

Do Time Series Foundation Models Outperform Traditional Macroeconomic Forecasting Methods?

A moderate-scale evaluation of foundation models against classical statistical benchmarks on 2,734 Bank of Israel macroeconomic series.

Ariel Mantzura & Guy Itzhaki — Bank of Israel

Core Questions

- 01 Can pretrained foundation models (TSFMs) outperform traditional statistical models — SARIMA and ETS — on macroeconomic data?
- 02 Does fine-tuning a TSFM on the target dataset improve accuracy over zero-shot forecasting?
- 03 Does clustering series by behavioral characteristics before fine-tuning yield further gains?

2,734

series evaluated

42

model instances

9

model families

1–12

month horizons

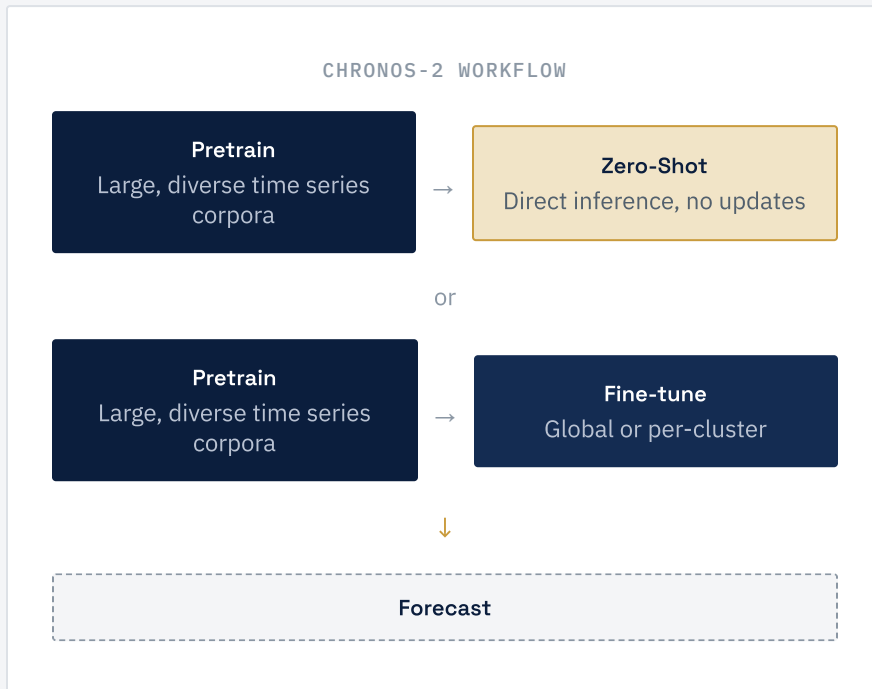
Time Series Foundation Models & Chronos-2

TSFMs learn general temporal patterns from large, diverse pretraining corpora, enabling zero-shot forecasting without task-specific training.

What is a TSFM?

Pretrained models that learn general temporal patterns from large and diverse time series collections. Once pretrained, they forecast previously unseen datasets with little or no additional training (zero-shot), and can optionally be fine-tuned on a specific dataset to improve accuracy.

Benchmarked alongside three other TSFMs — **Toto**, **TimesFM**, and **TiRex** — plus classical SARIMA and ETS baselines.



The Bank of Israel Time Series Database

2,700 monthly series screened from an initial pool of 5,000, spanning the principal areas of Israeli macroeconomic statistics since January 2000.

5,000 → **2,700**

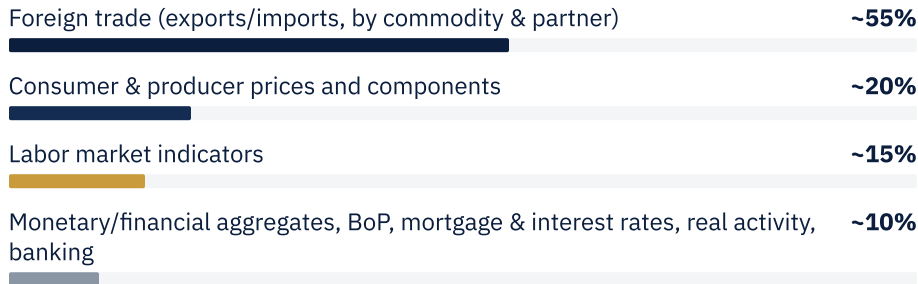
series after screening & filtering

Jan 2000 →

monthly frequency, raw levels

Series are highly heterogeneous — structural breaks, multi-frequency effects, seasonality, and policy-driven regime shifts.

Composition by Domain



Clustering Methodology

A Bayesian Gaussian Mixture model with a Dirichlet Process prior identified three stable, reproducible clusters of monthly economic series.

Behavioral feature vector

Trend strength

Seasonality

Autocorrelation

Entropy

Nonlinearity

Spikiness

Lumpiness

Skewness

Kurtosis

Cluster count selected via silhouette score, together with stability and cluster-size criteria. K-means and hierarchical Ward clustering independently converge on the same three-cluster partition — the clusters align strongly with metadata on measurement unit and data type, indicating the features capture meaningful structural properties.

3

stable clusters identified

0.157

silhouette score

0.798

cross-seed ARI
(reproducibility)

0.67

Cramér's V vs. metadata

PC1 + PC2 variance explained

52.5%

Evaluation Framework: Global, Cluster, and Zero-Shot Models

Cross-learning evaluation compares globally trained, cluster-specific, and zero-shot model variants across forecast horizons of 1–12 months.

MODEL TYPE

G — Global

Fine-tuned on the complete, heterogeneous dataset — a single model trained across all 2,734 series.

MODEL TYPE

Cluster

Fine-tuned independently within each of the three BGM-identified behavioral clusters.

MODEL TYPE

ZS — Zero-Shot

The pretrained model applied directly to unseen series, with no parameter updates.

Accuracy Metric

Hyndman MASE, calculated from the median forecast and scaled by the in-sample seasonal-naïve error.

2,734
series

42
model instances

9
model families

1–12
month horizons

Forecasting Performance by Horizon

Median forecast error — MASE by horizon (lower is better)

$$\text{MASE} = \frac{|Y - \hat{Y}|}{\text{mean seasonal-naive forecast error}}$$

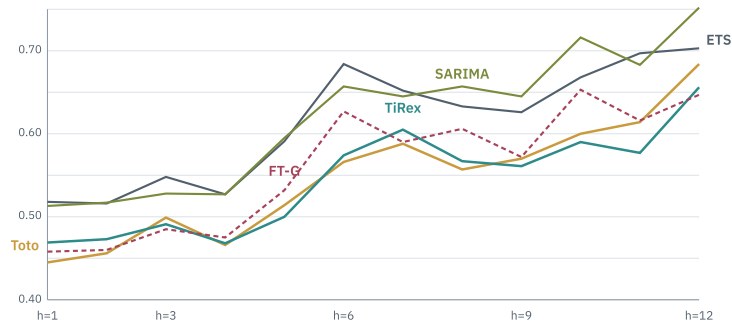
Model	h=1	h=3	h=6	h=12
toto2_zeroshot	0.4454	0.4993	0.5659	0.6845
zs-chronos2-nocl	0.4498	0.4954	0.6129	0.6537
chronos2-ft-g	0.4585	0.4845	0.6266	0.6471
chronos2-zs-cl-bgm	0.4601	0.4964	0.6070	0.6428
chronos2-ft-cluster-bgm	0.4618	0.4910	0.6158	0.6346
tirex_zeroshot	0.4692	0.4912	0.5738	0.6561
ets_auto	0.5177	0.5484	0.6838	0.7032
arima_auto	0.5129	0.5276	0.6569	0.7519
tft-g	0.5307	0.5492	0.6896	0.7783
dlinear-g	0.6370	0.9633	0.8677	0.8779
seasonal_naive	0.9172	0.8163	0.8016	0.9059

■ Best MASE in column · MASE = Hyndman scaled error, median forecast, seasonal-naive scaled — lower is better

Foundation models have lower error than classical baselines at every horizon. Lower MASE is better.

Key Findings Across Horizons

The best-performing model shifts from zero-shot foundation models at short horizons to cluster fine-tuned Chronos-2 at long horizons.



- Toto (zero-shot) — peaks early
- TiRex (zero-shot) — most consistent
- FT-G (fine-tuned) — tracks zero-shot
- ETS (classical) — worse at every horizon
- SARIMA (classical) — worst at long horizons

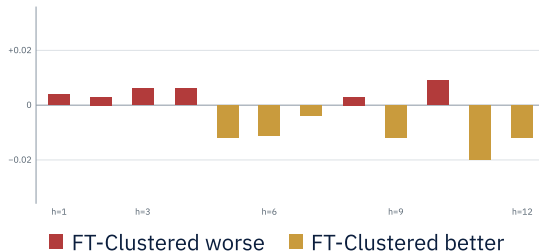
Does Clustering Help to Improve Forecast Results?

Chronos-2 **Fine-Tuned Global** (fine-tuned on all 2,734 series)

vs.

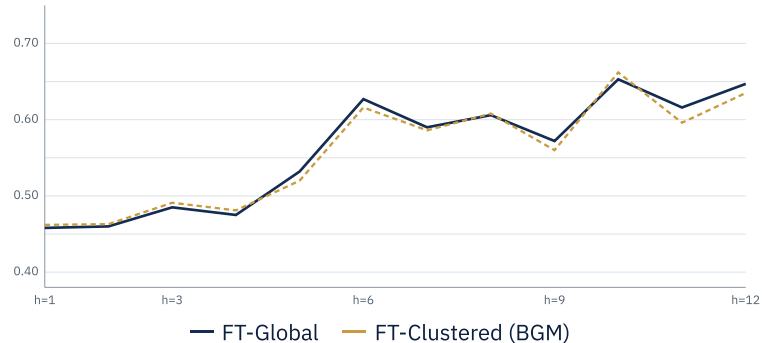
Chronos-2 **Fine-Tuned Clustered** (BGM, dividing 2,734 series into 3 clusters)

Difference (FT-Clustered – FT-Global)



Bars hover close to zero at every horizon, with a small negative gap re-emerging at h=12 — no consistent, reliable gain from cluster fine-tuning over the global model.

MASE by horizon — FT-Global vs. FT-Clustered



Is cluster fine-tuning worth it overall?

Cluster-conditional fine-tuning gives essentially no gain over a single global fine-tune on average (0.552 vs. 0.554). A small ~2% edge appears only at the 12-month horizon. In practice, a single model — often even zero-shot — is competitive everywhere.

Conclusions

- ▶ Foundation models — lower error than classical methods at every horizon.
- ▶ Fine-tuning barely helps — zero-shot and fine-tuned foundation models give similar results.
- ▶ Clustering does not help — only a small edge at the longer horizons (~2% at $h=12$).
- ▶ From-scratch DL models (TFT, DLinear) — clustering beat their own global version, but they remain far behind the foundation models.

0.538

best single model — tirex_zeroshot median MASE

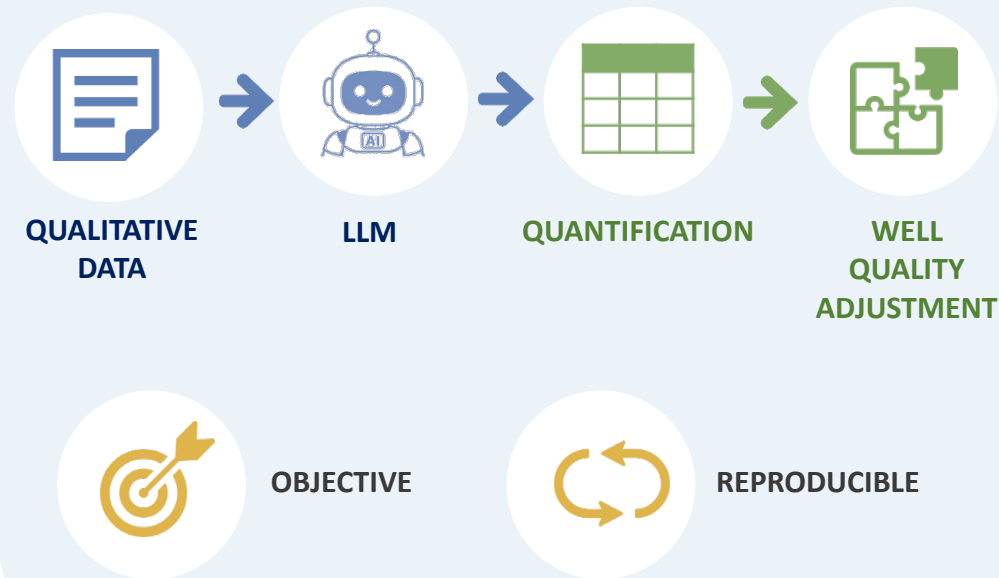
Central banks takeaway: strong forecasting accuracy is available off-the-shelf, zero-shot — no per-series tuning or clustering needed.

Quantifying Qualitative Information for Quality Adjustment Using LLMs

Reducing the Black-box Nature: A Case Study of Rice Cookers

Ayako Masujima
the Bank of Japan

Presented at the 13th biennial IFC conference
BIS, Basel 20-21 August 2026



Motivation and Our Approach

1. Problem



2. Our Solution



3. Benefits

Quality adjustment in price index compilation often faces the following challenges:



Many product features are described only in text or image (**qualitative** information)



Constructing **quantitative** explanatory variables from qualitative information is difficult

Manual evaluation is time-consuming and may lack objectivity.

We explored the use of **LLMs** to **quantify qualitative information**

To reduce the black-box nature of LLM-based analysis, we propose the following workflow.

A Two-Phase LLM Workflow

We separate the establishment of **evaluation criteria** from their **application** and use **LLMs** in **both phases**

Quantification
Criteria



Explanatory Variable
Data Construction

More refined quality adjustment

- Quantified qualitative information
- Improved quality adjustment using hedonic regression

Two-phase workflow ensures



Objectivity

Consistent criteria applied to all products



Reproducibility

Repeatable process

Two-Phase LLM Workflow

Phase 1: Quantification Criteria

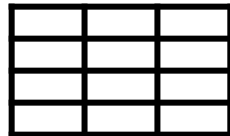
Extract candidate explanatory variables
representing quality characteristics



Establish quantification criteria
for each variable



Phase 2 : Explanatory Variable Data Construction



Evaluate all products consistently
and **construct the dataset**



- We use LLMs in both phases
- Our approach enables a systematic and reproducible dataset for hedonic quality adjustment

Case Study: Rice Cookers

Rice Cookers are a widely used appliance in Japan, and many of their features are qualitative. We examined how to quantify these features.

1 Why rice cookers?



White rice is the staple food in Japan.



Product quality is characterized by many qualitative features.

These features are often described in text and images, making quantitative evaluation difficult

2 Data Sources

Web-scraped data



Product information from e-commerce sites (65 products)

Official websites



Detailed specifications, feature descriptions and images from manufacturers



LLM

Information extraction and interpretation



Establish Criteria



Construct Explanatory Variable data

Transforming qualitative information into quantitative data

3 Example of Explanatory Variables and Criteria

Candidate variables



Cooking method
e.g. pressure IH, IH, microcomputer-heating technology



Sophistication of Pressure Control
Adaptive pressure control for rice type and room temperature



Texture-based cooking options
Availability to select rice textures (e.g. firm, regular, or soft)



Rice variety-specific cooking function
Availability of rice variety-specific cooking programs



User interface
e.g. monochrome or color display, button or touch panel

Example criteria

Sophistication of Pressure Control

0 Low

No pressure control

1 Medium

No keywords indicating advanced pressure control

2 High

Keywords indicating advanced pressure control (e.g. multi-stage pressure)



In this case study, we examined how well the LLM could: extract explanatory variables , establish quantification criteria and construct explanatory variable data

Results I : Explanatory Variable Dataset

In Phase1, we examined the relationship between the explanatory variables proposed by the LLM and product prices. In Phase 2, we confirmed the accuracy of the LLM response.

Explanatory Variables	Relationship with Price	Accuracy	Source	Why Did Accuracy Vary?
Cooking Method	◎	100%	Web-scraped data	 Easy to obtain and interpret ● Single source ● Simple interpretation Example: Cooking method 
Inner Pot Performance	○	100%		 Multiple sources ● Multiple webpages ● Interpretation required Example: Sophistication of pressure control
Power Consumption	△	100%		
Keep-warm Performance	△	100%		
Ease of Maintenance	○	100%		Manufacturers' official websites
Sophistication of Pressure Control	◎	83%	 Inference from unclear information ● Prior knowledge used Example: User Interface	
Texture-based Cooking Options	◎	63%		
Rice Variety-Specific Cooking Function	○	98%		
Learning Function	◎	98%	 Low prevalence ● High agreement by chance Example: Rice Variety-Specific Cooking Function	
User Interface	◎	86%		

◎: Strong, ○: Moderate, △: Weak relationship

Agreement rate is defined as the proportion of products for which the manual evaluation and the LLM-based evaluation produced the same result.

◎: Strong, ○: Moderate, △: Weak relationship

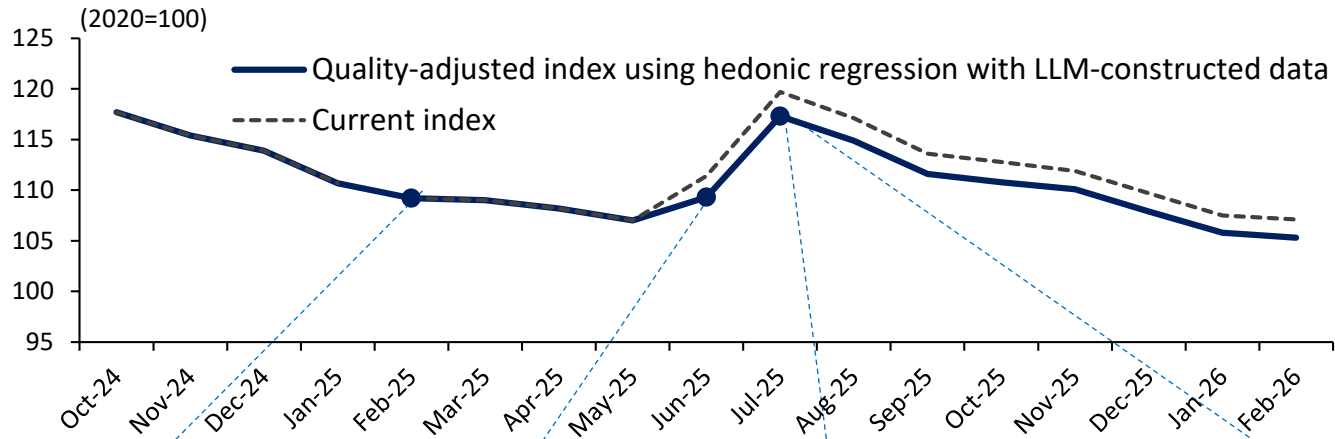
Agreement rate is defined as the proportion of products for which the manual evaluation and the LLM-based evaluation produced the same result.



- Most explanatory variables showed some relationship with price.
- Accuracy varies depending on how easy it is to obtain and interpret the information.

Results II : Impact on the Price Index

The LLM-constructed dataset is used for hedonic regression to produce the Producer Price Index item “Electronic Rice Cookers.”



Feb.2025

Feature Changes

- Lower power consumption
- Higher power consumption during keep-warm operation

Quality Changes

Hedonic	±0%
Conventional	±0%

Jun.2025

Feature Changes

- Learning Function

Quality Changes

Hedonic	+15.4%
Conventional	+4.4%

Jul.2025

Feature Changes

- Added firepower boost
- from monochrome touch panel → color touch panel
- Higher inner pot performance

Quality Changes

Hedonic	+5.4%
Conventional	+3.4%

Jul.2025

Feature Changes

- No substantial performance improvements

Quality Changes

Hedonic	±0%
Conventional	±0%



The LLM-constructed dataset makes it possible to reflect quality changes associated with model changes that could not be adequately captured in our conventional method.

Implications and Future Work

Implications

1



The relationship between explanatory variables and price showed a general correlation.

This suggests that both the extraction of explanatory variables by the LLM and the establishment of evaluation criteria worked well.

2



In dataset construction, accuracy varies depending on the difficulty of obtaining and interpreting the information.

Both accuracy and efficiency can be improved by combining confirmation by human evaluators with additional information collection.

3



Applying the hedonic method to the constructed dataset allowed us to capture quality changes more precisely than the current index.

This indicates that the proposed method can improve the accuracy of quality adjustment.

Future Work

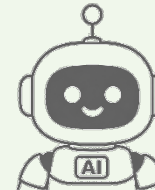
1



Validate the approach for other products.

This study was conducted only on rice cookers, so we plan to examine the validity of the proposed method for other products as well.

2



Examine differences across LLM models.

Since the evaluation was conducted using a single LLM, we plan to investigate how results vary across different models.

Conclusion

Key Takeaways



LLMs successfully **quantified qualitative information**.



The proposed **two-phase LLM workflow** enabled the creation of **objective** and **reproducible** explanatory variable dataset.



The resulting dataset **improved quality adjustment** through hedonic regression.

Overall Contribution



This study proposes a practical framework for **objective and reproducible LLM-based dataset construction**.



The proposed workflow has the potential to support dataset construction tasks **beyond quality adjustment in price statistics**.



NATIONAL BANK OF THE REPUBLIC OF NORTH MACEDONIA

Nowcasting Macedonian GDP with ChatGPT- Derived Economic Activity Score: Evidence from Machine Learning

Danica Unevska Andonova, Magdalena Petrovska,
Nedjati Kurtishi and Elena Bozhinovska Stojchevska

13th biennial IFC Conference

“Filling information gaps in central banking: effectively leveraging data and
innovative tools”

Basel, 20-21 August 2026

Motivation and goals

- ◆ In a small open economies, traditional indicators—like industrial production or retail sales—are essential, but they are often lagged or too narrow to capture the full picture of economic momentum.
- ◆ Central banks have long relied on qualitative surveys to fill these gaps, but until now, the narrative comments from analysts were treated as 'anecdotal'—useful for a speech, but too unstructured for a formal modelling.
- ◆ Integration of qualitative economic sentiment into macroeconomic forecasting using large language models (LLMs)
- ◆ Development of Economic Activity Score (EAS), as explanatory variable, from open-ended responses in the NBRNM's quarterly Inflation and GDP Expectations Survey by exploiting ChatGPT, transforming subjective text into objective, high-frequency metrics
- ◆ Application of Machine-learning (ML) models for GDP nowcasting which will enrich the forecast framework with frontier methods and techniques in forecasting
- ◆ Machine-learning methods seek to capture co-movements between high-frequency indicators and the target variables

Data

- ◆ Data span: Q1 2015 to Q1 2025 for Model 1 – Benchmark Model
Q1 2018 to Q1 2025 for Model 2 - Sentiment-Augmented Model
- ◆ Dependent variable: $\text{Dlog}(\text{Real GDP_SA})$
- ◆ Set of explanatory variables:
 - $\text{Dlog}(\text{Real GDP_SA})$ - up to 3 lags
 - $\text{Dlog}(\text{Foreign demand})$ - contemporaneous value and 1 lag
 - $\text{Dlog}(\text{Petroleum available for final consumption})$ - contemporaneous value and 1 lag
 - crisis dummy: 2017 Q2 (the domestic political crisis), 2020 Q2 (COVID-19 lockdown), 2021 Q2 (renewed pandemic restrictions)
 - $\text{D}(\text{Economic Activity Score})$ - contemporaneous value and 2 lags

From Words to Numbers (The ChatGPT EAS)

- ◆ We exploit the ability of Large Language Model (LLM) to understand **context, tone, and nuance within analyst narratives**
- ◆ We developed the **Economic Activity Score (EAS)** on a scale of -4 to +4 from open-ended responses in the NBRNM's Inflation and GDP Expectations Survey by exploiting ChatGPT
- ◆ It is a structured mapping of human expertise into a machine-readable format that capture 'latent' information that hasn't shown up in the hard data yet
- ◆ We utilize GPT-4, that allows the model to assign mathematical weights to different parts of a sentence
- ◆ By evaluating five distinct pillars—GDP momentum, risk balance, external demand, investment/consumption expectations, and labor market and inflation expectations spillovers to growth—we ensure that the score is not a 'vibe' (subjective impression), but a structured multi-dimensional assessment

The LLM as a Semantic Feature Extractor

- ♦ The AI uses a **Distance-to-Concept** logic. It measures how "close" the analyst's words are to some predefined anchor concepts: "Critical" (-4), "Neutral" (0), or "Peak" (+4) state.

Example:

- ♦ **GDP Outlook:** +2
- ♦ **Risks:** -1 (A bit of worry)
- ♦ **External environment:** +1 (EU is okay)
- ♦ **Investment/Consumption expectations:** +2 (Factories are buying machines)
- ♦ **Labor and inflation expectations spillover to growth:** 0 (Not mentioned/Neutral)
- ♦ **The Geometric Centroid:** By averaging these 5 coordinates (+2, -1, +1, +2, 0), the model identifies the "Center of Gravity". The Result is: $(+2-1+1+2+0)/5=0.8$
- ♦ **The raw sentiment:** +0.8
- ♦ **Ordinal EAS:** +1 – we apply a rounding rule to the "raw sentiment"

The Choice of Methodology: Why the Elastic Net?

- ♦ Frontier methods and techniques in forecasting, performing relatively well when capturing sharp turning points in GDP growth
- ♦ Small number of observations (quarterly data) and a relatively large number of potential predictors (lags of GDP, foreign demand, petroleum use, and several lags of the EAS)
- ♦ Standard Ordinary Least Squares (OLS) issues: Overfitting and possible Multi-collinearity, where our predictors are highly correlated
- ♦ Machine learning prevents this "fighting" between variables.
- ♦ We employ the **Elastic Net** (combination of Ridge and LASSO regressions), a regularized regression that minimizes the residual sum of squares plus a hybrid penalty term to prevent the size of the coefficients not to become too large:

$$\beta_{elastic\ net} = \operatorname{argmin} \left[\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij} \beta_j)^2 + Penalty \right]$$

$$Penalty(\beta_{elastic\ net}) = \lambda \left(\sum_{j=1}^p (1 - \alpha) \beta_j^2 + (\alpha) |\beta_j| \right)$$

n – number of observations, p – number of explanatory variables,
 λ – penalty hyperparameter, If $\lambda=0 \Rightarrow$ OLS estimation
 α - hyperparameter

Results – Forecast accuracy

- ♦ Model 2 has smaller RMSE compared to the benchmark Model 1, suggesting a slight improvement in one-step-ahead GDP forecast accuracy
- ♦ Both tests confirm that Model 2 is not statistically superior over benchmark model
- ♦ Model 2 tracks actual GDP annual rates more closely during both slower-growth periods (below 2%) and stronger expansions (above 3%), underscoring the value of incorporating changes in economic sentiment into short-term forecasting

Forecast Performance of Benchmark and Sentiment-Augmented Elastic Net Models

Model	RMSE
Model 1 (Benchmark Elastic Net)	0.686
Model 2 (Sentiment-Augmented Elastic Net)	0.615

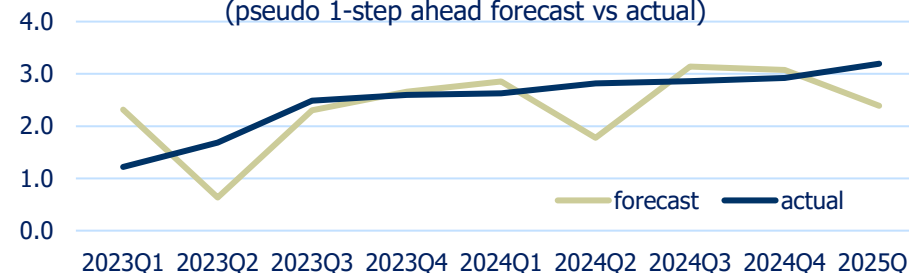
Note: Forecasts are generated using an expanding window approach. The evaluation covers the period from 2023 Q1 to 2025 Q1. Lower RMSE values indicate better predictive performance.

Statistical Comparison of Forecast Accuracy: DM and HLN Tests

Test	Statistic	p-value	Interpretation
Diebold–Mariano (DM)	0.41	0.69	Difference in forecast accuracy not statistically significant; cannot reject null of equal predictive accuracy.
Harvey–Leybourne–Newbold (HLN)	0.39	0.71	Small-sample correction confirms DM result; no evidence that either model is superior.

**Elastic net, benchmark model
GDP annual growth rates**

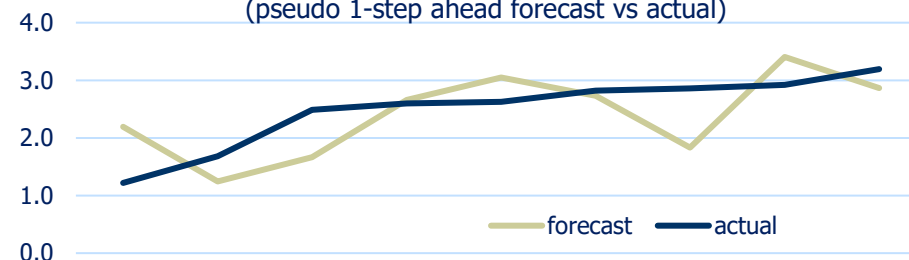
(pseudo 1-step ahead forecast vs actual)



Source: SSO and authors' calculations.

**Elastic net, model with EAS
GDP annual growth rates**

(pseudo 1-step ahead forecast vs actual)



Source: SSO and authors' calculations.

Conclusion

- ◆ We applied machine learning models for GDP nowcasting
- ◆ We demonstrated a novel approach in extracting macroeconomic sentiment from qualitative survey inputs using a standardized prompt with ChatGPT
- ◆ By leveraging Large Language Models, we have developed a systematic Economic Activity Score (EAS).
- ◆ This score was integrated into an Elastic net forecasting model for real GDP growth.
- ◆ Inclusion of the EAS resulted in a modest reduction in Root Mean Square Error (RMSE) compared to the baseline model
- ◆ While the improvement is not yet statistically significant, the study highlights the potential of combining structured LLM indicators with regularized regression to enhance real-time macroeconomic analysis in small open economies

Thank you for your attention!



Nighttime Satellite Radiance from Indonesia's Major Industrial Complexes

As a Proxy for Manufacturing Sector GDP Performance

Hanif Galih Pratama | Vita Rosiana Dewi

Statistics Department – Bank Indonesia

The views expressed in this presentation are those of the authors and do not necessarily reflect those of Bank Indonesia



Background & Motivation

The Challenge

Measuring economic activity in a timely and granular manner remains a key. While official GDP statistics are compiled with rigorous standards and released within a well-established timeline, their quarterly frequency limit the ability to capture high-frequency dynamics and heterogeneity.

The Opportunity

Satellite-based night-time light data (NTL) offer a promising complementary source of information. These data provide high-frequency, spatially consistent observations of human and industrial activity, independent of survey and administrative processes.

The Gap

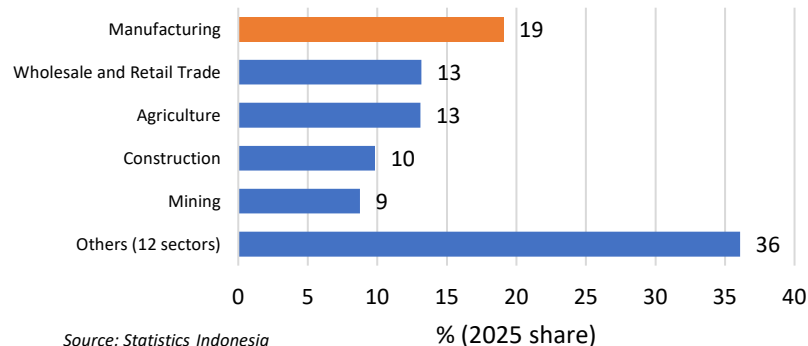
The application of night-time light data in Indonesia remains relatively limited. Existing studies tend to focus on aggregate or regional analysis, with less attention given to sector-specific dynamics such as manufacturing.

Moreover, there is limited work that links high-frequency satellite observations to official macroeconomic indicators within a consistent empirical framework.

Research Objective

Develop a nighttime light-based indicator from industrial complexes to track Indonesia's manufacturing GDP performance in near real-time.

INDONESIAN GDP BY SECTOR



Source: thetruesize.com

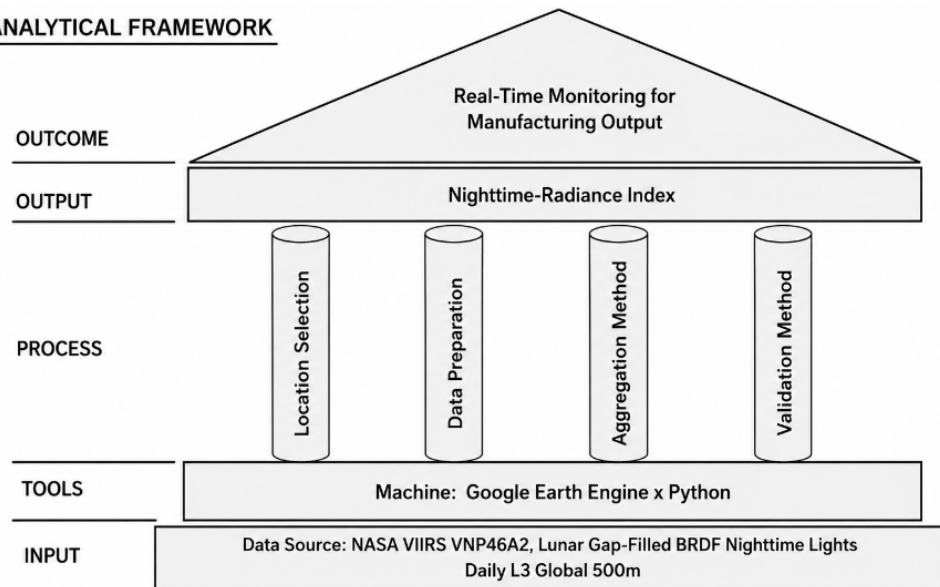


Framework

This study proposes a **replicable conceptual framework** to utilize nighttime light data through four main stages:

- Location selection of major industrial complexes
- Data preparation and cleaning;
- Index construction to national level aggregate;
- Statistical validation against official manufacturing GDP.

ANALYTICAL FRAMEWORK



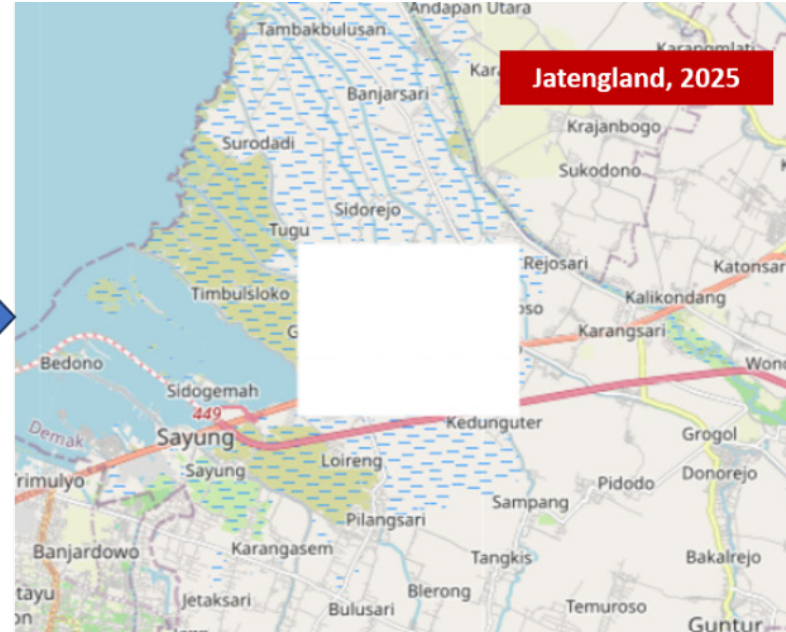
Selected Industrial Complexes

No	Industrial Complex	Lon	Lat
1	Pulogadung	106.91	-6.20
2	Karawang	107.27	-6.39
3	Jababeka	107.13	-6.28
4	Cikande	106.32	-6.19
5	Gresik	112.63	-7.11
6	Rungkut	112.76	-7.33
7	Sirie	112.75	-7.46
8	Ngoro	112.61	-7.56
9	Wijayakusuma	110.33	-6.97
10	Jatengland	110.55	-6.92

Selection criteria: spatial footprint > 500m and established long enough for consistent time series.

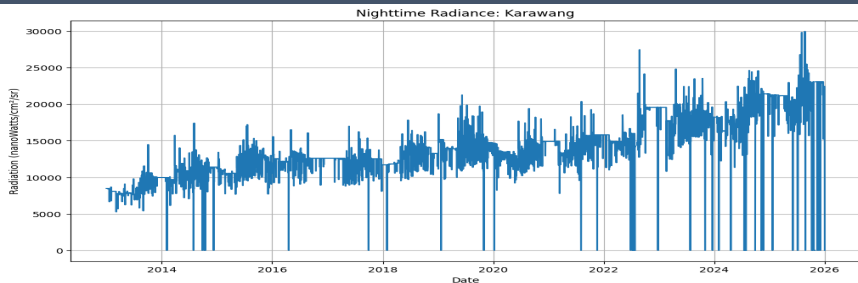
Nighttime Radiance — Satellite Imagery

Jatengland Industrial Park within 10 years of comparison, demonstrating satellite capability to capture industrial growth.

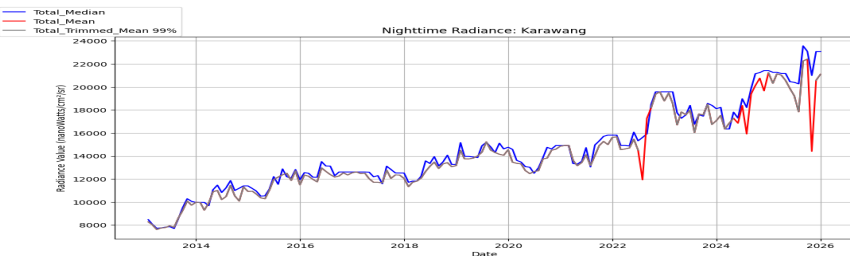




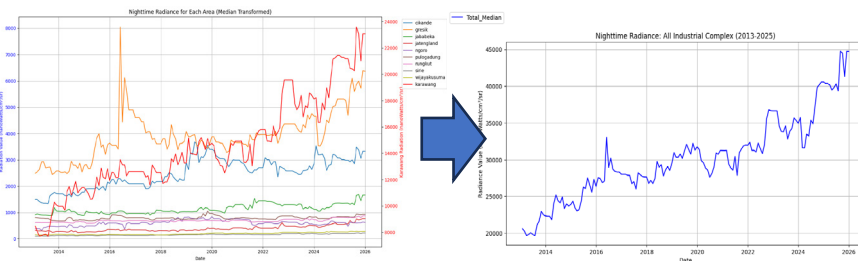
DAILY RADIANCE DATA (RAW)



MONTHLY AGGREGATED



ALL INDUSTRIAL COMPLEX (MONTHLY AGGREGATED)



Data Processing

Data Extraction (Daily)

The extraction process aggregates pixel-level radiance values within each predefined rectangular boundary.

$$NTL_{i,d}^{raw} = \sum_{p \in i} Radiance_{p,d}$$

Data Preparation

To address data gaps and reduce short-term noise, daily observations were aggregated into monthly frequency using arithmetic mean, the median, and the 95% trimmed mean -> median aggregation produce smoother data

$$NTL_{i,m}^{raw} = median_{d \in m}(NTL_{i,d}^{raw})$$

Data Aggregation

The monthly nighttime radiance values are subsequently aggregated across all selected industrial complexes using a sum-based approach to construct a national-level nighttime radiance measure.

$$NTL_m^{agg} = \sum_{i=m}^N NTL_{i,m}^{raw}$$

The aggregated monthly nighttime radiance series is transformed into an index using January 2013 as the base period

$$NRI_{it} = 100 \times \frac{NTL_{i,t}}{NTL_{i,0}}$$

Result and Validation



- The relationship between the Nighttime Radiance Index (NRI) and manufacturing GDP growth is examined using baseline and dynamic regression models, controlling for PMI and manufacturing electricity consumption.
- NRI is positively and significantly associated with manufacturing GDP growth in both models, even after accounting for conventional indicators and output persistence (Table 2).
- Including NRI improves the dynamic model's performance — higher Adjusted R² and lower RMSE & MAE (Table 3).
- The positive relationship remains robust across alternative aggregation methods (median, mean, trimmed mean) and different sample periods (Table 4).

Table 2. Summary of Regression Results

Variable / Statistic	Baseline Model	Dynamic Model
NRI	0.065** (0.0288)	0.063*** (0.0083)
PMI	0.349*** (0.0477)	0.398*** (0.0307)
Manufacturing Electricity	0.938 (0.6081)	0.192* (0.1041)
GDP Manufacturing (t-1)	— —	0.539*** (0.0534)
Adjusted R²	0.639	0.839
Residual Std. Error	—	—
Observations	32	31

Table 3. Model Comparison: Does NRI Improve the Dynamic Model?

Indicator	Dynamic without NRI	Dynamic with NRI
Adjusted R ²	0.830	0.839
AIC	106.97	106.09
BIC	114.14	114.69
RMSE (In-sample)	1.098	0.959
MAE (In-sample)	0.807	0.691
Newey–West significance of NRI	—	p < 0.001

**) estimation results based on Newey–West heteroskedasticity and autocorrelation consistent (HAC) standard errors*

Table 4. Robustness Check

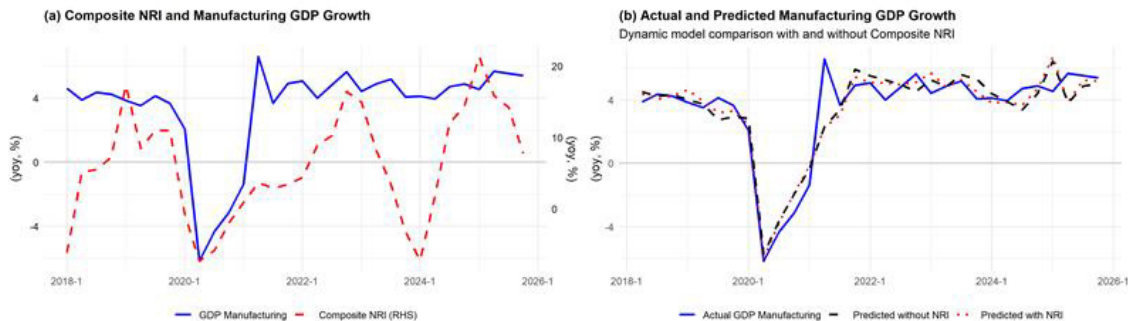
Monthly Aggregated by:	CORRELATION WITH GDP INDUSTRY SECTOR GROWTH			
	All Period	All Period Excl. Covid	Pre-Covid	Post-Covid
Median	0.45	0.36	0.29	0.52
Mean	0.40	0.30	0.27	0.43
Trimmed Mean	0.40	0.31	0.27	0.39

Chart Comparison

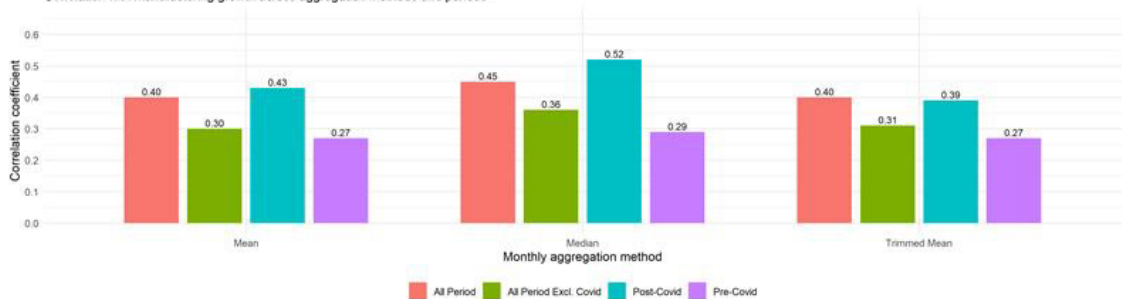


- Panel (a) shows that the Composite NRI broadly tracks the movement of manufacturing GDP growth over time, supporting its role as a complementary monitoring indicator.
- Panel (b) indicates that the dynamic model including NRI provides a closer fit to actual manufacturing GDP growth in several periods, consistent with the improvement in RMSE and MAE.
- Panel (c) confirms the positive correlation between NRI and manufacturing growth remains robust across alternative aggregation methods and sample periods
- Overall, the visual evidence supports the interpretation that NRI captures useful information on manufacturing activity

Visual Assessment of the Composite Nighttime Radiance Indicator (NRI)



(c) Robustness Check of the Composite NRI
Correlation with manufacturing growth across aggregation methods and periods



Notes: NRI = Nighttime Radiance Indicator. Panel (a) compares manufacturing GDP growth with the Composite NRI using a right-hand secondary axis. Panel (b) compares actual and fitted manufacturing GDP growth. Panel (c) presents correlator



Key Findings & Future Directions

KEY FINDINGS

1. The Composite Nighttime Radiance Indicator (NRI) contains meaningful information on Indonesia's manufacturing activity. Rather than replacing official statistics, the NRI serves as a useful high-frequency complementary indicator for real-time monitoring and nowcasting.
2. The NRI is positively and significantly associated with manufacturing GDP growth, even after controlling for PMI, manufacturing electricity consumption, and output persistence.
3. Including the NRI improves the dynamic model's fit and in-sample tracking performance (higher Adjusted R^2 , lower RMSE and MAE).
4. The positive relationship remains robust across alternative aggregation methods and sample periods.

FUTURE DIRECTIONS

1. Expand and refine the coverage of industrial complexes to improve the representativeness of the indicator.
2. Enhance spatial delineation of industrial areas by refining polygon geometries and applying spatial masking to minimize non-industrial light contamination.
3. Explore advanced machine learning approaches for constructing the NRI and evaluating its potential for real-time nowcasting and forecasting of manufacturing activity.