data series on key economic indicators. The machine learning techniques used include lasso, random forest and linear ensembles. We additionally compare the accuracy of the forecasts during and outside the Great Financial Crisis.

We find no statistically significant differences in forecasting accuracy whether with respect to the technique, the dataset used -small or big- or the time period.

CPB Discussion Paper

Andrei Dubovik, Adam Elbourne,
Bram Hendriks, Mark Kattenberg

Updated version November 2022

Doi: https://doi.org/ 10.34932/01mq-sn15

Forecasting World Trade Using Big Data and Machine Learning Techniques

Andrei Dubovik Adam Elbourne Bram Hendriks
Mark Kattenberg

November 17, 2022

Abstract

We compare machine learning techniques to a large Bayesian VAR for nowcasting and forecasting world merchandise trade. We focus on how the predictive performance of the machine learning models changes when they have access to a big dataset with 11,017 data series on key economic indicators. The machine learning techniques used include lasso, random forest and linear ensembles. We additionally compare the accuracy of the forecasts during and outside the Great Financial Crisis. We find no statistically significant differences in forecasting accuracy whether with respect to the technique, the dataset used—small or big—or the time period.

JEL: F17, C53, C55.
Keywords: world trade; forecasting; big data; machine learning; large BVAR.

1 Introduction

Accurate predictions of trade flows are vital for making economic forecasts for small open economies, since the latter depend heavily on trade and trade is often an early indicator of economic changes to come. The CPB World Trade Monitor (WTM) aggregates worldwide monthly data on merchandise trade for a sample of 81 countries.[1] Those 81 countries account for 99% of world trade (Ebregt, 2016) and, since the WTM is one of the first to report on world

[1]The WTM also aggregates data on industrial production, which we leave aside in this paper.

trade, it is therefore widely used. Even though timely, the WTM data still has a two-month lag so nowcasting and forecasting are necessary to bridge the gap between the data and today. These forecasts are especially useful in crises because simply projecting recent performance forward is unlikely to be informative. However, it is especially these periods that traditional time series techniques have the greatest difficulty in forecasting.

In recent decades more data became available and techniques were developed to extract useful information out of big datasets. In this paper we compare the out-of-sample forecasting performance of two approaches to big data: a state-of-the-art time series approach and commonly used machine learning algorithms, which are trained purely on forecast accuracy. Our results contribute to two issues: whether machine learning algorithms outperform traditional approaches in low dimensional data and whether there is useful information in the massive number of time series now available. In our application machine learning algorithms do not generally outperform traditional time series approaches and using a big dataset over a relatively small number of time series does not improve forecast accuracy either.

Our state-of-the-art time series model is the large Bayesian VAR that the CPB uses to forecast developments in world trade. Large BVARs have become a workhorse model for forecasting with big datasets. For example, the seminal article of Bańbura et al. (2010) showed that a large BVAR could successfully extract useful information from 131 time series and produce more accurate forecasts than models using fewer time series. Large BVARs also display comparable forecast accuracy to other leading techniques for extracting useful information from big datasets.[2]

These time series approaches are extensions of traditional approaches developed for situations where the forecaster has many observations in the time dimension, but relatively few series at their disposal. In contrast, machine learning methods thrive in high-dimensional settings where the signal-to-noise ratio is potentially low, but they typically require a lot of data in order to distinguish the signal from the noise (see Jean et al. 2016; Mullainathan and Spiess 2017; Athey and Imbens 2019). Additionaly, their flexibility allows them to handle nonlinearities in the data, which is beneficial in times of macroeconomic uncertainty or financial stress (Goulet Coulombe et al., 2022). However, whether machine learning methods are able to predict

[2]For more details on traditional methods of dealing with big data in macroeconomics, see Bok et al. (2018). For trade specific examples, Guichard and Rusticelli (2011) compare a large BVAR to a dynamic factor model to predict world trade with a similar dataset to this paper and conclude that both state-of-the-art time series methods produce similarly accurate forecasts. Cantú-Bazaldúa (2021) reports encouraging nowcast accuracy for trade flows from a dynamic factor model.

macro-economic series better than traditional methods when using limited datasets remains an open question.

In this paper we start by estimating our machine learning algorithms on identical data to our large BVAR, which uses 23 variables.[3] This exercise contributes to a growing literature that uses machine learning to predict real macro-economic variables from low-dimensional data (Ahmed et al., 2010; Jung et al., 2018; Richardson et al., 2018; Circlaeys et al., 2018; Chen et al., 2019; Milunovich, 2020). Several of the aforementioned studies have suggested that machine learning methods can be promising tools for forecasting economic time-series, but that they do not necessarily yield statistically better forecasts than the traditional econometric estimators.[4] Richardson et al. (2018) use machine learning techniques to nowcast New Zealand GDP and conclude that a linear ensemble of methods performs best. Interestingly for our application, they find that all of the machine learning techniques outperform a BVAR. Jung et al. (2018) use a number of machine learning techniques to forecast GDP for several countries. They find that these methods can outperform traditional statistical models, depending on the country and time period. Circlaeys et al. (2018) apply a neural network to a big dataset of bilateral trade flows. They conclude that a neural network is able to predict bilateral trade flows very well in a cross-sectional setting, but the gain from using neural nets becomes smaller once lagged dependent variables are included, which is a standard practice for traditional time series methods. Finally, Milunovich (2020) considers predictions for Australia's real house prices using machine learning, deep learning and traditional time-series models, and finds that a support vector regression together with simple mean forecast combinations are the best predictors in general. Our paper adds to the aforementioned literature by comparing the accuracy of large BVAR forecasts of the world merchandise trade to those from lasso, random forest and ensemble models over the period April 2006 to September 2019.

In addition to our comparisons using the 23 core variables, we present evidence on whether a significantly bigger dataset with 11,017 variables improves the accuracy of the machine learning algorithms. By doing both these steps we can separate the effect of the techniques from the expansion in the dataset size. Other authors have reported significant improvements in trade nowcasts and forecasts with bigger datasets. For example, Stamer (2021)

[3]Even though large BVARs have been shown to produce more accurate forecasts with much bigger datasets, unpublished research at CPB showed that increasing the number of variables beyond these 23 variables produces similar forecast accuracy for world trade. Hence the CPB large BVAR uses a medium sized dataset of 23 variables.

[4]Ahmed et al. (2010); Chen et al. (2019) consider the predictive performance of machine learning methods without comparing them to more conventional methods.

and Cerdeiro et al. (2020) report accurate nowcasts using container ship movement data and machine learning techniques. For forecasting further ahead, Kim (2020) reports better forecast performance for Korean exports using deep learning techniques than using traditional vector autoregressions and vector error-correction models. By contrast, our benchmark is a state-of-the-art large BVAR rather than traditional vector autoregressions and vector error-correction models. Moreover, since we use such a large dataset we pay particular attention to forecast performance during the 2007–2008 financial crisis, since it is possible that specific variables in the big dataset might be particularly helpful in forecasting turning points.

Our main finding is that neither machine learning model outperforms the BVAR over the whole sample. Moreover, adding bigger data does not significantly improve accuracy compared to the results of the BVAR on the pre-selected core data. Lastly, during the 2007–2008 financial crisis, the BVAR remains the best performing model. Our results can be viewed as a more cautious outlook on the usefulness of machine learning methods for macro-economic forecasts. While selected studies achieve better results with machine learning methods, such an improvement is not a given. Potentially, a further fine-tuning of the methods used in this paper would have allowed us to outperform the BVAR, but it would have also inadvertently introduced a bias towards our specific data.

The remainder of this papers is as follows. Section 2 describes the competing models. Section 3 describes the data. Section 4 reports the results. In Section 5 we discuss why machine learning methods might be failing to outperform the BVAR. Finally, Section 6 concludes.

2 Competing methods

We compare the forecast accuracy of four techniques before, during and after the Great Recession of 2008–2009. This section briefly introduces each of these techniques.

2.1 Large Bayesian Vector Autoregression

The large Bayesian Vector Autoregression (BVAR) follows the seminal work of Bańbura et al. (2010) and Giannone et al. (2015). As the number of parameters to be estimated in VAR can become very large, overfitting can be a significant problem for forecast accuracy. Analogous to shrinkage applied in ridge regression, the specification of the priors shrinks the parameters such that overfitting is reduced and out-of-sample forecast accuracy improves.

The large BVAR used in this paper follows de Wind (2015) and is estimated with a combination prior,[5] six lags and variables enter the large BVAR in log-levels. A Kalman filter is used to fill in the ragged edge. For multi-period forecasts, the BVAR forecasts are iterative: first a one-month-ahead forecast is made, then the one-month-ahead forecast is used to make the two-month-ahead forecast and so on. This contrasts to our machine learning specifications where direct forecasts are made: we use a separate model for each forecast horizon.

2.2 Model specifications for the machine learning methods

We adopt the same specification for all machine learning methods. Additionally, to deal with the ragged edge, we align the data on publication date. In particular, let y_t denote the world trade index, in logs, from period $t-2$ as published at the end of period t. Let x_t denote a vector of regressors that are available at the end of period t and let Δx_t denote the absolute growth of those regressors, also as available at the end of period t. The regressors included in x_t have varying publication lag, from zero to several months. We always include y_t among x_t. We use the following specification:

$$\mathbb{E}_t(y_{t+h} - y_t) = f_t^h(x_t, \Delta x_t, \Delta x_{t-1}, \Delta x_{t-2}), \tag{1}$$

where $f_t^h(\cdot)$ is estimated with a particular machine learning forecaster, namely lasso or random forest, using the data available up to and including period t. We estimate a separate model for each period t and time horizon h.

In the core data all the variables are log transformed. The big data has over 11 thousand variables and we log-transform those series that are always positive, and leave the other series as they are. As equation (1) shows, we use an error-correction model specification. Namely, the dependent variable is in differences and the regressors consist of levels together with lagged differences. This specification was chosen during a short trial at the start of the project and has not been revised since so as to avoid overfitting. We adopt a direct forecasting methodology for machine learning models. That is, we train a new model for each new forecast horizon.[6]

[5]The prior is a combination of three commonly used priors as in Sims and Zha (1998) and Giannone et al. (2015). They are the *Minnesota prior*, the *sum-of-coefficients prior* and the *no cointegration prior*.

[6]In principle, it is possible to adopt dynamic forecasting, but that raises a theoretical question about how to properly do cross-validation with dynamic forecasts, which falls outside of the scope of the present research.

Lasso

We use the R package "glmnet" to estimate the lasso models, following Friedman et al. (2010). We leave the cross-validation parameters at their default values with a 10-fold cross-validation for a lasso. Following Hastie et al. (2008) we set the value of the regularisation parameter λ to one standard deviation above the value that minimises the cross-validated root-mean-square error (RMSE).

Random Forest

We use the R package "randomForest" to estimate the random forest models, following Breiman (2001). Once again, we leave the package parameters at their default values. In particular, we estimate a random forest with 500 regression trees of final node size 5, where 10 randomly selected variables are used at each split.

Ensemble

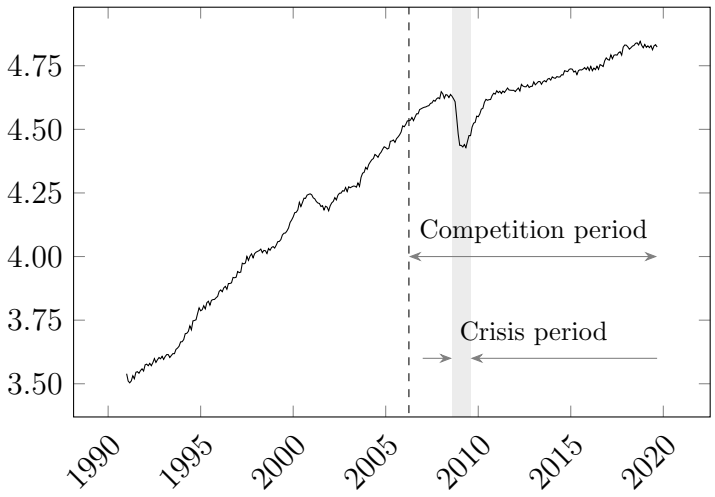
Previous research (see, for example, Hendry and Clements, 2004) has shown that linear combinations of forecasts from different models can produce more accurate forecasts than even the best individual model. For our ensemble we simply average the predictions by the machine learning models and the large BVAR, because more sophisticated methods to weight the predictions might lead to overfitting (see, for example, Claeskens et al., 2016).

3 Data

The time series of world trade is shown in Fig. 1. In the period before the Great Financial Recession, world trade typically grew at a steady rate with the exception of the early 2000s when there was a limited contraction. The decades of steady growth came to an abrupt end in the Great Financial Recession when world trade contracted by 13%, before partially recovering and then resuming growth at a slower rate than before the crisis. Arguably, the figure highlights two structural breaks. Firstly, the financial crisis represented a period of higher volatility with large changes in world trade flows. Secondly, the lower trend growth rate after the financial crisis is a type of change well-known to cause difficulties for traditional time series methods (Banbura and van Vlodrop, 2018).

For our out-of-sample forecasting competition we use two different datasets. Both datasets are single vintage: a key contribution of our analysis is investi-

Figure 1: World Trade Index (Log Scale)



Notes: The period starting in April 2006 (dashed line) is the competition period, i.e. it is the period in which the forecast errors are computed. The shaded area denotes the period between August 2008 and August 2009, used as the crisis period elsewhere in this paper.

gating whether a big dataset would improve forecast accuracy, and gathering real-time data vintages for 11,000 time series was deemed beyond the scope of this research. Moreover, many of the series involved are survey or financial data, which are not subject to revision. For an initial comparison between the large BVAR and the machine learning algorithms, we use a dataset of 23 regressors, comprising various economic and survey indicators (i.e. the core dataset). These regressors have been selected for the BVAR prior to the current research, with the selection similar to that in Guichard and Rusticelli (2011). The list of variables along with the respective data sources is shown in Table 5 in the appendix.

We then extend the core dataset with almost 11,000 additional monthly time series from Datastream. These series primarily consist of all key economic indicators and commodities that were available on a monthly basis for the period 1991–2019. A small additional number of series were also included manually, based on our expert judgement. To give an idea of the potential extra information in this big dataset, Table 1 reports the minimum number of factors required to explain at least 90% of the total variance (when using principal components decomposition). Whereas 4 factors are sufficient to explain over 90% of the variation in the core dataset, 33 are needed for the big dataset.

Table 1: Core and Big datasets

	Core	Big
Start	1991-03	1991-03
End	2020-04	2019-10
Number of variables	23	11,017
Number of observations	350	344
Average delay (months)	1.7	1.5
No. factors with 90% total variance	4	33

3.1 Competition rules

We evaluate the performance of the models based on a rolling out-of-sample RMSE where we use the data available up to and including period t to compute the prediction for period $t+h$. Formally,

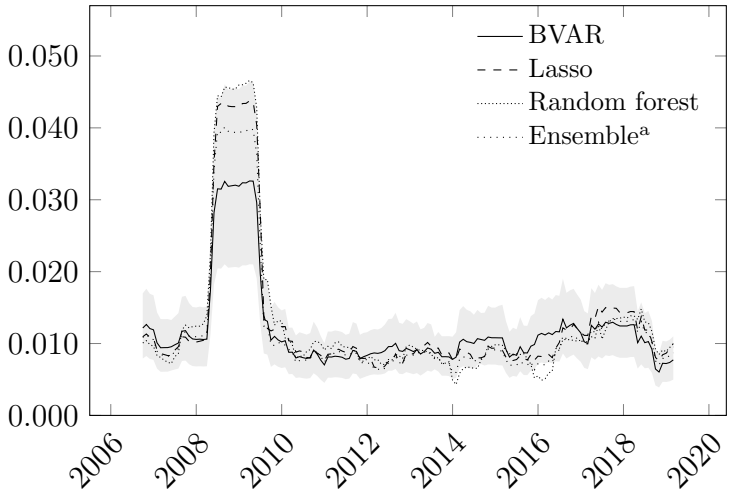
$$\text{RMSE}_h = \left(\frac{\sum_{t=t_0-h}^{T-h} \left(y_{t+h} - \mathbb{E}_t(y_{t+h}) \right)^2}{T - t_0 + 1} \right)^{\frac{1}{2}}, \tag{2}$$

where t_0 and T denote the start and end of a particular competition period. We compute $\mathbb{E}_t(y_{t+h})$ as $\mathbb{E}_t(y_{t+h}) = y_t + f_t^h(\cdot)$, where f_t^h is re-estimated for each separate period t and time horizon h using the data up to and including period t. When estimating f_t^h we use standard cross-validation, i.e. we do not take the time-series nature of the regressors into account.

There are a number of ways to compute out-of-sample errors for time-series forecasts, see, e.g. Bergmeier and Benitez (2012) for a discussion.[7] We have made our choice for a rolling RMSE so as to stay consistent with potential ex-post evaluations. That is, if we had implemented a specific machine learning method for the monthly CPB forecasts from the very beginning, then an ex-post evaluation of these forecasts would have been the same as how we compute the forecast errors with our approach.

[7]For instance, only the data before a given period, or both the data before and the data after a given period can be used to train a model. An error can be computed for a single period at a time or for a group of consecutive periods. Additional observations between a given period and the data used for training can be dropped, so as to help in breaking time-series dependencies. Et cetera.

Figure 2: Locally Averaged Rolling 2-month-ahead RMSE, Core Data



Notes: RMSE is averaged over a centred 13 month window. The shaded area is a 95% confidence interval around the BVAR RMSE.
[a] Average ensemble using core data BVAR, lasso, and random forest.

4 Results

We use three evaluation windows: the full sample, the financial crisis and the non-financial crisis part of our sample. Tables 2 to 4 contain the RMSEs of the competing models in those periods. The RMSEs are reported normalised to the RMSE of the large BVAR.

We use a two-sided Diebold and Mariano (1995) test to compare predictions by the machine learning algorithms to those by the BVAR. We run a separate test for each dataset (core or big) and for each forecast horizon, which results in testing 24 hypotheses for each of the evaluation windows. As we do not have *a priori* information that would suggest a hypothesis that a particular machine learning method might do better or worse at a specific forecasting horizon, we correct for multiple hypothesis testing using the Holm-Bonferroni method (Holm, 1979).

As shown in Table 2, over the full competition window, we cannot reject the hypothesis that the predictions by the machine learning methods are equal to those of the large BVAR. Looking at the absolute values of the RMSEs, Table 2 shows that the machine learning methods perform slightly worse compared to the BVAR, except for the big data ensemble at the 1 month horizon.

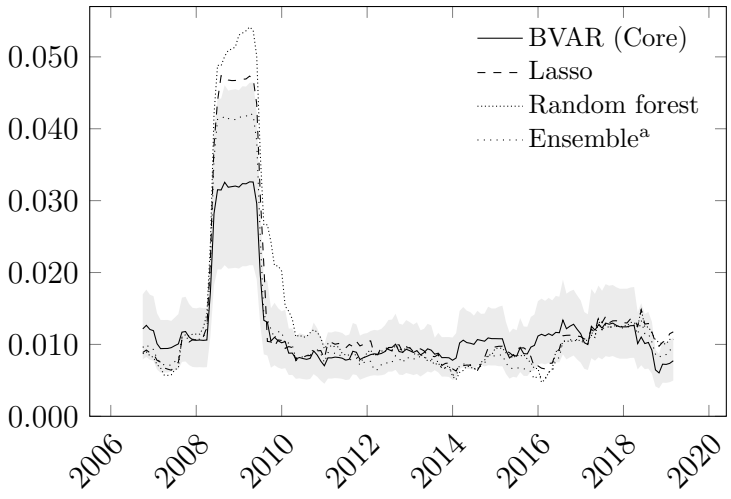
Since the WTM is published with a 2 month lag, to highlight variation over time at this horizon, we present a 2-month-ahead RMSEs calculated

Table 2: Relative RMSE (Apr 2006–Sep 2019)

Horizon	1	2	3	6
Core data				
BVAR	1.000	1.000	1.000	1.000
	(0.0103)	(0.0133)	(0.0169)	(0.0285)
Lasso	1.080	1.175	1.173	1.184
Random forest	1.059	1.210	1.298	1.246
Ensemble[a]	1.014	1.081	1.100	1.030
Big data				
Lasso	1.028	1.229	1.246	1.075
Random forest	1.111	1.348	1.476	1.608
Ensemble[b]	.991	1.105	1.152	1.111

Notes: Each number gives the RMSE of the respective method relative to the RMSE of BVAR. The original BVAR RMSEs are given in round brackets. Bold numbers indicate significance at 5% level according to the two-sided Diebold-Mariano test. None of the indicated differences remains significant when corrected for multiple hypotheses testing using the Holm-Bonferroni method.
[a] Average ensemble using core data BVAR, lasso, and random forest.
[b] Average ensemble using core data BVAR and big data lasso and random forest.

Figure 3: Locally Averaged Rolling 2-month-ahead RMSE, Big Data



Notes: RMSE is averaged over a centred 13 month window. The shaded area is a 95% confidence interval around the BVAR RMSE.
[a] Average ensemble using core data BVAR and big data lasso and random forest.

Table 3: Relative RMSE (Aug 2008–Aug 2009, i.e. crisis)

Horizon	1	2	3	6
Core data				
BVAR	1.000	1.000	1.000	1.000
	(0.0200)	(0.0324)	(0.0443)	(0.0829)
Lasso	1.222	1.341	1.324	.887
Random forest	1.191	1.421	1.517	**1.338**
Ensemble[a]	1.112	1.220	1.242	1.022
Big data				
Lasso	1.188	1.449	1.449	1.123
Random forest	1.395	1.647	1.764	**1.563**
Ensemble[b]	1.160	1.286	1.339	**1.201**

Notes: Each number gives the RMSE of the respective method relative to the RMSE of BVAR. The original BVAR RMSEs are given in round brackets. Bold numbers indicate significance at 5% level according to the two-sided Diebold-Mariano test. None of the indicated differences remains significant when corrected for multiple hypotheses testing using the Holm-Bonferroni method.
[a] Average ensemble using core data BVAR, lasso, and random forest.
[b] Average ensemble using core data BVAR and big data lasso and random forest.

over a rolling 13 month window for each model throughout our evaluation period in Fig. 2 and 3. They show clearly that the relative performance of the large BVAR is driven by its accuracy during the crisis. Whilst all models performed worse during the crisis, the accuracy of the machine learning techniques deteriorated significantly more than the BVAR. Table 3 shows that it's not just at the 2-month-ahead horizon that the BVAR performs relatively well during the crisis. At 1-, 2- and 6-months-ahead the BVAR almost always has the lowest RMSE. However, we stress that the difference in predictive performance is only exceeds the 5% critical value three times (and only when predicting six months ahead) using a two-sided Diebold-Mariano test. Importantly, none of the results are statistically significant once we control for multiple hypothesis testing using the Holm-Bonferroni correction.

In contrast, in both the pre-crisis and post-crisis periods the differences are much smaller. As shown in Table 4, outside the crisis the machine learning techniques often perform marginally better than BVAR. Based on a two-sided Diebold-Mariano test the difference in prediction performance exceeds the 5% critical value three times. Notably, it seems that especially the big data ensemble provides predictions that are better than the BVAR. Once again, however, these results should be interpreted carefully as none of the

Table 4: Relative RMSE (Apr 2006–Jul 2008, Sep 2009–Sep 2019)

Horizon	1	2	3	6
Core data				
BVAR	1.000	1.000	1.000	1.000
	(0.0090)	(0.0101)	(0.0118)	(0.0169)
Lasso	1.012	1.002	.955	1.644
Random forest	.997	.981	.964	1.024
Ensemble[a]	.968	.937	**.895**	1.046
Big data				
Lasso	.950	.989	.938	.967
Random forest	.962	1.004	1.018	1.700
Ensemble[b]	**.908**	.911	**.870**	.892

Notes: Each number gives the RMSE of the respective method relative to the RMSE of BVAR. The original BVAR RMSEs are given in round brackets. Bold numbers indicate significance at 5% level according to the two-sided Diebold-Mariano test. None of the indicated differences remains significant when corrected for multiple hypotheses testing using the Holm-Bonferroni method.
[a] Average ensemble using core data BVAR, lasso, and random forest.
[b] Average ensemble using core data BVAR and big data lasso and random forest.

differences are significant once we control for multiple hypothesis testing using the Holm-Bonferroni correction.

4.1 Analysis of forecast errors

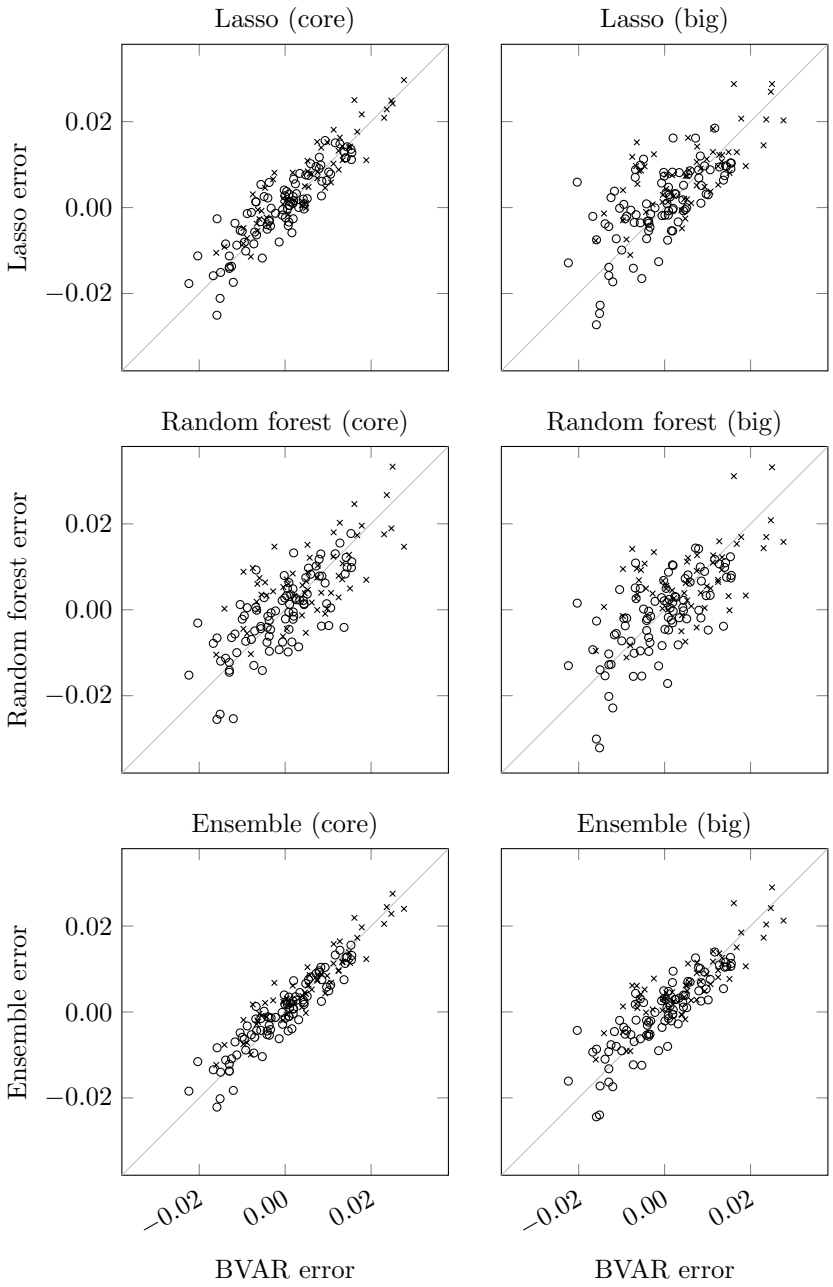
Figure 4 compares the machine learning errors to those from the large BVAR outside of the crisis. The 45 degree line indicates when both models give the same forecast. If the correlation represented by the scatter plot is shallower than 45 degrees, the machine learning models produced more accurate forecasts than the BVAR. As Figure 4 shows, that is not visible to the naked eye, which is not surprising given the similar RMSEs presented in Table 4.

Nonetheless, the figures do show that both the machine learning models and the large BVAR produce their largest forecast errors in similar periods. In Figure 4, periods when world trade is increasing are represented by circles, whilst crosses represent contractions. For all models, the largest overpredictions occur when world trade is contracting and the largest underpredictions occur when world trade is increasing. This could be caused by a number of issues. For example, if the largest errors are due to unforecastable shocks occurring after the forecasts are made, all models will make similar

forecast errors in these periods. Alternatively, it is also plausible that the machine learning algorithms might be failing to pick up signals that could have markedly improved the forecasts because of the relatively short time dimension of our data. As such, their forecasts are similar to the BVAR because the BVAR already embodies the clearest signals in the data.

Finally, the scatter plots for the ensemble models lie visibly closer to the 45 degree line than the underlying models. This implies the ensemble forecasts are more similar to the BVAR forecasts than the lasso and random forest forecasts are, which is not surprising since the BVAR is included in the ensemble.

Figure 4: Forecast Errors (Apr 2006–Jul 2008, Sep 2009–Sep 2019)



Notes: Circles denote errors that correspond to *increases* in the observed world trade index, crosses represent *decreases*.

5 Discussion

Our results suggest that we do not forecast world trade with more precision if we replace the forecasts of the large BVAR on the core dataset with the forecasts of machine learning methods on the big dataset. Possibly, this occurs because the machine learning methods overfit: they do not have enough observations to distinguish true signal from noise in the big data and therefore results are comparable to the results of the BVAR on the pre-selected core data. An alternative explanation would be that the big dataset contains many correlated variables that are important for forecasting world trade. In this case lasso might not select all the relevant predictors and the individual trees of the random forest would be "too similar." In this case, both the lasso and the random forest forecasts will have higher variance.

Let us consider multicollinearity first. If multicollinearity of the big dataset is problematic for the machine learning methods, we would expect that orthogonalising the regressors in the big dataset would improve the world trade forecasts. However, a study by Buck et al. (2021),[8] that uses the same datasets as we do, concludes that reducing the dimensionality of the big data with principal component analysis does not improve the forecasts by lasso and random forests on the big data. Their findings suggest that using big data does not improve world trade forecasts because the machine learning methods overfit, not because the big dataset contains many variables that are highly multicollinear.

Regarding overfitting, in a parallel study de Nerée tot Babberich et al. (2021)[9] also use the same dataset as we do and investigate whether lasso and random forest would have picked up informative series from the big dataset in case there were any. Their analysis relies on augmenting the big dataset with simulated informative series. They find that in general both lasso and random forest would have performed better on the big data if there was an informative additional predictor there, with lasso outperforming random forest in this regard. (Indeed, we see in Tables 3 and 4 that lasso outperforms random forest across the board when it comes to the big dataset.) However, the performance of both methods drops substantially if the simulated informative series are correlated with the existing regressors. So, while we cannot rule out that the big data contains additional information on the development

[8]The study by Buck et al. (2021) is an unpublished manuscript, which is available upon request. Two of the authors of the present paper were involved in the study by Buck et al. (2021) as supervisors.

[9]The study by de Nerée tot Babberich et al. (2021) is an unpublished manuscript, which is available upon request. Two of the authors of the present paper were involved in the study by de Nerée tot Babberich et al. (2021) as supervisors.

of world trade, the aforementioned result suggests that at least no clear, i.e. orthogonal, additional predictors exist in the big data as compared to the core data.

6 Conclusions

We have compared the nowcasts for merchandise world trade by a large Bayesian VAR using conventional data to the nowcasts by machine learning techniques using the same or big data. The machine learning techniques used include lasso, random forest and linear ensembles. We have used the CPB World Trade Monitor as the measure for world merchandise trade. We find that none of the machine learning models statistically outperforms the large Bayesian VAR when these models are trained on conventional data. Extending the set of explanatory variables from the 23 regressors included in the BVAR to some 11,000 additional data series on key economic indicators does not lead to a significant improvement in terms of forecasting accuracy either.

In particular, during the Great Financial Recession the large BVAR typically produced more accurate forecasts, with the machine learning methods having the RMSE anywhere between 2% and 76% higher than the BVAR (the exception being the lasso using the core dataset, which produced lower RMSE forecasts at the 6 month horizon). Outside of the crisis the machine learning methods produced marginally lower RMSEs than the BVAR: random forest was up to 4% better and lasso up to 6% better, depending on the dataset and the horizon. A linear ensemble including the BVAR was up to 13% better. However, these differences in forecast accuracy are not statistically significant once we control for multiple hypothesis testing using the Holm-Bonferroni correction.

References

Ahmed, N. K., Atiya, A. F., Gayar, N. E., and El-Shishiny, H. (2010). An empirical comparison of machine learning models for time series forecasting. *Econometric reviews*, 29(5-6):594–621.

Athey, S. and Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725.

Bańbura, M., Giannone, D., and Reichlin, L. (2010). Large bayesian vector auto regressions. *Journal of applied Econometrics*, 25(1):71–92.

Banbura, M. and van Vlodrop, A. (2018). Forecasting with bayesian vector autoregressions with time variation in the mean. Tinbergen Institute discussion paper 18-025/IV.

Bergmeier, C. and Benitez, J. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191(1):192–213.

Bok, B., Caratelli, D., Giannone, D., Sbordone, A. M., and Tambalotti, A. (2018). Macroeconomic nowcasting and forecasting with big data. *Annual Review of Economics*, 10(1):615–643.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Buck, R., Ten Harmsen van der Beek, L., Oosterbroek, G., and Reijnders, R. (2021). Improving machine learning forecasts with PCA-based preprocessing: A case study on world trade. Unpublished manuscript.

Cantú-Bazaldúa, F. (2021). Nowcasting global trade in goods and services. *Statistical Journal of the IAOS*, 37(1):259–277.

Cerdeiro, D. A., Komaromi, A., Liu, Y., Saeed, M., et al. (2020). World seaborne trade in real time: A proof of concept for building ais-based nowcasts from scratch. IMF working paper 20/57.

Chen, J. C., Dunn, A., Hood, K., Driessen, A., and Batch, A. (2019). Off to the races: A comparison of machine learning and alternative data for predicting economic indicators. In *Big Data for 21st Century Economic Statistics*. University of Chicago Press.

Circlaeys, S., Kanitkar, C., and Kumazawa, D. (2018). Bilateral trade flow prediction.

Claeskens, G., Magnus, J. R., Vasnev, A. L., and Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, 32(3):754–762.

de Nerée tot Babberich, D., Djibuti, A., Dijk, J., and Bon, T. (2021). Forecasting performance of lasso and random forest in high-dimensional setting. Unpublished manuscript.

de Wind, J. (2015). Technical background document for BVAR models used at CPB. CPB background document.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Ebregt, J. (2016). The CPB world trade monitor: Technical description. CPB background document.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1).

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2):436–451.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., and Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, 37(5):920–964.

Guichard, S. and Rusticelli, E. (2011). A dynamic factor model for world trade growth. OECD Economics Department Working Papers No. 874.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *Elements of statistical learning*. Springer.

Hendry, D. F. and Clements, M. P. (2004). Pooling of forecasts. *The Econometrics Journal*, 7(1):1–31.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 6(2):65–70.

Jean, N., Burke, M., Xie, M., Davis, W., Lobell, D., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Jung, J., Patnam, M., and Ter-Martirosyan, A. (2018). An algorithmic crystal ball: Forecasts based on machine learning. IMF working paper 18/230.

Kim, S. (2020). Macroeconomic and financial market analyses and predictions through deep learning. Bank of Korea working paper 2020-18.

Milunovich, G. (2020). Forecasting australia's real house price index: A comparison of time series and machine learning methods. *Journal of Forecasting*, 39(7):1098–1118.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometrics approach. *Journal of Economic Perspectives*, 31(2):87–106.

Richardson, A., Van Florenstein Mulder, T., and Vehbi, T. (2018). Nowcasting new zealand GDP using machine learning algorithms. CAMA working paper 47/2018.

Sims, C. A. and Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review*, pages 949–968.

Stamer, V. (2021). Thinking outside the container: A machine learning approach to forecasting trade flows. Working paper 2179.

Appendices

A The core dataset

Table 5 lists the variables in the core dataset. More details regarding the big dataset are available on request.

Table 5: Variables in the core data set

Variable description	Block of economy	Data source
World trade (*dependent variable*)	Goods trade volumes	CPB-WTM
Imports (advanced economies)	Goods trade volumes	CPB-WTM
Imports (emerging economies)	Goods trade volumes	CPB-WTM
Exports (advanced economies)	Goods trade volumes	CPB-WTM
Exports (emerging economies)	Goods trade volumes	CPB-WTM
World trade price	Goods trade prices	CPB-WTM
Import price (world)	Goods trade prices	CPB-WTM
Import price (advanced economies)	Goods trade prices	CPB-WTM
Fuels (HWWI)	Goods trade prices	CPB-WTM
Primary commodities excl. fuels (HWWI)	Goods trade prices	CPB-WTM
Industrial production excl. construction (world)	Industrial production	CPB-WTM
Industrial production excl. construction (adv. economies)	Industrial production	CPB-WTM
Retail trade (OECD countries)	Retail trade volumes	OECD
Retail trade (euro area)	Retail trade volumes	OECD
Composite leading indicator (OECD countries)	Composite leading indicators	OECD
Composite leading indicator (China)	Composite leading indicators	OECD
Ifo business climate (Germany)	Purchasing managers' indices	CESifo
ISM manufacturing (US)	Purchasing managers' indices	FRED
Brent oil price	Various	Refinitiv Datastream
Baltic exchange dry index	Various	Refinitiv Datastream
MSCI world index	Various	Refinitiv Datastream
World steel production	Various	Refinitiv Datastream
World semiconductor billings	Various	Refinitiv Datastream
WSTS Tech pulse index	Various	FRBSF