Journal of the Royal Statistical Society Series A: Statistics in Society, 2023, **00**, 1–19
https://doi.org/10.1093/jrsssa/qnad130

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society
A

Original Article

Incorporating short data into large mixed-frequency vector autoregressions for regional nowcasting

Gary Koop[1,2], Stuart McIntyre[1,2], James Mitchell[2,3], Aubrey Poon[2,4,5] and Ping Wu[1,2]

[1]Department of Economics, University of Strathclyde, Glasgow, UK
[2]Economic Statistics Centre of Excellence, London, UK
[3]Federal Reserve Bank of Cleveland, Cleveland, Ohio, USA
[4]School of Economics, University of Kent, Canterbury, UK
[5]School of Economics, Orebro University, Orebro, Sweden

Address for correspondence: Stuart McIntyre, University of Strathclyde, 199 Cathedral Street, Glasgow, G4 0QU, UK. Email: s.mcintyre@strath.ac.uk

Abstract

Interest in regional economic issues coupled with advances in administrative data is driving the creation of new regional economic data. Many of these data series could be useful for nowcasting regional economic activity, but they suffer from a short (albeit constantly expanding) time series which makes incorporating them into nowcasting models problematic. Regional nowcasting is already challenging because the release delay on regional data tends to be greater than that at the national level, and 'short' data imply a 'ragged edge' at both the beginning and the end of regional data sets, which adds a further complication. In this paper, via an application to the UK, we investigate various ways of including a wide range of short data into a regional mixed-frequency vector autoregression (MF-VAR) model. These short data include hitherto unexploited regional value-added tax turnover data. We address the problem of the two ragged edges by estimating regional factors using different missing data algorithms that we then incorporate into our MF-VAR model. We find that nowcasts of regional output growth are generally improved when we condition them on the factors, but only when the regional nowcasts are produced before the national (UK-wide) output growth data are published.

Keywords: Bayesian methods, factors, missing data, mixed-frequency data, regional data, vector autoregressions

JEL codes: C32, C53, E37

1 Introduction

Official sub-national data, as published by national statistical institutes, tend to be available at a lower frequency and are published more slowly than data for the nation as a whole. A case in point, and the motivating empirical example in this paper, is regional data for gross value added (GVA) for the UK regions.[1] While the Office for National Statistics (ONS) has long produced GVA data for the UK regions, until 2019 these data were produced at an annual frequency only and with a release delay of approximately 1 year.[2] In 2019, with the development of the quarterly regional

[1] Real GVA and real GDP are closely related concepts. GVA is in basic prices, while GDP is in market prices [i.e. GVA plus taxes (less subsidies) on products equals GDP].

[2] In this paper, we consider the 12 ITL1 regions, excluding the UK continental shelf. These 12 regions/devolved nations comprise North East England, North West England, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East England, South West England, Wales, Scotland, and Northern Ireland.

Received: April 21, 2023. **Revised:** October 23, 2023. **Accepted:** October 24, 2023

gross domestic product (GDP) data set, ONS production was sped up. Regional GVA then became available at a quarterly frequency back to 2012 and with a shorter, but still substantial, release delay of approximately six months. This situation is not unusual at a regional level.[3] There therefore remains a need, especially to support policy and business decisions made in real time, to produce more timely regional estimates—so-called 'nowcasts'.

The motivation for this paper is the observation that recent years have seen increasing availability, from both private and public sources, of higher frequency and more timely data, capturing various aspects of economic activity, at the regional and national levels. Ideally, when nowcasting, one would condition on all of these indicators—and let the data (the fit of the model) determine which indicators are most useful. But these new indicators are often available with only a limited historical time series. This impedes their inclusion in traditional nowcasting models. Table 1 illustrates the issue in the context of our UK application (but we emphasise that this issue arises for other variables and in other countries). From Table 1, we see that many regional indicators are available only from the 1990s onward; e.g. Labour Force Survey (LFS) data produced by the ONS, house price indices, and the Purchasing Managers' Index (PMI). But perhaps of most interest are some of the recently released data series based on administrative data that have not been previously used for regional nowcasting in the UK.

These administrative data include payroll employment information derived from the 'pay as you earn' (PAYE) tax system, which are available back to July 2014 only, and value-added tax (VAT) turnover data, which are available back to 2012. There are good reasons to think that these new 'short' indicators should be useful in regional nowcasting. Especially so when, like the PAYE and VAT data, they are derived from administrative data and, therefore, reflect the universe of individuals/firms covered by particular taxes. This provides a significant improvement upon survey-based measures of activity. The payroll employment data also benefit from being high frequency (monthly) and very timely (released within two weeks of the end of the month). The VAT data, while quarterly and released with a delay of around 5 months (this reflects the time taken to produce consistent sub-national aggregates from the underlying firm-level VAT returns), are a key input into the ONS's production of regional GDP data themselves. We should therefore expect VAT data to provide useful information when nowcasting regional GVA growth. But the short and differing lengths of these new data mean that questions arise over how to include them as indicators in a nowcasting model.

Researchers wishing to update their nowcasts using the latest available information are used to addressing what is commonly called, since Wallis (1986), the 'ragged edge' problem. Since different variables are released with different delays, variables with longer delays will have missing values at the end of the sample, whereas variables with shorter release delays will not. This leads to a ragged edge at the bottom of the spreadsheet. Here we have this issue, but we also have a similar problem occurring at the beginning of the sample. The question of how to address the ragged edge at *both* the beginning and the end of the sample in the context of a regional nowcasting model is thus the focus of this paper. Given that the regional GVA data themselves are available in some form back to 1966, there is potentially a large loss in estimation accuracy and efficiency if the model is simply estimated only over the common sample; in the context of Table 1, this would be from the 1990s onward. It is preferable to consider all available data across both the time-series and the cross-sectional dimensions. However, data with a short time span are difficult to incorporate into conventional time-series econometric models, which often work best with long samples.

We address this challenge by investigating ways of incorporating these 'short' data, and potentially a large number of 'short' indicators, into mixed-frequency regional nowcasting models. The specific class of model we consider is a mixed-frequency vector autoregression (MF-VAR). Mixed-frequency vector autoregressions are a popular nowcasting tool; for example, see Brave et al. (2019), Eraker et al. (2014), Schorfheide and Song (2015), and McCracken et al. (2021). Koop et al. (2020a, 2020b) extended MF-VAR models to the regional case by incorporating an additional measurement equation reflecting the fact that national GDP is the aggregation of

[3] In the European Union, for example, regional output data are available only annually (at the ITL2 level) and with a release delay of more than a year, while in the US state-level GDP data are available quarterly, but with a release delay of around 3 months—in contrast to data for the U.S. as a whole, which are released with a delay of 4 weeks.

Table 1. Regional indicators: short and long data

Variable	Tab	Description	Frequency	Geographic coverage	Time period	Typical release delay
1	UK Quarterly HPI	Nationwide Building Society House Price Index	Q	N	Q1 1967–Q1 2022	1 week
2	UKEMP	16–64 Employment Rate Labour Force Survey	Q	N	Q2 1992–Q4 2021	6 weeks
3	UKUnEMP	16+ Unemployment Rate Labour Force Survey	Q	N	Q2 1992–Q4 2021	6 weeks
4	UKMonthlyHousePrice	UK House Price Index	M	N	April 1968–Feb 2022	6 weeks
5	RegCBI	CBI Business Optimism index	Q	R	Q2 1958–Q1 2021	4 weeks
6	BusinessBirths	Business Births Geography Counts	Q	R	Q1 2017–Q1 2022	1 month
7	ConstructionOutput_TNH	Construction Output—Total new housing	Q	R (ex. NI)	Q1 1980–Q4 2021	6 weeks
8	ConstructionOutput_ANW	Construction Output—All New Work	Q	R (ex. NI)	Q1 1980–Q4 2021	6 weeks
9	ConstructionOutput_AW	Construction Output—All Work	Q	R (ex. NI)	Q1 1980–Q4 2021	6 weeks
10	Employment	16–64 Employment Rate Labour Force Survey	Q	R	Q2 1992–Q4 2021	6 weeks
11	Unemployment	16+ Unemployment Rate Labour Force Survey	Q	R	Q2 1992–Q4 2021	6 weeks
12	PublicEMP	Public-sector employment	Q	R	Q1 2008–Q4 2021	3 months
13	WorkforceJobs	Workforce jobs by region and industry	Q	R	Q1 1996–Q1 2022	3 months
14	Exports	Value of Exports	Q	R	Q1 2018–Q3 2021	4 months
15	Imports	Value of Imports	Q	R	Q1 2018–Q3 2021	4 months
16	RegCC	Claimant count rate	M	R	Apr 1974–Jan 2022	2 weeks
17	PayrollEMP	Payroll employment	M	R	July 2014–Jan 2022	2 weeks
18	PayrollPayMedian	Median payroll pay	M	R	July 2014–Jan 2022	2 weeks
19	HousePrice	House Price Index	M	R	Jan 1995–Mar 2022 (most regions, some start later)	6 weeks
20	PMIactivity	PMI activity measure (headline)	M	R	Jan 1997–Jan 2022 (most regions, some start later)	2 weeks
21	PMInewbus	New business measure	M	R	Jan 1997–Jan 2022 (most regions, some start later)	2 weeks
22	PMIoutbus	Outstanding business	M	R	Nov 1999–Jan 2022 (most regions, some start later)	2 weeks

(*continued*)

Table 1. Continued

Variable	Tab	Description	Frequency	Geographic coverage	Time period	Typical release delay
23	PMIcharges	Charges	M	R	Nov 1999–Jan 2022 (most regions, some start later)	2 weeks
24	PMIprices	Prices	M	R	Jan 1997–Jan 2022 (most regions, some start later)	2 weeks
25	PMIemploy	Employment	M	R	Jan 1997–Jan 2022 (most regions, some start later)	2 weeks
26	PMIfuture	Future orders	M	R	Jul 2012–Jan 2022 (most regions, some start later)	2 weeks
27	HousingRental	Private Housing Rental Prices	M	R	Jan 2005–Apr 2022 (most regions, some start later)	3 weeks
28	OvernightVisits	Number of overnight visits to the regions of the UK by area of residence	Q	R (selected regions)	Q1 2017–Q3 2021	5 months
29	Spending	Spending by overseas residents in regions of the UK by area of residence	Q	R (selected regions)	Q1 2009–Q3 2021	5 months
30	IncorporatedCompanies	Number of companies on the register, newly incorporated companies, and removals from the register.	Q	R (selected regions)	Q1 2011–Q1 2022	1 month
31	Scot_GDP	Scottish monthly GDP	M	R (Scot only)	Jan 2010–Mar 2022	2 months
32	Scot_LabourProductivity	Scottish Labour productivity	Q	R (Scot only)	Q1 1998–Q4 2019	5 months
33	Scot_RetailSales	Retail sales index for Scotland	Q	R (Scot only)	Q1 2008–Q1 2020	1 month
34	Scot_CSI	Scottish Consumer Sentiment Indicator	Q	R (Scot only)	Q2 2013–Q1 2022	1 month
35	NI_IOS	Northern Ireland Index of Services	Q	R (NI only)	Q1 2005–Q4 2021	3 months
36	NI_IOP	Northern Ireland Index of Production	Q	R (NI only)	Q1 2005–Q4 2021	3 months
37	NI_RSI	Northern Ireland Retail Sales Index	Q	R (NI only)	Q1 2014–Q4 2021	3 months
38	NI_Ports Traffic	Northern Ireland Ports Traffic	Q	R (NI only)	Q1 2009–Q4 2021	4 months
39	NI_ConstructionOutput	Construction output in Northern Ireland	Q	R (NI only)	Q1 2013–Q4 2021	3 months
40	VAT	VAT Turnover by ITL1/2/3 Regions	Q	R	Q1 2012–Q4 2021	5 months

Note. Q = Quarterly; M = Monthly; N = National; R = regional; SCOT = Scotland; NI = Northern Ireland; VAT = value-added tax; PMI = Purchasing Managers' Index.

regional GDP: they call this the 'cross-sectional constraint'. This paper sets out how to use this class of MF-VAR models when nowcasting using short data of the type seen in Table 1. We add to the MF-VAR regional factors leading to a mixed-frequency factor-augmented VAR (MF-FAVAR). At the beginning of the sample, the regional factors reflect information in only a few regional indicators. But as time passes and more variables become available, these data are also included in the factors.

In the statistical literature, there exist several ways of producing estimates of factors in the presence of missing data. In this paper, we investigate how well they work in the context of our regional nowcasting exercise given the mixed-frequency and ragged edge characteristics of the data shown in Table 1. Traditionally, missing data problems in factor-based macroeconomic analysis have been addressed using expectation maximisation principal components analysis (EMPCA); see Stock and Watson (2002). More recently, new approaches have been suggested, including the tall wide (TW) algorithm of Bai and Ng (2021) and the tall project (TP) algorithm of Cahan et al. (2023). These are all generic-factor estimation algorithms to handle missing data. We consider the relative performance of these different algorithms when dealing with our specific missing data configuration—the ragged edges at the beginning and end of the sample—and how to incorporate these factors into the MF-FAVAR.

It is worth stressing that these different factor extraction methods have the potential to be useful in a wide variety of regional nowcasting applications. They can be used to accommodate other evolving and ragged edge data features that might be evident for other variables and in other countries. This includes more (or less) timely and higher (lower) frequency regional output (GVA) data. Thus, they have the potential to overcome a major barrier to the more widespread use of new innovative data sources for regional nowcasting. Given the trend of increasing data availability at a regional level, and interest from policymakers in improving the spatial granularity of economic statistics, a regional nowcasting model such as ours, that is robust to different release delays, time-series lengths, and data frequencies, represents a valuable addition to the toolkit available to researchers. Via a UK application, we explore how our model performs in a pseudo-real-time nowcasting exercise and we test whether including these additional predictors improves the accuracy of our nowcasts of UK regional GVA.

The remainder of this paper is structured as follows. In Section 2, we describe in more detail the data we use in our empirical work. In Section 3, we set out the econometric methods used for nowcasting, including factor analysis of missing data. We explore the properties of three algorithms for the estimation of factors with ragged edge samples via a set of Monte Carlo exercises. Section 4 reports our empirical results, including a comparison of the factors produced across different methods and their performance in a pseudo-real-time nowcasting exercise. Section 5 concludes.

2 The data

In this section, we set out the data used in our model. We do this in two parts. First, we summarise the regional output growth data. Second, we describe the indicators that we incorporate into our model, via the calculation of regional factors, to nowcast quarterly regional output growth.

2.1 Regional output growth data

We combine official annual regional GVA growth data produced by the ONS since 1966 with quarterly regional GVA growth data once they become available. For the English regions and Wales, the ONS quarterly data date back to Q1 2012; Scottish government data date back to Q1 1998; and Northern Ireland Statistics and Research Agency data go back to Q1 2006. Pre-1998, the annual GVA data are available only in nominal terms, and we followed previous research and constructed a historical regional database for the UK by using a UK GDP deflator. For more details on the construction of this regional database, see Koop, McIntyre, et al. (2020) and Koop et al. (2020a, 2020b, 2022). In the absence of real-time data vintages for the regional GVA data, we consider only the latest-vintage GVA data.[4]

[4] In general, the release delay for the regional GVA data is 6 months, but the delay for the most recent data has been longer, meaning that at the time of writing the latest vintage is 2021Q4.

2.2 Regional indicators: short and long data

An increasing volume of regional economic data, from both official and unofficial sources, is becoming available. While we cannot claim that the data set we use in this paper represents complete coverage of the regional data available, it does represent a data set with features typical of regional economic data and it has coverage across many different types of economic data. For use in nowcasting, any data have to be released on a more timely basis than the target variable (in our case, regional GVA). This rules out annual indicators and any quarterly indicators that are released with too long a delay after the quarter to which they relate.[5] A number of other indicators exist, but we do not have access to them for a variety of reasons.[6] However, we should stress that the model we develop in this paper is capable of incorporating such data should they become available.

Having set out the criteria for data inclusion, Table 1 summarises the indicators that we ultimately include in the model. These indicators cover a range of time periods, with some measures having relatively short time spans (for example, payroll employment data that only go back to 2014), while others (for example, the house price information) cover the entire sample period. We also observe a range of release delays, with some indicators being released very quickly after the end of the reference period, and others being only slightly more timely than regional GDP itself. The indicators included in this model can be grouped into three main categories: measures of output, labour market data, and housing market indicators. Measures of output include data covering: construction sector output, retail sales, trade data, port traffic, tourism stays and spending, and business demographic information. As discussed, we also included output indices only available for Scotland and Northern Ireland (for example, the Northern Ireland Index of Services and the Northern Ireland Index of Production). At a quarterly frequency, there is also one survey of the business outlook, the CBI business optimism index available for all ITL1 regions, and a Scottish consumer sentiment indicator. A key monthly indicator of business activity is the PMI survey measure. This comprises separate indicators of activity that were each included separately: new business, outstanding business, charges, prices, employment, and future orders. For Scotland, there is also a monthly GDP measure.

This paper is the first to utilise VAT turnover aggregates at the regional level, provided to us by the ONS, in a pseudo-real-time nowcasting exercise. These data are disaggregated to the ITL3 (formerly NUTS3) regions of the UK, and are available on a quarterly basis from 2012Q1. For each ITL1 region, we use VAT data for that ITL1 region, as well as each of the ITL2 and ITL3 subregions within it, as separate variables used in the construction of the regional factors. Vintage data are available only from 2019Q4 (which, combined with the previously mentioned lack of real-time data for regional output growth, also explains our need to undertake a pseudo-real-time exercise). The typical publication lag of these data is 5 months after the end of the reference quarter. These data are produced by the ONS (although they are not typically available to researchers outside of the ONS). The data are aggregations of individual VAT returns from firms, which are cleaned by ONS to address any anomalies in the completion of these forms by businesses and the declared turnover data assigned to a quarter and a geographic location (reflecting an allocation to each ITL1/2/3 level). The issues involved in this data work by ONS are significant and challenging; see, for example, Labonne and Weale (2020).

Labour Force Survey data for the ITL1 regions of the UK are also incorporated. They are available with a release delay of around 6 weeks after the end of the quarter to which they relate. These data are released on a rolling 3-month basis and are updated each month. The LFS provides a range of measures of labour market activity, from which we select headline employment and unemployment, and public-sector employment. There are a small number of monthly labour market measures. These include data from the social security system (for example, the claimant count rate), as well as the HMRC real-time information data from the PAYE system on payroll employment and pay. Given the important role of the housing market in the economy (see Leamer, 2007),

[5] For example, in the UK there are local-level data on lending to small- and medium-size enterprises and also mortgage lending, but neither is more timely than our target measure of regional output.

[6] For example, the need for a commercial subscription (e.g. GfK consumer confidence data), the lack of data in aggregate form (e.g. ONS's Monthly Business Survey microdata and credit/debit card transactions data), or the data being privately held.

we also include two measures of changes in house prices. The first of these is monthly house price index data for each ITL1 region published by the UK government, as well as the quarterly UK house price index published by the Nationwide Building Society (which has the advantage of being more timely than the official data series). We also include information on rental prices for the private rental market.

Finally, we note that some of our regional variables are available at the monthly frequency but our econometric model is a quarterly model. Accordingly, we use the monthly observation for the final month of the quarter so that our nowcasts reflect the most recent information.

3 Econometric methods

3.1 Notation and data availability

We begin by describing some variable definitions, relationships, and notational conventions used in this paper.

- $t = 1, \ldots, T$ runs at a *quarterly* frequency.
- $r = 1, \ldots, R$ denotes the R regions in the UK.
- Y_t^{UK} is GVA for the UK in quarter t.
- $y_t^{\mathrm{UK}} = \log(Y_t^{\mathrm{UK}}) - \log(Y_{t-1}^{\mathrm{UK}})$ is the quarterly change (log difference) in GVA in the UK.
- Y_t^r is GVA for region r in quarter t. It is not observed before 2012 except for Scotland and Northern Ireland, where it is not observed before 1998 and 2006, respectively.
- $Y_t^{r,A} = Y_t^r + Y_{t-1}^r + Y_{t-2}^r + Y_{t-3}^r$ is annual GVA for region r. It is observed in quarter 4 of each year, but not in other quarters.
- $y_t^{r,A} = \log(Y_t^{r,A}) - \log(Y_{t-4}^{r,A})$ is annual GVA growth in region r. It is observed, but only in quarter 4 of each year. $\boldsymbol{y}_t^A = (y_t^{1,A}, \ldots, y_t^{R,A})'$ is the vector of annual GVA growth rates for the R regions.
- $y_t^r = \log(Y_t^r) - \log(Y_{t-1}^r)$ is the quarterly change in GVA in region r. It is not observed before 2012 except for Scotland and Northern Ireland, where it is not observed before 1998 in Scotland and 2006 in Northern Ireland. $\boldsymbol{y}_t^Q = (y_t^1, \ldots, y_t^R)'$ is the vector of quarterly GVA growth rates for the R regions.
- The link between the quarterly regional growth rates and their annual counterpart is referred to as the inter-temporal restriction and takes the form:

$$y_t^{r,A} = \frac{1}{4}y_t^r + \frac{1}{2}y_{t-1}^r + \frac{3}{4}y_{t-2}^r + y_{t-3}^r + \frac{3}{4}y_{t-4}^r + \frac{1}{2}y_{t-5}^r + \frac{1}{4}y_{t-6}^r. \tag{1}$$

  This restriction is not imposed in periods where quarterly regional GVA growth data are available.
- The link between the regional growth rates and the UK counterpart is referred to as the cross-sectional restriction and takes the form:[7]

$$y_t^{\mathrm{UK}} \approx \frac{1}{R}\sum_{r=1}^{R} y_t^r. \tag{2}$$

- $\boldsymbol{Z}_t^r$ is a vector containing k_r quarterly variables for region r. These are the 'short' data that start at differing times as described in Section 2.

[7] The cross-sectional restriction given here assumes growth rates are modelled as log differences; see Koop et al. (2020b) for the derivation of this restriction. As discussed in this paper, the constraint is approximate to reflect both the logarithmic approximation and the fact that regional output does not exactly sum to UK output because of the UK continental shelf.

3.2 The MF-FAVAR

To explain the structure of our MF-FAVAR, we begin with a structural VAR that relates a vector of N dependent variables, $\boldsymbol{y}_t$ to lags of the dependent variables:

$$\boldsymbol{B}\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{\varepsilon}_t, \tag{3}$$

where $\boldsymbol{x}_t$ is a vector containing p lags of $\boldsymbol{y}_t$. The errors, $\boldsymbol{\varepsilon}_t$, are assumed to be $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a diagonal matrix[8] and $\boldsymbol{B}$ is lower triangular with ones on the diagonal.[9]

In a conventional VAR, all of the dependent variables are simply observed variables. Note that $\boldsymbol{A}$ is an $N \times Np$ matrix and, thus, there are pN^2 VAR coefficients to be estimated. If N and/or p is large, VARs can be seriously over-parameterised. We investigated lag lengths of up to $p = 7$ (and our results are robust, so our main results set $p = 1$). Thus, our $\boldsymbol{A}$ matrix is big. Accordingly, we use Bayesian estimation methods that allow for prior shrinkage.

The MF-VAR is a VAR where $\boldsymbol{y}_t$ is no longer simply a set of *observed* variables. Instead, some of the elements in $\boldsymbol{y}_t$ are unobserved or latent variables. In particular, there are the unobserved high-frequency values of the low-frequency variables: the objects we wish to estimate. The latter are linked to the former via the inter-temporal restriction. We set $\boldsymbol{y}_t = (y_t^{\mathrm{UK}}, \boldsymbol{y}_t^{Q\prime})$, where $\boldsymbol{y}_t^{Q}$ contains the quarterly regional growth rates that we were seeking to estimate.

The MF-VAR can be set up as a state-space model, where the state equations are given by the VAR and the measurement equations are the inter-temporal and cross-sectional restrictions. Bayesian methods exist for posterior and predictive inference in such state-space models. For details, see Koop et al. (2020a, 2020b). We use variational Bayes (VB) methods instead of Markov chain Monte Carlo (MCMC) estimation, because of their computational advantages. Variational Bayes methods for the MF-VAR are developed in Gefang et al. (2020) and used by Koop et al. (2022). In this paper, we also use a computationally more efficient precision-based approach to estimate the states, instead of the Kalman filter; see Chan et al. (2023).

The MF-FAVAR we use is an MF-VAR, but with one final re-definition of $\boldsymbol{y}_t$. In particular, it sets $\boldsymbol{y}_t = (y_t^{\mathrm{UK}}, y_t^{Q\prime}, \boldsymbol{f}_t')$, where $\boldsymbol{f}_t = (\boldsymbol{f}_t^1, \ldots, \boldsymbol{f}_t^R)'$ is a vector of regional factors. It is possible to have more than one factor for each region and, accordingly, each of the $\boldsymbol{f}_t^r$ is a vector of n_f factors constructed using $\boldsymbol{Z}_t^r$ for $r = 1, \ldots, R$. These factors are observed at a quarterly frequency and, thus, can be treated as additional high-frequency observed variables in the MF-VAR.

3.3 Further discussion of the MF-FAVAR

Several issues issues relating to the econometric estimation of the model are worth elaborating on. These relate to our choice of MF-FAVAR instead of other possible approaches, the treatment of the estimated regional factors as additional variables in the MF-VAR and the high dimensionality of the model.

The alternatives to our MF-FAVAR would be either a dynamic factor model (DFM) or an MF-VAR involving *all* of the variables (so the regional output growth data and the short and long data). Even though large VAR methods have enjoyed great popularity and in many cases, for example in Banbura et al. (2010), been found to forecast better than DFMs, the enormous dimension of the MF-VAR that would arise and the large number of missing values we would have in our regional application would make estimation very difficult. In contrast, DFMs would be much easier to estimate. They can also be used with mixed-frequency and missing data and can be used both to interpolate and to nowcast; for example, see Angelini et al. (2006) on interpolation, and Banbura and Runstler (2011) and Frale et al. (2011) on nowcasting. Our MF-FAVAR is a compromise between these two approaches, enjoying some of the parsimony of the DFM and some of the forecasting advantages of the VAR. It also has the benefit of producing easy to interpret regional factors which depend only on data specific to that region.

It is also worth noting that the DFM differs from our MF-FAVAR in its treatment of the ragged edge at the end of the sample. That is, the DFM would interpolate the missing observations at the

[8] It is straightforward to extend this model to allow for parameter change by assuming that the diagonal elements follow stochastic volatility processes if parameter instability is a worry.

[9] Working in this structural VAR form does not restrict the reduced-form error covariance matrix and greatly simplifies computation, since estimation can proceed one equation at a time.

beginning of the sample using factor methods (as we do), but also use factor methods to fill in missing observations at the end of the sample—as arises when nowcasting given the ragged edge. This strategy is slightly different than what we do when nowcasting with our MF-FAVAR in that the ragged edge in the Z_t^r is handled using factor methods (namely one of TW, TP, or EMPCA) but the ragged edge in the regional GVA growth variables being nowcast is handled using the VAR methods involving the precision sampler.

Once the decision is made to use a factor method (either our MF-FAVAR or a DFM), the factors must be obtained in some way. In this paper, we use a two-step method where we first construct the factors using a method such as EMPCA and then plug them into the MF-FAVAR. The use of estimated factors in the MF-FAVAR raises the generated regressors issue. Frequentist theoretical results in Bai and Ng (2006) establish that when $\sqrt{T}/N \to 0$ the factors estimated at the first step can be treated as known when estimating factor-augmented regressions at the second step (when using ordinary least squares). Since we are using Bayesian methods and our focus is on nowcasting not parameter estimation, the applicability of the Bai and Ng (2006) results are not clear. But given that, in our application, $\sqrt{T} < N$ it is likely that the density nowcasts produced from our MF-FAVAR will be little impacted by the generated regressors issue. In our Monte Carlo experiments, provided in the online supplementary Appendix, we provide some indicative evidence that our constructed factors are providing reliable estimates of regional factors.[10]

The alternative to treating the estimated factors as known in the FAVAR is to treat them as latent states and write down a parametric state-space model governing both their determination and their relationship with the observed variables of interest. Such a joint MF-FAVAR model can then be estimated—in one-step—by frequentist or Bayesian methods. The advantages and disadvantages of the one-step and two-step approaches have been discussed in the literature. In their original paper, Bernanke et al. (2005) compared a specific Bayesian implementation of the one-step FAVAR estimator with a two-step estimator.[11] They concluded that the two-step approach is both computationally simple and less likely to be mis-specified since this approach is semi-parametric (that is, it does not involve making parametric assumptions about the factors). Given that interest in this paper is on nowcasting regional GVA and exploring the relative value of different algorithms designed to estimate factors from missing/short data, we therefore confine our attention to two-step methods. In this context, the issue of whether we are ignoring estimation uncertainty in 'true' regional factors is arguably subsidiary to the issue of whether alternative methods of constructing factors with short data improve nowcast performance in practice.

The next issue worthy of additional discussion arises from the fact that our MF-FAVAR is of very high dimension, including at a minimum $N = R \times n_f + 1$ dependent variables. This high dimension arises since, for each region, we are including regional output growth plus n_f factors and for the UK we include at least GVA growth. Even if we only include one factor for each of the 12 regions and no additional UK-wide variables in the model, we end up with a VAR of dimension $N = 25$, which is already quite large. In practice, the need to include more than one factor and/or additional UK variables means that most of our models are of a much higher dimension. Thus, we face the risk that our models are over-parameterised.

We partially surmount the over-parameterisation problem by working with a restricted version of the MF-FAVAR. This imposes the restriction that the equation for GVA growth for a particular region depends only on the regional factor for that region (as well as UK variables and lags of GVA growth for that region). In other words, each regional factor is specific to a region and does not appear in the equations for other regions. As a robustness check, we also estimated the unrestricted version of the model; see online supplementary Appendix B.

We also use Bayesian prior shrinkage as a way of avoiding over-parameterisation concerns. There exist a range of VAR priors and any of them could be used. In this paper, we use the popular adaptive Lasso (AL). This is a global-shrinkage prior that automatically chooses which coefficients should be shrunk to zero; see Zou (2006) for details. We have experimented with two versions of

[10] We note that there are alternative frequentist strategies for surmounting the generated regressors issue. For instance, bootstrap methods can be used; for example, see Goncalves and Perron (2014).

[11] Chan et al. (2023) provide a more general, and computationally more efficient, Bayesian framework for (one-step) estimation of state-space models with missing data.

the AL prior. One of these uses the AL on all the coefficients in the model (except for error variances, which are the diagonal elements of $\mathbf{\Sigma}$ for which we use relatively non-informative inverse-Gamma priors). In the other version, we use the AL for all coefficients except for the error covariances (that is, the parameters controlling the contemporaneous relationships between the variables in the model). For these, we adopt the asymmetric conjugate prior (ACP) of Chan (2022), which might be expected to have good properties for these parameters. We name our two priors AL and AL-ACP, respectively. In practice, we find that AL-ACP leads to slightly better forecasts and hence the results presented in the main body of the paper are for AL-ACP. In online supplementary Appendix B, we present the AL results.

In summary, our algorithm begins by constructing the factors, using methods described in the next subsection, and then includes them in the MF-FAVAR. We then use Bayesian methods to construct nowcasts of quarterly GVA growth for the UK regions.

3.4 Constructing regional factors with short data

In this section, we describe three methods for constructing factors from $\boldsymbol{Z}_t^r$, which, as seen in Table 1, is a matrix characterised by missing data at the beginning and end of the sample. These algorithms are denoted by the abbreviations EMPCA, TW, and TP. We now provide a brief description of each algorithm in turn, along with a summary of its properties. Full details are provided in Stock and Watson (2002) for EMPCA, in Bai and Ng (2021) for TW, and in Cahan et al. (2023) for TP.

All of these approaches are algorithms for filling in missing data where various patterns of missingness are possible. In the present paper, as discussed, we are especially interested in one particular pattern of missingness: the ragged edge at the beginning of the sample due to the short nature of many of the $\boldsymbol{Z}_t^r$ indicators. In order to explore the relative performance of the three algorithms across data sets with 'short' data, we simulate data sets with different properties (including sample size, number of variables, and degree of missingness) in a set of Monte Carlo experiments summarised in Section 3.4.4 below. We first summarise the three-factor estimation methods.

3.4.1 Expectation maximisation principal components analysis

Principal component analysis (PCA) is a popular method for estimating factors. It is a non-parametric method. The advantage of this is that it is less liable to specification error than parametric approaches. But this can also be a disadvantage, as sensible parametric assumptions can improve estimation accuracy (for example, in a macroeconomic context, factors might be expected to exhibit autoregressive behaviour and a parametric model that allows for such behaviour could improve estimation accuracy). As discussed, for example in Section 2.3 of Stock and Watson (2016), PCA exploits correlations across variables to produce estimates of the factors.

When data are missing, Stock and Watson (2002) combine PCA with methods to fill in missing data using an expectations–maximisation (EM) algorithm, leading to EMPCA. In our context, each regional factor, $\boldsymbol{f}_t^r$, is calculated using $\boldsymbol{Z}_t^r$ which contains missing values. Let $\hat{\boldsymbol{Z}}_t^r$ be a version of $\boldsymbol{Z}_t^r$ with these missing values replaced by estimates and $\hat{\boldsymbol{f}}_t^r$ denote estimates of the factors. Expectation maximisation principal components analysis is an iterative algorithm that uses PCA on $\hat{\boldsymbol{Z}}_t^r$ to produce $\hat{\boldsymbol{f}}_t^r$. Estimates of the missing values are produced as fitted values from a regression of $\hat{\boldsymbol{Z}}_t^r$ on $\hat{\boldsymbol{f}}_t^r$.

3.4.2 Tall wide

The TW algorithm of Bai and Ng (2021) is not an iterative algorithm and, thus, does not suffer from the computational issues that can afflict EMPCA. To understand TW, consider the $T \times k_r$ matrix of data for a specific region, $\boldsymbol{Z}_t^r$. Each column contains all the data for a variable. Some of the columns will not have any missing values for any time periods. These columns form what is called the 'tall' block; first undertake PCA using this block, produce factors $\hat{\boldsymbol{f}}_t^{r,\mathrm{Tall}}$ and factor loadings $\hat{\boldsymbol{\Lambda}}^{r,\mathrm{Tall}}$. Next consider the rows of $\boldsymbol{Z}_t^r$. Some of these rows will not have missing values and these rows will form the 'wide' block. PCA can be applied to this wide block to produce factors $\hat{\boldsymbol{f}}_t^{r,\mathrm{Wide}}$ and factor loadings $\hat{\boldsymbol{\Lambda}}^{r,\mathrm{Wide}}$. The TW algorithm is a simple method for combining $\hat{\boldsymbol{f}}_t^{r,\mathrm{Tall}}$, $\hat{\boldsymbol{\Lambda}}^{r,\mathrm{Tall}}$, $\hat{\boldsymbol{f}}_t^{r,\mathrm{Wide}}$, and $\hat{\boldsymbol{\Lambda}}^{r,\mathrm{Wide}}$ in an optimal way (using least squares regression methods) to produce a single regional factor.

Bai and Ng (2021) show the TW algorithm to have desirable asymptotic properties. These properties depend on the number of columns/rows in the tall/wide blocks. Note that if the tall block is very narrow (that is, few variables have data available for the full sample), then $\hat{f}_t^{r,\mathrm{Tall}}$ will be based on few variables and may be a poor estimate of the regional factor(s). Similarly, if the wide block is thin, then $\hat{f}_t^{r,\mathrm{Wide}}$ will only be available for a few observations and may produce poor estimates. In our regional nowcasting context, suppose, for instance, that the variable with the latest start date begins in 2010. In this case, $\hat{f}_t^{r,\mathrm{Wide}}$ would only be available for 2010 onward. Furthermore, all the observations pre-2010 will be discarded other than observations for variables in the tall block. We might therefore expect the TW algorithm to be sensitive to the choice of variables and that it would not work well if one (or a few) of the variables is available for a short period of time.

3.4.3 Tall project

The TP algorithm of Cahan et al. (2023) is similar to TW in that the tall block plays a key role and $\hat{f}_t^{r,\mathrm{Tall}}$ and $\hat{\boldsymbol{\Lambda}}^{r,\mathrm{Tall}}$ are key ingredients in the estimated regional factor(s). However, it surmounts the problem noted previously that occurs with the TW algorithm when one of the variables has a very short time and, thus, the wide block is thin. It does so by using auxiliary regressions for the observed values of each individual variable (other than those in the tall block) on the tall block factors. The auxiliary regression for variable i can be used to fill in the missing values for variable i, thus leading to $\hat{\boldsymbol{Z}}_t^r$ that does not have missing values. The regional factor is estimated using PCA on $\hat{\boldsymbol{Z}}_t^r$. With this algorithm, it is possible to iterate, but Cahan et al. (2023) show that asymptotically this is not necessary. In this paper, we do not iterate.

3.4.4 Monte Carlo evaluation of the three-factor algorithms

This section summarises the results from a set of Monte Carlo experiments designed to evaluate the EMPCA, TW, and TP algorithms when applied to data sets with ragged edges, to varying degrees, at the beginning of the sample. Full details of the data-generating process and the simulation exercise are in the online supplementary Appendix.

Expectation maximisation principal components analysis is an iterative algorithm that raises two computational issues: it is fundamentally slower than non-iterative methods and can fail to converge. Tall wide and TP are simpler algorithms, fast, and not subject to concerns about convergence. However, they are likely to be more sensitive to the number of variables without missing observations (that is, the size of the tall block). In this paper, where our application involves a substantial ragged edge at the beginning of the sample and our tall block potentially contains only a small number of variables, it is possible that these properties of the TW and TP algorithms will make it worthwhile to take on the larger computational burden of EMPCA. But the choice between the three algorithms is fundamentally an empirical issue.

To assess the precision of the estimates of the factors from the three algorithms, we conduct a set of Monte Carlo experiments. Data are generated from the factor model used in Banbura and Modugno (2014). We generate data for different sample lengths T, different sizes of the cross-sectional panel n, and different values for τ (which governs the degree of cross-correlation of the idiosyncratic component). Then we leave the first two simulated variables as complete: this is our tall block. We let two variables remain complete because in our regional nowcasting application we have two indicators with no missing data. For the remaining $(n-2)$ variables, we set a certain fraction of the data as missing. Given our interest in the ragged edge at the beginning of the sample, we place these missing data points at the beginning of the sample. We consider cases of 0%, 20%, 40%, 60%, and 80% of missing data. Then, using the simulated data, we estimate factors using the three algorithms.

To evaluate the precision of the factor estimates, we follow Banbura and Modugno (2014) and compute the trace R^2 of the regression of the estimated factors on the true ones. Table A1 in the online supplementary Appendix reports average trace statistics over 500 Monte Carlo replications for EMPCA, and Table A2 in the online supplementary Appendix reports the statistics for TW and TP. From these tables we see that the three algorithms have similar estimation accuracy. The estimates are less precise for small sample lengths ($T = 50$ vs. $T = 200$), small cross-sections ($n = 12$ vs. $n = 102$), a mis-specified model ($\tau > 0$ vs. $\tau = 0$) in small samples, and a large fraction of missing data. Table A1 in the online supplementary Appendix also shows that the EMPCA algorithm is slower than the other two approaches. But the additional computational burden is small. More significantly, however, there were some instances of convergence failures. This included samples where $T = 200$, $n = 50$, and 60% of the sample is missing: this matches the features of our regional nowcasting data set (seen in Table 1). Thus, we suspect that there might be convergence problems when using EMPCA in our application.

4 Regional nowcasting with the MF-FAVAR

4.1 Design of the nowcasting exercise and specification choices

In our regional nowcasting exercise, we will compare different versions of our MF-FAVAR to a MF-VAR that is identical (that is, same prior, same lag length choice, etc.), except that it does not include the regional factors extracted from the short (and longer) data summarised in Table 1. The MF-VAR is our 'benchmark' model. In terms of the indicators in the models, all include GVA growth for the 12 regions plus the UK as a whole, along with four quarterly UK macroeconomic predictors (inflation, interest rates, the exchange rate, and oil price inflation). Thus, our smallest model, the MF-VAR, includes 17 variables. Then in the MF-FAVARs we add a set of regional factors. Note that the factor for region r is based only on data from that region (i.e. $\boldsymbol{Z}_t^r$ contains data specific to region r, including sub-regional VAT data for region r). The variables are given in Table 1. The number of variables used to construct the factors will differ across regions due to data availability and the fact that the number of sub-regions varies across regions.

The MF-VARs differ in four ways:

1. The way the factors are calculated (EMPCA, TW, or TP);
2. The number of regional factors included (the choice of n_f);
3. Whether the VAT data are included in the calculation of the regional factors or not, and the assumed publication lag of the VAT data;
4. How the regional factors are selected.

The first three differences listed above should be clear, but Point (4) requires additional explanation. Expectation maximisation principal components analysis, TW, and TP all involve using PCA. PCA using a data set of k_r variables will produce k_r factors. These are typically ordered (and selected for inclusion in the VAR) according to the proportion of the variability in the data which they explain (that is, based on the eigenvalues of the sample covariance matrix). A small number of factors are typically chosen that account for most of the variance in the data; for example, see McCracken and Ng (2021). One of our models uses this approach. That is, we simply add into the VAR the $n_f \ll k_r$ factors with the highest eigenvalues. However, it is possible that the factors with the highest eigenvalues may not be the ones that are the most useful for nowcasting regional GVA growth. Accordingly, we also experiment with selecting for inclusion in the MF-VAR, not those factors with the highest eigenvalue but those with the highest (in-sample, updated recursively) correlation with UK GVA growth. In the body of the paper, we report results for a selection of these MF-FAVARs with a focus on the issue of factor construction. A complete set of results using all our models is given in online supplementary Appendix B. Overall, we find a high degree of robustness.

A final element to set out is the typical release calendar for the regional output data. We mimic this calendar in the design of our out-of-sample exercise. The current release schedule is presented in Figure 1. We choose to time the estimation of our model each quarter to coincide with the release of the UK growth rate for the previous quarter: this is approximately in the middle of the following quarter. From Figure 1, we can see that, given the release delays in producing estimates of regional output growth, there are three quantities of interest that can be produced using our model.

- 'Backcasts': An estimate of regional growth produced in quarter τ (say, 2020Q2) but which relates to quarter $\tau - 2$ (say, 2019Q4);
- 'Estimates': An estimate of regional growth produced in quarter τ (say, 2020Q2) but which relates to quarter $\tau - 1$ (say, 2020Q1);

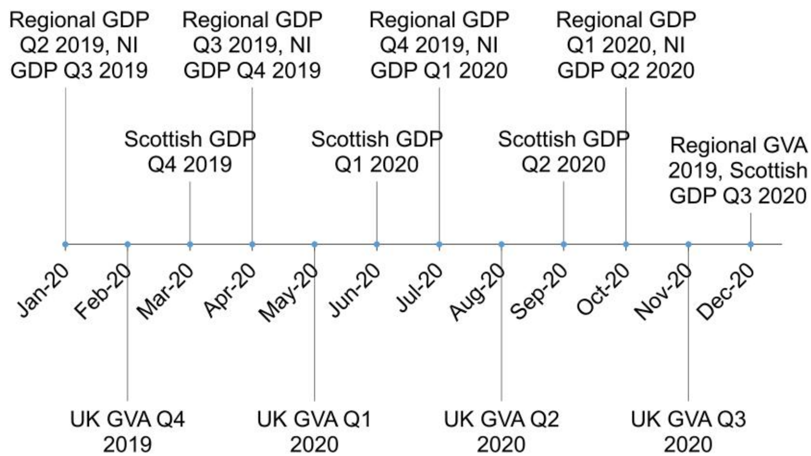

Figure 1. Typical release schedule for national and regional output data in the UK.

- 'Nowcasts': An estimate of regional growth produced in quarter τ (say, 2020Q2) but which relates to quarter τ (say, 2020Q2).

We produce 'nowcasts' and 'estimates' for each region every time we run the model, but given the release schedule, we only produce 'backcasts' for the English regions and Wales (because, continuing the example in 2020Q2, we already had estimates of Scottish and Northern Irish growth in 2019Q4, as these were released in March and April of 2020, respectively). Our 'nowcasts' are estimates of regional growth in the quarter for which we do not yet have official data for the UK as a whole. When producing our 'estimates' and 'backcasts', we already have the official estimate of UK growth for the corresponding quarter and these data are included in the model. In calculating the factors, we respect the release calendar for the predictors as set out in Table 1.

We use an expanding window of data to estimate the models used to produce our nowcasts, estimates, and backcasts. Aware of possible instabilities in nowcast performance (see Rossi, 2021 and Tables B19 and B20 in the online supplementary Appendix), we present indicative evidence, given the short sample, showing that the relative nowcast accuracy of our models is stable over time.

4.2 Nowcasting results: out-of-sample evidence

Having set out the timing convention for the production of our estimates, nowcasts, and backcasts we now present an evaluation of the performance of our models in estimating regional GVA growth. Our out-of-sample evaluation period is 2014Q2 through 2021Q4.[12] We present the results from a number of different model specifications, as described earlier. These include results from the following three models:

1. An MF-VAR model (our benchmark model with no short data);
2. An MF-FAVAR model;
3. An MF-FAVAR model, where the VAT data are not included, but all of the other regional (short) predictors are considered when estimating the factors.

In all cases, we evaluate the accuracy of the point and density nowcasts, estimates, and backcasts using the root mean square forecast error (RMSFE) and the continuous ranked probability score (CRPS), respectively. Lower values of each of these metrics indicate improved accuracy.

[12] Due to the lack of availability of real-time data vintages for our variables, our recursive out-of-sample analysis has to be 'pseudo'-real time. It is also worth noting that, although the focus of this paper is on nowcasting regional GVA, our model could also be used be used for short term forecasting of UK GVA.

Table 2 presents the RMSFE metrics by region for each model, namely, the benchmark MF-VAR and the three-factor-augmented MF-VAR models (MF-FAVAR), one for each of the three-factor estimation algorithms, produced using the AL-asymmetric conjugate prior. The RMSFE results for the factor-augmented VAR models are presented relative to those from the benchmark MF-VAR, such that ratios less than one indicate superior forecast accuracy relative to the benchmark. Several conclusions can be drawn from Table 2.

There is a clear pattern of accuracy improving as we move from our nowcast to our estimates to our backcasts. This reflects the accumulation of information that takes place between the production of each of these estimates over a given quarter. However, it is also notable that there is a much larger improvement in model performance as we go from the nowcast to the estimate, and a smaller improvement as we move from the estimate to the backcast. This is consistent with the finding in Koop et al. (2022), and reflects the fact that we know the aggregate (UK) estimate for that quarter when we produce our regional estimate (but not our regional nowcast). Crucially, it also reflects the presence of the additional measurement equation (which we refer to as the 'cross-sectional restriction') relating regional growth to aggregate national growth. Comparing the accuracy of the models including the additional short indicators set out earlier in this paper against the benchmark model, we generally see little or no improvement in the accuracy of the estimates or the backcasts. But there is an improvement in the accuracy of the nowcasts. Adding in the short data then helps. This conclusion holds across the different factor estimation algorithms (EMPCA, TW, and TP).

These conclusions hold when we evaluate the density nowcasts, estimates, and backcasts in Table 3 using the CRPS. Using Diebold and Mariano (1995) tests, we explore whether there are any statistically significant improvements in individual regions using the different MF-FAVAR models. For the density estimates and backcasts there are no statistically significant improvements, but the density nowcasts are almost always statistically significantly more accurate. This makes sense and is again consistent with the existing literature, in particular (Koop et al., 2022). It reflects the fact that the largest improvement in the accuracy of our estimates and backcasts comes from conditioning directly on the equivalent UK estimate (which itself reflects much of the information contained in the additional indicators). When we make the prediction of regional growth earlier and, as a result, we do not yet know the UK outturn for a given quarter, the additional indicators, as captured by the factors, lead to substantial and statistically significant improvements in the accuracy of our nowcasts.

Having compared the performance of our FAVAR model to our benchmark MF-VAR, a specific question arises about the role of VAT data relative to other (short) regional indicators in improving the accuracy of the regional backcasts, estimates, and nowcasts. In order to explore this issue we re-ran the MF-FAVAR without the VAT data. The results from this additional exercise are presented in the lower panels of Tables 2 and 3. These show results that are consistent with those presented above; in particular, they bear the same result relative to the MF-VAR benchmark model, and in some cases, the RMSFE/CRPS estimates are marginally better than those above. This is evidence that these VAT data are not adding significantly to our ability to nowcast regional GDP relative to our model with other regional indicators but no VAT data.

To check robustness to our modelling choices, in the online supplementary Appendix we present additional results covering cases where: (a) regional factors are not restricted to affect only their own region's output but are allowed to affect output in other regions too; (b) we use a different prior (the AL); (c) we use a different lag length in the VAR; (d) we include a different number of factors in the model (five rather than three); (e) we select which factors to include based on their correlations with UK output growth rather than on the size of their eigenvalues; (f) the regional VAT data are assumed to be available on a more timely basis than currently;[13] (g) the cross-sectional restrictions are switched off; (h) we re-estimate the VAR with the VAT data but dropping those regional indicators not available over the full sample;[14] and (i) we evaluate accuracy over different evaluation samples. Apart from (h), none of these modelling variants delivers consistent improvements in the accuracy of our different estimates relative to the results presented in the main

13 This is a counterfactual simulation exercise, designed to ascertain if the VAT data would be more useful if published more quickly than at present.

14 This simulation exercise, suggested by a referee, lets us conclude that consideration of the VAT data does not compensate for the omitted (short) regional variables. It sheds light on the trade-off between the number of regional variables and the information contained in the VAT data.

Table 2. RMSFE by region (multiplied by 100)

		NE	NW	York	EM	WM	EE	LON	SE	SW	WA	SCOT	NI	Average
Benchmark MF-VAR (RMSFE)														
	Nowcast	2.84	1.82	2.10	1.87	1.96	2.08	3.35	1.80	1.72	2.46	1.63	1.71	2.11
	Estimate	0.84	0.48	0.54	0.58	0.41	0.38	1.21	0.40	0.57	0.68	0.16	0.38	0.55
	Backcast	0.42	0.30	0.35	0.36	0.26	0.27	0.79	0.19	0.31	0.34	–	–	0.36
MF-FAVAR: model including VAT data (RMSFE ratios relative to benchmark)														
EMPCA	Nowcast	0.59	0.89	0.80	0.92	0.82	0.79	0.73	0.84	0.90	0.74	0.76	0.80	0.79
	Estimate	0.74	1.15	1.15	1.16	1.10	1.05	0.91	1.03	1.04	0.91	0.94	0.87	0.98
	Backcast	0.79	1.03	1.09	1.14	1.04	0.96	0.89	1.11	1.13	0.94	–	–	0.97
TW	Nowcast	0.60	0.89	0.80	0.90	0.82	0.79	0.73	0.85	0.90	0.74	0.75	0.80	0.79
	Estimate	0.74	1.15	1.15	1.14	1.10	1.05	0.91	1.05	1.04	0.93	0.94	0.87	0.98
	Backcast	0.79	1.03	1.09	1.14	1.08	0.93	0.89	1.16	1.13	0.94	–	–	0.97
TP	Nowcast	0.61	0.90	0.79	0.91	0.82	0.79	0.73	0.85	0.90	0.74	0.75	0.81	0.79
	Estimate	0.74	1.17	1.15	1.16	1.12	1.08	0.91	1.03	1.04	0.93	0.94	0.89	0.98
	Backcast	0.79	1.03	1.09	1.14	1.08	0.96	0.89	1.11	1.13	0.94	–	–	0.97
MF-FAVAR: without VAT data (RMSFE ratios relative to benchmark)														
EMPCA	Nowcast	0.61	0.87	0.79	0.90	0.82	0.78	0.67	0.87	0.90	0.76	0.75	0.82	0.78
	Estimate	0.69	1.10	1.11	1.10	1.10	1.05	0.80	1.05	0.98	0.97	0.94	0.92	0.96
	Backcast	0.74	1.01	1.03	1.08	1.04	0.93	0.80	1.21	1.06	0.97	–	–	0.94
TW	Nowcast	0.61	0.88	0.79	0.90	0.83	0.79	0.67	0.88	0.91	0.77	0.76	0.82	0.79
	Estimate	0.71	1.13	1.13	1.10	1.07	1.03	0.80	1.05	0.98	0.99	0.94	0.92	0.96
	Backcast	0.76	1.01	1.06	1.08	1.04	0.89	0.80	1.16	1.10	0.97	–	–	0.94
TP	Nowcast	0.61	0.87	0.78	0.90	0.82	0.78	0.66	0.88	0.90	0.77	0.75	0.81	0.78
	Estimate	0.69	1.10	1.11	1.09	1.10	1.03	0.80	1.08	0.98	0.99	0.94	0.89	0.96
	Backcast	0.71	1.01	1.03	1.08	1.04	0.93	0.80	1.16	1.06	0.97	–	–	0.94

Note. NE, North East England; NW, North West England; York, Yorkshire and the Humber; EM, East Midlands; WM, West Midlands; EE, East of England; LON, London; SE, South East England; SW, South West England; WA, Wales; SCOT; Scotland; NI, Northern Ireland; RMSFE, root mean square forecast error; MF-VAR, mixed-frequency vector autoregression; VAT, value-added tax; MF-FAVAR, mixed-frequency factor-augmented VAR; EMPCA, expectation maximisation principal components analysis; TW, tall wide; TP, tall project.
*denotes rejection of the null of equal forecast accuracy against the benchmark MF-VAR model at the 0.10 significance level using a two-sided (Diebold & Mariano, 1995) test.

Table 3. Average CRPS by region (multiplied by 100)

		NE	NW	York	EM	WM	EE	LON	SE	SW	WA	SCOT	NI	Average
Benchmark MF-VAR (CRPS)														
	Nowcast	1.21	0.87	0.99	0.91	0.89	0.91	1.23	0.84	0.86	1.07	0.61	0.72	0.93
	Estimate	0.43	0.29	0.31	0.33	0.27	0.26	0.51	0.26	0.32	0.37	0.11	0.20	0.30
	Backcast	0.23	0.17	0.19	0.19	0.16	0.15	0.28	0.14	0.18	0.20	–	–	0.19
MF-FAVAR: model including VAT data (CRPS ratios relative to benchmark)														
EMPCA	Nowcast	0.70*	0.93*	0.85	0.93	0.89*	0.87*	0.85*	0.88*	0.92*	0.82*	0.84*	0.82*	0.85
	Estimate	0.84	1.07	1.06	1.06	1.07	1.02	0.96	1.00	1.03	0.97	1.00	0.95	1.00
	Backcast	0.87	1.06	1.05	1.11	1.06	1.02	0.96	1.00	1.06	0.95	–	–	1.00
TW	Nowcast	0.72*	0.93*	0.85	0.92	0.88*	0.87*	0.85*	0.89*	0.92*	0.83*	0.82*	0.82*	0.85
	Estimate	0.84	1.07	1.06	1.06	1.07	1.01	0.96	1.00	1.03	0.97	1.00	0.95	1.00
	Backcast	0.87	1.06	1.05	1.11	1.06	1.01	0.96	1.00	1.06	0.95	–	–	1.00
TP	Nowcast	0.72*	0.94*	0.85	0.92	0.89*	0.88*	0.85*	0.89*	0.92*	0.83*	0.82*	0.82*	0.85
	Estimate	0.84	1.07	1.06	1.06	1.07	1.00	0.96	1.00	1.03	0.97	1.00	0.95	0.98
	Backcast	0.87	1.06	1.05	1.11	1.06	0.99	0.96	1.00	1.06	0.95	–	–	0.97
MF-FAVAR: without VAT data (CRPS ratios relative to benchmark)														
EMPCA	Nowcast	0.69*	0.89*	0.81*	0.88	0.84*	0.82*	0.80*	0.87*	0.88*	0.81*	0.80*	0.79*	0.82
	Estimate	0.74	0.97	0.97	0.97	0.96	0.88	0.88*	0.88	0.91	0.92	0.91	0.90	0.90
	Backcast	0.78	0.94	1.01	1.01	0.94	0.87	0.89	0.93	0.94	0.90	–	–	0.89
TW	NNowcast	0.70*	0.90*	0.81*	0.88	0.84*	0.82*	0.80*	0.88*	0.88*	0.81*	0.80*	0.79*	0.82
	Estimate	0.77	0.97	0.97	0.94	0.96	0.88*	0.88*	0.88	0.91	0.95	0.91	0.90	0.90
	Backcast	0.78	0.94	1.01	0.99	0.94	0.87	0.89	0.93	0.94	0.90	–	–	0.89
TP	Nowcast	0.69*	0.89*	0.80*	0.87	0.84*	0.81*	0.79*	0.88*	0.88*	0.81*	0.79*	0.79*	0.82
	Estimate	0.74	0.93	0.94	0.94	0.96	0.88*	0.88*	0.92	0.91	0.95	0.91	0.90	0.90
	Backcast	0.74	0.94	1.01	0.99	0.94	0.87	0.89	0.93	0.94	0.90	–	–	0.89

Note. NE, North East England; NW, North West England; York, Yorkshire and the Humber; EM, East Midlands; WM, West Midlands; EE, East of England; LON, London; SE, South East England; SW, South West England; WA, Wales; SCOT; Scotland; NI, Northern Ireland; CRPS, continuous ranked probability score; MF-VAR, mixed-frequency vector autoregression; VAT, value-added tax; MF-FAVAR, mixed-frequency factor-augmented VAR; EMPCA, expectation maximisation principal components analysis; TW, tall wide; TP, tall project.
*denotes rejection of the null of equal forecast accuracy against the benchmark MF-VAR model at the 0.10 significance level using a two-sided (Diebold & Mariano, 1995) test.

paper. Indeed most do not markedly change the average evaluation metric across regions. The only version of these results where the evaluation metrics do change more substantially is when we switch off the cross-sectional restriction. In this case, the accuracy of our estimates (across different models) clearly gets worse.

5 Conclusions

The twin factors of increasing interest in regional economic issues and advances in the availability of data (including administrative data) have led to the creation of new regional economic predictors and the promise of being able to better track regional economic activity. In the UK, we have seen a significant increase in regional economic data over the past decade and, on the administrative data side, this now includes payroll employment information from the PAYE tax system and VAT turnover data. While these data may provide an indication of what is happening to regional output, they are not a direct estimate of it. However, many of these data series could be useful for nowcasting regional GVA (GDP). A major barrier to doing so in practice is the relatively short time series that characterise these data that makes their use in many nowcasting models problematic.

In this paper, we investigate different ways of estimating large VAR models with mixed-frequency data when some of the indicators included in the VAR only have 'short' historical coverage. We capture the information in the short data by constructing regional factors from the short data using a well-known (EMPCA) and two more recently developed (TP and TW) algorithms. These algorithms estimate common factors from data that have missing data at both the beginning and the end of the sample. We then add these regional factors into a regional MF-VAR, so as to exploit the putative information in the short data.

We find that the differences across each of the three-factor extraction methods are small. It does not matter which algorithm is used to estimate factors from data sets characterised by a ragged edge at the beginning, as well as at the end, of the sample. When interest lies in producing 'backcasts' or 'estimates' of regional growth (using the nomenclature adopted in this paper), we find that there is little to be gained from incorporating these additional regional predictors into our model (relative to our relatively sophisticated benchmark model). When interest instead lies in producing (the more timely) 'nowcasts' of regional output growth, before official data for UK GDP become available, we do find that there are gains to conditioning regional nowcasts on the factors, irrespective of which algorithm is used to estimate the factors. This suggests that, in this case, there is some utility in the short data. But as time passes and information on aggregate (UK-wide) economic activity within the quarter becomes available, the value of the short regional indicators declines.

Acknowledgments

We thank two anonymous referees and an associate editor for their helpful comments. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Cleveland or the Federal Reserve System. Any views expressed are solely those of the authors and do not represent those of ESCoE, its partner institutions, or the ONS.

Conflict of interest: None declared.

Funding

This research has been funded by the Office for National Statistics (ONS) as part of the research programme of the Economic Statistics Centre of Excellence (ESCoE).

Data availability

Data and code for this paper are available here: https://github.com/pingwuecon/MF-FAVAR.git.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Angelini E., Henry J., & Marcellino M. (2006). Interpolation and backdating with a large information set. *Journal of Economic Dynamics and Control*, *30*(12), 2693–2724. https://doi.org/10.1016/j.jedc.2005.07.010

Bai J., & Ng S. (2006). Confidence intervals for diffusion index forecasts and inference for factor-augmented regressions. *Econometrica*, *74*(4), 1133–1150. https://doi.org/10.1111/j.1468-0262.2006.00696.x

Bai J., & Ng S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, *116*(536), 1746–1763. https://doi.org/10.1080/01621459.2021.1967163

Banbura M., Giannone D., & Reichlin L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, *25*(1), 71–92. https://doi.org/10.1002/jae.1137

Banbura M., & Modugno M. (2014). Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. *Journal of Applied Econometrics*, *29*(1), 133–160. https://doi.org/10.1002/jae.2306

Banbura M., & Runstler G. (2011). A look into the factor model black box: Publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, *27*(2), 333–346. https://doi.org/10.1016/j.ijforecast.2010.01.011

Bernanke B. S., Boivin J., & Eliasz P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, *120*(1), 387–422. https://doi.org/10.1162/0033553053327452

Brave S. A., Butters R. A., & Justiniano A. (2019). Forecasting economic activity with mixed frequency BVARs. *International Journal of Forecasting*, *35*(4), 1692–1707. https://doi.org/10.1016/j.ijforecast.2019.02.010

Cahan E., Bai J., & Ng S. (2023). Factor-based imputation of missing values and covariances in panel data of large dimensions. *Journal of Econometrics*, *233*(1), 113–131. https://doi.org/10.1016/j.jeconom.2022.01.006

Chan J. (2022). Asymmetric conjugate priors for large Bayesian VARs. *Quantitative Economics*, *13*(3), 1145–1169. https://doi.org/10.3982/QE1381

Chan, J. C. C., Poon, A., & Zhu, D. (2023). High-dimensional conditionally gaussian state space models with missing data.

Chan J. C. C., Poon A., & Zhu D. (2023). High-dimensional conditionally Gaussian state space models with missing data. *Journal of Econometrics*, *236*(1), 105,468. https://doi.org/10.1016/j.jeconom.2023.05.005

Diebold F. X., & Mariano R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144. https://doi.org/10.1080/07350015.1995.10524599

Eraker B., Chiu C. W. J., Foerster A. T., Kim T. B., & Seoane H. D. (2014). Bayesian mixed frequency VARs. *Journal of Financial Econometrics*, *13*(3), 698–721. https://doi.org/10.1093/jjfinec/nbu027

Frale C., Marcellino M., Mazzi G. L., & Proietti T. (2011). EUROMIND: A monthly indicator of the euro area economic conditions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *174*(2), 439–470. https://doi.org/10.1111/j.1467-985X.2010.00675.x

Gefang D., Koop G., & Poon A. (2020). Computationally efficient inference in large Bayesian mixed frequency VARs. *Economics Letters*, *191*, 109–120. https://doi.org/10.1016/j.econlet.2020.109120

Gonçalves S., & Perron B. (2014). Bootstrapping factor-augmented regression models. *Journal of Econometrics*, *182*(1), 156–173. https://doi.org/10.1016/j.jeconom.2014.04.015

Koop G., McIntyre S., & Mitchell J. (2020). UK regional nowcasting using a mixed frequency vector auto-regressive model with entropic tilting. *Journal of the Royal Statistical Society: Series A*, *183*(1), 91–119. https://doi.org/10.1111/rssa.12491

Koop G., McIntyre S., Mitchell J., & Poon A. (2020a). Reconciled estimates and nowcasts of regional output in the UK. *National Institute Economic Review*, *253*, R44–R59. https://doi.org/10.1017/nie.2020.29

Koop G., McIntyre S., Mitchell J., & Poon A. (2020b). Regional output growth in the United Kingdom: More timely and higher frequency estimates from 1970. *Journal of Applied Econometrics*, *35*(2), 176–197. https://doi.org/10.1002/jae.2748

Koop G., McIntyre S., Mitchell J., & Poon A. (2022). Using stochastic hierarchical aggregation constraints to nowcast regional economic aggregates. *International Journal of Forecasting*, https://doi.org/10.1016/j.ijforecast.2022.04.002

Labonne P., & Weale M. (2020). Temporal disaggregation of overlapping noisy quarterly data: Estimation of monthly output from UK value-added tax data. *Journal of the Royal Statistical Society, Series A*, *183*(3), 1211–1230. https://doi.org/10.1111/rssa.12568

Leamer E. E. (2007). *Housing is the business cycle* (Working Paper 13428). National Bureau of Economic Research. https://doi.org/10.3386/w13428

McCracken M. W., & Ng S. (2021). FRED-QD: A quarterly database for macroeconomic research. *Federal Reserve Bank of St Louis Review*, *103*(1), 1–44. https://doi.org/10.20955/r.103.1-44

McCracken M. W., Owyang M. T., & Sekhposyan T. (2021). Real-time forecasting and scenario analysis using a large mixed-frequency Bayesian VAR. *International Journal of Central Banking*, *17*(5), 1–41. https://doi.org/https://www.ijcb.org/journal/ijcb21q5a8.htm

Rossi B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature*, *59*(4), 1135–90. https://doi.org/10.1257/jel.20201479

Schorfheide F., & Song D. (2015). Real-time forecasting with a mixed-frequency VAR. *Journal of Business & Economic Statistics*, *33*(3), 366–380. https://doi.org/10.1080/07350015.2014.954707

Stock J., & Watson M. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, *20*(2), 147–162. https://doi.org/10.1198/073500102317351921

Stock J., & Watson M. (2016). Chapter 8—Dynamic factor models, factor-augmented vector autoregressions, and structural vector autoregressions in macroeconomics. In J. B. Taylor & H. Uhlig (Eds.), *Handbook of macroeconomics* (Vol. 2, pp. 415–525). Elsevier. https://doi.org/10.1016/bs.hesmac.2016.04.002

Wallis K. F. (1986). Forecasting with an econometric model: The 'ragged edge' problem. *Journal of Forecasting*, *5*(1), 1–13. https://doi.org/10.1002/for.3980050102

Zou H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. https://doi.org/10.1198/016214506000000735