

Ministry of Education and Science of Ukraine
National University “Kyiv-Mohyla Academy”
Faculty of Informatics
Mathematics Department

Master’s Thesis

educational level – master

on the topic: «**NOWCASTING WITH SHORT DATA USING A
MIXED-FREQUENCY VAR APPROACH**»

By: 2nd-year student,
of the educational program
113. “Applied Mathematics”

Zhuravlova Anastasiia

Supervisor: Svitlana Drin
Candidate of Physical and
Mathematical Sciences, Senior Lecturer

Reviewer: _____

The coursework was defended
with a grade

EC secretary _____
(signature)

« _____ » _____ 20__ p.

National University “Kyiv-Mohyla Academy”

Faculty of Informatics

Mathematics Department

Master’s Degree

Major: 113 Applied Mathematics

Educational and Scientific Program: “Applied Mathematics”

APPROVED

Head of the Department

“_____” _____, 20__

ASSIGNMENT

FOR MASTER’S THESIS OF STUDENT

Zhuravlova Anastasiia

1. Thesis Topic: “Nowcasting with short data using a mixed-frequency VAR approach”

Thesis Supervisor: Drin Svitlana, Candidate of Physical and Mathematical Sciences, Senior Lecturer

Approved by the order of the higher educational institution

dated “__” _____, 20__, Order No. ____

2. Deadline for thesis submission by the student:

3. Work Plan:

1. Analysis of nowcasting methods and their application to regional economic indicators.
2. Review of modern VAR models for forecasting economic activity.
3. Development of a nowcasting model for the Ukrainian regions.
4. Construction of an MF-VAR model for GRP of Kyiv city using real data.
5. Empirical evaluation of model accuracy based on data from Ukrainian regions.
6. Comparison of model results with traditional forecasting approaches.

Master's Thesis Preparation Schedule

No.	TASK DESCRIPTION	Deadline	Date of Supervisor's Review	Supervisor's Signature	Remarks
1.	Topic selection, approval by the department, and assignment of academic advisor. Agreement on the calendar schedule for thesis preparation. Student familiarisation with evaluation criteria.	October	08.10.2024		
2.	Literature review, study of sources, periodicals, academic publications, collection and synthesis of data.	December-January	13.01.2025		
3.	Drafting the thesis outline and coordinating with the supervisor.	January-March	17.03.2025		
4.	Setting up the research experiment and analyzing the obtained results.	March – May	04.04.2025		
5.	Interim progress check.	March	30.03.2024		
6.	Completion of full draft, submission of the first version to the supervisor.	February-May	08.05.2025		
	Chapter 1 (problem statement, theoretical foundations, literature review).	January	27.01.2025		
	Chapter 2 (analytical and research section).	April	04.04.2025		
	Chapter 3 (project and recommendation section).	April-May	08.05.2025		
7.	Final completion and formatting of the thesis according to guidelines; submission to the supervisor for feedback.	April – early June	03.06.2025		
8.	Submission of the thesis for academic integrity check at NaUKMA.	June	08.06.2025		
9.	Submission for external review.	June	09.06.2025		
10.	Preparation for thesis defense at the department: drafting presentation and visual materials.	12.06.2025	10.06.2025		
11.	Submission of thesis to the department along with all accompanying documents.	08.06.2025	08.06.2025		
12.	Public defense before the Examination Committee.	According to EC schedule	13.06.2025		

Schedule approved « ____ » _____ 20__

Academic Advisor _____
(Full Name)

Master's Thesis Author _____
(Full Name)

Contents

List of main symbols and abbreviations	3
Abstract	4
Inroduction	5
1 Theoretical part	10
1.1 VAR models	11
1.2 Factor extraction methods in the MF-FAVAR framework	17
1.3 Alternative models	20
1.4 Forecast evaluation metrics	22
2 Empirical application for London data	25
2.1 Forecasting regional GDP for London	25
2.2 Forecast evaluation and results	28
3 Forecasting regional GDP for Kyiv	30
3.1 Data collection and preprocessing for Kyiv	30
3.2 Data preparation and temporal disaggregation procedure	32
3.3 Forecast evaluation and model comparison	35
3.4 Robustness check with reduced dataset	38
3.5 Final comparative analysis of factor extraction Methods	39
3.6 Summary of Results and Method Selection Justification	40
Conclusions and future directions	42
References	46
Appendix	47

List of main symbols and abbreviations

VAR	Vector Autoregression
MF-VAR	Mixed-Frequency Vector Autoregression
MF-FAVAR	Mixed-Frequency Factor-Augmented VAR
MIDAS	Mixed Data Sampling
PCA	Principal Component Analysis
EMPCA	Expectation-Maximization PCA
TW	Tall-Wide model
TP	Tall-Projection method
BVAR	Bayesian Vector Autoregression
RMSE	Root Mean Squared Error
CRPS	Continuous Ranked Probability Score
SMAPE	Symmetric Mean Absolute Percentage Error
GDP	Gross Domestic Product
GRP	Gross Regional Product

Abstract

This thesis focuses on the problem of short-term economic forecasting with very limited data because of Russia's invasion of Ukraine. The invasion has created a gap in the release of statistical data, making traditional forecasting models less efficient, particularly in the regional forecasting scope.

The study uses RMSE, CRPS, and SMAPE metrics to evaluate the forecast's accuracy while measuring accuracy against traditional approaches. The results show that strong nowcasting models can be developed even with sparse data. This work contributes to the literature on economic monitoring during crises and provides a framework for real-time forecasting in post-conflict reconstruction.

Анотація

Ця робота присвячена проблемі короткострокового економічного прогнозування за умов обмеженого обсягу даних через вторгнення Росії в Україну. Через війну виникла перерва в оприлюдненні статистичних показників, що ускладнило використання традиційних моделей прогнозування, особливо у регіональному контексті.

У дослідженні здійснюється оцінка точності прогнозів за допомогою метрик RMSE, CRPS та SMAPE, а також порівняння результатів із традиційними методами. Результати показують, що ефективні моделі nowcasting можна розробити навіть при нестачі даних. Це дослідження робить вагомий внесок у літературу з моніторингу економіки під час криз та створює основу для оперативного прогнозування в умовах післяконфліктного відновлення.

Introduction

Relevance of the study

The socio-economic crisis unfolding in Ukraine due to the Russian Federation's attack on Ukraine is unprecedented. With the full-scale invasion on February 24, 2022, the so-called concept of a gradual, steady thrust toward sustainable economic growth was supplanted entirely by a need for harsh shifts motivated not by market considerations but by security needs, devastation, and the intensity of warfare.

As per the government, the Russian invasion led to Ukraine's GDP shrinking by 29.1% in 2022, marking the steepest drop in economic output since the country attained independence in 1991. It was also forecasted that economic contraction could be around 25-30% in specific scenarios. The war has taken a toll on retail, manufacturing, and logistics, while education, transport, and healthcare have shown some resilience. The consequences of the invasion also have an uneven pattern: frontline regions have suffered the most and are now operating in a state of utter devastation, while areas in the West, which had previously been safer, have turned into new logistical and financial hubs.

Significant changes across regions created a distinct need to measure economic activity in real-time for specific areas. Nevertheless, the enforcement of martial law and weakened institutions make it hard to collect and share official data. Early in the conflict, the State Statistics Service of Ukraine and most private research firms froze all activity. Therefore, more advanced macroeconomic data, such as consumption, investment, and industrial output, are currently unavailable, rendering traditional short-term forecasting methods unusable.

This crisis requires the creation of flexible models capable of handling sparse, irregularly spaced data with varying intervals of irregularly spaced series. One class of such models is Mixed-Frequency Vector Autoregression (MF-VAR). These models can incorporate high-frequency (monthly or weekly) and low-frequency (quarterly GDP) variables.

One of the most critical aspects of this research is using pre-war data on the region due to the lack of access to statistics for the duration of the ongoing state of war. This limitation does restrict the model's real-time applicability. Still, it

also provides valuable insights: it shows that, even with harsh data constraints, it is possible to build regionally targeted nowcasting models using MF-VAR and latent factor modelling techniques.

This thesis aims to construct a nowcasting model for gross domestic product (GDP) for regions of Ukraine (Kyiv as a case study) utilising high-frequency indicators and pre-invasion data. The research explores the potential of forecasting with minimal time series data while evaluating the effectiveness of several factor extraction techniques amid numerous unstable structural breaks and sparse data.

Object of the study: the process of nowcasting of regional economic indicators under limited, irregular, and mixed-frequency data conditions.

Research aim and objectives

This research aims to develop a nowcasting model for regional economic indicators in Ukraine by integrating short and high-frequency data into a Mixed-Frequency Vector Autoregression (MF-VAR) framework. The study seeks to improve forecast accuracy under conditions of limited data availability, using only pre-war regional data due to the ongoing restrictions on official statistics.

To achieve this aim, the study sets out the following specific objectives:

- Analyse current nowcasting methodologies and evaluate their relevance for regional economic forecasting.
- Review modern MF-VAR models and assess their applicability in contexts with limited and irregular data.
- Develop a methodological approach for incorporating short and high-frequency data into a nowcasting model tailored to Ukraine.
- Construct a region-specific nowcasting model using pre-invasion data, focusing on Kyiv as a case study.
- Empirically evaluate the model's forecasting performance based on available historical regional data.
- Compare the model's results with those of traditional forecasting approaches under structural instability and data scarcity conditions.

Subject of the study: the development and empirical evaluation of an MF-VAR model for forecasting gross regional product (GRP) based on pre-invasion high-frequency indicators, using Kyiv as a case study.

Scientific novelty of the research:

- adapting and applying mixed-frequency factor models to Ukrainian regional data;
- evaluating and comparing three modern factor extraction methods (EMPCA, TW, TP) in the context of data sparsity and structural instability;
- comparing the MF-VAR model with other more traditional models;
- demonstrating that accurate regional GRP forecasts are possible with highly limited open data.

Practical significance of the results

This method creates a model for monitoring Ukraine's regional economies during wartime and the post-war period. It can also be adapted for other regions or used in policymaking to support timely decisions even without complete data. The results demonstrate the utility of situational and practical forecasting based on freely available tools and datasets. This approach is a valuable resource for government institutions, researchers, and international partners.

Research methods

This thesis employs a combination of analytical approaches and numerical modelling techniques to construct a nowcasting model for regional economic indicators of Ukraine using short and mixed-frequency data. The core part of the implementation was conducted in Python and R. In Python, libraries such as pandas, numpy, sci-kit-learn, stats models, and matplotlib were used for data preprocessing, model development, and result visualization. A full modelling pipeline was also implemented in R, using packages such as tempdisagg (for temporal disaggregation), missMDA, vars, scoring rules, and ggplot2. Special focus was placed

on extracting latent factors from incomplete panels using EMPCA, Two-Way Shrinkage/Tall-Wide (TW), and Tall Projection (TP). An MF-FAVAR model was estimated for each factor extraction method, and forecast accuracy was assessed using the MAE, RMSE, SMAPE, and CRPS metrics.

Dissemination of results

The core results of this master's thesis were presented by Zhuravlova (2025) at the XIII All-Ukrainian scientific conference of young mathematicians, which took place on May 9, 2025, in Kyiv. The presentation outlined a practical approach to forecasting regional economic dynamics under limited data availability and heterogeneous data frequency constraints.

Structure of the thesis

Given the limited and varied datasets available, this thesis aims to address the problem of nowcasting regional economic indicators for Ukraine.

The first chapter describes the study's context. It starts with relevance, especially from the perspective of the disruption to economic data streams caused by the full-scale Russian invasion of Ukraine. The chapter also identifies the key research aim and tasks. It briefly captures the thesis's global outline and provides the actual structural overview.

In the second chapter, the study's outline is finalized with a deeper examination of its theoretical components. It elaborates on nowcasting principles in macroeconomics, describes the structure and advantages of MF-VAR models, and discusses challenges associated with short and irregular time series. Particular focus is given to the forecasting issue that arises when traditional statistical data is unavailable.

The third chapter outlines the application of the approach discussed in this thesis in the empirical context. It explains first the data set that contains the pre-invasion economic data for Ukraine by regions. It then explains the construction of the MF-VAR model with its integrations of short and high-frequency indicators and the application of various factor extraction methods. The chapter also

comprehensively assesses the model's predictive accuracy, including benchmarking against other forecasting techniques.

As for the final chapter, it provides an overview of the study's principal insights, synthesising the most important results while noting the study's impact. Also, it provides some other approaches related to the current methodology. It aims to expand the scope of work, such as moving on to the post-war situation and applying it to other regions.

The thesis is concluded with a list of references, which includes all the scientific literature and materials used in the course of writing the work.

1 Theoretical part

Nowcasting refers to the real-time assessment of the economy's current state using incomplete and asynchronous data streams designed to overcome delays in official statistics by incorporating high-frequency indicators Bańbura et al. (2013). As defined by Bańbura et al. (2011), nowcasting is a methodology that enables short-term forecasts of economic activity even before the release of official macroeconomic indicators. This feature is critically important during rapid change, crisis, or data scarcity.

In the context of regional economic analysis in Ukraine, where data availability is limited and time series structure is often disrupted due to war, modern approaches based on mixed-frequency frameworks have become particularly relevant. Koop et al. (2023) proposed the MF-VAR (Mixed-Frequency Vector Autoregression) model, which integrates short time series into regional forecasting models. Their approach addresses the so-called "ragged edge" problem both at the beginning and end of the sample, a common challenge in Ukraine, where regional GRP data is often only available annually and with significant delay.

This model incorporates factor extraction techniques (EMPCA, TW, TP) from a set of high-frequency indicators, which are then fed into a VAR system to construct an MF-FAVAR model. This framework underpins the empirical strategy of this thesis: for the city of Kyiv, I developed a nowcasting model of quarterly GRP growth, combining monthly, weekly, and annual time series, which were first aligned in frequency using the Denton-Cholette disaggregation method and reduced in dimensionality via factor analysis.

Simultaneously, the literature emphasises the need to formalise the processing of unaligned and asynchronous data streams in real-time. For example, Bańbura et al. (2013) stresses that dynamic factor models (DFM), as well as MIDAS and MF-VAR approaches, help automate expert judgment when handling large sets of economic indicators. These models are especially effective when data reporting is delayed or in periods of crisis, such as the war in Ukraine.

Also noteworthy are the contributions of Fosten and Greenaway-McGrevy (2022), who demonstrate the effectiveness of panel nowcasting models for multi-regional forecasting. While this thesis focuses solely on Kyiv, including cross-sectional

information may enhance the model's robustness in future research.

The current literature provides a solid theoretical foundation for constructing adaptive nowcasting models even under data irregularity, scarcity, and structural instability — conditions that define the economic reality in Ukraine's regions today.

1.1 VAR models

Classical Vector Autoregression (VAR) model

The classical Vector Autoregression (VAR) model, introduced by Sims (1980), is a foundational econometric framework designed to capture the dynamic interdependencies among multiple time series. It treats all variables as endogenous and models each variable in the system as a linear function of its past values and the past values of all other variables in the system. This structure is particularly valuable in macroeconomic and regional economic modelling, where feedback effects across indicators are common.

A VAR model of order p , denoted as VAR(p), for a K -dimensional vector of endogenous variables $\mathbf{y}_t \in \mathbb{R}^K$, takes the following form:

$$\mathbf{y}_t = A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + \cdots + A_p\mathbf{y}_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma), \quad (1)$$

where:

- \mathbf{y}_t — vector of all observed variables in the system at time t .
- A_1, A_2, \dots, A_p — coefficient matrices, each of size $K \times K$, where K is the number of variables. These matrices capture how past values of each variable influence current values.
- $\mathbf{y}_{t-1}, \dots, \mathbf{y}_{t-p}$ — lagged values of the variable vector. If the model order is $p = 2$, it incorporates the effect of the two previous time steps.
- ε_t — a shock term or innovation, which is unpredictable and modeled as white noise. It follows a normal distribution with zero mean and a positive definite covariance matrix Σ .

- p is the system's lag order.

Illustrative example (Kyiv case):

Suppose we observe three economic indicators: **Gross Regional Product (GRP) of Kyiv** — the target variable y_q , **Macroeconomic indicators** (e.g., consumer price index, wages, construction output), denoted as Macro_q and **Google Trends indicators**¹ (e.g., “Investment”, “Savings”, “Unemployment”), denoted as GT_q .

Let $K = 3$, so the vector of endogenous variables at quarter q is:

$$\mathbf{y}_q = \begin{bmatrix} y_q \\ \text{Macro}_q \\ \text{GT}_q \end{bmatrix}$$

We specify a VAR (1) model:

$$\mathbf{y}_q = A_1 \mathbf{y}_{q-1} + A_2 \mathbf{y}_{q-2} + \varepsilon_q, \quad \varepsilon_q \sim \mathcal{N}(0, \Sigma)$$

The coefficient matrix A_1 may look like:

$$A_1 = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} \end{bmatrix}$$

where:

- $a_{13}^{(1)}$ shows the effect of Google Trends indicators on GRP growth (behavioural influence),
- $a_{12}^{(1)}$ captures the effect of macroeconomic indicators on GRP growth,
- $a_{31}^{(1)}$ measures whether past GRP explains future search behaviour (feedback loop).

Each equation in the VAR corresponds to one element of \mathbf{y}_t and is estimated using Ordinary Least Squares (OLS), under the assumption of covariance stationarity. The VAR model allows for a fully data-driven specification of dynamic

¹Google Trends data are interest indices based on the frequency of user queries for selected keywords (e.g., “Investment”, “Savings”, “Unemployment”). Source: <https://trends.google.com>.

interactions without requiring structural economic theory as a prior. Its applications include point forecasting, impulse response analysis, and forecast error variance decomposition.

However, in practice, several limitations of the classical VAR framework become apparent:

- **Frequency homogeneity requirement:** all variables must be observed at the same frequency (e.g., monthly or quarterly), which is restrictive in contexts where low-frequency indicators (such as annual regional GDP) must be combined with high-frequency predictors.
- **High parameterisation in small samples:** with a growing number of variables and lags, the number of parameters to estimate increases quadratically, which poses risks of overfitting and weak out-of-sample performance, particularly in settings with limited historical data.
- **Sensitivity to missing data:** classical VAR models assume complete, regularly spaced observations, unrealistic in crisis-affected economies where data may be delayed, missing, or published irregularly.

These constraints are particularly relevant for regional nowcasting in Ukraine. In my case, for Kyiv, the gross regional product (GRP) is only available annually, with a short and irregular time series due to wartime disruptions. At the same time, various timely indicators — such as online activity, price indices, etc. — are available at higher frequencies.

These data characteristics violate key assumptions of the standard VAR model. Therefore, the classical VAR serves as a conceptual starting point in this thesis. At the same time, a more flexible Mixed-Frequency VAR (MF-VAR) framework is adopted to account for the temporal structure and missingness in real-world economic data. This extension is elaborated in the following section.

Mixed-Frequency VAR (MF-VAR) Model

The Mixed-Frequency Vector Autoregression (MF-VAR) model is an extension of the classical VAR framework that allows for the joint modelling of time series

observed at different frequencies, such as monthly, quarterly, and annual data, within a unified system. Initially proposed in works such as Forni and Marcellino (2013), Schorfheide and Song (2015), and Koop et al. (2023), MF-VAR models provide a powerful approach for nowcasting when key macroeconomic indicators are released with different periodicities.

Unlike traditional VAR models, which require all input variables to be sampled at the same frequency, MF-VAR can incorporate both high-frequency indicators (e.g., online activity, financial metrics) and low-frequency target variables (e.g., annual or quarterly regional GDP) without the need for aggressive interpolation or temporal aggregation.

A standard MF-VAR of lag order p takes the form:

$$\mathbf{y}_t = A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + \cdots + A_p\mathbf{y}_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma),$$

where:

- $\mathbf{y}_t \in \mathbb{R}^K$ is the vector of endogenous variables with mixed frequencies;
- $A_i \in \mathbb{R}^{K \times K}$ are the coefficient matrices for each lag;
- ε_t is a vector of white-noise disturbances with covariance matrix Σ .

To properly handle mixed-frequency data, the model may embed specialised transformations (e.g., MIDAS polynomials, Kalman filtering) or disaggregation methods (such as Denton–Cholette) to align observations across time.

Limitations of MF-VAR

While MF-VAR offers significant advantages over classical VAR in dealing with asynchronous data flows, it still suffers from several structural limitations:

- **No dimensionality reduction:** each high-frequency series is still treated as a separate variable. When the number of indicators grows, the model becomes highly parameterized and vulnerable to overfitting.
- **Sensitivity to noise and short series:** with limited historical data (often in regional applications like Ukraine), MF-VAR estimates can become unstable.

- **Lack of latent structure:** the model does not extract or exploit latent common trends or co-movements among indicators, reducing its ability to generalise in noisy or sparse inputs.

In my case study focused on the city of Kyiv, the gross regional product (GRP) is only available annually and often with substantial publication delays. These annual values were disaggregated into quarterly figures using Denton–Cholette methods to build a nowcasting system. Still, even with frequency harmonisation, including all high-frequency indicators directly in the model proved inefficient and unstable due to the small number of observations and large parameter space.

Motivation for MF-FAVAR

This thesis adopts an MF-FAVAR (Mixed-Frequency Factor-Augmented VAR) approach to address these challenges. MF-FAVAR extends MF-VAR by combining it with factor analysis. Instead of including every indicator as a separate regressor, the model first extracts a small number of latent dynamic factors that summarise the shared information across a large set of high-frequency indicators. These factors are then included in the VAR framework.

The benefits of MF-FAVAR include:

- **Dimensionality reduction:** Avoids overfitting by compressing many indicators into a few informative factors;
- **Robustness to missing and irregular data:** Factor models can naturally handle unbalanced panels;
- **Better generalisation:** The model is more robust in real-world economic settings with noise and volatility by capturing latent structure.

The MF-FAVAR framework used in this thesis is described in more detail in the following section.

MF-FAVAR model specification and illustration

The *Mixed-Frequency Factor-Augmented Vector Autoregression (MF-FAVAR)* model is an extension of the standard VAR framework that accommodates variables of

different sampling frequencies and reduces dimensionality by introducing latent factors. In this framework, observed macroeconomic indicators are assumed to be noisy manifestations of a few common, unobserved economic forces that evolve over time.

The model consists of two components: measurement equations and a state equation. Let $x_t^{(H)} \in \mathbb{R}^{n_H}$ denote the vector of high-frequency indicators (e.g., weekly or monthly CPI, Google Trends, sentiment indices), and $y_t^{(L)} \in \mathbb{R}^{n_L}$ represent low-frequency target variables (e.g., annual Gross Regional Product – GRP). The latent factors $f_t \in \mathbb{R}^r$ capture the common dynamics of both variables and evolve according to a VAR process.

The measurement equations are specified as:

$$x_t^{(H)} = \Lambda^{(H)} f_t + \eta_t^{(H)}, \quad y_t^{(L)} = \Lambda^{(L)} f_t + \eta_t^{(L)},$$

where:

- $\Lambda^{(H)} \in \mathbb{R}^{n_H \times r}$ is the factor loading matrix for high-frequency indicators;
- $\Lambda^{(L)} \in \mathbb{R}^{n_L \times r}$ is the factor loading matrix for low-frequency variables;
- $\eta_t^{(H)}, \eta_t^{(L)}$ are idiosyncratic error terms, uncorrelated with the latent factors.

The dynamics of the latent factors are governed by a VAR(p) process:

$$f_t = \Phi_1 f_{t-1} + \dots + \Phi_p f_{t-p} + \nu_t, \quad \nu_t \sim \mathcal{N}(0, Q),$$

where:

- $\Phi_i \in \mathbb{R}^{r \times r}$ are VAR coefficient matrices;
- $Q \in \mathbb{R}^{r \times r}$ is the innovation covariance matrix;
- ν_t is a white noise shock to the factors.

In our study, we consider the case of Kyiv, where annual GRP data are irregular and delayed due to wartime disruptions. At the same time, high-frequency proxies such as CPI, Google Searches, and social indicators are available weekly or monthly. Assume we choose $r = 2$ latent factors to capture the underlying

economic dynamics. These latent factors are extracted from the high-frequency panel using PCA or EMPCA methods, then used to nowcast GRP through:

$$\hat{y}_t^{(L)} = \Lambda^{(L)} \hat{f}_t.$$

To illustrate, suppose the high-frequency vector is

$$x_t^{(H)} = [Macro, GT]^\top \in \mathbb{R}^2,$$

| The extracted factors \hat{f}_t are then used to forecast GRP via the low-frequency measurement equation. Compared to the classical MF-VAR Schorfheide and Song (2020), which models each observable directly, MF-FAVAR reduces dimensionality and increases robustness, especially under limited data availability and short time series Bańbura et al. (2011). Thus, it offers a scalable solution for high-frequency regional nowcasting when data are sparse, asynchronous, or noisy.

1.2 Factor extraction methods in the MF-FAVAR framework

Within the MF-FAVAR framework, a critical step involves extracting latent factors f_t that synthesise information from a large set of regional indicators Z_t , many of which contain missing values, particularly at the beginning or end of the sample. Three main algorithms are commonly used to address this: **EMPCA**, **TW (Tall-Wide)**, and **TP (Tall-Projection)**. These methods are specifically designed to handle “ragged-edge” data and allow for extracting regional factors even when observations are incomplete Koop et al. (2023).

Principal Component Analysis

The PCA is the basic approach to identifying latent factors in factor-augmented models like MF-FAVAR. It efficiently reduces multivariate data’s dimensionality, transforming high-dimensional data sets into lower-dimensional ones while retaining as much relevant information as possible.

Originally proposed by Pearson (1901) and formalised by Hotelling (1933),

PCA converts potentially correlated variables into a different set of linearly uncorrelated variables known as principal components. These components are orthogonal by construction and ranked by the proportion of variance they capture from the original data.

The first principal component shows the most significant amount of variance. Each of the following components explains the remaining variability while ensuring it is uncorrelated with the previous ones.

Mathematically, consider a mean-centred data matrix $X \in \mathbb{R}^{n \times p}$, where n is the number of observations and p is the number of variables. The principal components relate to the eigenvectors of the covariance matrix:

$$C_X = \frac{1}{n-1} X^\top X$$

Solving the eigenvalue problem

$$C_X v = \lambda v$$

yields the directions v of maximum variance (principal components) and their associated eigenvalues λ , which represent the explained variance.

In practice, the computation of PCA often involves utilising the Singular Value Decomposition (SVD) of the data matrix:

$$X = U \Sigma V^\top$$

Here, the columns of V contain the principal components, and by selecting the first $k < p$ components, one obtains a reduced representation of the data via:

$$T = X V_k$$

The resulting matrix T represents the original data projected into a lower-dimensional space that captures the essential structure of the dataset. This approach is especially beneficial in economic applications involving high-frequency indicators with potential noise, missing values, or multicollinearity. In MF-FAVAR frameworks, PCA plays a central role in summarising large panels of economic indicators into a small number of interpretable latent factors that drive

the system’s dynamics Jolliffe (2002).

Although classical PCA works well for complete datasets, it struggles when some values are missing, as is often the case with high-frequency economic indicators or regional data that are delayed or partially available. Data gaps are common in real-world applications like nowcasting regional GRP, especially at the edges of the time series. Several improved PCA versions, such as EMPCA, Tall-Wide, and Tall-Projection, have been developed to handle this. These methods allow us to extract meaningful latent factors even when the input panel is incomplete, making them more suitable for mixed-frequency and ragged-edge data situations Cahan et al. (2023).

Expectation-Maximisation PCA

The EMPCA method is an iterative extension of traditional Principal Component Analysis (PCA), adapted to accommodate missing values. The missing values in the data matrix Z are initially imputed using mean or heuristic methods. Then, PCA is applied to the imputed matrix to obtain estimates of the factors \hat{F} and loadings $\hat{\Lambda}$, and the missing values are re-estimated as the product $\hat{\Lambda}\hat{F}^\top$. This process repeats until convergence. Although EMPCA is flexible and suitable for arbitrary missingness patterns, it is computationally intensive and may fail to converge when a large portion of the data is missing Stock and Watson (2002).

Tall-Wide PCA

The TW algorithm, proposed by Bai and Ng (2021), avoids iterations by dividing the data matrix into two subsets: a “tall block” of variables with complete time series and a “wide block” of time periods with complete observations across variables. PCA is performed separately on both subsets—first on the columns without missing values (to obtain f_t^{tall}), and then on the rows without missing values (yielding f_t^{wide}). These two factor estimates are then optimally combined using regression. This method is computationally efficient, but its performance is sensitive to the number of variables in the tall block; with too few, the resulting factor estimates may be unreliable.

Tall-Projection

The TP method, developed by Cahan et al. (2023), improves on TW by addressing the issue of short or sparse variables that limit the wide block. Like TW, it begins with PCA on the tall block to estimate f_t^{tall} . Then, for each variable with missing data, an auxiliary regression on the estimated tall factors is used to impute missing entries. Once all missing values are filled, a final PCA is conducted on the complete matrix to extract the regional factors. This approach is more robust in cases with uneven coverage across indicators, though it also depends heavily on the quality of the tall block.

1.3 Alternative models

In addition to EMPCA, TW, and TP, which are specifically designed to deal with ragged-edge structures in high-frequency data, it was also necessary to consider a broader set of techniques due to the severe limitations of regional economic datasets. These datasets often exhibit short time series, structural gaps, or inconsistent update cycles, particularly in wartime or crisis-affected regions. I explored several complementary approaches to address these challenges: Bayesian PCA (BPCA) for probabilistic factor estimation under uncertainty and SVD-based imputation (SVDI) for iteratively reconstructing missing entries. These methods provide additional robustness and flexibility when standard factor extraction techniques struggle under real-world regional data constraints.

Bayesian Principal Component Analysis

The BPCA is a probabilistic extension of the classical PCA method, developed to address the problem of missing data in a principled and flexible manner. Building upon the framework of Probabilistic PCA (PPCA) introduced by Tipping and Bishop (1999), BPCA treats the parameters of the PCA model—including the loading matrix W , mean vector μ , and noise variance σ^2 —as random variables, assigning them prior distributions. This fully Bayesian formulation enables the computation of posterior distributions over all unknown quanti-

ties using Bayes' theorem, rather than relying on point estimates as in traditional or EM-based PCA.

The core generative model in BPCA assumes that each observed data vector $t \in \mathbb{R}^p$ is generated from a lower-dimensional latent vector $x \in \mathbb{R}^k$ via a linear Gaussian model:

$$t = Wx + \mu + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Unlike PPCA, which estimates parameters by maximizing likelihood, BPCA introduces Gaussian priors for the columns of W and implements a hierarchical prior structure that enables *Automatic Relevance Determination* (ARD). This mechanism induces sparsity in the latent space by shrinking irrelevant components of W toward zero, allowing the model to automatically infer the effective dimensionality without requiring a predefined number of factors.

One of BPCA's key advantages is its natural ability to handle missing data. Instead of imputing missing values with single-point estimates, BPCA marginalizes over their possible values using the posterior distribution. Posterior inference is typically performed using an iterative procedure similar to the Expectation-Maximisation (EM) algorithm, alternating between latent variable estimation and updating the posterior of model parameters.

This makes BPCA especially suitable for high-dimensional economic datasets that often exhibit ragged edges or regional incompleteness. In MF-FAVAR frameworks, BPCA provides a robust alternative to EMPCA by supporting uncertainty quantification, latent dimensionality estimation, and more reliable inference under data sparsity.

Singular Value Decomposition Imputation

The SVDI method is an iterative algorithm designed to estimate missing entries in data matrices by leveraging their low-rank structure. Originally introduced in the context of gene expression analysis by Troyanskaya et al. (2001), it has since become a widely used baseline technique in many areas, including economic modelling.

The core assumption of SVDI is that the underlying data can be well-approximated by a low-rank matrix. This makes it particularly useful in high-dimensional ap-

plications like MF-FAVAR, where economic indicators often exhibit strong interdependence and can be explained by a few latent components.

The algorithm proceeds iteratively:

- **Initialisation:** all missing values in the data matrix X are initially replaced with preliminary estimates, commonly the column-wise means. This yields a complete matrix $\hat{X}^{\text{complete}}$ to start the iteration.
- **SVD step:** Singular Value Decomposition is applied to the current version of the matrix:

$$\hat{X}^{\text{complete}} = U\Sigma V^\top$$

- **Low-rank approximation:** only the top k singular values are retained, producing a low-rank approximation:

$$\hat{X}^{\text{approx}} = U_k \Sigma_k V_k^\top$$

- **Update step:** the missing entries in $\hat{X}^{\text{complete}}$ are updated using the corresponding values from \hat{X}^{approx} , while the observed values remain unchanged.

This loop continues until convergence, typically when the changes in imputed values fall below a set threshold.

SVDI is conceptually similar to EM-PCA in that both rely on iterative low-rank reconstructions. Its main strength lies in computational simplicity and reasonable accuracy under the assumption of a strong latent structure. However, unlike probabilistic methods such as BPCA, SVDI does not provide uncertainty estimates and may be less robust to complex missingness patterns. Despite this, SVDI remains a practical choice for imputing incomplete panels of economic indicators, especially when speed and scalability are key concerns.

1.4 Forecast evaluation metrics

To assess forecast performance precisely, I utilise four widely recognised metrics: *Root Mean Squared Error (RMSE)*, *Continuous Ranked Probability Score (CRPS)*, *Mean Absolute Percentage Error (MAPE)*, and *Symmetric Mean Absolute Percentage Error (SMAPE)*. These metrics evaluate different aspects of

predictive accuracy, covering both point and probabilistic forecasts, and are particularly suitable for economic time series data Gneiting and Raftery (2007); Hyndman and Koehler (2006).

Root Mean Squared Error

The RMSE measures how big the forecast errors are, giving more importance to larger mistakes. It is calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2},$$

where y_t and \hat{y}_t are the observed and predicted values at time t , respectively. The RMSE is measured in the same units as the predicted variable and is particularly responsive to extreme outliers, making it valuable in situations where large prediction errors are especially problematic.

Continuous Ranked Probability Score

In contrast, the CRPS assesses the accuracy of probabilistic forecasts. It compares the entire predictive cumulative distribution function $\hat{F}(z)$ against the realized value y , and is defined as:

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

For normally distributed forecasts $\hat{F} = \mathcal{N}(\mu, \sigma^2)$, CRPS admits a closed-form solution, enabling efficient computation. Lower CRPS values indicate better probabilistic calibration and sharpness of the forecast.

Mean Absolute Percentage Error

The MAPE expresses forecast error as a percentage of the actual values:

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right|.$$

MAPE is intuitive and easy to interpret, especially for business and policy audiences. However, it is undefined when $y_t = 0$, and small denominators can distort it.

Symmetric Mean Absolute Percentage Error

To mitigate these issues, we also compute the SMAPE, which symmetrically scales the absolute error using the average of predicted and observed values:

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^n \frac{|\hat{y}_t - y_t|}{(|\hat{y}_t| + |y_t|)/2}$$

SMAPE is bounded between 0% and 200% and is more robust in the presence of zero or near-zero values.

These measures clearly show how well the model works. They examine accuracy, possible errors, and how well the model shows uncertainty in its forecasts.

2 Empirical application for London data

2.1 Forecasting regional GDP for London

The empirical analysis starts with a specific examination of the London area, chosen for its better data accessibility, timeliness, and comprehensiveness compared to other regions in the UK. London offers the most reliable and frequent release of regional economic metrics, usually published with little delay. These characteristics make it an ideal candidate for validating the implementation of the Mixed-Frequency Factor-Augmented Vector Autoregression (MF-FAVAR) framework before applying it to data-constrained environments such as Ukrainian regions.

Objective and model overview

The goal of this analysis is to nowcast London's quarterly real Gross regional product (GDP) using a **Mixed-Frequency Factor-Augmented Vector Autoregression (MF-FAVAR)** model. The implementation was carried out in Python using the `pandas`, `scikit-learn`, `statsmodels`, and `properscoring` libraries.

Data loading and preprocessing

Two main data sources were used:

- **Quarterly GDP data** (`Quarterly.csv`) for London, obtained from the UK Office for National Statistics (ONS);
- **Monthly and quarterly predictor variables** (`Predictors_*.csv`), including a broad range of indicators spanning labour market activity, economic output, and housing conditions.

The complete set of indicators used in the forecasting model is presented in Table 2.1. Only variables with publication delays no longer than the delay for regional GDP were retained.

Table 2.1: Regional Indicators: Short and Long Data
(Adapted from Koop et al., 2023)

#	Variable	Description	Freq.	Coverage	Time period	Delay
1	UKQuarterlyHPI	Nationwide Building Society House Price Index	Q	N	Q1 1967–	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 Government UK House Price Index	M	N	Apr 1968–	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	Total New Housing Output	Q	R (ex. NI)	Q1 1980–Q4 2021	6 weeks
8	ConstructionOutput_ANW	All New Work Output	Q	R (ex. NI)	Q1 1980–Q4 2021	6 weeks
9	ConstructionOutput_AW	All Work Output	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	Jul 2014–Jan 2022	2 weeks
18	PayrollPayMedian	Median Payroll Pay	M	R	Jul 2014–Jan 2022	2 weeks
19	HousePrice	House Price Index	M	R	Jan 1995–Mar 2022	6 weeks
20	PMIActivity	PMI Activity Measure (Headline)	M	R	Jan 1997–Jan 2022	2 weeks
21	PMINewbus	New Business Measure	M	R	Jan 1997–Jan 2022	2 weeks
22	PMIOutbus	Outstanding Business	M	R	Nov 1999–Jan 2022	2 weeks
23	PMICharges	Charges	M	R	Nov 1999–Jan 2022	2 weeks
24	PMIPrices	Prices	M	R	Jan 1997–Jan 2022	2 weeks
25	PMIEmploy	Employment	M	R	Jan 1997–Jan 2022	2 weeks
26	PMIFuture	Future Orders	M	R	Jul 2012–Jan 2022	2 weeks
27	HousingRental	Private Housing Rental Prices	M	R	Jan 2005–Apr 2022	3 weeks
28	OvernightVisits	Overnight Visits to UK by Region	Q	R (selected)	Q1 2017–Q3 2021	5 months
29	Spending	Overseas Residents Spending by UK Region	Q	R (selected)	Q1 2009–Q3 2021	5 months
30	IncorporatedCompanies	Newly Incorporated and Removed Firms	Q	R (selected)	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	NI Index of Services	Q	R (NI only)	Q1 2005–Q4 2021	3 months
36	NI_IOP	NI Index of Production	Q	R (NI only)	Q1 2005–Q4 2021	3 months
37	NI_RSI	NI Retail Sales Index	Q	R (NI only)	Q1 2014–Q4 2021	3 months
38	NI_PortsTraffic	NI Ports Traffic	Q	R (NI only)	Q1 2009–Q4 2021	4 months
39	NI_ConstructionOutput	NI Construction Output	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

To ensure the feasibility of real-time nowcasting, only predictors that are published promptly were included. Indicators such as VAT turnover, which suffer from long publication lags (typically 5+ months), were excluded. Monthly variables were aggregated to quarterly frequency using within-quarter averages.

Notably, the model incorporates both regional and national-level indicators. Although the forecasting target is regional, national indicators (e.g., UK-level employment or house prices) are used to enhance factor quality. Koop et al. (2023) noted that national indicators are timelier, more stable, and less sparse than regional series. Their inclusion improves the robustness of factor extraction methods, especially in high-dimensional settings with missing or noisy regional data, and allows the model to capture macroeconomic trends that influence regional dynamics.

Latent factor extraction

Three dimensionality-reduction techniques, EMPCA, TW, and TP, described before, were applied to the predictor panel.

Additionally, a ridge-regression-based **BVAR** model was used as a benchmark, operating directly on lagged values without factor extraction. The extracted factors (excluding BVAR) were merged into a consolidated panel, forming the input for the forecasting stage.

Forecasting model construction

Forecasts were generated using a **one-step-ahead** approach, where each future value is predicted based solely on past data. Two modelling strategies were employed:

- A **VAR**(1) model on the factor-augmented panel for MF-FAVAR specifications;
- A **ridge regression** model as a simplified BVAR proxy using lagged GVA values directly.

Forecast accuracy evaluation

Three performance metrics were used to evaluate the models:

- **RMSFE (Root Mean Squared Forecast Error)** – measures the average magnitude of forecast error;
- **CRPS (Continuous Ranked Probability Score)** – captures both accuracy and uncertainty of probabilistic forecasts;
- **MAPE (Mean Absolute Percentage Error)** – expresses error as a percentage of the observed value.

2.2 Forecast evaluation and results

To evaluate the predictive performance of the MF-FAVAR model for the London region, we applied four forecasting methods: EMPCA, TW, TP, and a benchmark Bayesian VAR approximation (BVAR). The accuracy of each method was assessed using two complementary metrics: Root Mean Squared Forecast Error (RMSFE) and Continuous Ranked Probability Score (CRPS). Lower values of both metrics indicate higher forecast accuracy and better uncertainty calibration, as summarised in Table 2.2.

Table 2.2: Forecast accuracy metrics across methods for London (lower = better)

Method	RMSFE	CRPS (%)
EMPCA	0.104	6.3
TW	0.119	7.3
TP	0.161	9.4
BVAR	0.160	70.4

Among the tested approaches, **EMPCA** demonstrated the best performance, achieving the lowest RMSFE (0.11) and CRPS (6.25%), outperforming both classical factor extraction strategies (TW and TP) and the ridge-regression-based BVAR baseline. This result highlights the effectiveness of the Expectation-Maximisation approach in handling missing values and noisy high-dimensional economic indicators, which are prevalent in regional forecasting scenarios.

Results visualization

A time series plot was produced to compare actual and predicted GVA values across the test period. The best-performing model’s forecasts were visualised alongside a **95% confidence interval**, constructed from the forecast error standard deviations to reflect prediction uncertainty.

Figure 2.1 presents the forecast trajectory for the average GRP for London under the best-performing model (EMPCA). The solid blue line shows the actual observed values, while the dashed orange line indicates the model’s one-step-ahead forecasts. The shaded area represents the 95% predictive interval derived from the model’s forecast error variance. The plot visually confirms that the EMPCA-based forecast tracks the underlying trend well while maintaining realistic confidence bounds.

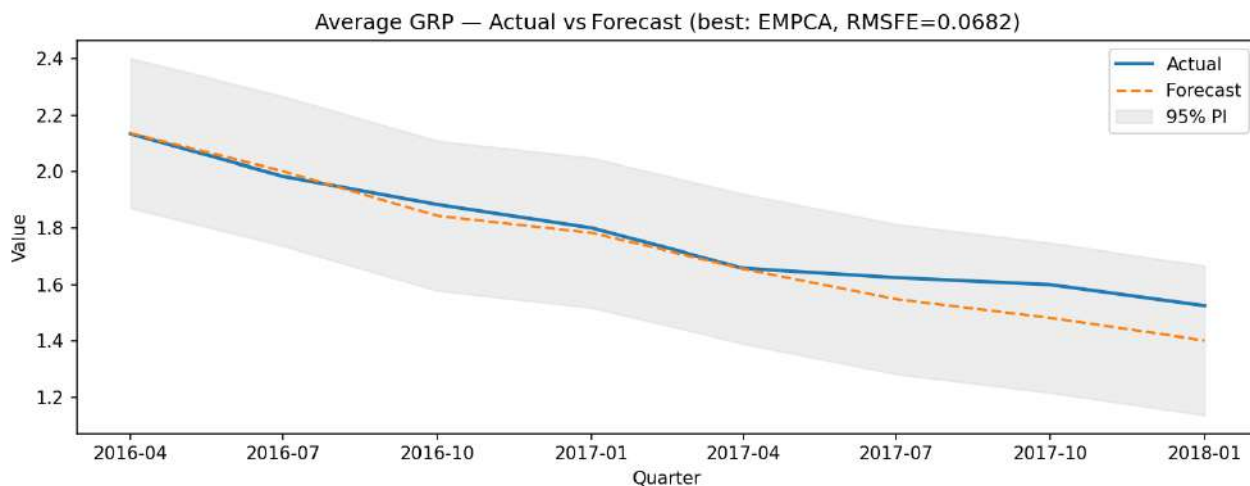


Figure 2.1: Average GRP — Actual vs Forecast (Best: EMPCA, RMSFE = 0.1041)

Compared to alternatives, EMPCA’s superior performance suggests that it is well-suited for nowcasting applications involving mixed-frequency and incomplete economic data. The low CRPS value further indicates the model’s ability to generate sharp and well-calibrated probabilistic forecasts—an important property for policy-relevant decision-making.

3 Forecasting regional GDP for Kyiv

3.1 Data collection and preprocessing for Kyiv

To illustrate the practical application of the proposed MF-FAVAR approach for regions in Ukraine, we focus on the region of Kyiv, which provides the most comprehensive set of regional economic indicators available in the country. Data availability was a key factor in choosing this region, as a wide variety of both annual and monthly series are published at the regional level. (Table 3.1)

Importantly, the regional gross product (GRP) for Kyiv is available only at an annual frequency, which determined the choice of yearly data as the primary target variable for forecasting. The GRP data were retrieved from the official database of the UkrStat. The published dataset covers the period from 2004 to 2021. Based on the archive structure and publication dates of recent years, GRP figures are typically released with a delay of approximately 12–15 months after the end of the reporting year.

The annual dataset includes indicators such as household income, disposable income, wages, social benefits, expenditures, retail goods purchases, income taxes, social contributions, consumption of non-profit institutions serving households (NPISH), gross capital formation of households, number of immigrants and emigrants, subway passengers, number of employees, capital investments, and volume of industrial output. However, some of these series—migration data, employment numbers, capital investments, and production volumes—were too short to be useful in time series modelling and were excluded from the analysis.

In addition, we collected monthly data that included the consumer price index (CPI), disaggregated CPI by categories (e.g., food and beverages, alcohol and tobacco, clothing, housing, furnishings, transport, communication, recreation, education, restaurants and hotels), total industrial output, volume of construction production, and average wages. Nevertheless, several CPI subcategories contained very short periods and were removed to reduce noise in the dataset.

Table 3.1: Overview of regional economic indicators for Kyiv

Variable	Description	Frequency	Time Period	Release Lag	Source
income	Household income	Yearly	2001–2021	16M	(UkrStat, 2021d)
real_income	Disposable income	Yearly	2001–2021	16M	(UkrStat, 2021b)
salary_1	Wages	Yearly	2001–2020	14M	(UkrStat, 2021g)
sochelp	Social benefits	Yearly	2001–2020	16M	(UkrStat, 2021e)
expenses	Expenditures	Yearly	2001–2020	16M	(UkrStat, 2021c)
buying_goods	Retail goods purchases	Yearly	2001–2020	16M	(UkrStat, 2021e)
taxes	Income taxes	Yearly	2001–2020	16M	(UkrStat, 2021e)
socpay	Social contributions	Yearly	2001–2020	16M	(UkrStat, 2021e)
consume_expenses	NPISH consumption	Yearly	2008–2020	16M	(UkrStat, 2021e)
gross_kapital	Household capital formation	Yearly	2008–2020	16M	(UkrStat, 2021e)
metro_passangers	Subway passengers	Yearly	1995–2020	16M	(UkrStat, 2021f)
hired_workers	Hired employees	Yearly	2017–2018	16M	(UkrStat, 2018)
investments	Capital investments	Yearly	2019–2022	16M	(UkrStat, 2023)
realised_goods	Realized output	Yearly	2015–2016	10M	(UkrStat, 2016)
consumer_price_index	CPI (total)	Monthly	2001–2021	1M	(MinFin, 2021)
in_products	CPI: Food	Monthly	2020	1M	(UkrStat, 2020)
in_alco	CPI: Alcohol and tobacco	Monthly	2020	1M	(UkrStat, 2020)
in_clothes	CPI: Clothing	Monthly	2020	1M	(UkrStat, 2020)
in_appartments	CPI: Housing	Monthly	2020	1M	(UkrStat, 2020)
in_house_goods	CPI: Furnishings	Monthly	2020	1M	(UkrStat, 2020)
in_transport	CPI: Transport	Monthly	2020	1M	(UkrStat, 2020)
in_communication	CPI: Communication	Monthly	2020	1M	(UkrStat, 2020)
in_culture	CPI: Recreation and culture	Monthly	2020	1M	(UkrStat, 2020)
in_education	CPI: Education	Monthly	2020	1M	(UkrStat, 2020)
in_restaurants	CPI: Restaurants and hotels	Monthly	2020	1M	(UkrStat, 2020)
building	Construction output	Monthly	2012–2021	3M	(UkrStat, 2021a)
salary_2	Wages (monthly)	Monthly	2016	1M	(UkrStat, 2021g)
apple_gt	GT: “Apple”	Weekly	2020–2025	1D	(Trends, 2025)
economy_gt	GT: “Economy”	Weekly	2020–2025	1D	(Trends, 2025)
export_gt	GT: “Export”	Weekly	2020–2025	1D	(Trends, 2025)
gucci_gt	GT: “Gucci”	Weekly	2020–2025	1D	(Trends, 2025)
inflation_gt	GT: “Inflation”	Weekly	2020–2025	1D	(Trends, 2025)
investment_gt	GT: “Investment”	Weekly	2020–2025	1D	(Trends, 2025)
mercedes_gt	GT: “Mercedes”	Weekly	2020–2025	1D	(Trends, 2025)
saving_gt	GT: “Savings”	Weekly	2020–2025	1D	(Trends, 2025)
unemployment_gt	GT: “Unemployment”	Weekly	2020–2025	1D	(Trends, 2025)

Nevertheless, traditional economic indicators alone do not fully capture rapid economic sentiment and consumer behaviour changes, especially under crisis conditions. To address the problem of limited data, we augmented the dataset using Google Trends. We extracted monthly search intensity for terms potentially correlated with economic conditions and household behaviour in Kyiv, such as: *apple*, *economy*, *export*, *Gucci*, *inflation*, *investment*, *Mercedes*, *saving*, and *unemployment*. These keywords were selected to reflect consumption patterns, perceptions of economic conditions, and lifestyle indicators that may have predictive power for regional GDP dynamics. (Figure 3.1)

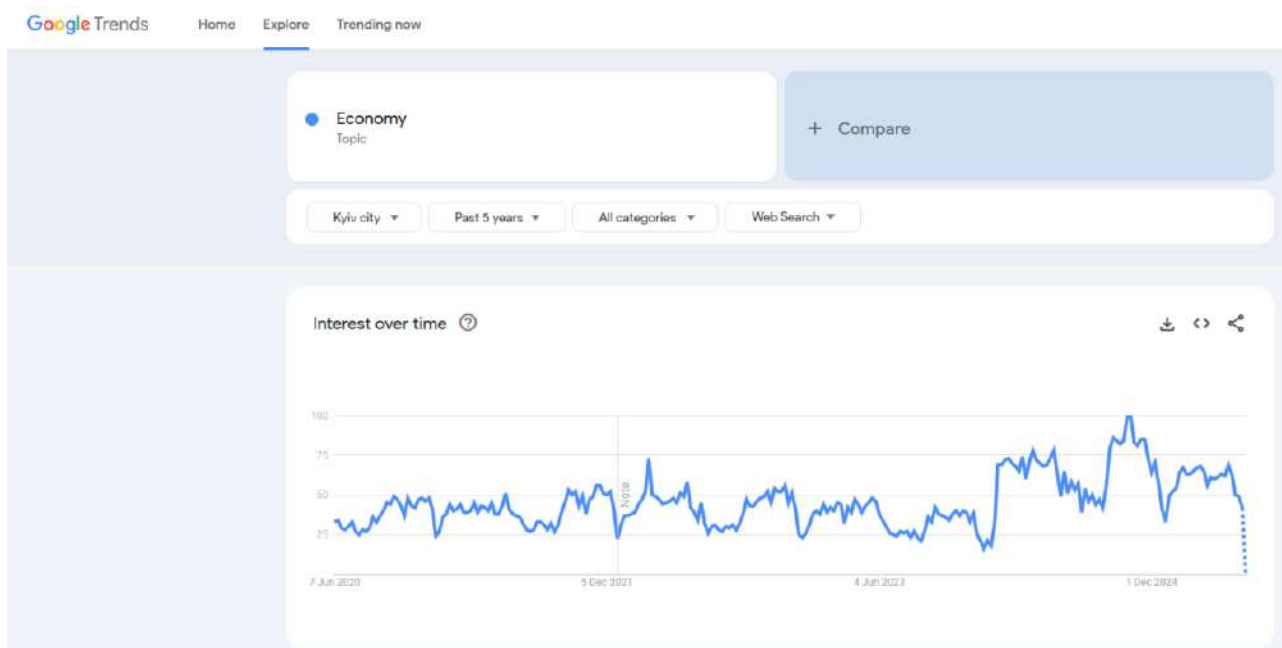


Figure 3.1: Example of the Search topic from Google Trends

This combined dataset—consisting of traditional economic statistics and high-frequency digital indicators—formed the basis for estimating the MF-FAVAR model and evaluating its performance in forecasting Kyiv’s regional gross product (GRP).

3.2 Data preparation and temporal disaggregation procedure

To initiate the construction of a mixed-frequency factor-augmented VAR (MF-FAVAR) model for nowcasting Kyiv’s regional economic activity, we begin by importing the necessary computational packages in R. These include libraries for

data manipulation (`dplyr`, `tidyr`, `zoo`, `lubridate`), time series modelling (`vars`, `tempdisagg`, `bvarsv`), imputation and factor extraction (`missMDA`, `pcaMethods`), and forecast evaluation (`scoringRules`).

As part of preprocessing, we define auxiliary functions for handling missing values, removing nearly constant columns, and implementing factor extraction methods—particularly the Tall Projection (TP) method, adapted from Cahan et al. (2023), and the Two-Window (TW) method. These approaches are designed for extracting latent factors from unbalanced mixed-frequency panels.

Cleaned and pre-aggregated datasets representing different temporal resolutions were then imported:

- Weekly and monthly indicator series relevant to Kyiv’s economy;
- Yearly macroeconomic indicators;
- Annual Gross Regional Product (GRP) data for Kyiv.

Since our target variable is the quarterly GRP growth rate, we disaggregated the annual GRP series to quarterly frequency using the Denton–Cholette method, implemented via the `tempdisagg` package. This method interpolates quarterly values such that the resulting series:

- Sums to the original annual values, and
- Exhibits smooth temporal evolution, minimizing distortions in growth patterns.

Mathematically, the procedure solves:

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

where:

- $y_{t,s}$ are the quarterly values to estimate,
- Y_t are known annual GRP values,
- $x_{t,s}$ is an optional indicator series (not used here),

- Δ^d is the differencing operator (typically $d = 1$),
- λ is a scaling parameter,
- $m = 4$ (quarters per year).

The result is a smooth quarterly GRP series consistent with annual totals and aligned with higher-frequency predictors.

Next, all high-frequency predictors (weekly and monthly indicators) were aggregated quarterly to match the disaggregated target. Quarterly sums were calculated for flow variables (e.g., wages, retail turnover). Quarterly averages were used for stock variables (e.g., price indices and search trends). Annual indicators were propagated across quarters using forward-fill, consistent with standard mixed-frequency practices.

The resulting design matrix and target variable were defined as:

- Y : log-differenced quarterly GRP values;
- X : standardized matrix of aggregated indicators, with missing values imputed via PCA-based methods.

Initially, three methods were applied to extract latent factors from X : Expectation-Maximisation PCA (EMPCA), Tall-Wide PCA (TW), and Tall Projection PCA (TP). However, due to sparse and irregularly observed variables, two additional imputation strategies were added:

- Bayesian PCA (BPCA), which estimates posterior distributions of latent structures;
- Singular Value Decomposition Impute (SVDI), which uses iterative matrix completion via SVD.

Using the extracted factors, MF-FAVAR models of selected lag order p were estimated to forecast quarterly GRP growth. A time-based train-test split was used: approximately 80% of the data for training and 20% for testing. Forecasts from each factor extraction + model specification pair were then evaluated comparatively.

3.3 Forecast evaluation and model comparison

Only three of five factor extraction methods considered produced valid forecasts. The Two-Window (TW) and tall projection (TP) methods failed due to structural limitations in the dataset. Specifically, the TW method could not identify sufficient overlapping windows with complete observations. In contrast, the TP method could not proceed due to insufficient fully observed columns for the initial factor estimation step, as highlighted by internal diagnostic warnings.

Consequently, the evaluation focused on the three robust PCA-based imputation strategies: Expectation-Maximisation PCA (EMPCA), Bayesian PCA (BPCA), and Singular Value Decomposition Imputation (SVDI).

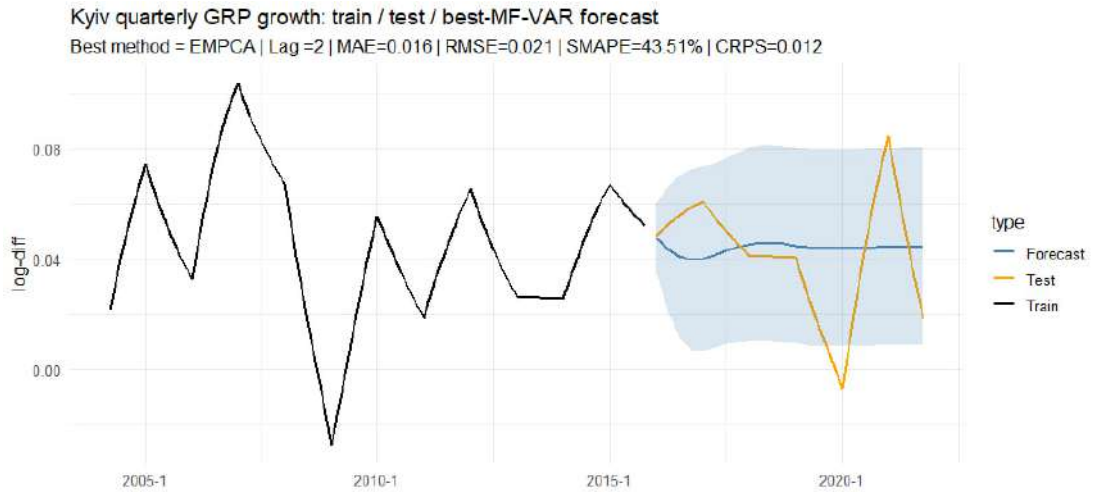
Forecast accuracy under varying horizons

We examined predictive accuracy across multiple forecast horizons to evaluate how model performance responds to structural disruptions such as the COVID-19 pandemic and the onset of war. The results are presented in Table 3.2 and Figure 3.2. Three scenarios were considered:

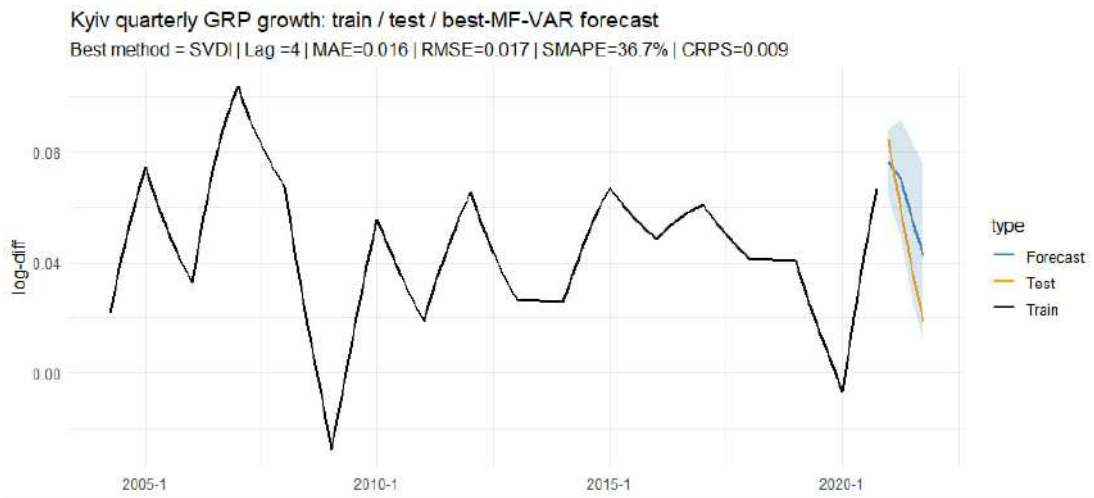
- **Long horizon (24 quarters):** EMPCA slightly outperformed BPCA in terms of RMSE and CRPS, indicating its superior long-term stability.
- **Medium horizon (4 quarters):** SVDI yielded the lowest MAE and CRPS, suggesting higher short- to mid-range responsiveness.
- **Very short horizon (2 quarters):** EMPCA delivered the most accurate forecasts, with minimal error across all metrics, making it ideal for near-term nowcasting.

Table 3.2: Forecast accuracy by horizon

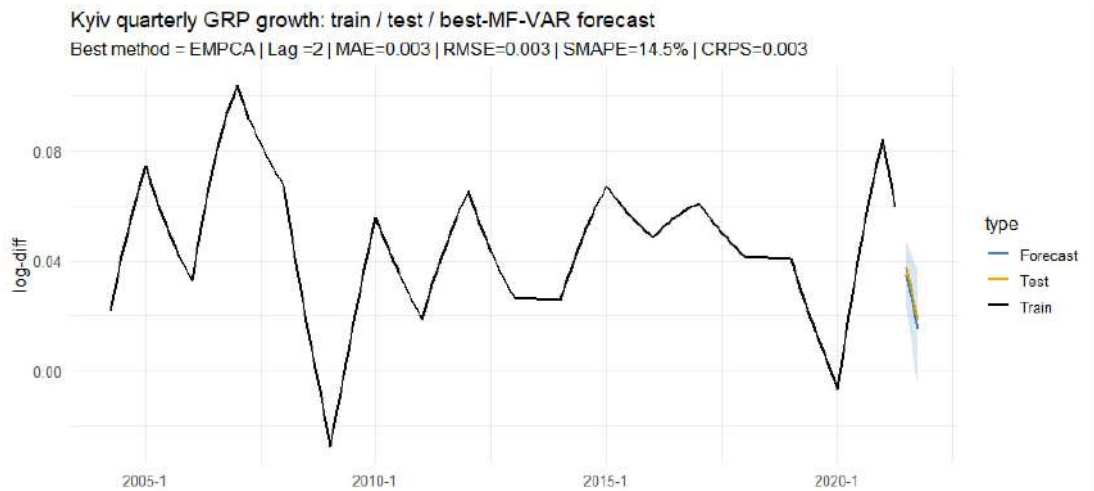
Horizon	Method	Lag	MAE	RMSE	CRPS
24 qtrs	EMPCA	2	0.0158	0.0208	0.0116
	BPCA	2	0.0159	0.0208	0.0117
	SVDI	4	0.0255	0.0319	0.0194
4 qtrs	SVDI	4	0.0155	0.0167	0.0093
	BPCA	2	0.0211	0.0237	0.0134
	EMPCA	2	0.0221	0.0247	0.0141
2 qtrs	EMPCA	2	0.0033	0.0033	0.0034
	BPCA	2	0.0033	0.0034	0.0034
	SVDI	4	0.0053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



(a) Long horizon (24 quarters): EMPCA outperforms in long-run accuracy



(b) Medium horizon (4 quarters): SVDI provides best performance



(c) Very short horizon (2 quarters): EMPCA yields minimal forecast error

Figure 3.2: Forecast performance across three evaluation windows. Each panel compares observed (Train/Test) and forecasted GRP growth along with prediction intervals.

These conclusions indicate that different methods of factor extraction may be optimal at different forecast horizons. EMPCA maintains accuracy in short- and long-range forecasts, while SVDI excels in mid-term periods. Thus, EMPCA seems most appropriate for high-frequency nowcasting and responsive systems in which short-term forecast accuracy is crucial.

3.4 Robustness check with reduced dataset

As a robustness check, we repeated the modelling procedure on a reduced dataset with a more selective choice of predictors. Specifically, all **weekly indicators were excluded** due to their sparse coverage and potential to introduce high-frequency noise. Only the most consistent and economically meaningful variables were retained from the monthly indicators: `consumer_price_index`, `realised_goods`, `building`, and `salary_2`. The annual indicators were reduced to a core set of 12 variables: `income`, `real_income`, `salary_1`, `sochelp`, `expenses`, `buying_goods`, `taxes`, `socpay`, `consume_expenses`, `gross_kapital`, and `metro_passangers`.

Using this streamlined dataset, we re-estimated the MF-FAVAR model with three factor extraction methods: **EMPCA**, **TP**, and **TW**. The Two-Window (TW) method failed again due to insufficient overlap in fully observed blocks, consistent with earlier findings. However, both EMPCA and TP successfully completed and produced forecasts. Their performance is summarised in Table 3.3.

Table 3.3: Forecast accuracy on reduced dataset

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

The EMPCA method once again outperformed the others across all accuracy metrics. Interestingly, the simplified variable selection improved EMPCA’s forecast error compared to the full dataset, indicating that removing noisy or weakly informative predictors can enhance model robustness. The TP method worked, but its forecasting quality was not good, as shown by high RMSE and SMAPE values. This suggests that careful variable selection is important for improving

nowcast accuracy, especially when data is limited.

3.5 Final comparative analysis of factor extraction Methods

Three key methods, EMPCA, TW, and TP, were considered to address the dimensionality reduction task under conditions of incomplete mixed-frequency panel data. Choosing the best method is critically important because it affects the quality of the extracted factors, which impacts the forecasting model’s accuracy.

This section compares these methods based on three key criteria:

- **Algorithmic complexity and computational cost** — the theoretical and practical runtime of the algorithm.
- **Reliability and convergence risk** — the method’s ability to generate results under challenging structural data issues (e.g., missing values, irregular structure).
- **Forecasting accuracy and robustness** — the method’s performance on short and irregular time series, measured using forecast error metrics.

The analysis is based on both the theoretical properties of the algorithms and the empirical results obtained during this study.

EMPCA Method

Computational complexity: EMPCA has higher computational costs than non-iterative alternatives due to its iterative nature. A single iteration has complexity $O(n \cdot k^2)$, and the total cost is $O(T \cdot n \cdot k^2)$, where T is the number of iterations. However, this was not a limiting factor in this study, as the dataset sizes were relatively small.

Reliability: In theory, convergence is not always guaranteed when the amount of missing data is substantial. Nevertheless, testing proved the method reliable: it worked well with both the Kyiv and London datasets and effectively handled complex missing data patterns.

Accuracy: The main advantage of EMPCA in this study was its forecast accuracy. EMPCA achieved the lowest error metrics (RMSE, CRPS) in most test scenarios, particularly for London, and in both long-term (24 quarters) and very short-term (2 quarters) forecasts for Kyiv. This demonstrates its ability to extract informative signals from noisy data.

TW Method

Computational complexity: Being non-iterative, TW is the fastest method, with complexity estimated as $O(n \cdot k^2)$.

Reliability: The main theoretical weakness of TW lies in its rigid structural data requirements. This limitation was confirmed in practice: the method failed on the Kyiv dataset due to the inability to find the required fully observed blocks. Thus, despite its speed advantage, the method proved unsuitable for the task.

Accuracy: Forecast accuracy for Kyiv could not be evaluated, as the method failed to produce results.

TP Method

Computational complexity: This method is also non-iterative and fast, though slightly slower than TW due to the added regression step.

Reliability: Empirical results were mixed. The method failed on the full Kyiv dataset but succeeded on a reduced (cleaned) version. This makes it more robust than TW, but it is still sensitive to the quality of the “supporting” variables used for projection.

Accuracy: Even in the scenario where TP returned results, its accuracy was notably lower than EMPCA (RMSE = 0.0236 vs. RMSE = 0.0198). This questions its usefulness in tasks where forecast precision is critical.

3.6 Summary of Results and Method Selection Justification

The analysis supports a well-grounded conclusion about the suitability of each method in this research context. Table 3.4 summarises the results.

Table 3.4: Comparison of factor extraction methods

Criterion	EMPCA	TW	TP
Computational cost	Moderate. Iterative, but manageable for small datasets.	Low. Fastest method, but unusable due to data gaps.	Low. Slightly slower than TW due to the regression step.
Reliability	High. Robust to missing and irregular data.	Low. Failed due to strict block structure requirement.	Medium. Worked on reduced data only.
Forecast accuracy	High. Best results across all test cases.	N/A (forecast not available).	Moderate. Consistently less accurate than EMPCA.

Summary. While TW and TP offer theoretical speed advantages, their strict structural requirements make them unreliable for short and irregular series, such as those typical in Ukrainian regional statistics during crisis periods.

Despite higher computational costs, EMPCA demonstrated the best reliability and forecast accuracy combination. It was the only method that consistently worked with complex datasets and generated the most accurate predictions. Therefore, this thesis chose EMPCA as the primary method for building the nowcasting models.

Conclusions and future directions

Conclusions

As a result of this qualification thesis, we performed a comparative analysis of key factor extraction methods — EMPCA, TW, TP — tailored for incomplete mixed-frequency panel data. The evaluation included theoretical considerations and empirical testing, with preliminary model development performed on London data and applied thereafter to Ukrainian regional datasets, particularly in Kyiv.

In addition to the three core structural methods, two additional PCA-based imputation techniques — BPCA (Bayesian PCA) and SVDI (SVD Imputation) — were also analysed. Although less computationally intensive than EMPCA, they demonstrated varying levels of robustness and accuracy across different forecast horizons.

The outcomes of this study demonstrated that EMPCA stood out as the only method capable of stably handling structurally incomplete and irregular datasets. Despite higher computational costs due to its iterative EM-based design, EMPCA consistently achieved the lowest forecast error metrics (RMSE, CRPS) across nearly all test scenarios. It showed superior performance for both long-term (24 quarters) and very short-term (2 quarters) prediction tasks, outperforming alternatives even under sparse and noisy conditions.

SVDI demonstrated competitive results in medium-range (4 quarters) settings, suggesting its suitability when prioritising rapid estimation. BPCA performed comparably to EMPCA on longer horizons but was less effective in short-term accuracy. These results — summarised in Table 3.3 — indicate that the optimal forecasting method may vary depending on the temporal scope of prediction. EMPCA, however, proved to be the most consistently accurate and robust approach overall.

Therefore, EMPCA was chosen as the primary method for building the now-casting models developed in this thesis, due to its reliable convergence behaviour, strong predictive performance, and ability to manage missingness in mixed-frequency datasets.

Future Directions

The current research provides a strong foundation for further improvements to nowcasting frameworks using mixed-frequency data. Several directions for future investigation are outlined below:

- **Incorporating national-level indicators:** Expanding the model to include macroeconomic signals such as inflation rates, unemployment figures, and monetary policy variables could increase explanatory power and better capture top-down effects on regional dynamics.
- **Integrating alternative and unconventional data:** Future iterations of this model could benefit from non-traditional sources like social media sentiment (e.g., Twitter), labor market platforms (e.g., Work.ua job postings or employment flows), or survey data from research agencies such as Kantar or Gradus, which provide insights into consumer behavior, business confidence, or social resilience.
- **Exploring mixed-frequency machine learning techniques:** Recent developments in deep learning for mixed-frequency data — including Mixed-Frequency LSTMs and ML-MIDAS models — offer the potential to improve forecast performance by capturing nonlinear interactions and higher-order dynamics. For instance, (Xu et al., 2023) show that neural network-based models outperform traditional methods in several economic prediction tasks involving heterogeneous temporal resolutions.
- **Developing adaptive nowcasting systems:** Another promising line of research involves designing dynamic nowcasting pipelines that update in real time using streaming data and adaptive learning frameworks, allowing models to react to abrupt changes such as crises, policy shocks, or structural breaks in the economy.

These directions could significantly enhance mixed-frequency nowcasting models' scalability, reliability, and accuracy, particularly in regions like Ukraine, where data availability is often limited, fragmented, or irregular due to external disruptions or crisis conditions.

References

- Bai, J. and Ng, S. (2021). Matrix completion with cross-sectional and serial dependence. *Journal of Econometrics*, 222(1):413–430.
- Bañbura, M., Giannone, D., Modugno, M., and Reichlin, L. (2013). Nowcasting and the real-time data flow. Working Paper Series 1564, European Central Bank.
- Bañbura, M., Giannone, D., and Reichlin, L. (2011). Nowcasting. *The Oxford Handbook of Economic Forecasting*.
- Cahan, D., Foerster, A., Sarte, P.-D., and Watson, M. (2023). Nowcasting with ragged-edge data: A projection-based approach. *Federal Reserve Working Paper*.
- Forni, C. and Marcellino, M. (2013). A survey of econometric methods for mixed-frequency data. *Advances in Econometrics*, 31:1–45.
- Fosten, J. and Greenaway-McGrevy, R. (2022). Panel data nowcasting. *Econometric Reviews*, 41(7):675–696.
- Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441.
- Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer Series in Statistics. Springer-Verlag, New York.
- Koop, G., McIntyre, S., Mitchell, J., Poon, A., and Wu, P. (2023). Incorporating short data into large mixed-frequency vars for regional nowcasting. *Federal Reserve Bank of Cleveland Working Paper*, (23-09).
- MinFin (2021). Inflation index. <https://index.minfin.com.ua/ua/economy/index/inflation/kiev/>. Official data.

- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572.
- Schorfheide, F. and Song, D. (2015). Real-time forecasting with a mixed-frequency var. *Journal of Business Economic Statistics*, 33(3):366–380.
- Schorfheide, F. and Song, D. (2020). Real-time forecasting with a (standard) mixed-frequency var during a pandemic. This Version: July 8.
- 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.
- Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.
- Trends, G. (2025). Google trends. <https://trends.google.com>. Accessed 2025-05.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001). Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.
- UkrStat (2016). Realized goods archive. https://ukrstat.gov.ua/operativ/operativ2011/pr/orp_reg/orp_reg_u/arh_orp_reg_u.html. Official data.
- UkrStat (2018). Employment publications. https://www.ukrstat.gov.ua/druk/publicat/kat_u/publ6_u.htm. Official data.
- UkrStat (2020). Cpi categories archive. https://ukrstat.gov.ua/operativ/operativ2020/ct/iscR/iscR_u/arh_iscR20m_u.htm. Official data.
- UkrStat (2021a). Construction statistics. https://www.ukrstat.gov.ua/operativ/operativ2021/bud/ovb_reg/ovb_reg_u/arh_ovb_reg_2021u.htm. Official data.
- UkrStat (2021b). Disposable income archive. https://ukrstat.gov.ua/operativ/operativ2008/gdn/dvn_ric/dvn_ric_u/arh_dn_reg_u.html. Official data.

- UkrStat (2021c). Household expenditures dataset. <https://data.gov.ua/dataset/d83c20f1-c2a1-4fb8-96f7-4ec9788ab41b/resource/a59f294d-2342-436f-8002-bd7dd7dff89c>. Official data.
- UkrStat (2021d). Household income archive. https://ukrstat.gov.ua/druk/publicat/Arhiv_u/03/Arch_dv.htm. Official data.
- UkrStat (2021e). Social expenditures archive. http://ukrstat.gov.ua/druk/publicat/Arhiv_u/01/Arch_zor_zb.htm. Official data.
- UkrStat (2021f). Transport statistics. https://www.ukrstat.gov.ua/operativ/menu/menu_u/tr.htm. Official data.
- UkrStat (2021g). Wages by region. https://ukrstat.gov.ua/operativ/operativ2005/gdn/reg_zp_m/reg_zpm_u/arh_zpm_u.htm. Official data.
- UkrStat (2023). Capital investment statistics. https://ukrstat.gov.ua/operativ/operativ2023/ibd/kin_reg/kin_ved_new_u.htm. Official data.
- Xu, Q., Wang, Z., Jiang, C., and Liu, Y. (2023). Deep learning on mixed frequency data. *Forecasting*, 5(3):406–424. Accessed 2025-06.
- Zhuravlova, A. D. (2025). Nowcasting regional economic indicators of ukraine using short and mixed-frequency data. In *XIII All-Ukrainian Scientific Conference of Young Mathematicians*, Kyiv. NTUU KPI.
- [title=References]

Appendix

ПОВУДОВА МОДЕЛІ ПОТОЧНОГО ПРОГНОЗУВАННЯ ДЛЯ РЕГІОНАЛЬНИХ ЕКОНОМІЧНИХ ПОКАЗНИКІВ УКРАЇНИ З ВИКОРИСТАННЯМ КОРОТКИХ ДАНИХ ТА ЗМІШАНОЇ ЧАСТОТНОСТІ VAR

А.Д. ЖУРАВЛЬОВА

У задачі nowcasting економічної активності ми маємо справу з даними, які надходять з різною частотою — наприклад, щоквартальні індикатори y_t та щомісячні або щотижневі короткі індикатори z_t^r , що доступні лише для окремих регіонів r і лише з певного моменту часу $T_0^r < T$. Типовою проблемою є так званий «рваний край» (ragged edge), коли на момент прогнозування t деякі змінні вже спостерігаються, а інші ще ні, а також ситуація, коли деякі регіональні дані мають дуже коротку історію спостережень.

Ці труднощі виникають у багатьох країнах із затримками в офіційній статистиці та стають особливо критичними при прогнозуванні таких агрегованих показників, зокрема валової доданої вартості (ВДВ), яка є основою для обчислення ВВП. Для обробки такого типу даних використовуються моделі змішаної частотності типу MF-VAR, які можуть комбінувати різночастотні ряди, заповнювати пропущені значення та враховувати інформацію з коротких індикаторів за допомогою побудови латентних факторів.

У даній роботі розглядається модель nowcasting на базі змішано-частотної факторної VAR (MF-FAVAR), що враховує короткі та нерівномірно наявні регіональні економічні дані України. Нехай $t = 1, \dots, T$ — індекс часу, $r = 1, \dots, R$ — регіони, де один з них — місто Київ. Позначимо: Y_t^{UA} — ВДВ України, $y_t^{UA} = \log Y_t^{UA} - \log Y_{t-1}^{UA} - \bar{\pi}$ квартальне зростання. Для регіону r : Y_t^r — регіональна ВДВ, $y_t^r = \log Y_t^r - \log Y_{t-1}^r$.

Річне значення та річне зростання визначаються як:

$$Y_t^{r,A} = Y_t^r + Y_{t-1}^r + Y_{t-2}^r + Y_{t-3}^r, \quad y_t^{r,A} = \log Y_t^{r,A} - \log Y_{t-4}^{r,A}.$$

Обмеження:

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

$$y_t^{UA} \approx \frac{1}{R} \sum_{r=1}^R y_t^r.$$

VAR-модель:

$$By_t = Ax_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \Sigma)$$

де x_t — лаги y_t , B — нижньотрикутна матриця. У MF-FAVAR розширення: $y_t = (y_t^{UA}, y_t^Q, f_t)$, де f_t — вектор регіональних факторів із коротких індикаторів Z_t^r .

Методи побудови факторів:

- **EMPCA** — ітеративний метод з PCA та EM для заповнення пропусків;
- **TW (Tall-Wide)** — неітеративний метод, що показав найкращу якість прогнозу;
- **TP (Tall-Project)** — регресійне заповнення пропусків на основі <tall блоку.

Для регуляризації використовуються Adaptive Lasso. Для порівняння також була оцінена модель BVAR:

$$y_t = \Phi_1 y_{t-1} + \dots + \Phi_p y_{t-p} + \varepsilon_t,$$

що показала найкращу точність за метрикою MSCE серед усіх моделей.

Запропонована модель може бути використана для оперативного планування на регіональному рівні, що дозволить уряду та бізнесу приймати швидкі та обґрунтовані рішення. Також модель може бути адаптована для аналізу економічної активності в інших країнах із подібними обмеженнями у доступності даних.

ЛІТЕРАТУРА

- [1] Gary Koop, Stuart McIntyre, James Mitchell, Aubrey Poon, and Ping Wu. *Incorporating Short Data into Large Mixed-Frequency VARs for Regional Nowcasting* // Federal Reserve Bank of Cleveland Working Paper Series — May 2023
- [2] Jack Fosten, Ryan Greenaway-McGrevy. *Panel data nowcasting*. // *Econometric Reviews*, 41:7, 675-696 — 2022.
- [3] Mihnea Constantinescu, Kalle Kappne, Nikodem Szumilo. *The Warcast Index: Estimating Economic Activity without Official Data during the Ukraine War in 2022* // NBU Working Papers — March 2024

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ», Київ, УКРАЇНА
 Email address: a.zhuravlova@ukma.edu.ua