Online Appendix

Nowcasting GDP using machine learning methods

by

Dennis Kant Andreas Pick Jasper de Winter

A.1 Data appendix

Table A.2
Macroeconomic time series of various economic indicators transformed into growth rates

No.	predictor	transformation				source	start	end	link	publ. delay
		ln.	dif.	fil.	sa.					
I. Production (N = 21)										
1	Av. daily prod. - Prod. industries	1	1	3	NSA	CBS	jan 1965	jan 2019	link	2
2	Av. daily prod. - Industry	1	1	3	NSA	CBS	jan 1965	jan 2019	link	2
3	Ind. prod. - Cap. goods industry	1	1	3	NSA	ECB	jan 1970	jan 2019	STS.M.NL.W.PROD.NS0050.3.000	2
4	Cons. exp. - Households, dom. cons.	1	1	3	NSA	CBS	feb 1977	jan 2019	link	2
5	Ind. prod. - Manufacture tobacco	1	1	3	NSA	ECB	jan 1965	jan 2019	STS.M.NL.W.PROD.2C1200.3.000	2
6	Ind. prod. - Manufacture wearing apparel	1	1	3	NSA	ECB	jan 1970	jan 2019	STS.M.NL.W.PROD.2C1400.3.000	2
7	Ind. prod. - Manufacture motor vehicles/(semi-)trailers	1	1	3	NSA	ECB	jan 1985	jan 2019	STS.M.NL.W.PROD.2C2900.4.000	2
8	Ind. prod. - Manufacture other transport equipment	1	1	3	NSA	ECB	jan 1970	jan 2019	STS.M.NL.W.PROD.2C3000.3.000	2
9	Ind. prod. - Manufacturing	1	1	3	SA	ECB	dec 1979	jan 2019	STS.M.NL.Y.PROD.2C0000.3.000	2
10	Ind. prod. - Manufacture of textiles	1	1	3	SA	ECB	jan 1980	jan 2019	STS.M.NL.Y.PROD.2C1300.3.000	2
11	Ind. prod. - Printing/reproduction of recorded media	1	1	3	SA	ECB	jan 1980	jan 2019	STS.M.NL.Y.PROD.2C1800.3.000	2
12	Ind. prod. - Constr.	1	1	3	SA	ECB	jan 1985	jan 2019	STS.M.NL.Y.PROD.2F0000.3.000	2
13	Ind. prod. - MIG capital goods ind.	1	1	3	SA	ECB	jan 1970	jan 2019	STS.M.NL.Y.PROD.NS0050.3.000	2
14	Belgium, Retail trade excl. fuel, motor vehicles/cycles	1	1	3	SA	ECB	jan 1970	jan 2019	STS.M.BE.Y.TOVV.NS4701.3.000	2
15	Germany, Total ind. (excl. constr.)	1	1	3	SA	ECB	jan 1965	jan 2019	STS.M.DE.Y.PROD.NS0020.3.000	2
16	Germany, Retail trade excl. fuel, motor vehicles/cycles	1	1	3	SA	ECB	jan 1968	jan 2019	STS.M.DE.Y.TOVV.NS4701.3.000	2
17	Spain, Total ind. (excl. constr.)	1	1	3	SA	ECB	jan 1965	jan 2019	STS.M.ES.Y.PROD.NS0020.3.000	2
18	France, Total ind. (excl. constr.)	1	1	3	SA	ECB	jan 1965	jan 2019	STS.M.FR.Y.PROD.NS0020.3.000	2
19	France, Retail trade excl. fuel, motor vehicles/cycles	1	1	3	SA	ECB	jan 1970	jan 2019	STS.M.FR.Y.TOVV.NS4701.3.000	2
20	Italy, Total ind. (excl. constr.)	1	1	3	SA	ECB	jan 1965	jan 2019	STS.M.IT.Y.PROD.NS0020.3.000	2
21	Germany, Total ind.	1	1	3	SA	ECB	mrt 1978	jan 2019	STS.M.DE.Y.PROD.NS0010.3.000	2
II. Surveys (N = 36)										
22	Prod. conf. - Headline	0	1	3	SA	ES	jan 1985	feb 2019	link	1

Table continued on next page

Macroeconomic time series of various economic indicators transformed into growth rates (*continued)*

No.	predictor	transformation				source	start	end	link	publ. delay
		ln.	dif.	fil.	sa.					
23	Constr. conf. - Headline	0	1	3	SA	ES	jan 1985	feb 2019	link	1
24	Constr. conf. - Building development past 3 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
25	Constr. conf. - Evolution current overall order books	0	1	3	SA	ES	jan 1985	feb 2019	link	1
26	Constr. conf. - Employment expect. next 3 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
27	Ind. conf. - Headline	0	1	3	SA	ES	jan 1985	feb 2019	link	1
28	Ind. Confidence - Production trend observed in recent months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
29	Ind. Confidence - Assessment of order-book levels	0	1	3	SA	ES	jan 1985	feb 2019	link	1
30	Ind. Confidence - Assessment of stocks of finished products	0	1	3	SA	ES	jan 1985	feb 2019	link	1
31	Ind Confidence - Production expectations for the months ahead	0	1	3	SA	ES	jan 1985	feb 2019	link	1
32	Ind. Confidence - Employment expectations for the months ahead	0	1	3	SA	ES	jan 1985	feb 2019	link	1
33	Cons. conf. - Headline	0	1	3	SA	ES	jan 1985	feb 2019	link	1
34	Cons. conf. - Financial situation over last 12 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
35	Cons. conf. - Financial situation over next 12 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
36	Cons. conf. - General economic situation over last 12 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
37	Cons. conf. - General economic situation over next 12 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
38	Cons. conf. - Unemployment expectations over next 12 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
39	Cons. conf. - Major purchases at present	0	1	3	SA	ES	jan 1985	feb 2019	link	1
40	Cons. conf. - Major purchases over next 12 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
41	Cons. conf. - Savings at present	0	1	3	SA	ES	jan 1985	feb 2019	link	1
42	Cons. conf. - Savings over next 12 months	0	1	3	SA	ES	jan 1985	feb 2019	link	1
43	Cons. conf. - Statement on financial situation of household	0	1	3	SA	ES	jan 1985	feb 2019	link	1
44	BNB-indicator, gross-index	0	1	3	SA	BNB	jan 1985	feb 2019	link	1
45	Belgium, Cons. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
46	Germany, Cons. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
47	France, Cons. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
48	Italy, Cons. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
49	Belgium, Ind. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
50	Germany, Ind. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
51	Italy, Ind. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1

Table continued on next page

Macroeconomic time series of various economic indicators transformed into growth rates (*continued)*

No.	predictor	transformation				source	start	end	link	publ. delay
		ln.	dif.	fil.	sa.					
52	United Kingdom, Ind. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
53	United Kingdom, Cons. confidence	0	1	3	SA	ES	jan 1985	feb 2019	link	1
54	Ind. Confidence - (CBS definition)	0	1	3	SA	CBS	jan 1985	feb 2019	link	1
55	Ind. Confidence - Prod. expect, months ahead (CBS definition)	0	1	3	SA	CBS	jan 1985	feb 2019	link	1
56	Ind. Confidence - Ass. of order-book levels (CBS definition)	0	1	3	SA	CBS	jan 1985	feb 2019	link	1
57	Ind. Confidence - Ass. of stocks of fin. products (CBS definition)	0	1	3	SA	CBS	jan 1985	feb 2019	link	1
III. Financial (N = 8)										
58	Loans to the private sector	1	1	3	NSA	ECB	dec 1982	jan 2019	BSI.M.NL.N.M.A20.A.I.U2.2200.Z01.E	2
59	M1	1	2	3	NSA	ECB	jan 1980	jan 2019	BSI.M.NL.N.V.M10.X.1.U2.2300.Z01.E	2
60	M3 (money in circulation inclusive)	1	2	3	NSA	ECB	jan 1970	jan 2019	BSI.M.NL.N.V.M30.X.1.U2.2300.Z01.E	2
61	Interest rate (short term) - euro	0	1	3	NSA	DNB	nov 1984	feb 2019	link	1
62	Loans on mortgage (nominal rate 5 to 10 years mortgage)	0	1	3	NSA	ECB	jan 1980	jan 2019	link	2
63	Interest rate (long term)	0	1	3	NSA	DS	jan 1965	mrt 2019	NLGBD10	0
64	Share index, AEX	1	1	3	NSA	DS	jan 1983	mrt 2019	AMSTEOE	0
65	Share index, Amsterdam Midkap-index	1	1	3	NSA	DS	jan 1983	mrt 2019	AMSMKAP	0
IV. Prices (N = 14)										
66	Exchange rate, US-Dollar per Euro	0	1	3	NSA	ECB	jan 1965	feb 2019	link	1
67	Housing price	1	2	3	NSA	CBS	jan 1976	feb 2019	link	1
68	Consumer-price index, total CPI, all households	1	2	3	NSA	CBS	jan 1965	feb 2019	link	1
69	Consumer-price index, underlying inflation	1	2	3	NSA	CBS	jan 1976	feb 2019	link	1
70	World market commodity prices, overall	1	2	3	NSA	HWWI	sep 1978	feb 2019	link	1
71	World market commodity prices, industrial materials	1	2	3	NSA	HWWI	sep 1978	feb 2019	link	1
72	World market commodity prices, agric. & ind. materials	1	2	3	NSA	HWWI	sep 1978	feb 2019	link	1
73	World market commodity prices, metals	1	2	3	NSA	HWWI	sep 1978	feb 2019	link	1
74	World market commodity prices, energy-components	1	2	3	NSA	HWWI	sep 1978	feb 2019	link	1
75	Producer prices, total intermed. & fi. products (dom. market)	1	2	3	NSA	CBS	jan 1981	jan 2019	link	2
76	Producer prices, consumer goods (dom. market)	1	2	3	NSA	ECB	jan 1976	jan 2019	STS.M.NL.N.PRIN.NS0080.3.000	2
77	Producer prices, intermediate goods (dom. market)	1	2	3	NSA	ECB	jan 1976	jan 2019	STS.M.NL.N.PRIN.NS0040.3.000	2
78	Producer prices, intermediate & final products (for. market)	1	2	3	NSA	CBS	jan 1981	jan 2019	link	2

Table continued on next page

Macroeconomic time series of various economic indicators transformed into growth rates (*continued)*

No.	predictor	transformation				source	start	end	link	publ. delay
		ln.	dif.	fil.	sa.					
79	Producer prices, energy (dom. market)	1	2	3	NSA	ECB	jan 1980	jan 2019	STS.M.NL.N.PRIN.NS0090.3.000	2
V. Other (N = 4)										
80	Unemployment	0	1	3	SA	ES	jan 1983	feb 2019	link	1
81	Issued vehicle registration certificates	1	1	3	NSA	RAI	jan 1965	feb 2019	link	1
82	Bankruptcies	1	1	3	NSA	CBS	jan 1965	feb 2019	link	1
83	Hourly wages (collective labour agreement), industry	1	1	3	NSA	CBS	jan 1972	feb 2019	link	1
Quarterly variables (N = 1)										
84	Gross domestic product (GDP)	1	1	3	SA	CBS	1970Q1	2018Q4	link	3

Note: The table presents the transformations of the monthly series that are used for estimation of forecasting models. Transformation: ln.: 0 = no logarithm, 1 = logarithm; dif.: degree of differencing 1 = first difference, 2 = second difference; fil.: moving average filter of degree n; sa: SA = seasonally adjusted at the source, NSA = not seasonally adjusted, adjusted with X12-ARIMA; source: CBS = Statistics Netherlands, BNB = National Bank of Belgium, DNB = National Bank of the Netherlands, DS: Datastream, ECB: European Central Bank, ES = Eurostat, HWWI = Hamburg Institute of International Economics, RAI = RAI Association; start: Starting year and month of the series, end: Final year and month of the series; code: link = link to the data, code = Refinitiv code (only available for subscribed users) or restricted data portal ECB; publ. delay: publication delay of the series in months.

A.2 Details of the nowcasting models

Dynamic factor model

Consider a vector of n stationary monthly series $\boldsymbol{x}_m = (x_{1,m}, \ldots, x_{n,m})'$, with monthly time index $m = 1, 2, \ldots, T_m$, which have been standardized to have zero mean and unit variance. The dynamic factor model is

$$\begin{aligned} \boldsymbol{x}_m &= \boldsymbol{\Lambda} \boldsymbol{f}_m + \boldsymbol{\xi}_m, \quad \boldsymbol{\xi}_m \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\xi) \\ \boldsymbol{f}_m &= \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{f}_{m-i} + \boldsymbol{B} \boldsymbol{\eta}_m, \quad \boldsymbol{\eta}_m \sim N(\boldsymbol{0}, \boldsymbol{I}_q) \end{aligned}$$

where $\boldsymbol{f}_m$ is a $q \times 1$ vector of factors, $\boldsymbol{\Lambda}$ is a $n \times q$ matrix of factor loadings, $\boldsymbol{A}_i$ is a $q \times q$ matrix of coefficients, and $\boldsymbol{\xi}_m$ and $\boldsymbol{\eta}_m$ are $n \times 1$ and $q \times 1$ vectors of disturbances.

The latent monthly GDP growth, y^*_m, is related to the common factors through

$$y^*_m = \boldsymbol{\lambda}' \boldsymbol{f}_m + \varepsilon_m, \quad \varepsilon_m \sim \mathcal{N}(0, \sigma^2_\varepsilon)$$

where $\boldsymbol{\lambda}$ is the vector of loading coefficients of the factors on latent GDP growth. The observed quarterly GDP growth series, y_t, with quarterly time index $t = 1, 2, \ldots, T_q$, is then

$$y_t = (y^*_{3t} + y^*_{3,t-1} + y^*_{3,t-2})/3$$

The aggregation for the quarterly GDP growth implies that y_m is in terms of 3-month growth rates. The state space form contains the monthly quarterly GDP growth in the third month of the respective quarter with the remaining observations treated as missing

$$y_m = \begin{cases} y_{3t}, & t = 1, 2, \ldots, T \\ \text{unobserved}, & \text{otherwise} \end{cases}$$

The literature estimates the matrix of factor loadings, $\boldsymbol{\Lambda}$, via a static principal components analysis applied to a balanced sub-sample of the data, where observations in periods with missing data are discarded. In our data set, however, only the rows of the first few observations are discarded, which are missing as the result of vertical alignment due to publication lags. The static principal component analysis also gives sample estimates of the common factors.

The number of common factors and the number of lags in the vector autoregressive process need to be specified. We take the equally weighted average of forecasts over a range of values with the maximum value of q and p set to six, see Kuzin et al. (2013) and Jansen et al. (2016) for similar choices.

Mixed-data sampling factor-augmented model

The MIDAS model of Ghysels et al. (2007) has been adapted for nowcasting by Marcellino and Schumacher (2010). The purpose of MIDAS is to jointly model predictors of different frequency, in this case quarterly GDP and monthly economic indicators. In the factor-augmented MIDAS model, factors are extracted at the monthly frequency and then linked to lower frequency GDP growth. The model for h-period ahead GDP growth, y_{t+h}, in this model is

$$y_{t+h} = \alpha + \boldsymbol{\beta}' C(\boldsymbol{\theta}) \boldsymbol{f}_t^{(3)} + \varepsilon_{t+h}$$

where α is a scalar, $\boldsymbol{f}_t^{(3)}$ the skip-sampled factors extracted from the monthly indicators , where the superscript three indicates the skip sampling of monthly indicators to quarterly frequency. Various specifications of nonlinear weighting schemes $C(\boldsymbol{\theta})$ can be employed to parsimoniously parameterize the coefficients (Ghysels et al., 2007).

The mixed-data sampling model is estimated with ordinary or nonlinear least squares for the unrestricted or restricted model. The restricted model uses the exponential Almon lag and the unrestricted model uses skip sampling. We obtain estimates of the factors via principal components on a skip sampled data set including lags of the predictors.

Regularization techniques

We use the least absolute shrinkage and selection operator (LASSO) and the elastic net in this paper. We also obtained results for ridge regression and adaptive LASSO. However, the results were strictly dominated by the LASSO and elastic net and for brevity we therefore omit these results.

The LASSO of Tibshirani (1996) performs both regularization and predictor selection by imposing an ℓ_1 penalty in the estimation of the coefficients. As the response predictor and predictors are of different frequencies, the mixed-data sampling approach of Section 2.1 is employed using skip sampling for the monthly predictors.

Given a sample of length N consisting of n covariates $x_m := (x_{1,m}, x_{2,m}, \ldots, x_{n,m})$, $\forall m \in \{1, 2, \ldots, T_m\}$, one obtains the parameter estimates optimizing the penalized loss function

$$\min_{\beta_0, \boldsymbol{\beta}} \|\boldsymbol{y}_h - \beta_0 \iota_{T-h} - \boldsymbol{x}^{(3)} \boldsymbol{\beta}\|_2^2 \quad \text{subject to } \|\boldsymbol{\beta}\|_1^2 \leq \lambda$$

where $\mathbf{y_h}$ is GDP growth for forecasting horizon h, ι_N an $N \times 1$ vector of ones, $\mathbf{x^{(3)}}$ a matrix of the skip-sampled versions of $x_{i,m}$, $\boldsymbol{\beta}$ a vector of coefficients, $\|\cdot\|_p$ denotes the ℓ_p, and ι_{T-h} an $T - h \times 1$ vector of ones and λ determines the extent of regularization. The optimal regularization parameter, λ, is determined via cross-validation.

The elastic net of Zou and Hastie (2005) imposes a combination of ℓ_1 and ℓ_2 penalties. Similar to the LASSO, the ℓ_1 norm selects parameters by shrinking some to zero but it also shrinks the remaining coefficients towards zero through the use of the ℓ_2 norm. The elastic net regression is

$$\min_{\beta_0,\boldsymbol{\beta}} \|\boldsymbol{y}_h - \beta_0 \iota_{T-h} - \boldsymbol{x}^{(3)}\boldsymbol{\beta}\|_2^2 \quad \text{subject to } \alpha\|\boldsymbol{\beta}\|_1^2 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2 \leq \lambda$$

α determines the relative extent of regularization performed by both norms and is determined via cross-validation.

For both, LASSO and elastic net, we determine hyperparameters via leave-one-out cross-validation. In this scheme, one observation in the estimation sample is removed from the estimation and then predicted. This is repeated 500 times and the hyperparameters that minimise the average square forecast loss are chosen.

Random subspace regression

Model averaging has been shown to reduce the MSFE. Based on this observation, Elliott et al. (2013) introduce complete subset regression, where forecasts are constructed from all combinations of k predictors out of the variable pool. The forecasts are then averaged. If, however, the predictor pool is large the number of combinations of k predictors is prohibitively large. A solution is to take R randomly chosen subsets of predictors. Boot and Nibbering (2019) show that this approximates the complete subset regression for mildly large R, such as $R = 1000$.

In the nowcasting context, the regression is

$$y_{t+h} = \boldsymbol{x}_t^{(3)}\boldsymbol{R}\boldsymbol{\beta}_{\boldsymbol{R}} + \varepsilon_{\boldsymbol{R},t+h}$$

where $\boldsymbol{R}$ is an $K \times k$ random selection matrix, $\boldsymbol{\beta}_R$ the associated $k \times 1$ vector of coefficients. More specifically, $\boldsymbol{R}$ is a random selection matrix that selects random sets of k predictors out of the total available n predictors, that is, it is a matrix of zeros except for k elements: the j,l-th element, which is unity if the l-th predictor in the random subset regression is the j-th predictor.

A tuning parameter of this method is the size of each predictor subset, k. Theoretical results by Boot and Nibbering (2019) suggest that k should be chosen relatively large at about 30. The experience of Pick and Carpay (2022) suggests that smaller k can deliver more precise forecasts. We initially experimented with different choices of k up to 30 and our experience confirms that smaller choices of k deliver better nowcasts. As a result we average nowcasts over those obtained using $k = 2, 3, 4, 5$.

An alternative to selecting predictors would be to combine the predictors with random weights. Boot and Nibbering (2019) discuss this option and name it random projection. In place of a selection matrix, random projection uses a random weighting matrix, that calculates k predictors that are

weighted averages of the n predictors. For Gaussian random projections, the weights are drawn from a normal distribution and each entry of the matrix $\boldsymbol{R}$ is independently and identically distributed as

$$\big[R\big]_{i,j} \sim \mathcal{N}(0,1), \quad 1 \leq i \leq n, \quad 1 \leq j \leq k$$

Multiple realizations of the random matrix R are drawn and the resulting forecasts are averaged. Again, the choice predictor k needs to be determined. Again, we average nowcasts over those obtained using $k = 2, 3, 4, 5$.

Random forest

The random forest forecast averages the forecasts of multiple decision trees. To grow a decision tree, the space of predictor values is partitioned with the aim of minimizing the in-sample squared error. At each partition, the algorithm chooses a split based on one of the predictors that realizes the largest decrease in squared error. Hence, the split of a skip-sampled predictor that is minimizing the cost function is chosen at each node, i.e.

$$C = \sum_{R_g} \sum_{t_j \in R_g} (\bar{y}_{R_g} - y_j)^2$$

where C is the cost to be minimized, R_g for $g \in [1, \ldots, G]$ is the set of partitioned responses, $\bar{y}_{R_g}$ is the average GDP realization within cluster R_g and y_j is the j^{th} element of partition R_g. A tree is therefore a nonlinear combination of the predictors and allows for a nonlinear underlying GDP.

Trees are designed to have a high degree of independence of each other by randomly drawing a subset of predictor predictors and a subset of observations to grow any given tree. Averaging the forecasts from the trees in the random forest therefore minimizes the variance of the average forecast.

In order to reduce overfitting, each estimation sample is divided in a training and a validation set. The share of the training set in all estimation samples is varied such that ω of the estimation sample is assigned to the training set, with $\omega \in \{0.6, 0.7, 0.8, 0.9\}$ and choose $\kappa \in [1, 2, \ldots, 249]$ skip-sampled predictors to split the tree. Subsequently, a validation set is used to measure the performance for each κ. We use the prediction of the 400 trees that resulted in the lowest prediction error in the training set.