Forecasting and Nowcasting with Text as Data

Module 2 – Session 3: Mixed-frequency methods

Renato Vassallo

May, 2026

Barcelona School of Economics

All views expressed here and any remaining errors are my own.

Where we are

So far:

- **Session 1:** from text to signals (embeddings, zero-shot)
- **Session 2:** from signals to decisions (few-shot, evaluation, applications)

Today: the forecasting layer that sits on top of any signal (text or otherwise):

- how to combine **quarterly targets** with **monthly indicators** without throwing information away;
- when **nonlinear ML** pays for itself (and when it does not);
- does adding text-based features help?

Code: `session3/`

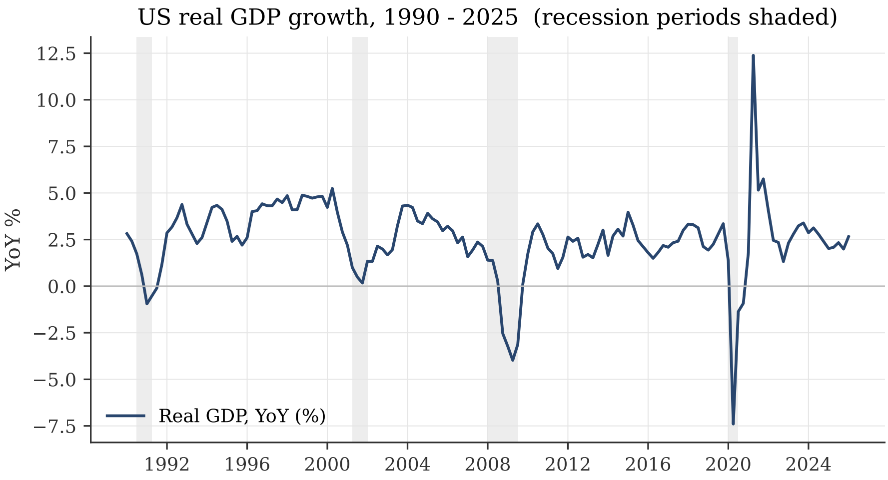
Today's narrative

We will progressively unfold one story:

1. **Why mixed frequency matters** – ragged edge, publication delays.
2. **Simple baselines** – AR, bridge equation, expanding-window.
3. **MIDAS** – parsimonious mixed-frequency regression.
4. **Machine learning** – where nonlinearity pays.
5. **Text + ML** – adding the signals from Session 1 & 2.
6. **Closing**.

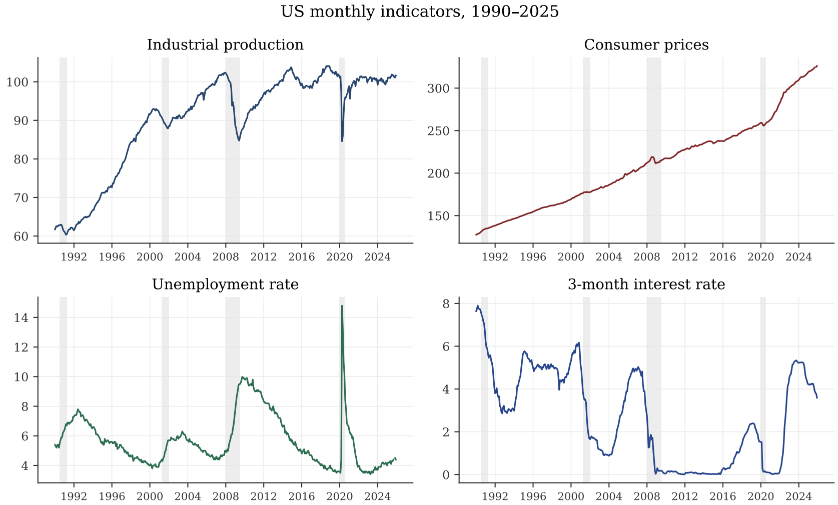
Why mixed frequency matters

The problem in one chart



COVID-Q2 of 2020 prints near −7.4%. Crisis quarters are visually *discontinuous* from normal times, foreshadowing why linear models will struggle.

...but most data arrives in between



Each panel moves with the cycle, but at *different speeds*, with *different leads/lags*. Averaging them to quarterly frequency throws information away on purpose.

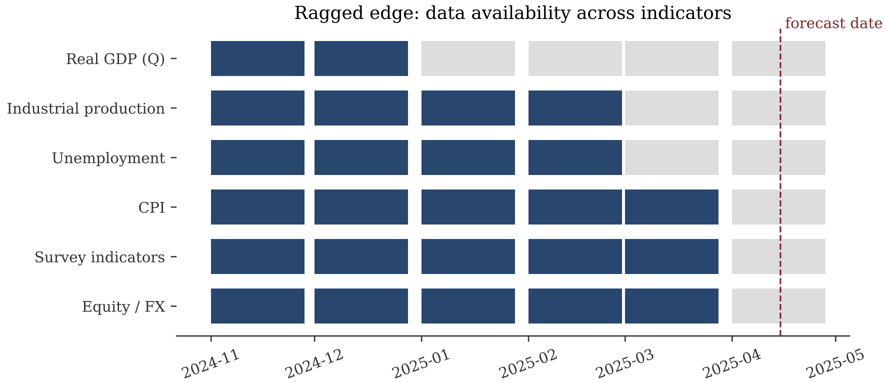
Frequency mismatch

- **Target:** quarterly real GDP growth $y_t^{(Q)}$
- **Predictors:** industrial production, CPI, unemployment, rates, surveys, prices
- Each predictor has its own frequency, publication delay, and revision history

Three options:

- average monthly $\rightarrow$ quarterly (bridge; loses information)
- drop predictors that don't align (MIDAS; doesn't have to)
- pretend they all align (don't)

The ragged edge



At any forecast date, different indicators are at different stages of release. Real-time forecasting is built on top of this picture.

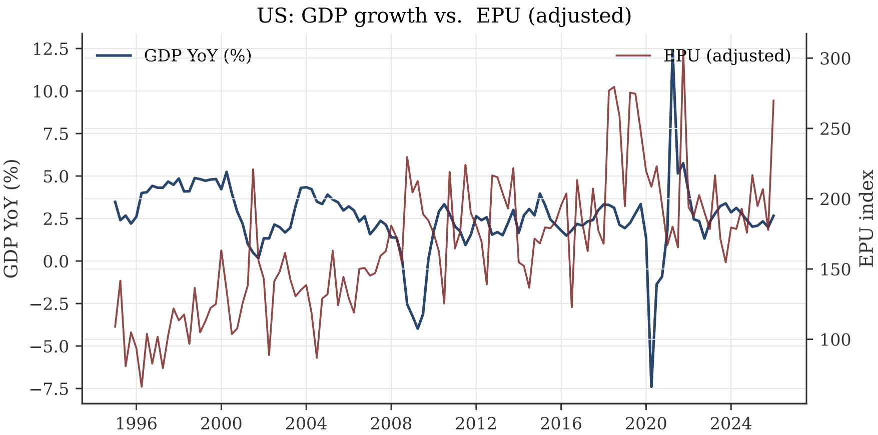
No two forecasters in the world have exactly the same dataset on the same day.

Three forecasting tasks

- **Nowcast**: estimate the current quarter as data accumulates
- **Short-horizon forecast**: 1–4 quarters ahead
- **Recession risk**: classification, probability of contraction in the next year

Different losses, different audiences, different evaluation metrics.

A teaser: text and growth



US real GDP growth (blue) and the EconAI *adjusted* EPU index (red).

Simple baselines

The forecasting protocol

Expanding-window, one-step-ahead forecasts:

- For each origin date t in the test sample:
  - refit the model on data through $t-1$;
  - predict y_t;
  - record $\hat{y}_t - y_t$.
- Aggregate the errors: RMSE, MAE, etc.

Always include trivial benchmarks like random walk ($\hat{y}_t = y_{t-1}$) or historical mean. If your fancy model does not beat them, do not deploy it.

For the rest of the session: expanding window, refit every quarter, $h = 1$.

Start simple

Univariate AR benchmark

$$y_t^{(Q)} = \alpha + \sum_{p=1}^{P} \phi_p \, y_{t-p}^{(Q)} + \varepsilon_t$$

- Fit on past quarters of GDP growth, predict the next one
- Surprisingly hard to beat
- Always include it: if your fancy model loses to AR, the model is not the problem

The AR benchmark is the smell test.

Bridge models

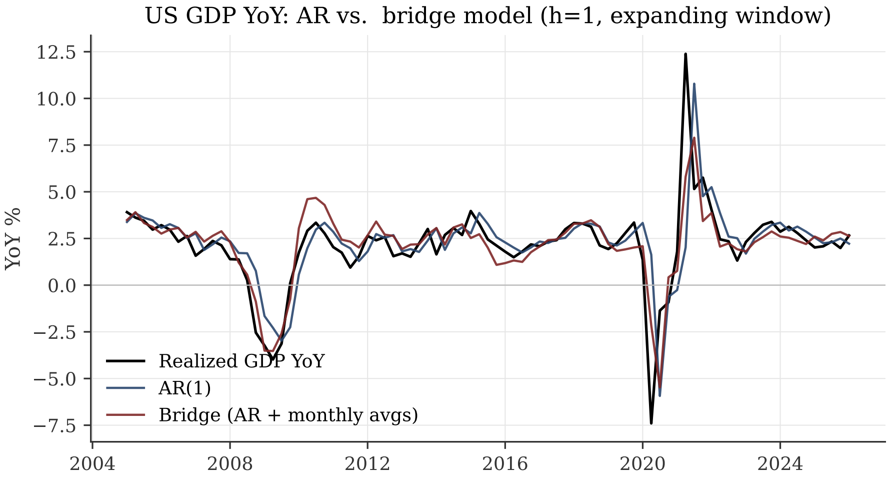
Idea: bring monthly information in by averaging it within the quarter.

$$y_t^{(Q)} = \alpha + \sum_{p=1}^{P} \phi_p \, y_{t-p}^{(Q)} + \beta \, \bar{x}_t^{(M\to Q)} + \varepsilon_t$$

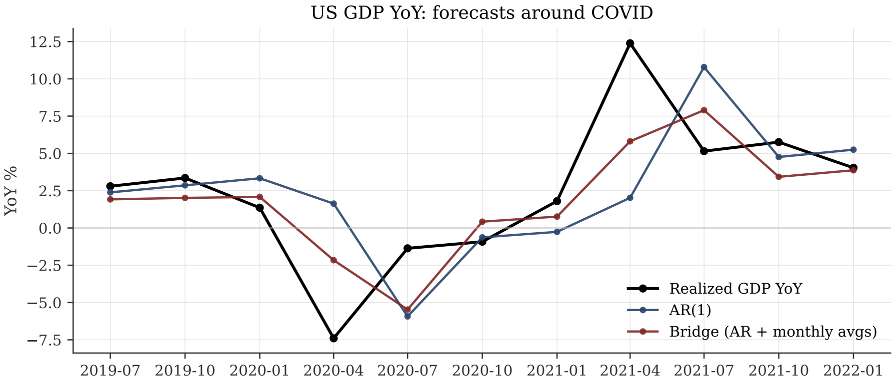
- One coefficient per monthly indicator – simple
- Uses real-time information coarsely
- Implicit assumption: all three months matter equally

Throws away timing.

Comparing AR vs. bridge – US

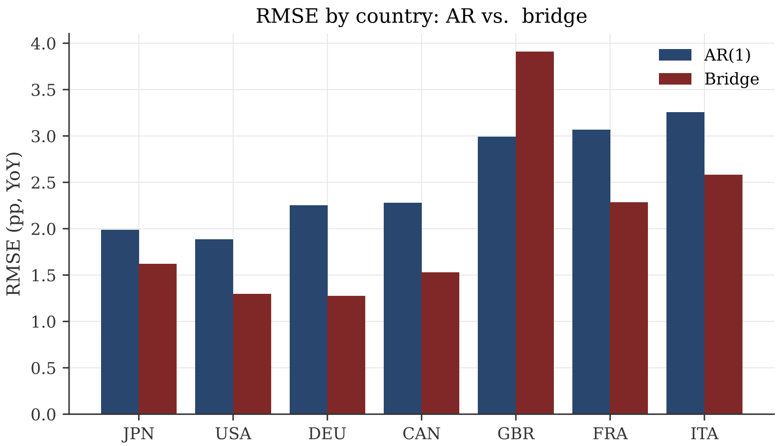


Zoom around COVID: linear models break



Linear forecasts under-shoot the crash and the rebound. This is the visual signature of a regime-switch problem.

Across the panel



- Adding monthly information **usually** helps
- Gains are heterogeneous: small-open economies vs. large advanced economies

What did we leave on the table?

- Bridge **averages** the three months equally
- But the third month of the quarter *is closer* to next quarter than the first
- Real-time information has **structure** that simple averaging hides

We need a model that lets recent observations speak louder.

MIDAS

MIDAS: the basic idea

- Bridge equation uses **equal weights** on monthly lags.
- MIDAS replaces the equal weights by a **parameterised weight function** learned from the data:

$$y_t^{(Q)} = \alpha + \beta \sum_{k=1}^{K} w_k(\theta)\, x_{t-k/m}^{(M)} + \varepsilon_t$$

- Lag $k = 1$ is the *most recent* monthly observation before the target quarter.
- Parameter vector θ control the weight shape: front-loaded, hump, back-loaded, flat.
- Parsimony: $K = 12$ lags, 12 parameters in U-MIDAS; only 2 in Beta-MIDAS.

MIDAS engineering (I)

Raw monthly table

Date	GDP	IP
Jan	–	IP_1
Feb	–	IP_2
Mar	GDP_{Q1}	IP_3
Apr	–	IP_4
May	–	IP_5
Jun	GDP_{Q2}	IP_6
Jul	–	IP_7
Aug	–	IP_8
Sep	GDP_{Q3}	IP_9

GDP is only observed once per quarter

keep only GDP rows

MIDAS estimation sample

GDP	IP_t	IP_{t-1}	IP_{t-2}
GDP_{Q1}	IP_3	IP_2	IP_1
GDP_{Q2}	IP_6	IP_5	IP_4
GDP_{Q3}	IP_9	IP_8	IP_7

Each quarter becomes one row; monthly observations enter as distributed lags

$$\mathrm{GDP}_q = \alpha + \sum_{j=0}^{2} w_j(\theta)\, IP_{q,j} + \varepsilon_q$$

MIDAS engineering (II)

One quarterly observation

GDP_{Q2} → IP_6 IP_5 IP_4

monthly lags for the quarter

Instead of estimating many unrelated lag coefficients, MIDAS estimates a small set of weight parameters θ

MIDAS restriction

$w_0(\theta)$ $w_1(\theta)$ $w_2(\theta)$

weighted monthly signal
$\sum_j w_j(\theta) IP_{q,j}$

→ nonlinear least squares estimate θ

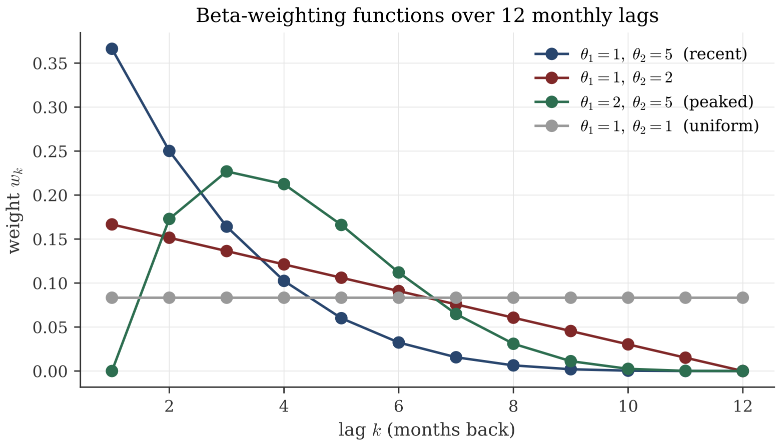
$$\min_{\alpha,\beta,\theta} \sum_q \left(GDP_q - \alpha - \beta \sum_{j=0}^{K} w_j(\theta) IP_{q,j} \right)^2$$

Why restrict the weights?

- Unrestricted MIDAS = one coefficient per monthly lag $\rightarrow$ **overfits**
- Restricted MIDAS = a parametric shape $w_k(\theta)$ $\rightarrow$ regularization through economic priors
- The Beta function is flexible enough to nest most reasonable shapes

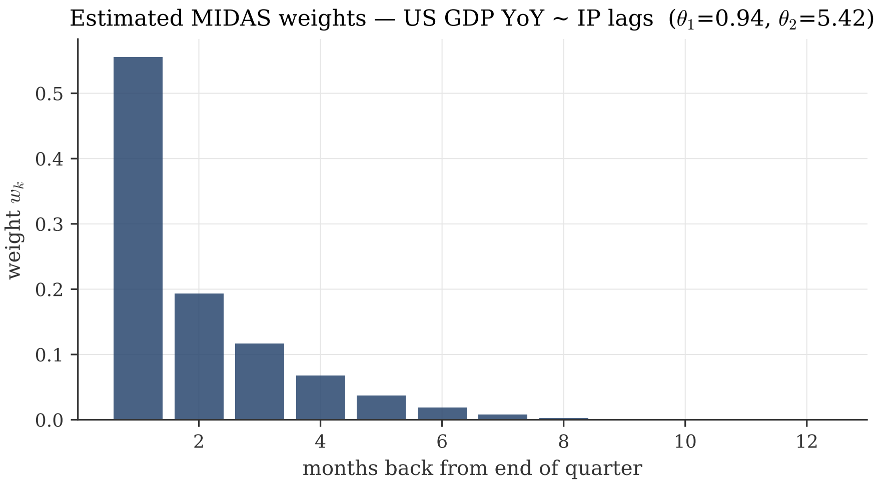
$$w_k(\theta_1, \theta_2) \propto \left(\frac{k}{K}\right)^{\theta_1 - 1}\left(1 - \frac{k}{K}\right)^{\theta_2 - 1}$$

Beta polynomial: typical shapes



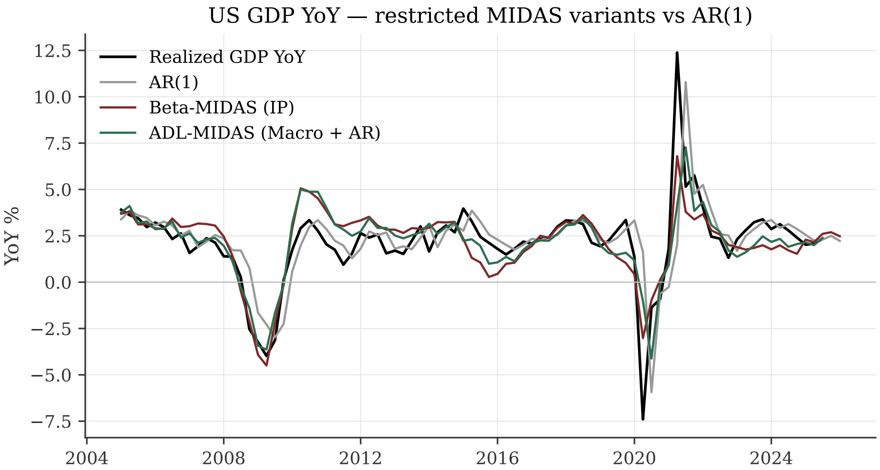
- Different (θ_1, θ_2) give very different stories about *which lags matter*
- Estimated jointly with (α, β) by nonlinear least squares

US: estimated weights

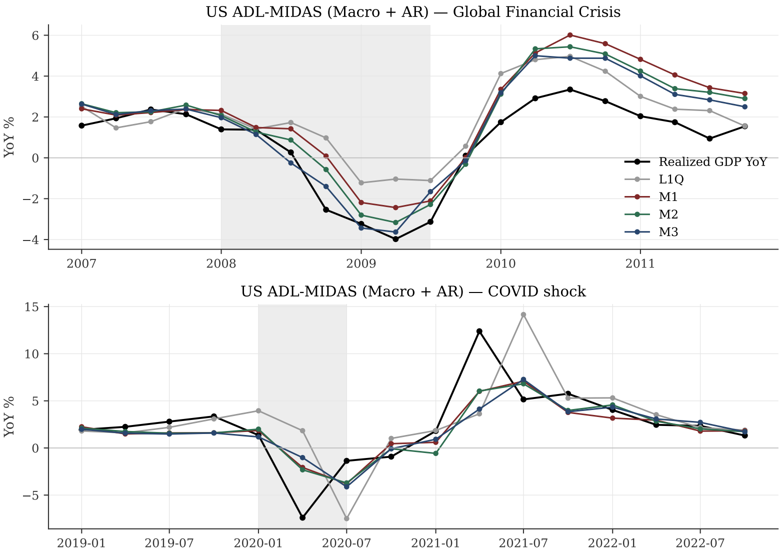


Beta-MIDAS for US GDP $\sim$ monthly IP. Estimated weights are sharply front-loaded: recent months matter most, decaying smoothly to zero around 8-9 months back.

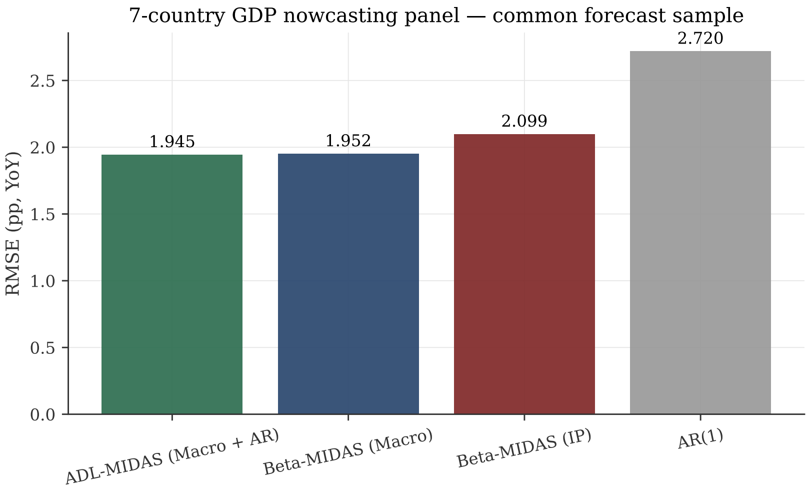
US: out-of-sample forecast



US: vintage nowcasts around crises



Pooled RMSE across MIDAS variants



Where MIDAS struggles

- MIDAS is linear in the high-frequency lags
- In crises, the relationship between leading indicators and growth *changes*:
  - **nonlinear** thresholds (financial stress, mobility shocks)
  - **interactions** between predictors that didn't matter in normal times
- Linear models cannot represent this without exploding parameter counts

Time to let the model find structure on its own.

Machine learning

Why machine learning here

- MIDAS is **linear in the predictors**; macro relationships often aren't
- Tree-based models capture:
  - nonlinear thresholds (e.g. growth collapses when financial stress crosses a level)
  - interactions (rates $\times$ uncertainty, FX $\times$ trade openness)
  - regime dependence without an explicit regime variable
- Cost: less interpretable, more hyperparameters, easier to overfit

Use ML where flexibility pays for itself, not because it's there.

Random Forest: intuition

- Many regression **trees**, each on a bootstrap sample, averaged
- Each tree splits the predictor space into rectangles
- Splits chosen to maximize within-rectangle homogeneity
- Averaging reduces variance without much bias

In macro: the forest discovers which lags interact without you specifying it.

Gradient Boosting: intuition

- Build trees **sequentially**: each one fits the residuals of the previous
- Shrink each step (learning rate) and stop early
- XGBoost (Chen and Guestrin, 2016) adds regularization and second-order updates

Cost: more sensitive to hyperparameters than RF; reward is usually worth it on macro panels.

Setup

Predictors (per country, lags 1–4):

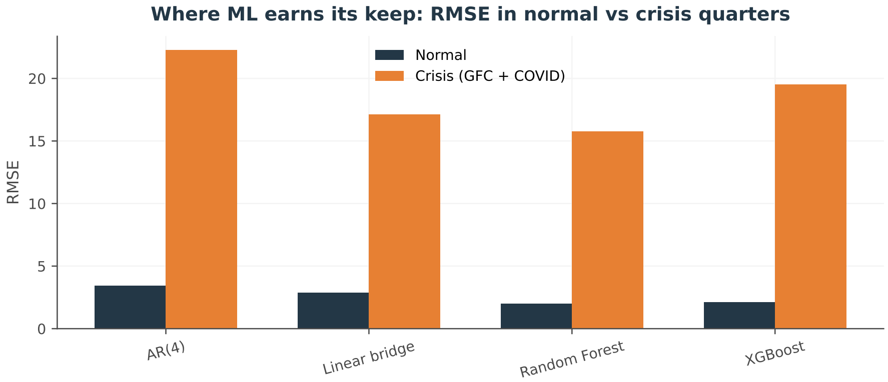
- GDP YoY (autoregressive)
- industrial production YoY, CPI YoY, unemployment, 3-month rate

Models:

- Linear regression (benchmark)
- Random Forest (200 trees, default depth)
- XGBoost (300 trees, depth 4, $\eta = 0.05$)

Evaluation: expanding window, refit each quarter, $h = 1$, OOS from 2005Q1.

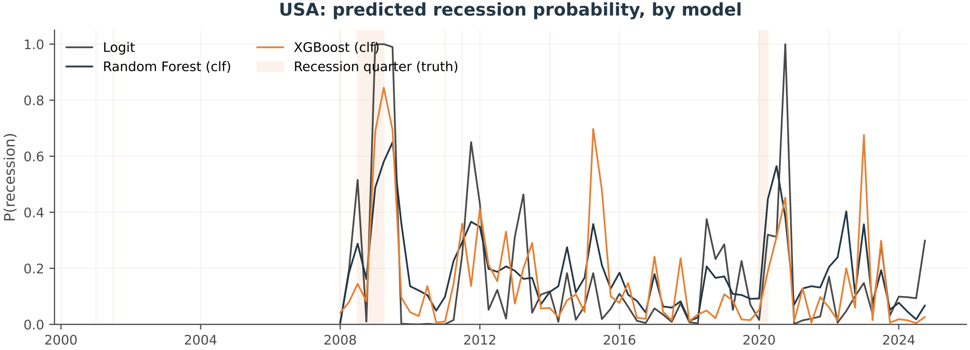
Where ML earns its keep – normal vs crisis



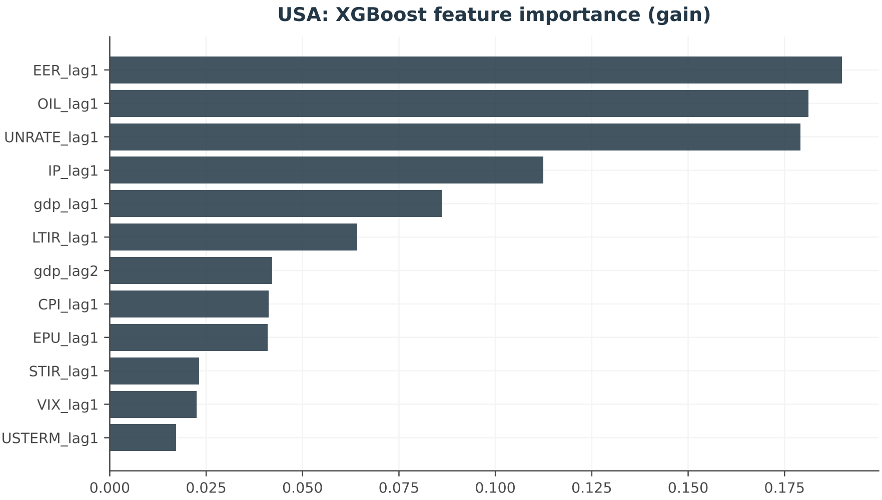
Decomposition: ML cuts the error in crisis quarters meaningfully, while doing roughly the same as the linear bridge in normal times. There is no free lunch – ML buys you tail performance.

Recession-risk classification

- Re-frame the problem as binary: $y_t = 1$ iff $\Delta \log \mathrm{GDP}_t < 0$.
- Logit, RF, XGBoost on the same feature set.
- Use threshold-free metrics (ROC-AUC, AP) because the positive class is rare.



Feature importance (XGBoost gain)



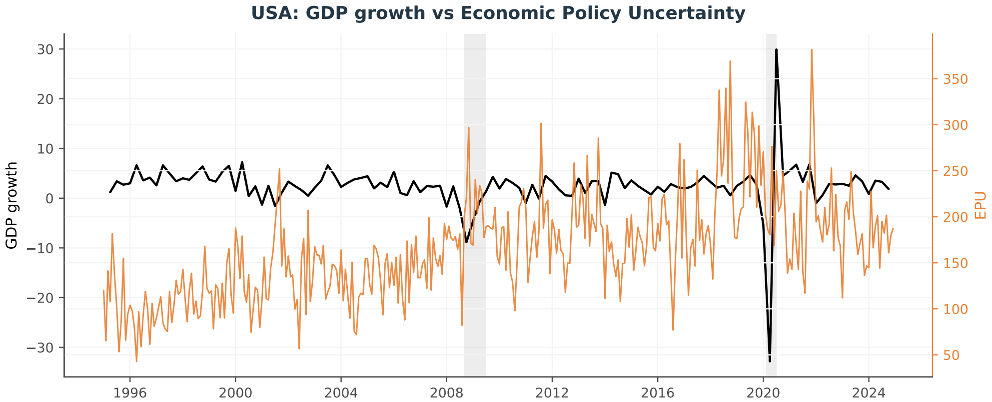
Lagged GDP and unemployment dominate, as expected. Term-spread (`USTERM`) is informative – consistent with the yield-curve recession-prediction literature. EPU shows up too: setting up the next section.

Adding text

From text signals to forecast inputs

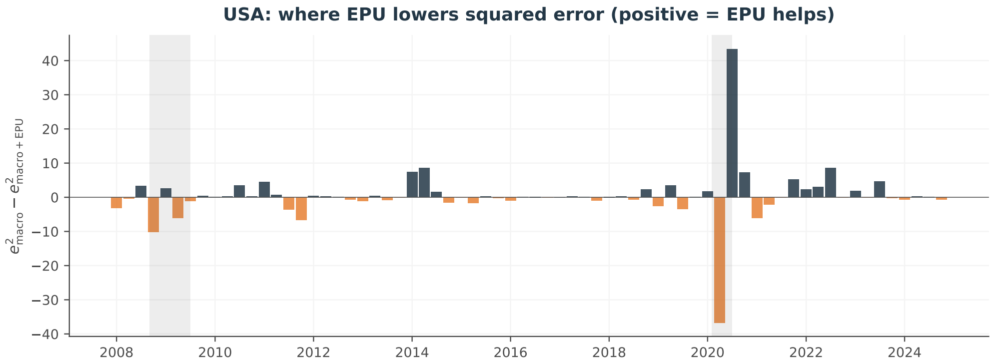
- We use the adjusted **Economic Policy Uncertainty (EPU)** index of Brochet et al. (2025).
- Three questions:
  1. Does adding a text-derived feature lower forecast error?
  2. *When* does it help?
  3. *How* does the model use it?
- The XGBoost setup from before, two feature sets: macro-only vs macro + EPU.

EPU and GDP move together at turning points



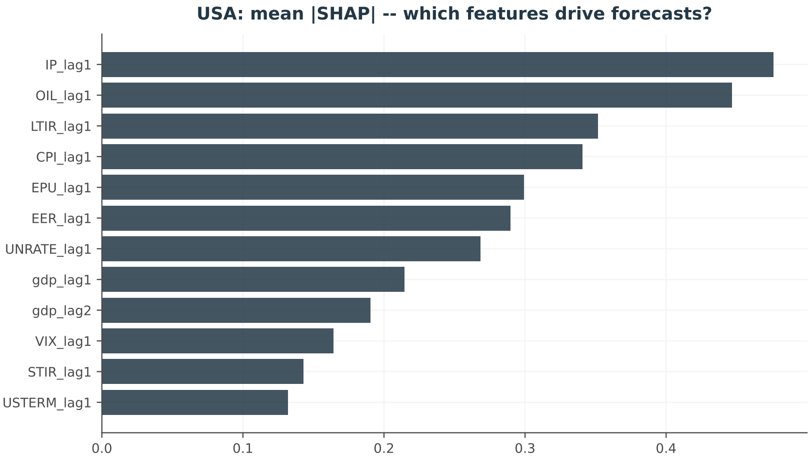
EPU spikes at GFC, eurozone, COVID. The visual hypothesis: uncertainty leads activity.

Quarter-by-quarter gain from adding EPU



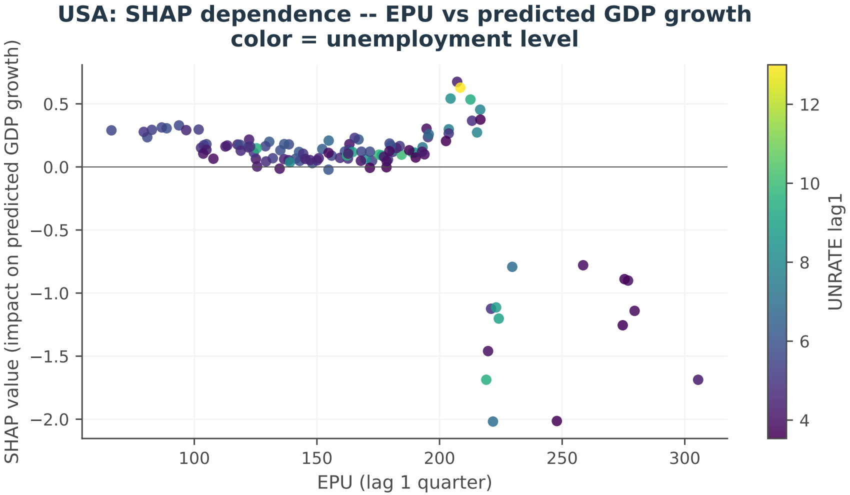
$e^2_{\text{macro}} - e^2_{\text{macro+EPU}}$ – positive bars are quarters where text helps. The biggest gains cluster at uncertainty episodes, as the theory predicts.

What the model is doing: mean $|\mathrm{SHAP}|$



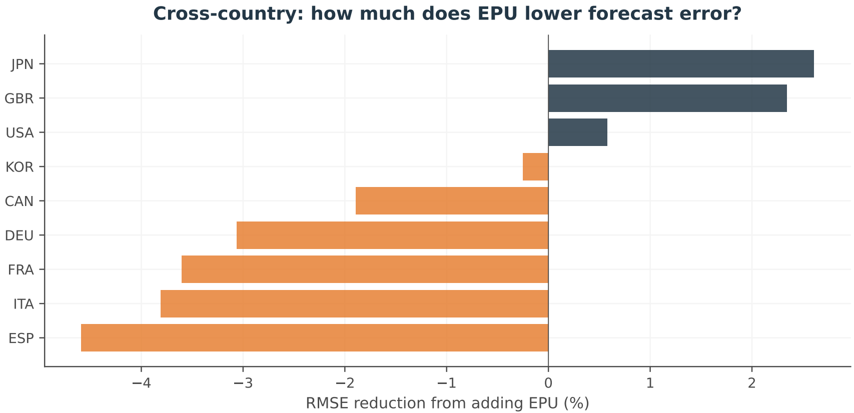
EPU_lag1 is a top-five feature. Its average impact on predictions is comparable to IP or CPI.

The interaction effect: SHAP dependence



High EPU pulls the prediction down, but the strongest negative impact comes from *high EPU* **and** *elevated unemployment*. The model has learned an interaction that a linear model could not represent without explicit cross terms.

Cross-country: does text help everywhere?



RMSE reduction from adding EPU, by country. Magnitude depends on the country's news-EPU coverage and the size of recent uncertainty episodes.

Closing

Take-home messages

- Mixed frequency is the **rule**, not the exception
- MIDAS gives you shape-aware linear forecasts at very low parameter cost
- ML's gains live almost entirely **in crises**
- Text features pay off most when **uncertainty is high**
- Evaluate *per period* and *per regime*, not just on average
- Threshold choice is **half** the recession-risk problem

The hard part isn't the model. It's the rest.

Appendix

Data sources

Variable	Frequency	Source
Real GDP	Q	FRED / OECD QNA
Industrial production	M	FRED / OECD MEI (`<ISO3>PROINDMISMEI`)
CPI, all items	M	FRED / OECD MEI (`<ISO3>CPIALLMINMEI`)
Unemployment rate	M	FRED / OECD (`LR*TTTT*M156S`)
3-month interbank rate	M	FRED / OECD (`IR3TIB01*M156N`)
Recession indicator	M	FRED / OECD recession dummies
EPU (adjusted)	M	EconAI BrochetMuellerRauh2025

Country panel (10): USA, GBR, DEU, FRA, ITA, ESP, CAN, JPN, AUS, MEX. Sample: 1990–2025.

MIDAS – a touch more detail

The Beta-MIDAS regression solves

$$\min_{\alpha,\beta,\theta_1,\theta_2} \sum_t \left(y_t - \alpha - \beta \sum_{k=1}^{K} w_k(\theta_1, \theta_2)\, x_{t,k} \right)^2$$

with

$$w_k(\theta_1, \theta_2) = \frac{f(k/K; \theta_1, \theta_2)}{\sum_{j=1}^{K} f(j/K; \theta_1, \theta_2)}, \quad f(u; \theta_1, \theta_2) = u^{\theta_1 - 1}(1-u)^{\theta_2 - 1}.$$

- Solved by Nelder-Mead, no derivatives
- Identification needs $\theta_2 > 1$ if you want recency loading
- U-MIDAS = unrestricted weights w_k, OLS, used as a sanity check

What we didn't cover

- **Unbalancedness**
- **Dynamic factor models** for very high-dimensional panels (Bok et al., NY Fed nowcast)
- **Kalman filtering / state-space MIDAS** for principled ragged-edge handling
- **Neural mixed-frequency** architectures (TFT, mixed-freq RNNs, transformers)
- **Quantile / pinball** regression for full distributional forecasts
- **Density combinations** of competing models

Pick one for your final project.

References i

- Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring Economic Policy Uncertainty. *QJE*, 131(4), 1593–1636.
- Ghysels, E., Sinko, A., & Valkanov, R. (2007). MIDAS Regressions: Further Results and New Directions. *Econometric Reviews*, 26(1), 53–90.
- Andreou, E., Ghysels, E., & Kourtellos, A. (2013). Should Macroeconomic Forecasters Use Daily Financial Data and How? *Journal of Business & Economic Statistics*, 31(2), 240–251.
- Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *NIPS*.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *KDD*.
- Kalamara, E., et al. (2022). Making Text Count: Economic Forecasting Using Newspaper Text. *Journal of Applied Econometrics*, 37(5), 896–919.
- Mueller, H., & Rauh, C. (2022). The Hard Problem of Prediction for Conflict Prevention. *JEEA*, 20(6), 2440–2467.

Thank you

Questions?

`session2/` is yours – open the notebooks and play.