Empirical Economics (2021) 61:667–692
https://doi.org/10.1007/s00181-020-01891-7

Horizon confidence sets

Jack Fosten[1] · Daniel Gutknecht[2]

Received: 7 May 2019 / Accepted: 20 May 2020 / Published online: 11 June 2020

Abstract

This paper introduces a new statistical procedure to discriminate between competing forecasting models at different forecast horizons. Unlike existing tests, which eliminate a model from *all* horizons if dominated according to some loss measure, our methodology identifies an ‘optimal’ set of models *at each* horizon, retaining a model that performs well at a given horizon even if dominated at others. While our method is especially useful in applications to long-term forecasting as well as short-term nowcasting, it can also be applied in wider settings like the comparison of forecasting models across countries. We conduct a small Monte Carlo study to investigate the finite sample properties and apply our procedure to nowcasting US real GDP growth and its subcomponents, comparing a factor-based nowcasting method to a naïve benchmark. Unlike existing methods, ours can formally detect the point in the quarter at which the factor method beats the benchmark or vice versa.

Keywords Nowcasting · Multiple model comparison · Model confidence set · Bootstrap

JEL Classification C12 · C22 · C52 · C53

This paper was previously circulated under the title ‘Model Confidence Sets for Nowcast Procedures’. We are grateful for the comments of Tim Robinson and Viet Nguyen as well as participants of the 2018 Econometric Society Australasian Meeting (ESAM), which have helped to improve this paper. Our thanks go to Shaoni Nandi for the excellent research assistance on this project.

✉ Jack Fosten
jack.fosten@kcl.ac.uk

Daniel Gutknecht
Gutknecht@econ.uni-frankfurt.de

[1] King’s Business School, King’s College London, London WC2B 4BG, UK

[2] Faculty of Economics and Business, Goethe University Frankfurt, 60629 Frankfurt am Main, Germany

1 Introduction

Forecasters are often interested in the performance of econometric models at forecasting different horizons into the future. However, as soon as we compare two or more competing models across many horizons, challenging questions arise. Is there a single best model at every horizon? Do models all perform equally well at all horizons? Is there a different optimal model at different horizons? Many empirical studies in different settings have suggested that the best predictive model does, indeed, change with horizon. For instance, in long-run exchange rate forecasting, the survey of Rossi (2013) finds that univariate models dominate at short horizons while models with economic fundamentals are best at long horizons. In short-term nowcasting, where horizons are typically thought of in terms of days before the release of gross domestic product (GDP) data, studies such as Banbura et al. (2013) find that big data methods like factor models only dominate when the daily nowcast horizon is small enough to allow a lot of relevant information to become available.

In this paper, we propose a statistical procedure which addresses the above questions, and obtains the collection of models which dominate at different horizons with a given level of confidence. We coin the term *Horizon Confidence Set* to denote the collection of 'optimal horizon-specific models'. Our approach is based on the model confidence set (MCS) procedure developed by Hansen et al. (2011), but is modified to operate over multiple horizons on the same set of models. Specifically, we compute Diebold–Mariano t-statistics (Diebold and Mariano 1995) for equal predictive ability (EPA) to compare two competing models in each of the horizons. Then we propose an elimination rule based on the maximal t-statistic which removes a model from a specific horizon if its p-value falls below the nominal level. Unlike existing procedures, our methodology therefore does not operate as a 'horse race'-type test identifying only the dominant model overall, but retains the optimal set of models across all horizons. It does so while guarding against the multiple testing problem which occurs by testing across horizons. In the 'Appendix', we generalize our procedure to allow more than two models at each horizon.

Besides the multi-horizon context of forecasting and nowcasting, our procedure can also be applied to other settings: for instance, replacing 'horizon' with 'country', the methodology allows the comparison of two competing models across countries, retaining possibly a different model for different countries. Alternatively, taking the exchange rate forecasting example again, dating back to Mark (1995) it has become custom to compare the predictive ability of the different exchange rate models for different currency pairs with the US dollar. Our method could therefore be used to perform this cross-currency comparison, instead of the multi-horizon aspect.

The horizon confidence set procedure proposed in this paper differs from the original MCS procedure of Hansen et al. (2011) in two important ways. Firstly, one could consider directly applying the MCS procedure to *all* model-horizon pairs jointly. However, this procedure would potentially eliminate all models from a single horizon and would also involve computing unfair comparisons of, say, model A at horizon x to model B at horizon y. Secondly, one might consider applying the MCS procedure to *each* horizon independently. However, by not guarding against the multiple testing issue across horizons, this may produce too many false positives and provide the researcher with a sparser set of models than is statistically justified. Such ‘sparsity’ cannot arise in our case as models are removed at a given horizon only when the performance is worse relative to comparisons from other horizons. Moreover, note that standard Bonferroni-type methods are typically not advisable in many of the settings we consider as they become too conservative when the number of horizons is large.

Our method also differs from other procedures which are in principle applicable to multiple models at multiple horizons, such as the MCS procedure based on the concept of *uniform* and *average* superior predictive ability (SPA) proposed by Quaedvlieg (2020). In his procedure, Quaedvlieg (2020) aims to detect the model(s) which either strictly dominate the competitor models (uniform SPA) or which exhibit the best average performance (average SPA) across all horizons. Though of separate interest, in the case where the ‘optimal’ set of models changes across horizons, uniform SPA would fail to provide a conclusive answer, while average SPA may lead us to retain all models in all horizons, even if models are dominated at specific horizons. On the contrary, our procedure is able to potentially identify this changing pattern of ‘optimal’ models across horizons.[1]

We will apply our methodology to short-term nowcasting, which we consider to be a leading case. Our method is complementary to various existing nowcasting papers which have tended to shut down one of the two channels of multiple testing in model evaluation. On the one hand, some studies have performed tests for nowcast evaluation on a pair of models at single nowcast horizons (for example Giannone et al. 2016), whereas other studies focus on the performance of a single nowcasting model across many nowcast horizons (Fosten and Gutknecht 2020). To further add to the literature of nowcasting, we outline how our method can be helpful in performing nowcast combination at different release dates and demonstrate how this nowcast combination approach using our confidence set output can be used to test nowcast monotonicity (see also Banbura et al. 2013; Aastveit et al. 2014; Knotek and Zaman 2017).

More specifically, our empirical application looks at the factor model method used by Bok et al. (2018) in making the *New York Fed Staff Nowcasts*. We extend their analysis to consider nowcasts of the five subcomponents of US real GDP as well as aggregate GDP. We compare the nowcasts of this factor method to a simple autoregressive benchmark across the different GDP subcomponents. This builds on existing empirical nowcasting studies, including Marcellino and Schumacher (2010), Banbura et al. (2013), Luciani and Ricci (2014), Foroni and Marcellino (2014), Aastveit et al. (2014, 2017), Foroni et al. (2015), Antolin Diaz et al. (2017), Kim and Swanson (2018) and McCracken et al. (2019). As a preview of the results, our procedure does not find any evidence of substantial differences between the factor method and the benchmark for aggregate GDP growth or consumption growth. On the other hand, in subcomponents like investment and government spending, we are able to determine the point in the nowcast period at which the factor method beats the benchmark, or vice versa. This finding demonstrates how our method improves over the use of average

[1] To give an example, we believe that short-termist nowcast users would prefer the method which produces the best nowcast at the time they need it, rather than one which performs well on average across all horizons.

or uniform SPA which would have forced us to reject or retain the models across all horizons.

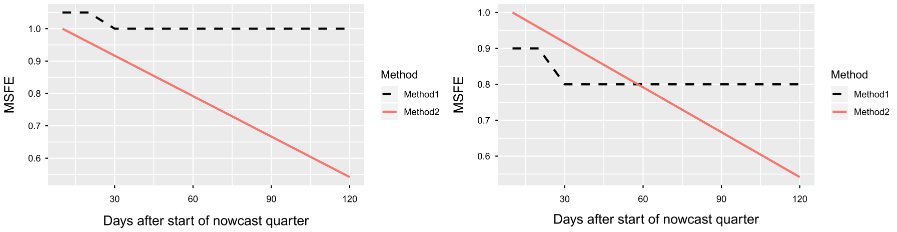
The rest of the paper is divided as follows. Section 2 describes the horizon confidence set procedure, and Sect. 3 provides some further uses of this methodology for practitioners. Section 4 contains a small Monte Carlo study to assess the procedure's small sample behaviour, while Sect. 5 is the empirical application to nowcasting US real GDP. Finally, Sect. 6 concludes the paper. Additional figures and the extension to multiple models are given in the Appendix.

2 The horizon confidence set

In what follows, we outline the set-up and details of our horizon confidence set approach. We are interested in predicting a target variable y_t, for which we have observations $t = 1, \ldots, T$. As stated above, for tractability we will take the simplest possible modelling set-up where we wish to compare $M = 2$ models, which we collect into the set $\mathcal{M}^0 = \{1, 2\}$, over a set of $h = 1, \ldots, H$ different horizons. In the 'Appendix', we will set out how this extends so that $\mathcal{M}^0$ contains more than two models. We note that these horizons can be as in the traditional multi-step forecasting sense where we make forecasts of y_{t+h} at increasing horizons for $h = 1, \ldots, H$. Alternatively, in the near-term nowcasting literature, where we nowcast y_t (at a quarterly horizon of zero), the term horizon usually refers to daily, weekly or monthly horizons throughout the nowcast quarter at which we make predictions before the release date of the target variable.

In order to compare the predictions from different methods, we will use the losses of each model computed at each of the H horizons. We define the loss $L^h_{i,t}$ to be the loss of model $i \in \mathcal{M}^0$ at horizon h in period t. With squared error loss, for instance, we obtain the commonly used mean squared forecast error (MSFE) losses. In order to illustrate this set-up, Fig. 1 shows two examples to visualize the kind of loss behaviour one might observe in practice.

The examples in Fig. 1 mimic a typical case in nowcasting, where we compare a big data factor method (Method 2) which is regularly updated at many nowcast horizons,



(a) Example 1 **(b)** Example 2

Fig. 1 Illustrative example of MSFE

with an autoregressive model (Method 1) which is only updated once in the quarter.[2] We will explore this type of comparison in our empirical illustration later. Figure 1 is useful to show how different patterns of optimal models can arise, even in a two-model multi-horizon setting. On the one hand, in Example 1 in Fig. 1a we see that Method 2 dominates Method 1 in all horizons and we want a statistical procedure which is capable of detecting this. On the other hand, in Example 2 in Fig. 1b we see that Method 2 only dominates Method 1 in the second half of the horizons. In this case, we want a procedure which can formally detect at which point Method 2 improves over Method 1.

We define $d^h_{ij,t} = L^h_{i,t} - L^h_{j,t}$ to be the loss differential between models i and j in time period t and horizon h.[3] Moreover, we define its expectation to be $\mu^h_{ij} = \mathrm{E}[d^h_{ij,t}]$. Note that we generally expect the pairs of loss differentials to be correlated across time t since one (if not both) of the models in $\mathcal{M}_0$ will be (dynamically) misspecified. At each horizon $h = 1, \ldots, H$, the horizon confidence set is defined as:

$$\mathcal{M}^*_h \equiv \left\{ i \in \mathcal{M}^0 \colon \mu^h_{ij} \leq 0 \quad \forall j \in \mathcal{M}_0 \right\}. \tag{1}$$

This gives us the identity of the model or models at each horizon which are weakly dominant. The collection of horizon confidence sets $\{\mathcal{M}^*_h\}^H_{h=1}$, where each $\mathcal{M}^*_h$ is determined according to Eq. (1), fully describes which models should be used throughout the whole sequence of horizons. Note that this differs from the MCS procedures proposed in Quaedvlieg (2020) and Hansen et al. (2011), which can be compared best in terms of the differing null hypothesis being tested (see below). Thus, our idea is to arrive at a (different) subset of $\mathcal{M}^0$ at each horizon by eliminating any model which is inferior to the other model at that same horizon. This is accomplished by testing a sequence of null hypotheses $H_{0,h}$, $h = 1, \ldots, H$, given by:

$$H_{0,h} \colon \mu^h_{ij} = 0 \tag{2}$$

Here, each $\mathcal{M}_h$, $h = 1, \ldots, H$, is a subset of $\mathcal{M}^0$, and may in fact be different across h depending on which model survives elimination in specific horizons in the sequential procedure we outline further below. As mentioned above, this differs conceptually from Hansen et al. (2011), whose direct translation into a multi-horizon framework (HLN-M) would be based on testing the following null:

$$H^{\text{HLN-M}}_0 \colon \mu^{hl}_{ij} = 0 \quad \text{for all} \quad i, j \in \mathcal{M}, h, l \in \{1, \ldots, H\},$$

where $\mu^{hl}_{ij} \equiv \mathrm{E}[L^h_{i,t} - L^l_{j,t}]$ and $\mathcal{M}$ denotes a generic subset of $\mathcal{M}_0$ containing models *across* horizons $1, \ldots, H$. By contrast, the concepts of uniform and average

[2] We depict Method 2 as downward-sloping as the majority of studies which nowcast real GDP growth find that MSFE declines as we approach the publication date when a suitable big data method is used (see for example, Banbura et al. 2013 or Fosten and Gutknecht 2020).

[3] Note that we maintain the more general i, j notation (rather than 1 and 2) to facilitate the outline to multiple models in the Appendix.

SPA (uSPA and aSPA, respectively) as defined by Quaedvlieg (2020) condense to testing the nulls:

$$H_0^{\text{uSPA}}\colon \mu_{ij}^h = 0 \quad \text{for all} \quad i, j \in \mathcal{M}_h, h \in \{1, \dots, S\},$$

and

$$H_0^{\text{aSPA}}\colon \sum_{h=1}^{H} \omega_h \mu_{ij}^h = 0 \quad \text{for all} \quad i, j \in \mathcal{M},$$

where ω_h, $h = 1, \dots, H$ denote predetermined weights. Thus, the main conceptual difference between our approach and the multi-horizon version of Hansen et al. (2011) as well as uSPA and aSPA of Quaedvlieg (2020), respectively, is that our *horizon-specific* tests which are carried out focusing exclusively on comparisons at *each* horizon h, while both Hansen et al. (2011) and Quaedvlieg (2020) compare models *across* horizons dates h and l, for $h, l \in \{1, \dots, H\}$. As outlined in the Introduction, restricting ourselves to 'within-horizon' tests avoids unfair comparisons of say model 1 at horizon h with model 2 at horizon l. It also allows to retain models that outperform at specific horizons, but underperform at others (and could thus be eliminated by concepts such as uSPA or aSPA). Finally, note that an alternative application of Hansen et al. (2011) could be to apply their procedure to each horizon separately. This amounts to testing $H_{0,h}$, $h = 1, \dots, H$, for each h as in our case. However, this independent procedure does not take account of the issue of multiple testing which occurs when we compare models across horizons. We therefore expect this method to over-reject the null, whereas we expect our method to be better able to guard against over-rejections. Our Monte Carlo simulations in Sect. 4 further explore this point.

An alternative way of writing our null hypotheses in Eq. (2) is to write a horizon-stacked version of Hansen et al. (2011):

$$\mathbf{H}_{0,H}\colon \begin{array}{l} \mu_{ij}^1 = 0 \\ \vdots \\ \mu_{ij}^H = 0 \end{array} \tag{3}$$

The alternative hypothesis, denoted by $\mathbf{H}_{A,H}$, is that $\mu_{ij}^h \neq 0$ for at least some $i, j \in \mathcal{M}_h$, $h \in \{1, \dots, H\}$. To implement the horizon-specific MCS procedure, we require an equivalence test and a corresponding elimination rule (see Hansen et al. 2011). Unlike in the original paper, however, both must be adapted so that they operate on specific horizons. Let $\boldsymbol{\delta}_{H,\mathcal{M}}$ be the equivalence test which is used to test the hypothesis $\mathbf{H}_{0,H}$ for any $\mathcal{M}_h \subset \mathcal{M}^0$ and $h = 1, \dots, H$, and let $\boldsymbol{e}_{H,\mathcal{M}}$ be the elimination rule which removes an object i from one of the sets $\mathcal{M}_h$, $h \in \{1, \dots, H\}$. More specifically, for $\boldsymbol{\delta}_{H,\mathcal{M}}$, a natural mapping of the null hypothesis is:

$$\mathcal{T}_{H,\mathcal{M}} = \max_{1 \leq h \leq H} \mathcal{T}_h, \tag{4}$$

where $\mathcal{T}_h$ is an $(H \times 1)$ vector with elements $\mathcal{T}_h \equiv \max_{i,j\in\mathcal{M}_h} |t_{ij}^h|$. Here, t_{ij}^h can either denote Diebold–Mariano t-statistics (Diebold and Mariano 1995) of the form:

$$t_{ij}^h = \frac{\bar{d}_{ij}^h}{\sqrt{\widehat{V}(\bar{d}_{ij}^h)}},$$

where $\bar{d}_{ij}^h$ and $\widehat{V}(\bar{d}_{ij}^h)$ are the estimated mean loss differential of models i, j at horizon h and its estimated variance (see below), or simply a non-studentized statistic $t_{ij}^h = \bar{d}_{ij}^h$.

The use of non-studentized statistics (along with an appropriate bootstrap procedure) has been seen in papers such as White (2000). Moreover, standard HAC estimators for the variance may suffer from size distortions in small samples unless appropriately corrected (see Coroneo and Iacone 2019, and references therein). On the other hand, as argued by Romano and Wolf (2005) in the context of superior predictive ability testing, studentization may have favourable properties in terms of improving the power in finite samples. We therefore proceed in a general way allowing for t_{ij}^h to be either a studentized or a non-studentized statistic at this stage, although we will restrict ourselves to non-studentized statistics later on in the empirical section.

In order to construct $\bar{d}_{ij}^h$ and $\widehat{V}(\bar{d}_{ij}^h)$, respectively, we use a sample of T time series observations. However, while $\bar{d}_{ij}^h$ and $\widehat{V}(\bar{d}_{ij}^h)$ typically depend on parameters which need to be estimated, we abstract from the parameter estimation problem in this context to avoid additional notation. From a technical perspective, this may be motivated by the fact that when the sample is split into sub-samples of sizes R and P, where R is the number of observations retained for parameter estimation and P the number of observations used for pseudo-out-of-sample forecasts, a condition such as $\lim_{T\to\infty} (P/R) = 0$ is sufficient to make parameter estimation error in the Diebold–Mariano test negligible asymptotically, or that the in-sample and out-of-sample loss function are the same (cf. West 1996).[4]

At each period $t = 1, \ldots, T$, we define $\bar{d}_{ij}^k$ as:

$$\bar{d}_{ij}^k = \frac{1}{T}\sum_{t=1}^{T} d_{ij,t}^k,$$

while an appropriate and consistent HAC-type estimator $\widehat{V}(\bar{d}_{ij}^h)$ can be used for the long-run variance of $d_{ij,t}^h$ (see Newey and West 1987) to account for the (potential) serial correlation in the loss differential due to model misspecification.

Intuitively, the test statistic in (4) picks out the largest gap across all pairwise model comparisons for each horizon given the surviving set of models in each $\mathcal{M}_h$, $h \in \{1, \ldots, H\}$, and then chooses the largest such deviation across horizons. Letting

[4] Note that when models are nested and the set-up of West (1996) is not directly applicable, this argument still holds when restricting oneself to models estimated through ordinary least squares (cf. Clark and McCracken 2005). This rules out certain model classes estimated for instance via nonlinear least squares such as MIDAS nowcasting models with a nonlinear weighting function for the lags, whereas unrestricted MIDAS models are permitted (as used in the empirical illustration in Sect. 5).

$\mathcal{H}$ denote the set of all horizons $1, \ldots, H$, the procedure is completed by using the following elimination rule:

$$\boldsymbol{e}_{H,\mathcal{M}} = \arg \max_{i \in \mathcal{M}_h ; h \in \mathcal{H}} \max_{j \in \mathcal{M}_h} t_{ij}^h. \tag{5}$$

This rule eliminates model i from the horizon h, identified by taking the arg max over both h and i.[5] We are now ready to state the horizon confidence set algorithm:

1. Initially set $\mathcal{M}_h = \mathcal{M}^0$ for all $h = 1, \ldots, H$
2. Test $\mathbf{H}_{0,H}$ using $\boldsymbol{\delta}_{H,\mathcal{M}}$ at level α
3. If $\mathbf{H}_{0,H}$ is accepted, let $\{\widehat{\mathcal{M}}_{h,1-\alpha}^*\}_{h=1}^H = \{\mathcal{M}_h\}_{h=1}^H$, otherwise use $\boldsymbol{e}_{H,\mathcal{M}}$ to eliminate an object from the relevant $\mathcal{M}_h$, leaving the remaining $\{\mathcal{M}_j\}_{j \neq h}$ as they were, and repeat from Step 2.

We refer to $\widehat{\mathcal{M}}_{h,1-\alpha}^*$ as the estimated horizon confidence set, and the collection $\{\widehat{\mathcal{M}}_{h,1-\alpha}^*\}_{h=1}^H$ determines the full set of models to be used at every horizon. Given that the equivalence test $\mathcal{T}_{H,\mathcal{M}}$ and the elimination rule $\mathbf{e}_{H,\mathcal{M}}$ adhere to the definition of coherency of Hansen et al. (2011), the asymptotic validity of $\boldsymbol{\delta}_{H,\mathcal{M}}$ and $\boldsymbol{e}_{H,\mathcal{M}}$ follows by arguments similar to Hansen et al. (2011).

Since the asymptotic distribution of the test based on Diebold–Mariano statistics depends on unknown nuisance parameters, a bootstrap procedure will be used for inference. More specifically, letting $\mathbf{d}_{ij,t}$ denote an H-dimensional vector containing $d_{ij,t}^h$, $h = 1, \ldots, H$ as elements, we can construct bootstrap samples as follows. For each $b = 1, \ldots, B$:

- Re-sample blocks of length l from $\mathbf{d}_{ij,t}$, $t = 1, \ldots, T$, $i, j \in \mathcal{M}_h$, with replacement using the moving block bootstrap of Künsch (1989). Call these draws $\mathbf{d}_{ij,t}^b$ with elements $d_{ij,t}^{hb}$, $h = 1, \ldots, H$.
- Construct $\bar{d}_{ij}^{hb} = \frac{1}{T}\sum_{t=1}^T d_{ij,t}^{hb}$, $h = 1, \ldots, H$, and, when the studentized version of t_{ij}^h is used, the bootstrap variance is estimated according to Götze and Künsch (1996) and Gonçalves and White (2004):

$$\widehat{V}_{ij}^{hb} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{l}\left\{\sum_{t=1}^{l}\left(d_{ij,(q-1)l+t}^{hb} - \bar{d}_{ij}^{hb}\right)\right\}^2,$$

where Q denotes the number of blocks and q is the corresponding counter (for simplicity, assume that $T = Q \cdot l$).

[5] There are cases in nowcasting with more than two models, as set out in the Appendix, where it is possible that some of the models are not updated at every horizon (e.g. if a nowcasting model contains only one monthly variable, it is potentially only updated at three horizons). They therefore have fixed losses over a set of horizons $\mathbf{h} \subset \mathcal{H}$. This makes it possible that the maximal Diebold–Mariano statistic is identical for multiple horizons between infrequently updated models. In this case, the elimination rule is adapted to eliminate the weaker model from all horizons $\mathbf{h} \subset \mathcal{H}$ in which the maximal statistic occurs and not just a single horizon.

- Construct the statistic:

$$t_{ij}^{hb} = \frac{\bar{d}_{ij}^{hb} - \bar{d}_{ij}^{h}}{\sqrt{\widehat{V}_{ij}^{hb}}} \quad \text{or} \quad t_{ij}^{hb} = \bar{d}_{ij}^{hb} - \bar{d}_{ij}^{h}, \quad h = 1, \ldots, H,$$

  and obtain:

$$\mathcal{T}_{H,\mathcal{M}}^{b} = \max_{1 \leq h \leq H} \max_{i,j \in \mathcal{M}_h} |t_{ij}^{hb}|.$$

Finally, construct the α-critical value from the empirical distribution $\{\mathcal{T}_{H,\mathcal{M}}^{1}, \ldots, \mathcal{T}_{H,\mathcal{M}}^{B}\}$, say $c(\alpha)$. Rejection occurs when $\mathcal{T}_{H,\mathcal{M}} > c(\alpha)$.

3 Extensions to the horizon confidence set

There are several ways in which the horizon confidence set might be extended and used for further analysis. In this section, we shine particular light on two such extensions. Firstly, we might want to perform model averaging at the various horizons, thereby extending the original suggestion of Hansen et al. (2011) to the multi-horizon setting. Secondly, in the nowcasting case we might use these results to test for nowcast monotonicity, which has become a common criterion used to check whether nowcasting methods improve as we add information (Banbura et al. 2013).

3.1 Model averaging

The horizon confidence set procedure outlined above gives rise to a set of models which are to be used at different horizons. Asymptotically speaking, one should in principle be 'indifferent' between the model(s) included at a given horizon. Depending on the scenario, this could potentially lead to cases where individual models move in and out as the horizon changes and the number of models used in each horizon could change repeatedly. If there is more than one optimal model at different horizons, it may be operationally preferable to just form averages constructed across the different $\widehat{\mathcal{M}}_{h,1-\alpha}^{*}$, $h = 1, \ldots, H$. For example, for every model i and horizon h, one could form simple averaging weights from the non-eliminated models as follows:

$$\widehat{w}_{ih} = \frac{\mathbb{I}\{i \in \widehat{\mathcal{M}}_{h,1-\alpha}^{*}\}}{|\widehat{\mathcal{M}}_{h,1-\alpha}^{*}|} \tag{6}$$

where $|\widehat{\mathcal{M}}_{h,1-\alpha}^{*}|$ denotes the number of models in $\widehat{\mathcal{M}}_{h,1-\alpha}^{*}$ and the indicator function $\mathbb{I}\{\cdot\}$ returns a value of 1 if model i is included in $\widehat{\mathcal{M}}_{h,1-\alpha}^{*}$ and zero otherwise. The nowcast combination is then calculated as:

$$\widehat{\widehat{y}}_{ht} = \sum_{i=1}^{M} \widehat{w}_{ih} \widehat{y}_{iht}, \tag{7}$$

where $\widehat{y}_{iht}$ is method i's prediction at horizon h in quarter t. In the case of nowcasting, this procedure gives an alternative to recent nowcast averaging procedures such as in Aastveit et al. (2018) where the weights are estimated directly. We also note that the resulting nowcasts retain the same asymptotic 'optimality' as the individual models from the collection $\{\mathcal{M}_h^*\}_{h=1}^H$ as they are just a linear combination of optimal models. Unlike using the individual models themselves, however, this simplification of the horizon confidence set not only allows us to reduce the complexity of the nowcasting procedure, but also to perform further specification tests such as the monotonicity test outlined next.

3.2 Monotonicity testing

When looking at nowcasting, in addition to the selection of relevant nowcast models, an important consideration in the empirical literature is whether or not nowcast methods are monotonically improving as we add information, Banbura et al. (2013) is one example. In the presence of more than one model, however, there is little guidance on how to perform monotonicity tests, with the recently proposed test of Fosten and Gutknecht (2020) being established in the single-model case. One appeal of the averaging approach from the previous subsection is that it results in a single nowcast from the different models and allows us to perform this monotonicity test as if it were a single model. More specifically, suppose one constructs the nowcast combination $\widehat{\widehat{y}}_{ht}$ as in Eq. (7) and the nowcast errors $\widehat{\widehat{\varepsilon}}_{ht} = y_t - \widehat{\widehat{y}}_{ht}$. Then, these errors can be used to construct a monotonicity test to assess whether the losses from these multiple models are (on average) declining over the horizon $1, \ldots, H$, i.e. as the nowcast period approaches the publication date of the target variable.

4 Monte Carlo simulation

In this section, we report the results of Monte Carlo simulations to investigate the properties of the horizon confidence set procedure developed in Sect. 2. In order to assess the performance, we will compare the results to those where the MCS procedure of Hansen et al. (2011) is applied independently to each horizon where we expect the rejection rate to be too high even if models have equal predictive ability.[6]

We set up our Monte Carlo design to be similar in spirit to the related simulation studies in Hansen et al. (2011) and Quaedvlieg (2020). As such, we will directly generate the losses $L_{i,t}^h$ which gives us the flexibility to freely change various aspects like which of the models appear in the true model set across horizons. This is impor-

[6] We are grateful to two anonymous referees for suggesting we explore this comparison through simulation.

tant given that our set-up has an additional dimension to the aforementioned studies, specifically the operation of the elimination rule across $h = 1, \dots, H$,

We focus on the two-model set-up described above with $M = 2$, and for models $i = 1, 2$, we generate a sample of T observations of the losses across horizons $\mathbf{L}_{i,t} = [L^1_{i,t}, \dots, L^H_{i,t}]'$ according to the following data generating process (DGP):

$$\mathbf{L}_{i,t} = \boldsymbol{\theta}_i + \mathbf{e}_{i,t} \tag{8}$$

$$\mathbf{e}_{i,t} = \delta \mathbf{e}_{i,t-1} + \boldsymbol{\Sigma}^{1/2} \boldsymbol{\epsilon}_t \tag{9}$$

where δ is a scalar parameter which controls the time series dependence of the losses, $\boldsymbol{\epsilon}_t \sim i.i.d.N(0, \mathbf{I}_H)$ is an $H \times 1$ random draw from a multivariate standard normal distribution and $\boldsymbol{\Sigma}$ is used to control the dependence of model i's losses across horizons. In the nowcasting setting, we indeed expect a reasonable amount of correlation of the losses for a given model across data release dates. We set $\boldsymbol{\Sigma}$ to be the Toeplitz matrix with elements $\boldsymbol{\Sigma}_{i,i'} = \rho^{|i-i'|}$ so that cross-horizon dependence is only governed by the single parameter ρ.

The important parameter vector $\boldsymbol{\theta}_i = [\theta^1_i, \dots, \theta^H_i]'$ controls the behaviour of the mean of the loss differential because $E[L^h_{i,t}] = \theta^h_i$, and therefore in the terminology of the null hypotheses formulated in Eq. (2), we have that $\mu^h_{i,j} = E[L^h_{i_t} - L^h_{j,t}] = \theta^h_i - \theta^h_j$ for models i, j and for horizons $h = 1, \dots, H$. In this two-model case, for $\left[\theta^h_1, \theta^h_2\right]$ we use the following specification:

$$[\theta^h_1, \theta^h_2] = \begin{cases} [\widetilde{\theta}, 0] & \text{if}\, h \leq H/2 \\ [0, \widetilde{\theta}] & \text{if}\, h > H/2 \end{cases} \tag{10}$$

where $\widetilde{\theta}$ is a scalar parameter. When $\widetilde{\theta} = 0$, the models have equal loss at every $h = 1, \dots, H$ and we do not expect the average rejection rate of the horizon confidence set procedure to be larger than α. When $\widetilde{\theta} > 0$, the models do not have equal loss: model 1 has lower loss in the second half of the blocks when $h > H/2$, whereas model 2 has lower loss in the first half of the horizons ($h \leq H/2$). This mimics the nowcasting setting described above where one model may have better predictive ability for earlier data releases but be beaten by another model at other points in the data flow. Clearly, the larger the $\widetilde{\theta}$, the greater the the average loss differential and we expect the rejection rate to increase towards unity.

We consider a variety of values for this DGP set-up. We let the number of horizons be $H \in \{2, 4\}$ which gives a small but representative example of the set-up above. We let $\widetilde{\theta} \in \{0, 0.1, 0.2, 0.5\}$ to give a range for the loss differential behaviour, and we consider values of $\rho \in \{0, 0.5\}$ and $\delta = 0.2$ for the time series and cross-sectional dependence. The sample sizes we consider are $T \in \{100, 200, 500\}$. We use $B = 400$ bootstrap replications to derive the critical values at each step of the MCS procedure and perform $K = 1000$ Monte Carlo replications. The nominal size is set to be $\alpha = 0.1$.

The results of these simulations are displayed in Table 1. The results compare the average rejection rate for our method relative to the method where we independently apply the MCS procedure of Hansen et al. (2011) to each horizon. We see that, in the

Table 1 Average rejection rates across $K = 1000$ Monte Carlo simulations

$\widetilde{\theta}$	$T = 100$		$T = 200$		$T = 500$	
	$H = 2$	$H = 4$	$H = 2$	$H = 4$	$H = 2$	$H = 4$
$\rho = 0.5$						
Horizon confidence set						
0	0.188	0.104	0.151	0.108	0.149	0.100
0.1	0.231	0.172	0.290	0.224	0.451	0.391
0.2	0.424	0.334	0.600	0.535	0.901	0.877
0.5	0.940	0.932	0.998	0.997	1.000	1.000
Independent MCS						
0	0.237	0.208	0.196	0.213	0.197	0.202
0.1	0.290	0.288	0.348	0.343	0.484	0.515
0.2	0.474	0.463	0.624	0.630	0.898	0.895
0.5	0.940	0.938	0.997	0.997	1.000	1.000
$\rho = 0$						
Horizon confidence set						
0	0.162	0.095	0.152	0.107	0.155	0.101
0.1	0.230	0.165	0.274	0.204	0.463	0.385
0.2	0.435	0.342	0.606	0.528	0.896	0.883
0.5	0.941	0.934	0.995	0.997	1.000	1.000
Independent MCS						
0	0.225	0.212	0.210	0.217	0.216	0.215
0.1	0.288	0.286	0.334	0.341	0.514	0.501
0.2	0.486	0.469	0.647	0.637	0.902	0.896
0.5	0.948	0.939	0.995	0.997	1.000	1.000

case of $\widetilde{\theta}$ where the models have equal loss, our method delivers a rejection rate close to the nominal size. This rejection rate improves with both H and T. On the other hand, if we independently apply the Hansen et al. (2011) procedure to each horizon, the rejection rate is too high with greater than 20% rejection rate.

As we increase $\widetilde{\theta}$, we see that the rejection rate moves closer to 100% and this rejection rate improves with the sample size as seen in the $T = 500$ case. We note that the results are not very sensitive to changing the ρ parameter which controls the degree of correlation of the losses across blocks. Overall, we find that our procedure is better able to control the rejection rate than an independent procedure across horizons, as is expected.

5 Empirical application

In order to illustrate our methodology, we use the example of nowcasting quarterly aggregate US real GDP growth and its subcomponents. We will focus on comparing the performance of factor-based methods, which use the common component from a data set of macroeconomic series, to the predictions of a naïve autoregressive benchmark. This kind of comparison of factor methods to a univariate benchmark has become standard in the factor model nowcasting literature (see Anesti et al. 2019 for a recent example). Our method will be able to formally detect at which points in the nowcast period the univariate benchmark is dominated (if at all). We will report results for aggregate GDP growth and five subcomponents (consumption, investment, government spending, imports and exports). This will shed more light on recent analyses of GDP subcomponent nowcasting, including Antolin Diaz et al. (2017) and Fosten and Gutknecht (2020). In what follows, we will describe the data and empirical set-up before presenting the results.

5.1 Data and set-up

In predicting quarterly real GDP and its subcomponents, we will follow the approach of Bok et al. (2018) who document the *New York Fed Staff Nowcast* procedure based on the factor model nowcasting methodology of Giannone et al. (2008). The data series and their transformations to stationarity are described at length in Bok et al. (2018). They construct a parsimonious database of the series most widely followed by market participants, only focusing on the headline series and not disaggregates. The data set comprises mostly monthly variables related to production, employment, consumption and consumer sentiment, housing, trade and so on. We update the data set using the FRED Economic Data service, starting in 1985M1 as in Bok et al. (2018) and ending in 2020M2, with the final data on the quarterly GDP series being in 2019Q3.[7] We remove some variables which do not have sufficient data for the out-of-sample analysis detailed below, which results in a total of $N = 25$ series being used.[8]

As is customary in nowcasting studies, we keep track of the calendar of releases of all predictors, which dictates the nowcast horizons in the nowcast updating procedure. We will make nowcasts which are updated at intervals of 10 days from the start of the nowcast quarter up until day 20 of the first month of the following quarter, which is just before when GDP is first released by the Bureau of Economic Analysis (BEA). This gives a total of $H = 11$ nowcast updates per quarter observed, each of which corresponds, respectively, to days 10, 20, 30, …, 110 after the beginning of the reference nowcast quarter. Therefore, the last two nowcast updates are actually backcasts.

The factor-based method we use can be described as follows: we denote $y_{i,t}$ as the monthly or quarterly variables in the data set, where $i = 1, \dots, N$ and $N = 25$ as above. We assume that there exists one latent factor, f_t, which drives the co-movement amongst the $y_{i,t}$ series as follows:

$$y_{i,t} = \mu_i + \lambda_i f_t + \varepsilon_{i,t}, \tag{11}$$

[7] See https://fred.stlouisfed.org/. Data last accessed: 08/04/20.

[8] We treat the real GDP series (either the aggregate or one of the five sub-components) as the 'target variable' and only use that variable in the data set. In other words, we do not use all six real GDP series in the data set at one time. This is simply because all these series are released at the same time, so they are not useful for making timely nowcasts of each other.

where λ_i is the factor loading for variable i and $\varepsilon_{i,t}$ is an idiosyncratic error term. As in Bok et al. (2018), we only use one global factor in this relatively small data set. We do not mimic their use of additional local block factors due to the lack of data availability in our initial estimation window in the pseudo-out-of-sample experiment described below. We therefore have one factor, which we treat as fixed across all estimation windows.[9]

The model in Eq. (11) is specified at the monthly frequency, with the quarterly series treated as a filtered monthly series with missing observations. For these quarterly series, the aggregation from a latent monthly growth rate to the quarterly growth rate is dealt with using the method of Mariano and Murasawa (2003). In order to cast the model into state space form, the factor and idiosyncratic disturbances are state variables which are both assumed to follow AR(1) processes with normal innovations:

$$f_t = \alpha f_{t-1} + u_t \tag{12}$$

$$\varepsilon_{i,t} = \rho_i \varepsilon_{i,t-1} + v_{i,t} \tag{13}$$

for $i = 1, \ldots, N$, where u_t and $v_{i,t}$ are i.i.d. normal processes. Equations (11)–(13) jointly form the state space model which is estimated using the Kalman smoother and the expectation maximization (EM) algorithm (see Giannone et al. 2008; Doz et al. 2011, 2012; Bańbura and Modugno 2014 for full details of this procedure).

Having estimated the factor model and obtained the nowcasts for aggregate real GDP growth and the five subcomponents, we will compare the predictions to that of a simple $AR(1)$ model as this is the most commonly used benchmark model in the literature.[10] For these quarterly target series, the $AR(1)$ method only has one distinct release date in the nowcast period, and so we only observe two distinct nowcasts throughout the prediction period.

To generate the nowcasts and nowcast errors, we will split the sample into $T = R + P$ observations and use the pseudo-out-of-sample procedure as in West (1996). We will use the rolling scheme as suggested by Hansen et al. (2011), but we will compare the results to those of the recursive scheme which is widely used in empirical studies. We therefore start using data on the first R quarterly observations to estimate the models. The first predictions are made of quarter $R + 1$ where we begin adding information released at the beginning of the quarter and we update the nowcasts $H = 11$ times every 10 days until day 20 of the first month of the next quarter, just before the GDP data are released. Then, the sample is expanded by one quarter and the procedure is repeated, adding one quarter at a time, until the end of the sample. We start making nowcasts in 1994Q1 which results in $P = 103$ out-of-sample evaluation periods. For the nowcast error loss function, we will consider both the absolute value function $L(e) = |e|$ and the squared error loss function $L(e) = e^2$, giving rise to test statistics involving mean absolute error (MAE) and mean squared forecast error

[9] Results were also run for larger numbers of global factors, but these typically resulted in worse performance.

[10] We note that, since the GDP release for the previous quarter occurs at around day 28 of the nowcast period, the $AR(1)$ nowcast from day 1 through to day 28 is a two-step ahead prediction, whereas the nowcast from day 28 onwards is a one-step ahead prediction.

(MSFE).[11] As pointed out in Sect. 2, the tests are constructed using non-studentized statistics.

For the horizon confidence set procedure, we will construct these at the 75% confidence level which, although using a somewhat high nominal level, is common practice in most empirical work using MCS including that of Hansen et al. (2003, 2011). Finally, the number of bootstrap repetitions is $B = 400$, where we chose the block length as the estimated optimal AR length across the loss differential series.[12]

5.2 Horizon confidence set results

To gain a preliminary insight into the performance of the competing nowcast methods, Figs. 2 and 3 graph the evolution of the MAE and MSFE throughout the $H = 11$ horizons of the nowcast prediction period. These graphs are for the rolling scheme, whereas the corresponding figures for the recursive scheme are in the Appendix. From these sets of charts, we can see that these six different target variables produce rather different behaviour in terms of the loss differentials between the factor model and the AR(1) benchmark across different nowcast horizons. This gives us a good mixture of settings in which to apply the horizon confidence set procedure.

In the case of aggregate GDP, we see that there is no clear ‘winner’ between the factor method and the AR(1) benchmark uniformly across nowcast horizons in terms of MAE or MSFE. Looking at the scale of the MAE/MSFE, the two models indeed appear to deliver very close nowcast error losses, which will be formally assessed by our horizon confidence set procedure. For the cases of consumption, government spending and exports, we find that the AR(1) model provides lower MAE and MSFE than the factor method, whereas for investment the AR(1) model is worse over all nowcast horizons. The loss differential appears to be much larger in the case of MSFE in Fig. 5, especially for investment and government spending. In the case of imports, there is a less clear differential between the two models.

In order to make more formal statements about the performance of the models in Figs. 2 and 3, we now perform the horizon confidence set procedure. Figures 4 and 5 are graphical depictions of the final estimated collection $\{\widehat{\mathcal{M}}^*_{h,1-\alpha}\}_{h=1}^{H}$. Looking at the main case, aggregate GDP growth, we find no evidence that either model outperforms the other at any of the nowcast horizons. This seems to confirm the graphical evidence from above which shows there to be only a small loss differential on average. The same picture holds in the case of consumption and imports, where we do not see a rejection of any model over any of the nowcast horizons.

Looking at the other variables, focusing first on the MSFE results in Fig. 5, for investment we see that the horizon MCS procedure removes the AR(1) model in every nowcast horizon after the first 2 months of the nowcast quarter. This implies that one might consider averaging the nowcasts from the factor and AR(1) models only up to day 60 of the nowcast period and only use the factor model nowcasts thereafter. For government spending, the reverse is true: both the factor method and the AR(1)

[11] The case of non-differentiable loss functions and MAE statistics in the context of out-of-sample evaluation is considered by McCracken (2000).

[12] This implementation is adapted from the `MCS` package in `R`.

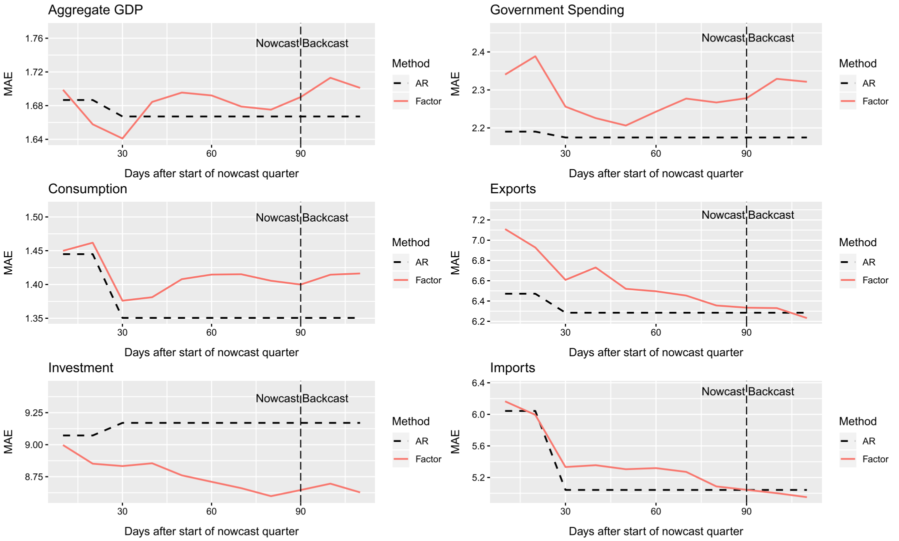


Fig. 2 MAE by nowcast horizon—GDP subcomponents

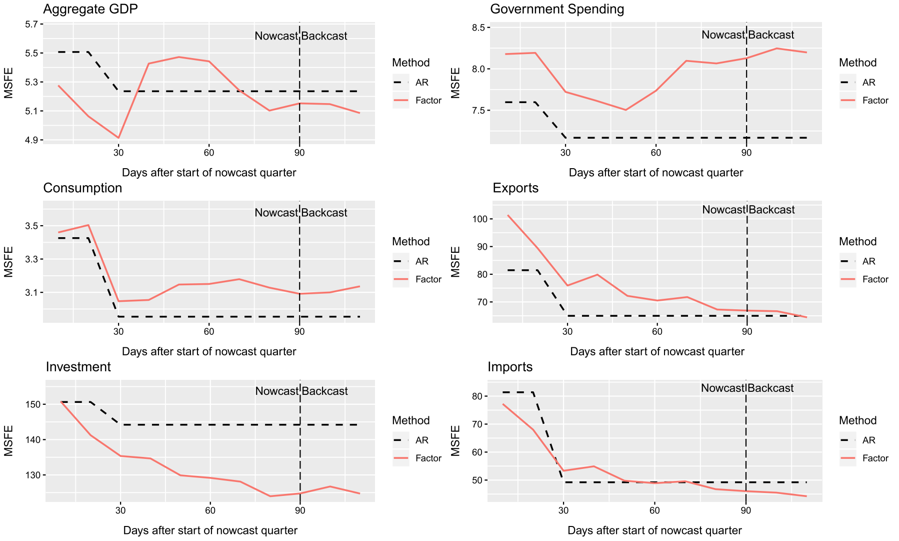


Fig. 3 MSFE by nowcast horizon—GDP subcomponents

method remain in the horizon confidence set until the end of the second month of the nowcast quarter, after which only the AR(1) model should be used. Notably, in the case of MAE, Fig. 4 shows fewer rejections for investment and government spending, which reinforces the earlier comment analysing the graphical evidence in Figs. 2 and 3. In the case of exports, we also see a handful of rejections of the factor method in the earlier part of the nowcast period.

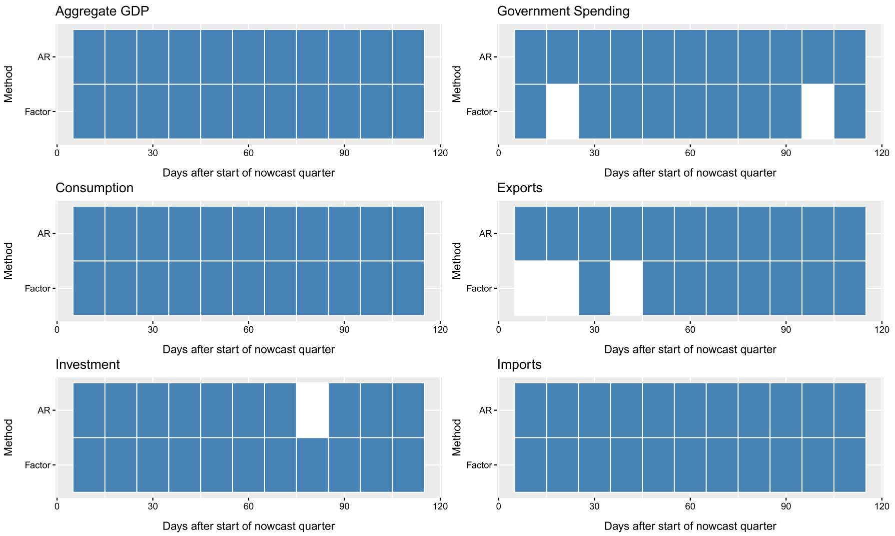


Fig. 4 Horizon MCS results—MAE loss

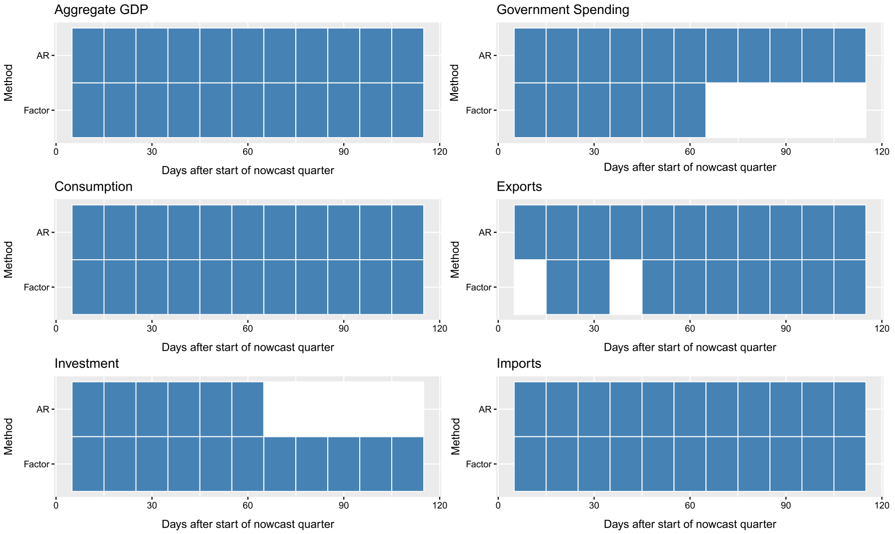


Fig. 5 Horizon MCS results—MSFE loss

The fact that we see a variety of different features in these horizon confidence sets highlights the usefulness of our method in making decisions about the use of different nowcast models across multiple nowcast horizons. In the case of aggregate GDP, consumption and imports, our method justifies the use of model averaging of these two methods. On the other hand, for investment and government spending, model averaging only makes sense up to a certain point in the nowcast period, after which one of the two models is dominated. This kind of pattern would have remained undetected

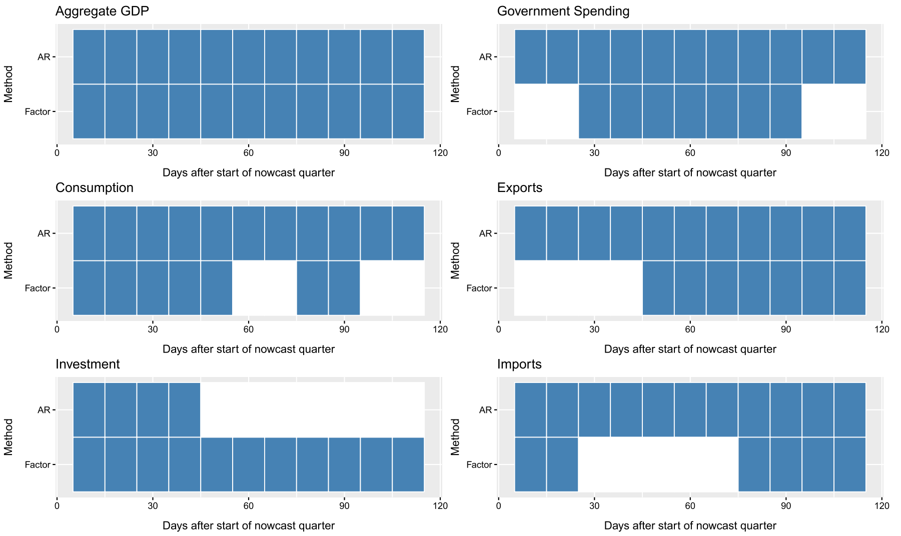


Fig. 6 Independent horizons MCS results—MAE loss

using the concept of uSPA or aSPA outlined in Quaedvlieg (2020), where we would have to reject or retain a given model across *all* horizons.

5.3 Comparison with the independent horizons MCS

In this section, we compare the results of our horizon confidence set procedure to the case where we treat horizons independently and run the MCS procedure of Hansen et al. (2011) separately at each of the horizons. Figures 6 and 7 display the equivalent results to those in the previous section. We firstly see that there are many more rejections using this procedure than using our procedure. This could be likened to the results of Sect. 4 where simulation evidence suggested that the independent method has a high rejection rate even if the models have equal predictive ability. Our method also tends to produce more stable results with less fluctuation of models in and out of the confidence set across horizons than in Figs. 6 and 7.

To give an example of this switching behaviour, looking at the results for aggregate GDP under MSFE loss in Fig. 7, we see that there the AR(1) model is removed in the second and third nowcast periods, whereas the factor method is removed in the fifth. This seems at odds with the graphical evidence in Fig. 3 where the small crossings in the MSFE profiles are of very small magnitude relative to the scale.

5.4 Nowcast averaging and monotonicity test results

We next aim to shed some light about the performance of nowcast averaging in terms of monotonicity, when weights are constructed using the horizon confidence set results or if a simple equal weights scheme is used. We first construct nowcast averages

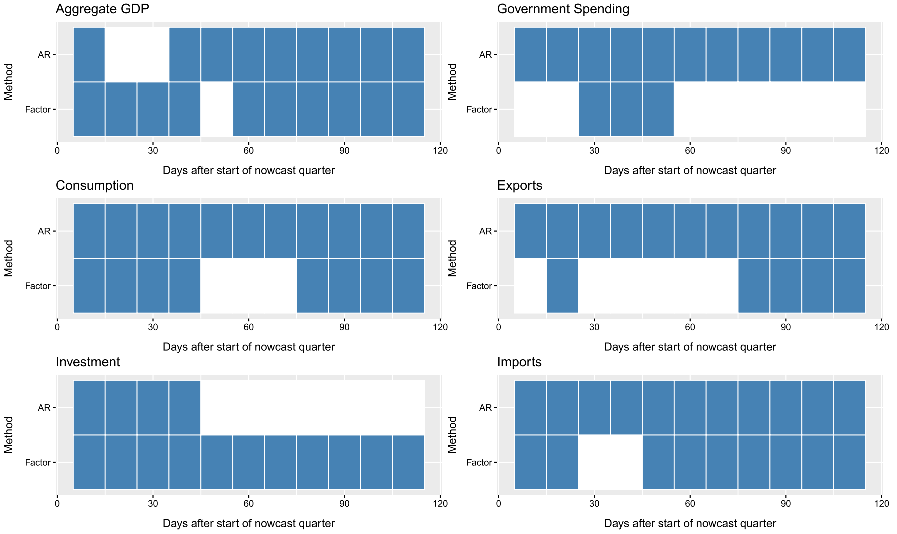


Fig. 7 Independent horizons MCS results—MSFE loss

$\widehat{\overline{y}}_{kt}$ in Eq. (7) using the two models from the horizon confidence set procedure (we denote this method 'MCS_AVE' in the following figures and tables). Clearly, these weights will fluctuate between $(1, 0)$, $(0, 1)$ and $(1/2, 1/2)$ across all the nowcast horizons. This gives a single combined nowcast at each horizon which can be used to test monotonicity using the recent procedure of Fosten and Gutknecht (2020). We will also compare the MSC_AVE method to that of using a simple average across the models in all nowcast horizons (denoted simply 'AVE').

Focusing on the MSFE results, Fig. 3 shows that, at least graphically, there is evidence of non-monotonicity in the MSFE profiles of the factor method for aggregate GDP, whereas cases like investment appear to be monotonically declining. Looking at these two cases of aggregate GDP and investment, the MSFE of the nowcast combination is shown in Fig. 8 which uses the MCS weights derived from the results displayed in Fig. 5. For aggregate GDP, the results coincide as no model is removed at any horizon. For investment, the MSFE of the MCS_AVE method is slightly lower than that of the simple average across horizons.

To assess these two cases formally, we obtain results of the monotonicity test of Fosten and Gutknecht (2020) for the null hypothesis of nowcast monotonicity of the averaged MSFE profiles in Fig. 8. These results are provided in Table 2. We find no evidence to reject the null hypothesis of nowcast monotonicity for either of the two series. This indicates that, while we do see graphical evidence of non-monotonicity for aggregate real GDP growth, these movements are very small and not statistically different from the case of a flat MSFE profile. For investment, we see that the p-value reduces slightly for the MCS_AVE version of the test relative to the AVE version. Both results demonstrate that a simple average across all models is equally capable of producing monotonically declining MSFE, even it is slightly outperformed by MCS_AVE. Overall, this indicates that the method of nowcast averaging is capable of producing

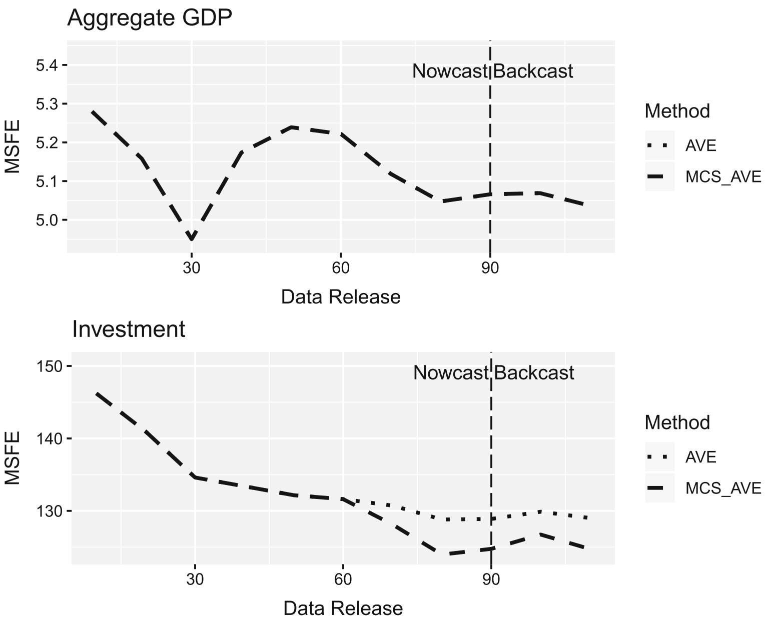


Fig. 8 MSFE of the nowcast combination

Table 2 Bootstrap monotonicity test with multiple models (MSFE loss)

	κ	U^*	50%	90%	95%	p-value
AVE						
Aggregate GDP	55	2.935	2.638	8.160	9.874	0.433
Investment	55	10.832	49.654	241.325	296.483	0.938
MCS_AVE						
Aggregate GDP	55	2.935	2.638	8.160	9.874	0.433
Investment	55	28.006	69.008	309.643	377.224	0.730

κ represents the number of moment inequalities used in the monotonicity test and $\widehat{U}^*$ is the test statistic of Fosten and Gutknecht (2020) along with the bootstrap 50th, 90th and 95th percentiles and the one-sided p-value

nowcasts of GDP and its subcomponents which are monotonically improving as the time horizon shrinks until the publication date of GDP.

6 Conclusion

In this paper, we have proposed a methodology which allows researchers to determine which models to use at different horizons, which we call the *Horizon Confidence Set*. We build on the MCS procedure introduced by Hansen et al. (2011), adapting it to the multi-horizon set-up. In both long-term forecasting and near-term nowcasting, which are the main applications of our methodology, we argue that we need a method capable of retaining models which are better at some horizon but removing them at horizons where they underperform. Our proposed method sequentially eliminates the worst models in each horizon, resulting in a potentially changing set of ‘optimal’ models across different horizons. This provides an advantage over existing multi-horizon approaches which look for uniform or average model superiority (Quaedvlieg 2020). Our method also has advantages over naïvely applying the MCS procedure independently across horizons.

To facilitate the practical applicability of our methodology, we discuss how model combination can be performed on the basis of the horizon confidence sets. This is similar in spirit to various existing nowcast combination studies (including Kuzin et al. 2013; Aastveit et al. 2018) and also allows to conduct formal monotonicity tests as recently proposed by Fosten and Gutknecht (2020).

Finally, we apply our methodology using the factor model methodology employed by Bok et al. (2018) in making the *New York Fed Staff Nowcasts*. However, we extend their analysis of nowcasting the aggregate US real GDP growth rate to that of the five GDP subcomponents. Comparing this factor method to a naïve autoregressive benchmark, our procedure shows the point in the quarter at which the factor method beats the benchmark model or vice versa. This finding is novel and could have potentially remained undetected by existing tests for uniform or average SPA. On the other hand, when no model is dominant at any horizon our method is capable of reaching this conclusion, and model averaging is considered in these cases. Our results are also more stable than a simple independent horizons application of Hansen et al. (2011). We therefore deem our procedure a useful and complementary tool to existing methods in the literature which can yield new insights into the comparison of multi-horizon forecasts and nowcasts.

Appendix

Multiple model comparison

This subsection outlines the extension of our procedure to allow for the comparison of more than two models. To start, note that the definitions of $\mathcal{M}_0$, $\mathcal{M}_h^*$, and $\mathcal{M}_h$, $h = 1, \ldots, H$ remain the same as in the text, but for the fact that these sets may now contain more than just two models so that $M > 2$. In fact, the null hypothesis for each $h = 1, \ldots, H$ is given by:

$$H_{0,h}^{M}\colon \mu_{ij}^{h} = 0 \quad \text{for all } i, j \in \mathcal{M}_h. \tag{14}$$

The procedure works as follows:

1. Set $\mathcal{M}_h = \mathcal{M}^0$ for each $h = 1, \ldots, H$.
2. Test the null hypothesis $H_{0,h}^M$ in (14) using the testing procedure below at level $\widetilde{\alpha}$.
3. If $H_{0,h}^M$ is not rejected, set $\{\widehat{\mathcal{M}}_{h,1-\alpha}^*\}_{h=1}^H = \{\mathcal{M}_h\}_{h=1}^H$, otherwise use $\boldsymbol{e}_{H,\mathcal{M}}$ to eliminate an object from the relevant $\mathcal{M}_h$, leaving the remaining $\{\mathcal{M}_j\}_{j \neq h}$ as they were, and repeat from Step 2.

The main difference between the two-model set-up and the more general case lies in the need to account for the fact that different $t_{i,j}^h$ may have different critical values as the maximum horizon h may not be the same across model pairs i, j. To mitigate this problem, we follow Quaedvlieg (2020) and propose a double bootstrap procedure, which takes these different critical values into account. More specifically, the procedure is based on the maximum of the re-centred t-statistics $(t_{i,j}^h - c_{i,j}(\alpha))$, where $c_{i,j}(\alpha)$ is the first round critical value for model pair i, j as outlined in the following algorithm:

1. For each given pair $i, j \in \mathcal{M}_h$, $h = 1, \ldots, H$, compute $t_{i,j}^h$ and $c_{i,j}(\alpha)$ using the algorithm from the main text.
2. Define $t_{\max} = \max_{i,j \in \{\mathcal{M}_h\}_{h=1}^H} \left(\max_{1 \leq h \leq H} t_{i,j}^h - c_{i,j}(\alpha) \right)$, i.e. the test statistic $t_{i,j}^h$ furthest away from its corresponding critical value across all possible model pairs i, j.
3. For each bootstrap sample $\mathbf{d}_{ij,t}^b$, $b = 1, \ldots, B$, obtained in 1., execute the following steps:
   (a) For each pair $i, j \in \{\mathcal{M}_h\}_{h=1}^H$, apply the algorithm from 1. treating the bootstrap sample $\mathbf{d}_{ij,t}^b$, $t = 1, \ldots, T$, as the actual sample to obtain $c_{i,j}^b(\alpha)$.
   (b) Compute the bootstrap t-statistics:

$$t_{\max}^b = \max_{i,j \in \{\mathcal{M}_h\}_{h=1}^H} \left(\max_{1 \leq h \leq H} t_{i,j}^{hb} - c_{i,j}^b(\alpha) \right)$$

4. Obtain the p-value as:

$$\frac{1}{B} \sum_{b=1}^{B} 1\{t_{\max} < t_{\max}^b\}.$$

As pointed out by Quaedvlieg (2020), to obtain reasonable p-values one can follow Hansen et al. (2011) and impose that a p-value for a model cannot be lower than any previously eliminated model and that the last remaining model at each horizon $h = 1, \ldots, H$ obtains a p-value of one by convention.

7.2 Additional figures

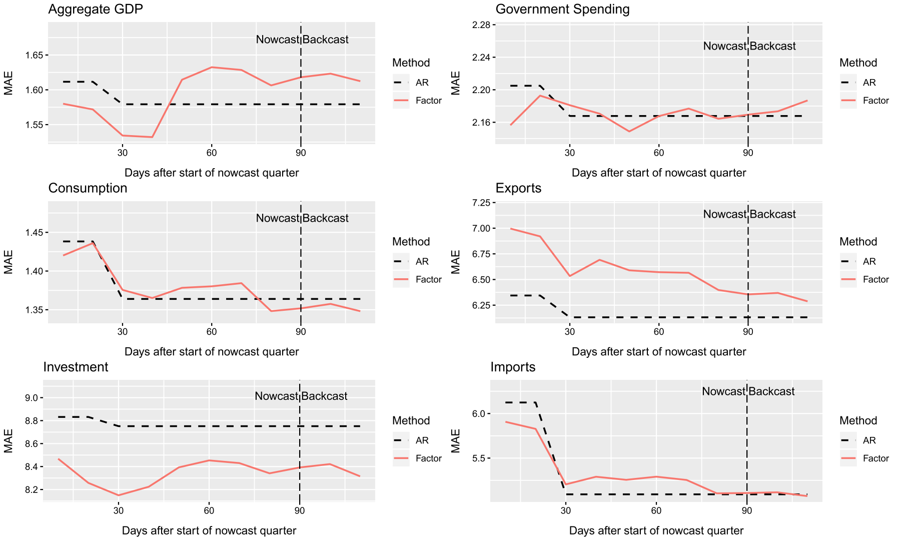


Fig. 9 MAE by nowcast horizon—recursive

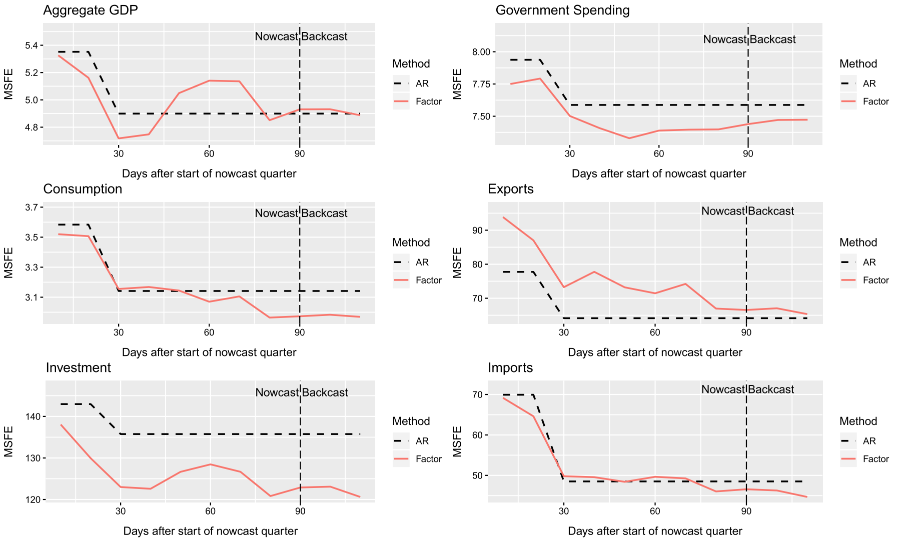


Fig. 10 MSFE by nowcast horizon—recursive

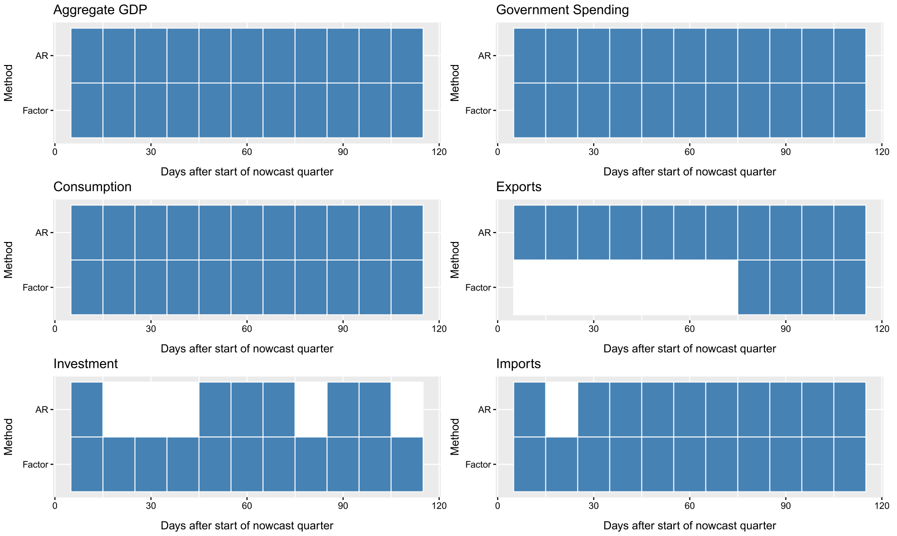


Fig. 11 Horizon MCS results—recursive—MAE loss

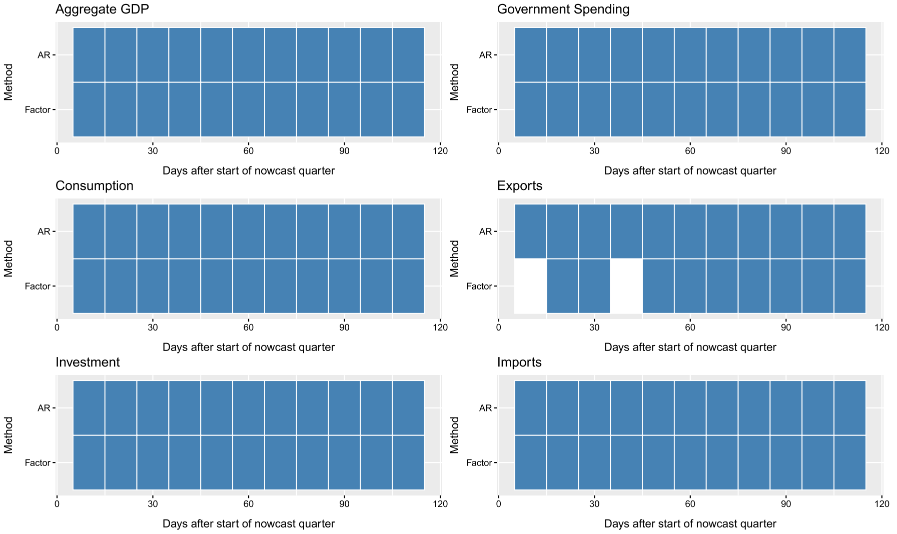


Fig. 12 Horizon MCS results—recursive—MSFE loss

References

Aastveit KA, Gerdrup KR, Jore AS, Thorsrud LA (2014) Nowcasting GDP in real time: a density combination approach. J Bus Econ Stat 32(1):48–68

Aastveit KA, Foroni C, Ravazzolo F (2017) Density forecasts with MIDAS models. J Appl Econ 32(4):783–801

Aastveit KA, Ravazzolo F, Van Dijk HK (2018) Combined density nowcasting in an uncertain economic environment. J Bus Econ Stat 36(1):131–145

Anesti N, Galvao AB, Miranda-Agrippino S (2019) Uncertain Kingdom: nowcasting GDP and its revisions. Working Paper

Antolin Diaz J, Drechsel T, Petrella I (2017) Tracking the slowdown in long-run GDP growth. Rev Econ Stat 99(2):343–356

Bańbura M, Modugno M (2014) Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. J Appl Econ 29(1):133–160

Banbura M, Giannone D, Modugno M, Reichlin L (2013) Now-casting and the real-time data flow. In: Elliott G, Timmermann A (eds) Handbook of economic forecasting, vol 2A. North-Holland, Amsterdam, pp 195–236

Bok B, Caratelli D, Giannone D, Sbordone AM, Tambalotti A (2018) Macroeconomic nowcasting and forecasting with big data. Annu Rev Econ 10(1):615–643

Clark TE, McCracken MW (2005) Evaluating direct multistep forecasts. Econ Rev 24(4):369–404

Coroneo L, Iacone F (2019) Comparing predictive accuracy in small samples using fixed-smoothing asymptotics. Working Paper

Diebold FX, Mariano RS (1995) Comparing predictive accuracy. J Bus Econ Stat 13(3):253–263

Doz C, Giannone D, Reichlin L (2011) A two-step estimator for large approximate dynamic factor models based on Kalman filtering. J Econ 164(1):188–205

Doz C, Giannone D, Reichlin L (2012) A quasi-maximum likelihood approach for large, approximate dynamic factor models. Rev Econ Stat 94(4):1014–1024

Foroni C, Marcellino M (2014) A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. Int J Forecast 30(3):554–568

Foroni C, Marcellino M, Schumacher C (2015) Unrestricted mixed data sampling (MIDAS): MIDAS regressions with unrestricted lag polynomials. J R Stat Soc Ser A (Stat Soc) 178(1):57–82

Fosten J, Gutknecht D (2020) Testing nowcast monotonicity with estimated factors. J Bus Econ Stat 38(1):107–123

Giannone D, Reichlin L, Small D (2008) Nowcasting: the real-time informational content of macroeconomic data. J Monet Econ 55(4):665–676

Giannone D, Monti F, Reichlin L (2016) Exploiting the monthly data flow in structural forecasting. J Monet Econ 84:201–215

Gonçalves S, White H (2004) Maximum likelihood and the bootstrap for nonlinear dynamic models. J Econ 119(1):199–219

Götze F, Künsch HR (1996) Second-order correctness of the blockwise bootstrap for stationary observations. Ann Stat 24(5):1914–1933

Hansen PR, Lunde A, Nason JM (2003) Choosing the best volatility models: the model confidence set approach. Oxf Bull Econ Stat 65:839–861

Hansen PR, Lunde A, Nason JM (2011) The model confidence set. Econometrica 79(2):453–497

Kim HH, Swanson NR (2018) Methods for backcasting, nowcasting and forecasting using factor-MIDAS: with an application to Korean GDP. J Forecast 37(3):281–302

Knotek ES, Zaman S (2017) Nowcasting US headline and core inflation. J Money Credit Bank 49(5):931–968

Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17(3):1217–1241

Kuzin V, Marcellino M, Schumacher C (2013) Pooling versus model selection for nowcasting GDP with many predictors: empirical evidence for six industrialized countries. J Appl Econ 28(3):392–411

Luciani M, Ricci L (2014) Nowcasting Norway. Int J Centr Bank 10(4):215–248

Marcellino M, Schumacher C (2010) Factor MIDAS for nowcasting and forecasting with ragged-edge data: a model comparison for German GDP. Oxf Bull Econ Stat 72(4):518–550

Mariano RS, Murasawa Y (2003) A new coincident index of business cycles based on monthly and quarterly series. J Appl Econ 18(4):427–443

Mark N (1995) Exchange rates and fundamentals: evidence on long-horizon predictability. Am Econ Rev 85(1):201–218

McCracken MW (2000) Robust out-of-sample inference. J Econ 99(2):195–223

McCracken M, Owyang MT, Sekhposyan T (2019) Real-time forecasting with a large, mixed frequency, Bayesian VAR. Mimeo

Newey WK, West KD (1987) A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55(3):703–708

Quaedvlieg R (2020) Multi-horizon forecast comparison. J Bus Econ Stat (forthcoming)

Romano JP, Wolf M (2005) Stepwise multiple testing as formalized data snooping. Econometrica 73(4):1237–1282

Rossi B (2013) Exchange rate predictability. J Econ Lit 51(4):1063–1119

West KD (1996) Asymptotic inference about predictive ability. Econometrica 64(5):1067–1084

White H (2000) A reality check for data snooping. Econometrica 68(5):1097–1126

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.