WORKING PAPER 1/2026 (STATISTICS)

Real-Time Nowcasting of Kyiv's Regional GRP Using Google Trends and Mixed-Frequency Data

Svitlana Drin and Anastasiia Zhuravlova

ISSN 1403-0586

Örebro University School of Business
SE-701 82 Örebro, Sweden

Real-Time Nowcasting of Kyiv's Regional GRP Using Google Trends and Mixed-Frequency Data

Svitlana Drin[1,2] and Anastasiia Zhuravlova[2]

[1]School of Business, Örebro University, 70182 Örebro, Sweden
[2]Department of Mathematics, National University of Kyiv-Mohyla Academy, 04070 Kyiv, Ukraine

January 2, 2026

Abstract

Timely assessment of regional economic activity in Ukraine is severely constrained by institutional and data-related limitations. Official regional gross regional product (GRP) statistics are available only at low frequency, are published with substantial delays, and, in the post-2022 period, are further affected by disruptions to statistical production caused by martial law. At the same time, a growing set of potentially informative regional indicators derived from administrative records and official short-term statistics is available at higher frequencies but only over short and heterogeneous time spans. These features make the direct application of standard regional nowcasting models infeasible.

This paper develops a mixed-frequency factor-augmented vector autoregressive framework tailored to the Ukrainian data environment and designed to incorporate short and incomplete regional indicators into the nowcasting of regional GDP. The model explicitly exploits the hierarchical structure of Ukrainian regional statistics by combining information from quarterly and annual measures of economic activity and by linking regional dynamics to national output developments. Short regional indicators are summarised through latent regional factors extracted using missing-data factor estimation techniques that are robust to ragged edges at both the beginning and the end of the sample.

The proposed framework is implemented using Ukrainian macro-regional aggregates constructed from official data published by the State Statistics Service of Ukraine. Particular attention is paid to the treatment of labour market indicators, housing price dynamics, and other short-term variables that exhibit discontinuities or limited availability. A pseudo-real-time nowcasting exercise shows that conditioning regional GDP nowcasts on factor information derived from short regional data improves predictive performance when contemporaneous national GDP estimates are not yet available. Once national aggregates are released, the marginal informational contribution of regional short-term indicators diminishes.

Overall, the results demonstrate that mixed-frequency factor-augmented VAR models provide a coherent and empirically viable framework for regional GDP nowcasting in Ukraine. The approach is particularly well suited to data environments

characterised by short samples, publication delays, and institutional disruptions, and thus offers a valuable tool for real-time regional economic monitoring in periods of heightened uncertainty.

Keywords: MF-FAVAR, FAVAR, Nowcasting, EMPCA, GRP, Google Trends.

1 Introduction

The literature on real-time economic monitoring has expanded rapidly over the past two decades, driven by the growing demand for timely assessments of economic activity and the increasing availability of high-frequency data. A central challenge in this literature arises from the mismatch between the low frequency and delayed publication of key macroeconomic aggregates, such as gross domestic product (GDP) or GRP, and the need for policy-relevant information in real time. These constraints are particularly pronounced at the regional level, where statistical coverage is typically weaker and publication delays are longer than for national aggregates.

Early methodological foundations were laid by the development of dynamic factor models for forecasting with large panels of macroeconomic indicators. The seminal contribution of Stock and Watson (2002) demonstrated that a small number of latent factors, extracted using principal components, can capture a substantial share of common variation across a large set of predictors, thereby addressing the curse of dimensionality in forecasting problems. This framework was subsequently extended by Bernanke et al. (2005), who embedded latent factors into vector autoregressive systems, giving rise to the factor-augmented VAR (FAVAR). The FAVAR approach allows high-dimensional information sets to affect macroeconomic dynamics in a parsimonious manner and has become a cornerstone of modern nowcasting and policy analysis. A comprehensive treatment of dynamic factor models and FAVAR specifications is provided by Stock and Watson (2016).

A parallel strand of the literature has focused on the explicit modelling of mixed-frequency data, reflecting the fact that many informative indicators are observed at monthly or higher frequencies, while target variables such as GDP or GRP are typically available only quarterly or annually. Mixed-frequency VAR models offer a coherent framework for handling such data structures. Contributions by Schorfheide and Song (2015) and Schorfheide and Song (2020) demonstrate the effectiveness of mixed-frequency VARs for real-time forecasting under ragged-edge data conditions, including during periods of heightened economic uncertainty. A systematic overview of mixed-frequency VAR and Mixed Data Sampling (MIDAS) models is provided by Foroni et al. (2018), who document both theoretical developments and empirical applications.

Regional nowcasting introduces additional challenges beyond those encountered at the national level. Subnational output data are often released with substantial delays and at low frequency, while the available set of regional indicators is typically fragmented, heterogeneous, and characterised by short historical coverage. Recent work by Koop et al. (2023) addresses these issues by proposing mixed-frequency VAR and MF-FAVAR frameworks specifically tailored to regional applications. Their approach combines inter-temporal aggregation constraints, cross-sectional links between regional and national output, and factor extraction from short regional indicator panels. A key insight of their analysis is that even indicators with limited historical availability can improve regional nowcasts, particularly when contemporaneous national output data are not yet observed.

A critical technical challenge in factor-based mixed-frequency models concerns the treatment of missing observations and ragged-edge panels. Early approaches, including expectation–maximisation principal components analysis, were already discussed in Stock and Watson (2002). More recent contributions have proposed matrix completion and projection-based methods with stronger theoretical guarantees. In particular, Bai and Ng (2021) develop factor estimation techniques that explicitly account for cross-sectional and serial dependence in incomplete panels, while Cahan et al. (2023) introduce projection-based methods designed to handle ragged-edge datasets commonly encountered in real-time forecasting.

In parallel, the nowcasting literature has increasingly incorporated digital data sources as proxies for real-time economic activity. Choi and Varian (2012) show that Google Trends indicators can capture contemporaneous movements in economic conditions and improve short-term forecasts of macroeconomic variables. Such digital indicators are especially valuable in environments where official statistics are delayed, incomplete, or disrupted, as they reflect real-time behaviour related to consumption, labour markets, and expectations.

The evaluation of nowcasting performance requires appropriate accuracy measures for both point and density forecasts. Standard forecast error metrics are reviewed by Hyndman and Koehler (2006), while Gneiting and Raftery (2007) formalise strictly proper scoring rules, including the Continuous Ranked Probability Score, which has become a standard tool for evaluating density forecasts in macroeconomic applications.

This paper builds on the above literature but focuses on a fundamentally different empirical setting. Ukraine represents a data-constrained and institutionally disrupted environment, where regional GRP statistics are available only at low frequency, many short-term indicators have short and fragmented histories, and standard survey-based data collection has been suspended since 2022. These features raise important questions about the applicability and performance of MF-FAVAR models when the cross-section of regional indicators is limited, the time dimension is short, and structural breaks are pronounced.

The contribution of this study is fourfold. First, it extends the MF-FAVAR framework of Koop et al. (2023) to a severely data-scarce regional environment by tailoring the model structure to the institutional realities of Ukrainian regional statistics. Second, it integrates insights from the factor and missing-data literature to extract robust regional signals from short and ragged-edge indicator panels. Third, it provides the first systematic MF-FAVAR-based nowcasting framework for Ukrainian regional GDP, with a particular focus on the Kyiv region. Fourth, it evaluates the role of high-frequency digital indicators, including Google Trends data, in improving regional nowcasts under extreme data limitations.

Overall, this study positions itself at the intersection of the literatures on factor models, mixed-frequency VARs, and regional nowcasting. While existing contributions demonstrate the effectiveness of these methods in data-rich environments, this paper shows how they can be adapted and operationalised in settings characterised by limited data availability, structural breaks, and institutional disruption. As such, the results are relevant not only for Ukraine, but also for other economies facing similar challenges in regional economic measurement.

The remainder of the paper is organised as follows. Section 2 presents the methodological framework. Section 3 describes the data sources and preprocessing procedures.

Section 4 introduces the MF-FAVAR model, and Section 5 discusses factor extraction under short and ragged-edge data. Section 7 concludes.

2 Methodology

This section outlines the methodological framework used for nowcasting regional economic activity under severe data constraints. The approach is based on mixed-frequency vector autoregressive models and their factor-augmented extensions, specifically designed to accommodate short time series, irregular data availability, and ragged-edge structures that characterise regional statistics in Ukraine.

2.1 Baseline Vector Autoregression

We begin with the classical vector autoregressive model introduced by Sims (1980), which serves as a benchmark for modelling dynamic interactions among macroeconomic variables. Let

$$y_t \in \mathbb{R}^K$$

denote a vector of endogenous variables observed at time t. A VAR of order p is defined as

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \cdots + A_p y_{t-p} + \varepsilon_t, \tag{1}$$

where $A_i \in \mathbb{R}^{K \times K}$ are coefficient matrices and $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$ is a vector of innovations.

While VAR models are flexible and fully data driven, their application in the present context is limited by three factors: the requirement that all variables be observed at the same frequency, the rapid growth of parameters relative to sample size, and their sensitivity to missing observations. These limitations are particularly severe for Ukrainian regional data, where GRP is observed only annually and with substantial publication delays, while other indicators are available at higher frequencies.

2.2 Mixed-Frequency VAR Framework

To address the frequency mismatch, we adopt a Mixed-Frequency Vector Autoregression (MF-VAR) framework, which allows variables sampled at different temporal resolutions to be modelled jointly (Foroni and Marcellino, 2013; Schorfheide and Song, 2015; Koop et al., 2023). Let

$$y_t = \begin{pmatrix} y_t^{(L)} \\ y_t^{(H)} \end{pmatrix}$$

denote the stacked vector of low-frequency variables $y_t^{(L)}$ (e.g. annual or quarterly GRP) and high-frequency variables $y_t^{(H)}$ (e.g. monthly or weekly indicators).

The MF-VAR retains the VAR(p) structure in (1), but incorporates temporal aggregation constraints and data alignment procedures to link low- and high-frequency observations. In the Kyiv application, annual GRP is disaggregated to quarterly frequency using the Denton–Cholette method (Denton, 1971; Cholette, 1984), ensuring consistency between aggregated quarterly values and observed annual totals.

Although MF-VAR models resolve the frequency mismatch, they remain vulnerable to over-parameterisation when the number of high-frequency indicators is large relative to the available sample size. This motivates the use of factor-augmented representations.

2.3 Mixed-Frequency Factor-Augmented VAR

To achieve dimensionality reduction and enhance robustness, we employ a Mixed-Frequency Factor-Augmented VAR (MF-FAVAR), following Koop et al. (2023). The MF-FAVAR model assumes that a large set of observed indicators is driven by a small number of latent common factors.

Let

$$x_t^{(H)} \in \mathbb{R}^{N_H}$$

denote the vector of high-frequency indicators and

$$y_t^{(L)} \in \mathbb{R}^{N_L}$$

the vector of low-frequency target variables. The measurement equations are given by

$$x_t^{(H)} = \Lambda^{(H)} f_t + \eta_t^{(H)}, \tag{2}$$

$$y_t^{(L)} = \Lambda^{(L)} f_t + \eta_t^{(L)}, \tag{3}$$

where $f_t \in \mathbb{R}^r$ is a vector of latent factors, $\Lambda^{(H)}$ and $\Lambda^{(L)}$ are loading matrices, and $\eta_t^{(H)}$, $\eta_t^{(L)}$ are idiosyncratic components.

The latent factors evolve according to a VAR(p) process,

$$f_t = \Phi_1 f_{t-1} + \cdots + \Phi_p f_{t-p} + \nu_t, \qquad \nu_t \sim \mathcal{N}(0, Q). \tag{4}$$

This structure allows the MF-FAVAR to capture common economic dynamics while substantially reducing the number of parameters relative to a standard MF-VAR. In the Kyiv case, the extracted factors summarise information from macroeconomic indicators and alternative high-frequency proxies, including Google Trends series.

2.4 Factor Extraction under Ragged-Edge Data

A defining feature of the Ukrainian regional dataset is the presence of missing observations and ragged edges at both the beginning and the end of the sample. Since factor-augmented mixed-frequency models rely on latent factors extracted from a large and potentially unbalanced indicator panel, factor estimation must explicitly accommodate incomplete data. In this study, latent factors are obtained using alternative factor extraction and imputation strategies for ragged-edge panels, including EMPCA, TW, TP, BPCA, and SVDI. The formal objective functions, identification restrictions, and method-specific implementation details are provided in Section 5.

2.5 Nowcasting and Forecast Evaluation

Nowcasts of quarterly GRP growth are generated using the estimated MF-FAVAR model in a pseudo–real–time setting. Forecast performance is evaluated using both point and density-based metrics. Point forecast accuracy is assessed using RMSFE and SMAPE,

while predictive densities are evaluated using the Continuous Ranked Probability Score (CRPS), following Gneiting and Raftery (2007) and Hyndman and Koehler (2006).

This comprehensive evaluation framework allows us to compare alternative factor extraction methods across different forecast horizons and to assess their suitability for real-time regional nowcasting under extreme data limitations.

3 Data Collection and Preliminary Modelling Framework

To illustrate the empirical implementation of the proposed MF-FAVAR framework, we focus on the Kyiv region, which offers the most comprehensive and internally consistent set of regional economic indicators currently available in Ukraine. The choice of this region is primarily driven by data availability, as Kyiv provides a relatively rich collection of annual and monthly series compared to other regions. An overview of the indicator set is reported in Table 1.

A key feature of the Kyiv dataset is that gross regional product (GRP) is observed exclusively at an annual frequency. This institutional constraint determines the choice of annual GRP as the observable target variable in the forecasting exercise. GRP data are retrieved from the official database of the State Statistics Service of Ukraine (UkrStat)[1]. The published series covers the period from 2004 to 2021.

Based on the archive structure and historical release patterns, regional GRP figures are typically published with a delay of approximately 10–16 months after the end of the reporting year. This substantial publication lag motivates the use of mixed-frequency methods and the introduction of latent quarterly GRP dynamics inferred from higher-frequency regional indicators.

3.1 Regional Focus and Target Variable

To illustrate the empirical implementation of the proposed MF-FAVAR framework, we focus on the Kyiv region, which provides the richest and most consistent set of regional economic indicators currently available in Ukraine. Let $r \in \mathcal{R}$ denote a region, with $r = \text{Kyiv}$ in the baseline analysis.

The primary target variable is GRP. Let

$$Y_{r,t}^{(A)}$$

denote annual GRP for region r in calendar year t. Official GRP data are published by the State Statistics Service of Ukraine and are available only at an annual frequency, with publication delays of approximately 10–16 months. The available sample covers the period $t = 2004, \ldots, 2021$.

Following the mixed-frequency literature, we assume that annual GRP is an aggregation of an unobserved quarterly process

$$y_{r,q}^{(Q)}, \qquad q = 1, \ldots, 4,$$

[1] Archived regional macroeconomic indicators are available from the State Statistics Service of Ukraine: `https://www.ukrstat.gov.ua/operativ/menu/menu_u/nac_r.htm`.

such that

$$Y_{r,t}^{(A)} = \sum_{q=1}^{4} y_{r,4(t-1)+q}^{(Q)}.$$

The quarterly series $y_{r,q}^{(Q)}$ is latent and inferred within the MF-VAR and MF-FAVAR frameworks using higher-frequency information.

Table 1: Overview of regional economic indicators for the Kyiv region

Description	**Frequency**	**Sample period**	**Release lag**	**Source**
GRP, Kyiv region (target variable)	Annual	2004–2021	16 months	State Statistics Service of Ukraine
Household income	Annual	2001–2021	16 months	SSSU
Disposable income	Annual	2001–2021	16 months	SSSU
Wages	Annual	2001–2020	14 months	SSSU
Household expenditures	Annual	2001–2020	16 months	SSSU
Retail goods purchases	Annual	2001–2020	16 months	SSSU
CPI (total)	Monthly	2001–2021	1 month	SSSU
Construction output	Monthly	2012–2021	3 months	SSSU
Wages	Monthly	2016–2021	1 month	SSSU
Google Trends indicators (weekly)				
Search interest: "Apple"	Weekly	2020–2025	1 day	Google Trends
Search interest: "Economy"	Weekly	2020–2025	1 day	Google Trends
Search interest: "Inflation"	Weekly	2020–2025	1 day	Google Trends
Search interest: "Unemployment"	Weekly	2020–2025	1 day	Google Trends

Table 1 provides an overview of all regional indicators used in the empirical analysis. The table summarises the target variable, annual and monthly macroeconomic indicators, and high-frequency digital indicators, together with their sampling frequency, sample coverage, publication lag, and data source. This information set reflects the heterogeneous and ragged-edge data environment motivating the use of mixed-frequency and factor-augmented modelling techniques.

3.2 Temporal Disaggregation of Annual GRP

Since annual GRP is observed only at a yearly frequency, quarterly GRP dynamics are recovered through temporal disaggregation using the Denton–Cholette method. In its general form, the Denton–Cholette procedure solves

$$\min_{\{y_{t,s}\}} \sum_{t,s} \left(\Delta^d (y_{t,s} - \lambda x_{t,s})\right)^2 \quad \text{subject to} \quad \sum_{s=1}^{m} y_{t,s} = Y_t, \tag{5}$$

where $y_{t,s}$ denotes quarterly GRP in quarter s of year t, Y_t is observed annual GRP, $x_{t,s}$ is an optional high-frequency indicator, λ is a scaling parameter, d is the order of differencing, and $m = 4$ is the number of quarters per year.

In the present application, no single quarterly indicator is available over the full sample period that could reliably proxy regional GRP dynamics. We therefore employ the indicator-free version of the Denton–Cholette method by setting $x_{t,s} = 0$, which reduces (5) to a pure smoothing problem that penalises excessive variation in quarterly growth rates while strictly preserving annual totals.

3.3 Annual Regional Indicators

Let

$$\mathbf{X}_{r,t}^{(A)} \in \mathbb{R}^{N_A}$$

denote a vector of annual regional indicators observed for region r in year t. The annual dataset includes household income, disposable income, wages, social benefits, consumption expenditures, retail trade turnover, tax payments, social contributions, employment, capital investment, industrial output, transportation usage, and migration flows.

All annual variables are aligned with $Y_{r,t}^{(A)}$ and treated as low-frequency covariates. These variables provide structural information on regional economic activity but are insufficient for real-time monitoring due to their low frequency and publication delays.

3.4 Monthly and Quarterly Indicators

To enrich the information set, we collect higher-frequency indicators observed at monthly or quarterly frequencies. Let

$$\mathbf{X}_{r,m}^{(M)} \in \mathbb{R}^{N_M}$$

denote a vector of monthly indicators for region r at month m. These include the consumer price index (CPI), disaggregated CPI components, industrial production, construction output, and average nominal wages.

Monthly indicators are mapped to the quarterly timeline using standard aggregation rules. For a monthly variable $x_{r,m}^{(M)}$, the corresponding quarterly series is defined as

$$x_{r,q}^{(Q)} = \frac{1}{3} \sum_{m \in q} x_{r,m}^{(M)},$$

where the summation is taken over the three months belonging to quarter q.

3.5 Digital Indicators: Google Trends

Traditional macroeconomic indicators may fail to capture rapid shifts in expectations and behaviour, particularly under crisis conditions. To address this limitation, we augment the dataset with digital indicators derived from Google Trends.

Let

$$\mathbf{G}_{r,m} \in \mathbb{R}^{N_G}$$

denote a vector of Google Trends indices for region r at month m. Each element measures the normalised search intensity for a specific keyword, scaled to the interval $[0, 100]$. The selected keywords (e.g., *investment*, *savings*, *inflation*, *unemployment*, *export*) are chosen to proxy consumption behaviour, labour market conditions, and economic sentiment.

As with other monthly indicators, Google Trends series are aggregated to quarterly frequency and treated as short and potentially noisy predictors that enter the model exclusively through the latent factor structure.

3.6 Stacked Data Representation

For each quarter q, we define the stacked vector of observed variables as

$$\mathbf{z}_{r,q} = \begin{pmatrix} y_{r,q}^{(Q)} \\ \mathbf{X}_{r,q}^{(Q)} \\ \mathbf{G}_{r,q} \end{pmatrix},$$

where $y_{r,q}^{(Q)}$ is latent, while $\mathbf{X}_{r,q}^{(Q)}$ and $\mathbf{G}_{r,q}$ are partially observed and subject to ragged-edge missingness.

3.7 Classical VAR Representation

In a fully observed setting, a classical VAR(p) model for $\mathbf{z}_{r,q} \in \mathbb{R}^K$ would take the form

$$\mathbf{z}_{r,q} = A_1 \mathbf{z}_{r,q-1} + \cdots + A_p \mathbf{z}_{r,q-p} + \boldsymbol{\varepsilon}_{r,q}, \qquad \boldsymbol{\varepsilon}_{r,q} \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{6}$$

where $A_j \in \mathbb{R}^{K \times K}$ are coefficient matrices capturing dynamic interdependencies across variables.

Each element $a_{ik}^{(j)}$ of A_j measures the marginal effect of the k-th variable at lag j on the i-th variable at time q. Own-lag effects ($i = k$) describe persistence, while cross-effects ($i \neq k$) capture spillovers between economic indicators and behavioural proxies.

Illustrative Kyiv example. To clarify the interpretation of the coefficient matrices in (6), consider an illustrative VAR system for the Kyiv dataset with three endogenous blocks: (i) quarterly GRP growth, denoted by y_q, (ii) a macroeconomic indicator block (e.g., CPI, wages, construction output), denoted by Macro_q, and (iii) an index summarising Google Trends activity, denoted by GT_q. Let

$$\mathbf{y}_q = \begin{bmatrix} y_q \\ \text{Macro}_q \\ \text{GT}_q \end{bmatrix}, \qquad \mathbf{y}_q = A_1 \mathbf{y}_{q-1} + A_2 \mathbf{y}_{q-2} + \varepsilon_q, \qquad \varepsilon_q \sim \mathcal{N}(0, \Sigma).$$

In this setting, selected elements of A_1 have the following interpretation: (i) $a_{12}^{(1)}$ captures the effect of lagged macroeconomic conditions on current GRP growth, (ii) $a_{13}^{(1)}$ measures the marginal contribution of lagged Google Trends activity to current GRP growth, and (iii) $a_{31}^{(1)}$ reflects potential feedback from lagged GRP growth to current search behaviour.

3.8 Limitations of Classical VAR

In the present application, the assumptions underlying (6) are violated. First, variables are observed at mixed frequencies. Second, the effective sample size is small relative to the dimensionality of the system. Third, the dataset exhibits pervasive missing observations due to publication lags and wartime disruptions.

These features render classical VAR models unstable and motivate the use of mixed-frequency and factor-augmented approaches.

3.9 Mixed-Frequency VAR Motivation

The Mixed-Frequency VAR (MF-VAR) model extends (6) by allowing $\mathbf{z}_{r,q}$ to contain variables observed at different frequencies. Low-frequency variables, such as annual GRP, are linked to latent quarterly components via temporal aggregation constraints, while high-frequency indicators enter directly at the quarterly level.

In the Kyiv application, annual GRP is disaggregated into quarterly latent values using the Denton–Cholette method, ensuring that the sum of quarterly estimates matches observed annual totals. However, when the number of high-frequency indicators is large, MF-VAR models remain vulnerable to over-parameterisation and noise amplification.

These limitations motivate the introduction of latent factors, leading to the MF-FAVAR framework developed in the subsequent section.

4 The MF-FAVAR Model

This section presents MF-FAVAR model that constitutes the core econometric framework of the analysis. The model is designed to jointly handle mixed-frequency observations, short time series, and ragged-edge data structures that characterise regional economic indicators in Ukraine.

4.1 Notation and Data Structure

Let $t = 1, \ldots, T$ denote the time index at the highest frequency used in the model, which is quarterly. Let

$$y_t^{(L)}$$

denote the low-frequency target variable, corresponding to quarterly GRP growth, which is not directly observed. Instead, GRP is available only at an annual frequency. Let

$$Y_\tau^{(A)}, \qquad \tau = 1, \ldots, T_A,$$

denote observed annual GRP, where each annual observation is related to the underlying quarterly process via the aggregation constraint

$$Y_\tau^{(A)} = \sum_{t \in \mathcal{T}(\tau)} y_t^{(L)}, \tag{7}$$

with $\mathcal{T}(\tau)$ denoting the set of quarters belonging to year τ.

Let

$$\mathbf{x}_t^{(H)} \in \mathbb{R}^{N_H}$$

denote the vector of high-frequency indicators observed at the quarterly level. This vector includes transformed monthly macroeconomic variables and aggregated Google Trends indicators. The panel $\{\mathbf{x}_t^{(H)}\}$ is unbalanced and characterised by missing observations and heterogeneous release lags.

4.2 Measurement Equations

The MF-FAVAR model assumes that both the latent low-frequency target variable and the high-frequency indicators are driven by a small number of latent common factors. Formally, the measurement equations are given by

$$y_t^{(L)} = \boldsymbol{\lambda}_y^\top \mathbf{f}_t + \varepsilon_t^{(y)}, \tag{8}$$

$$\mathbf{x}_t^{(H)} = \Lambda_x \mathbf{f}_t + \boldsymbol{\varepsilon}_t^{(x)}, \tag{9}$$

where:

- $\mathbf{f}_t \in \mathbb{R}^r$ is a vector of latent common factors with $r \ll N_H$;
- $\boldsymbol{\lambda}_y \in \mathbb{R}^r$ and $\Lambda_x \in \mathbb{R}^{N_H \times r}$ are factor loading matrices;
- $\varepsilon_t^{(y)}$ and $\boldsymbol{\varepsilon}_t^{(x)}$ are idiosyncratic components.

The idiosyncratic disturbances are assumed to satisfy

$$\mathbb{E}[\varepsilon_t^{(y)}] = 0, \qquad \mathbb{E}[\boldsymbol{\varepsilon}_t^{(x)}] = \mathbf{0},$$

and to be weakly cross-sectionally correlated, capturing measurement error and variable-specific noise.

4.3 Factor Dynamics

The latent factors evolve according to a vector autoregressive process of order p,

$$\mathbf{f}_t = \Phi_1 \mathbf{f}_{t-1} + \Phi_2 \mathbf{f}_{t-2} + \cdots + \Phi_p \mathbf{f}_{t-p} + \boldsymbol{\nu}_t, \qquad \boldsymbol{\nu}_t \sim \mathcal{N}(\mathbf{0}, Q), \tag{10}$$

where $\Phi_1, \ldots, \Phi_p \in \mathbb{R}^{r \times r}$ are autoregressive coefficient matrices and Q is a positive definite covariance matrix.

This dynamic specification captures persistent common movements across regional indicators and allows information from short and noisy series to be pooled efficiently through the factor structure.

4.4 State-Space Representation and Mixed Frequencies

Equations (8)–(10), together with the aggregation constraint (7), define a state-space representation of the MF-FAVAR model. The latent state vector is given by $\mathbf{f}_t$, while observed variables enter the system through the measurement equations and the annual aggregation constraint.

Mixed-frequency aspects arise naturally from the fact that $y_t^{(L)}$ is latent at the quarterly level and observed only through annual aggregates, whereas $\mathbf{x}_t^{(H)}$ is partially observed at the quarterly level with ragged-edge missingness. This structure allows the model to exploit all available information from high-frequency indicators while respecting the annual aggregation constraint for GRP.

5 Factor Extraction under Short and Ragged-Edge Data

This section describes the estimation of latent factors in the MF-FAVAR model when the high-frequency indicator panel is characterised by short time series, missing observations, and ragged-edge structures. The focus is on factor extraction methods that are well suited to data-constrained regional environments and that remain consistent with the notation introduced in Section 4.

5.1 Incomplete High-Frequency Panel

Let

$$\mathbf{X}^{(H)} = \begin{pmatrix} \mathbf{x}_1^{(H)\prime} \\ \mathbf{x}_2^{(H)\prime} \\ \vdots \\ \mathbf{x}_T^{(H)\prime} \end{pmatrix} \in \mathbb{R}^{T \times N_H}$$

denote the stacked panel of high-frequency indicators. Due to heterogeneous release lags and limited historical availability, $\mathbf{X}^{(H)}$ is an unbalanced matrix with missing entries at both the beginning and the end of the sample. Let $\Omega \subset \{1, \dots, T\} \times \{1, \dots, N_H\}$ denote the set of observed entries.

The measurement equation in (9) implies the approximate factor structure

$$x_{t,i}^{(H)} = \boldsymbol{\lambda}_i^\top \mathbf{f}_t + \varepsilon_{t,i}^{(x)}, \qquad (t,i) \in \Omega, \tag{11}$$

where $\boldsymbol{\lambda}_i$ is the i-th row of Λ_x.

5.2 Estimation Objective

Factor extraction is based on minimising the reconstruction error over observed entries,

$$\min_{\{\mathbf{f}_t\},\, \Lambda_x} \sum_{(t,i)\in\Omega} \left(x_{t,i}^{(H)} - \boldsymbol{\lambda}_i^\top \mathbf{f}_t \right)^2, \tag{12}$$

subject to standard normalisation conditions,

$$\frac{1}{T} \sum_{t=1}^{T} \mathbf{f}_t \mathbf{f}_t^\top = I_r, \qquad \Lambda_x^\top \Lambda_x \text{ diagonal.}$$

These restrictions ensure identification of the factor space up to sign and ordering.

5.3 Expectation–Maximisation Principal Components

The baseline factor extraction method is EMPCA, which iteratively alternates between:

- imputing missing entries of $\mathbf{X}^{(H)}$ using current estimates of $\mathbf{f}_t$ and Λ_x;
- updating factor estimates by applying principal components to the completed panel.

EMPCA converges to a local minimum of (12) and is computationally efficient even when the proportion of missing observations is high. Its iterative nature allows information from short and partially observed series to contribute to factor estimation without discarding variables with limited availability.

5.4 Alternative Methods for Ragged-Edge Panels

To assess robustness, we consider alternative approaches that differ in how they exploit the observed data structure.

Tall–Wide and Tall–Projection Methods. These approaches construct factor estimates from subsamples of $\mathbf{X}^{(H)}$ with sufficiently dense coverage. Let $\mathbf{X}^{(TW)}$ denote a tall–wide submatrix with minimal missingness. Factors are estimated from $\mathbf{X}^{(TW)}$ and projected onto the full dataset to recover $\mathbf{f}_t$ for all t. Projection-based variants further reduce sensitivity to noise by conditioning on observed blocks only.

Bayesian and SVD-Based Imputation. Bayesian principal components analysis treats $\mathbf{f}_t$ and Λ_x as random variables and integrates over missing observations using posterior distributions. Truncated singular value decomposition (SVD) provides a low-rank approximation to $\mathbf{X}^{(H)}$ by solving a regularised version of (12). These methods offer complementary perspectives on uncertainty and regularisation but are computationally more demanding.

5.5 Choice of the Number of Factors

The number of factors r plays a central role in balancing explanatory power and parsimony. In data-rich settings, information criteria or eigenvalue-based rules are commonly used. In the present application, the effective sample size is limited and the indicator panel is unbalanced. We therefore select r by combining information criteria with sensitivity analysis, evaluating forecast performance across alternative factor dimensions.

5.6 Integration with the MF-FAVAR Model

Once estimated, the factor series $\{\hat{\mathbf{f}}_t\}_{t=1}^{T}$ enter the MF-FAVAR model as latent states governing the dynamics of both the target variable $y_t^{(L)}$ and the high-frequency indicators. Factor uncertainty is propagated through the state-space representation described in Section 4, allowing nowcasts of GRP to reflect both data incompleteness and parameter uncertainty.

The next section presents the empirical results for the Kyiv region, including pseudo-real-time nowcasting experiments and forecast evaluation under alternative factor extraction strategies.

6 Empirical Results

This section presents the empirical application of the proposed MF-FAVAR framework to the Kyiv regional dataset. The analysis evaluates the model's ability to nowcast quarterly

GRP growth under severe data constraints, irregular publication lags, and short effective samples. Particular attention is paid to the comparative performance of alternative factor extraction methods across different forecast horizons and to robustness with respect to indicator selection.

6.1 Model implementation and evaluation design

Quarterly GRP growth for Kyiv is defined as the log-difference of the temporally disaggregated quarterly GRP series,

$$Y_t = \Delta \log(\mathrm{GRP}_t).$$

The MF-FAVAR model is estimated in a pseudo-real-time setting using a chronological split of the sample. Approximately 80% of the observations are used for model estimation, while the remaining 20% constitute the out-of-sample evaluation period.

Let $\hat{Y}_{t|t-h}$ denote the h-step-ahead forecast produced at time $t-h$. Forecast accuracy is assessed using complementary point and density-based metrics. Point forecast performance is evaluated using the Root Mean Squared Forecast Error (RMSFE),

$$\mathrm{RMSFE}(h) = \sqrt{\frac{1}{N_h} \sum_{t \in \mathcal{T}_h} \left(Y_t - \hat{Y}_{t|t-h} \right)^2},$$

while density forecast accuracy is measured using the Continuous Ranked Probability Score (CRPS),

$$\mathrm{CRPS}(\hat{F}, Y) = \int_{-\infty}^{\infty} \left(\hat{F}(z) - \mathbb{1}\{Y \leq z\} \right)^2 dz.$$

To complement scale-dependent measures, percentage-based accuracy is summarised using the Symmetric Mean Absolute Percentage Error (SMAPE).

6.2 Feasibility of factor extraction methods

Five factor extraction and imputation strategies are initially considered: Expectation–Maximisation Principal Components Analysis (EMPCA), Tall–Wide (TW), Tall Projection (TP), Bayesian PCA (BPCA), and Singular Value Decomposition Imputation (SVDI). Due to the extreme sparsity and heterogeneity of the Kyiv dataset, the TW and TP methods fail to produce feasible factor estimates when applied to the full indicator panel. Specifically, TW cannot identify sufficiently large overlapping blocks of fully observed data, while TP lacks an adequate tall block for reliable initialisation.

Consequently, the empirical comparison focuses on the three methods that yield valid forecasts: EMPCA, BPCA, and SVDI. Each method produces a distinct factor representation, which is subsequently embedded into the MF-FAVAR system.

6.3 Forecast accuracy across horizons

Forecast performance is evaluated across three policy-relevant horizons: a long horizon of 24 quarters, a medium horizon of 4 quarters, and a short horizon of 2 quarters. Table 2 reports forecast accuracy across methods and horizons.

Table 2: Forecast accuracy by horizon

Horizon	Method	Lag	MAE	RMSE	CRPS
24 Q	**EMPCA**	2	0.0158	0.0208	0.0116
	BPCA	2	0.0159	0.0208	0.0117
	SVDI	4	0.0255	0.0319	0.0194
4 Q	EMPCA	2	0.0221	0.0247	0.0141
	BPCA	2	0.0211	0.0237	0.0134
	SVDI	4	0.0155	0.0167	0.0093
2 Q	**EMPCA**	2	0.0033	0.0033	0.0034
	BPCA	2	0.0033	0.0034	0.0034
	SVDI	4	0.0053	0.0054	0.0042

At the long forecast horizon ($h = 24$ quarters), EMPCA marginally outperforms BPCA in terms of both RMSE and CRPS, indicating superior long-run stability when extracting persistent regional signals from short and noisy indicator panels. In contrast, SVDI performs substantially worse at this horizon, suggesting limited ability to capture low-frequency dynamics.

At the medium horizon ($h = 4$ quarters), SVDI achieves the lowest MAE and CRPS among the considered methods. This result indicates that SVD-based imputation is particularly effective in capturing short- to medium-term fluctuations, likely due to its emphasis on low-rank reconstruction of the indicator matrix.

At the very short horizon ($h = 2$ quarters), EMPCA consistently delivers the most accurate forecasts across all evaluation metrics. Forecast errors are minimal and predictive distributions are tightly concentrated, highlighting the suitability of EMPCA for real-time nowcasting applications where near-term accuracy is critical.

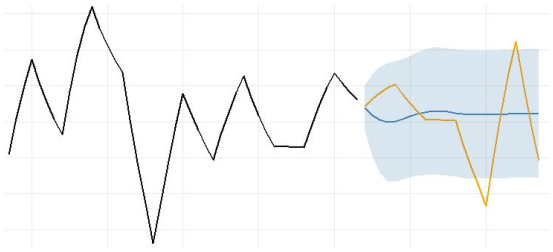
6.4 Dynamic forecast interpretation

Figure 1 provide a dynamic illustration of forecast performance at long and medium horizons. At the long horizon, EMPCA generates smoother forecast paths and narrower predictive intervals, reflecting superior long-term stability. At the medium horizon, SVDI more effectively captures short-to-medium-term fluctuations, consistent with its superior performance in Table 2.

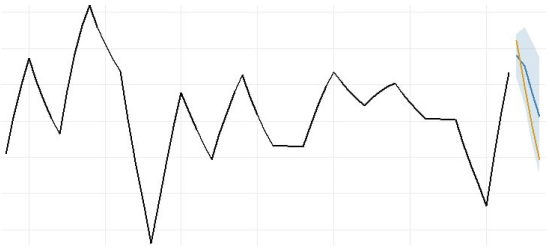
Together, the tabular and graphical evidence confirms that no single factor extraction method is uniformly optimal across horizons. However, EMPCA provides the most robust overall performance, delivering strong results at both short and long horizons.

6.5 Robustness check with reduced indicator set

As an additional robustness check, the MF-FAVAR model is re-estimated using a reduced and more homogeneous set of predictors. Weekly indicators are excluded due to extreme sparsity, while the monthly and annual datasets are restricted to the most economically interpretable and consistently observed variables. Let $\tilde{X}_t \subset X_t$ denote this reduced panel.



(a) Long-run forecast accuracy over 24 quarters: EMPCA exhibits the lowest RMSE and CRPS, indicating superior long-term stability.



(b) Medium-term forecast accuracy over 4 quarters: SVDI yields the best performance in terms of MAE and CRPS.

Figure 1: Forecast accuracy of the MF-FAVAR model across different horizons for the Kyiv region.

Under this streamlined configuration, the Tall–Projection (TP) method becomes computationally feasible, while the Tall–Wide (TW) approach continues to fail due to insufficient overlap in fully observed blocks. Forecast accuracy for EMPCA and TP is reported in Table 3. EMPCA again outperforms TP across all evaluation metrics:

$$\text{RMSFE}_{\text{EMPCA}} < \text{RMSFE}_{\text{TP}}, \qquad \text{CRPS}_{\text{EMPCA}} < \text{CRPS}_{\text{TP}}.$$

Table 3: Forecast accuracy on reduced indicator set

Method	Lag	MAE	RMSE	SMAPE (%)	CRPS
EMPCA	2	**0.0152**	**0.0198**	**40.6**	**0.0114**
TP	3	0.0212	0.0236	59.9	0.0140

Interestingly, forecast accuracy under EMPCA improves relative to the full dataset, highlighting the importance of careful variable selection in data-constrained environments.

6.6 Summary and implications

Overall, the empirical results demonstrate that MF-FAVAR models can deliver reliable regional GRP nowcasts for Kyiv despite extreme data limitations. Among the considered factor extraction methods, EMPCA emerges as the most robust and versatile approach, performing consistently well across short and long horizons. These findings underscore the importance of factor extraction methods that are resilient to missing data, short samples, and structural breaks when conducting regional nowcasting under crisis conditions.

7 Conclusions

This paper develops and empirically evaluates a mixed-frequency factor-augmented vector autoregressive (MF-FAVAR) framework for real-time nowcasting of regional gross regional product (GRP) under extreme data constraints. Using the Kyiv region as a case study, the analysis demonstrates that meaningful short-term regional economic monitoring remains

feasible even when official statistics are sparse, irregular, and subject to long publication delays, as has been the case in Ukraine since 2022.

The proposed methodology explicitly addresses three core challenges that characterise crisis-affected regional data environments. First, it accommodates severe frequency mismatch between the target variable, observed only annually, and a diverse set of higher-frequency indicators. Second, it accounts for short and fragmented time series with pronounced ragged-edge structures at both the beginning and end of the sample. Third, it mitigates over-parameterisation and noise amplification through factor-based dimensionality reduction tailored to incomplete panels.

From a methodological perspective, the paper extends the MF-FAVAR framework of Koop et al. to a setting with substantially weaker data availability and stronger structural instability. By integrating temporal disaggregation via the Denton–Cholette method with latent factor extraction under missing data, the framework provides a coherent state-space representation linking annual GRP outcomes to quarterly latent dynamics and high-frequency predictors. This adaptation is essential in institutional environments where standard mixed-frequency VAR models become unstable or infeasible.

A central empirical contribution of the study is the systematic comparison of alternative factor extraction and imputation methods in a regional nowcasting context. The results show that Expectation–Maximisation Principal Component Analysis (EMPCA) consistently delivers the most robust and accurate nowcasts across short and long forecast horizons. Its iterative structure allows it to effectively exploit partially observed indicators and to remain stable under pronounced structural breaks. Bayesian PCA and SVD-based imputation perform competitively at specific horizons, particularly in medium-term forecasting, but exhibit weaker performance in near-term nowcasting. Tall-Wide and Tall-Projection methods, while theoretically appealing, fail in the Kyiv application due to insufficient overlap in fully observed blocks, highlighting important practical limitations of these approaches in highly sparse datasets.

The empirical findings further indicate that careful variable selection is critical in data-constrained environments. Reducing the indicator set by excluding weakly informative or excessively sparse series improves forecast accuracy and stability, particularly when combined with EMPCA-based factor extraction. This result underscores that more data are not necessarily better when indicator quality and coverage are uneven.

From a substantive perspective, the analysis confirms that high-frequency digital indicators, such as Google Trends, can provide valuable supplementary information for regional nowcasting when traditional statistics are delayed or unavailable. While their marginal contribution diminishes once national aggregates become available, these indicators significantly enhance short-horizon nowcasts and improve real-time situational awareness during periods of heightened uncertainty.

Overall, the study contributes to the literature on regional nowcasting, mixed-frequency modelling, and crisis-time economic monitoring by demonstrating how MF-FAVAR models can be operationalised in environments characterised by short samples, institutional disruption, and structural breaks. Although the empirical application focuses on Kyiv, the methodological insights are directly relevant for other regions and countries facing similar constraints.

Future research may extend the proposed framework in several directions. These include incorporating cross-regional dependence through multi-region factor structures, integrating additional alternative data sources such as mobility or transaction-based in-

dicators, and exploring nonlinear or machine-learning-based mixed-frequency extensions. More broadly, the results suggest that flexible, factor-based mixed-frequency models offer a powerful and transparent tool for real-time regional economic analysis in crisis-affected economies.

References

Bai, J. and Ng, S. (2021). Matrix completion with cross-sectional and serial dependence. *Journal of Econometrics*, 222(1):413–430.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (favar) approach. *Quarterly Journal of Economics*, 120(1):387–422.

Cahan, D., Foerster, A., Sarte, P.-D., and Watson, M. W. (2023). Nowcasting with ragged-edge data: A projection-based approach. *Federal Reserve Working Paper*.

Choi, H. and Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88:2–9.

Cholette, P. A. (1984). Adjusting sub-annual series to yearly benchmarks. *Survey Methodology*, 10:35–49.

Denton, F. T. (1971). Adjustment of monthly or quarterly series to annual totals: An approach based on quadratic minimization. *Journal of the American Statistical Association*, 66(333):99–102.

Foroni, C. and Marcellino, M. (2013). A survey of econometric methods for mixed-frequency data. *Advances in Econometrics*, 31:1–45.

Foroni, C., Marcellino, M., and Schumacher, C. (2018). A survey of mixed frequency var and midas models. *Journal of Economic Surveys*, 32(6):1876–1899.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Koop, G., McIntyre, S., and Mitchell, J. (2023). Incorporating short data into large mixed-frequency vars for regional nowcasting. *Journal of the Royal Statistical Society: Series A*.

Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency var. *Journal of Business and Economic Statistics*, 33(3):366–380.

Schorfheide, F. and Song, D. (2020). Real-time forecasting with a (standard) mixed-frequency var during a pandemic. *Journal of Business and Economic Statistics*.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2016). Dynamic factor models, factor-augmented vars, and the analysis of monetary policy. *Handbook of Macroeconomics*, 2:415–525.