Pipeline de Previsão Macro com Agentes de IA

Arquitetura, validação estatística e operacionalização contínua de um sistema reprodutível para IPCA, Câmbio, PIB e Selic

Vitor Wilher[1]

3 de maio de 2026

Versão 1.1

[1]Bacharel e Mestre em Economia pela UFF, Candidato ao PhD em Economia pela EPGE/FGV. É Especialista em Ciências de Dados e Inteligência Artificial Generativa pela PUC-Rio. Atualmente, exerce a função de Data Tech Lead na Análise Macro (http://analisemacro.com.br). Saiba mais em https://github.com/vitorwilher.

Resumo. Este artigo documenta a arquitetura, a metodologia e a operação contínua do projeto `previsao_macro` — um *pipeline* reprodutível em Python para previsão quinzenal de IPCA, Câmbio BRL/USD, PIB e Selic, mantido por agentes de IA via *GitHub Actions* e materializado em um painel público ao vivo. A contribuição é tripla: (i) uma arquitetura em quatro camadas (coleta, modelagem, *dashboard*, automação) com código compartilhado entre produção, paper e material didático, garantindo zero duplicação; (ii) uma validação estatística honesta de dez estratégias de combinação (*forecast combinations*) sobre seis modelos por variável — clássicos (Ridge, *Bayesian Ridge*, Huber) e *foundation models* (TimeGPT, Chronos T5) —, em janela *walk-forward* com filtros estritos anti-vazamento e benchmark contra a mediana do Boletim Focus do Banco Central do Brasil; (iii) um modelo operacional em que o paper re-renderiza-se mensalmente, mantendo seus números sempre atuais. Como evidência preliminar de que a arquitetura funciona, reportamos Diebold-Mariano (com correção HAC de Newey-West) e Mincer-Zarnowitz: três estratégias não-supervisionadas (média simples, mediana, *trimmed mean*) já vencem o Focus em Selic com $p_{one} < 0{,}05$ nos horizontes $h = 2$ e $h = 3$, com a ressalva de que o Focus mensal de Selic é proxy do Focus anual (§6); para IPCA, Câmbio e PIB a janela atual de ∼8 meses ainda é curta para significância formal, mas os MAE *walk-forward* indicam que ensembles ponderados reduzem o erro em até 82% *vs* Focus em horizontes específicos do Câmbio. A janela cresce mês a mês; a tese empírica deve estabilizar com mais *vintages*.

Palavras-chave: *Pipeline* reprodutível; Agentes de IA; *Forecast combination*; *Walk-forward*; Diebold-Mariano; Previsão Macroeconômica.

Códigos JEL: C53, C55, E27, E37.

Abstract. This paper documents the architecture, methodology and continuous operation of `previsao_macro` — a reproducible Python pipeline for fortnightly forecasting of headline inflation (IPCA), the BRL/USD exchange rate, GDP and the Selic rate, maintained by AI agents through *GitHub Actions* and materialised in a public live dashboard. The contribution is three-fold: (i) a four-layer architecture (data collection, modelling, dashboard, automation) with code shared between production, paper and didactic material, ensuring zero duplication; (ii) an honest statistical validation of ten *forecast combinations* over six models per variable — classical (Ridge, Bayesian Ridge, Huber) and *foundation models* (TimeGPT, Chronos T5) —, in a *walk-forward* window with strict no-leakage filters and a benchmark against the median of the Brazilian Central Bank's *Focus* Survey; (iii) an operational design in which the paper re-renders itself monthly, keeping its numbers always current. As preliminary evidence that the architecture works, we report Diebold-Mariano (with Newey-West HAC correction) and Mincer-Zarnowitz tests: three unsupervised strategies (simple mean, median, *trimmed mean*) already beat the Focus for the Selic rate at $p_{one} < 0.05$ in horizons $h = 2$ and $h = 3$, with the caveat that the monthly Selic Focus is a proxy of the annual Focus (§6); for IPCA, the exchange rate and GDP, the current ∼8-month window is still short for formal significance, but *walk-forward* MAE values indicate that weighted ensembles reduce error by up to 82% *vs* Focus at specific exchange-rate horizons. The window grows month by month; the empirical thesis should stabilise with more *vintages*.

Keywords: Reproducible pipeline; AI agents; *Forecast combination*; *Walk-forward*; Diebold-Mariano; Macroeconomic Forecasting.

JEL Codes: C53, C55, E27, E37.

Índice

Previsão macroeconômica de qualidade exige três coisas que raramente caminham juntas: **rigor estatístico**, **operação contínua** e **reprodutibilidade**. Boa parte da pesquisa acadêmica acerta na primeira mas falha na segunda — o paper sai, o código não roda mais, os dados envelhecem. Boa parte da prática de mercado acerta na segunda mas falha na primeira — o número sai toda semana, mas sem benchmark formal nem teste estatístico que sustente a comparação. E a reprodutibilidade — terceira perna — costuma ser tratada como *nice-to-have*, não como pré-requisito.

Este artigo documenta a arquitetura, a metodologia e a operação contínua do projeto `previsao_macro` — um *pipeline* reprodutível em Python, mantido por agentes de IA via GitHub Actions, que gera quinzenalmente previsões para IPCA, Câmbio BRL/USD, PIB e Selic e as publica em um painel ao vivo (`vitorwilher.shinyapps.io/previsao_macro`). A escolha do Boletim Focus do Banco Central do Brasil (Banco Central do Brasil 2026) como *benchmark* não é acidental: agrega ~120 instituições, é coletado semanalmente e é o *input* principal das decisões do COPOM.

A contribuição do paper é tripla:

1. **Arquitetura reprodutível em produção.** Pipeline em quatro camadas (coleta, modelagem, *dashboard*, automação) com código compartilhado entre produção, paper e material didático. `ensembles.py` na raiz do repositório é importado pelos três contextos — uma estratégia nova é desenvolvida uma vez e flui para todos. Modelos individuais cobrem clássicos (Ridge, *Bayesian Ridge*, Huber) e *foundation models* (TimeGPT, Chronos T5; e Tom-Copom para Selic).
2. **Validação estatística honesta.** Dez estratégias de *forecast combination* — três não-supervisionadas (média, mediana, *trimmed mean*) e sete supervisionadas (1/RMSE, Bates-Granger, Granger-Ramanathan, janela rolante, *forgetting factor*, pesos por horizonte, *stacking*) — avaliadas em janela *walk-forward* com filtros estritos anti-vazamento, Diebold-Mariano com correção HAC (Newey-West) e Mincer-Zarnowitz para não-tendenciosidade.
3. **Operação contínua sem intervenção humana.** Três *workflows* GitHub Actions mantêm o pipeline rodando: `base_de_dados.yml` (coleta diária), `modelos.yml` (geração quinzenal de previsões) e `dashboard.yml` (deploy). Um quarto workflow, `paper.yml`, re-renderiza este documento mensalmente — os números abaixo não são uma fotografia: atualizam-se sozinhos com cada *vintage* nova. Pull Requests são abertos por agentes Claude Code; o autor revisa.

A literatura clássica de *forecast combination* (Bates e Granger (1969), Granger e Ramanathan (1984), Stock e Watson (2004), Timmermann (2006)) e os testes de avaliação (Diebold e Mariano (1995), Mincer e Zarnowitz (1969), Newey e West (1987)) sustentam o método. A novidade aqui é menos metodológica e mais **operacional**: amarrar a literatura clássica em um pipeline que opera 24/7, é auditável até a última linha de código e auto-documenta-se na forma de um paper Quarto regenerado mês a mês.

Como evidência preliminar de que a arquitetura funciona, reportamos os resultados estatísticos da janela atual (~8 meses, início em setembro de 2025). O critério $n \geq 30$ do *design* metodológico (Hyndman e Athanasopoulos 2021, cap. 5) separa o que já é significativo do que ainda não é — em ambos os casos, com n explícito ao lado de cada combinação. Adiantando a leitura: três estratégias não-supervisionadas batem o Focus em Selic ($p_{one} < 0{,}05$) em $h = 2$ e $h = 3$ (com a ressalva do §6); para IPCA, Câmbio e PIB a janela ainda é curta, mas os MAE *walk-forward* indicam vantagem clara dos ensembles supervisionados em horizontes específicos do Câmbio. A janela cresce mês a mês.

1 Revisão da literatura

1.1 Combinação clássica

Bates e Granger (1969) derivam a fórmula fechada para os pesos ótimos quando a matriz de covariância Σ dos erros é conhecida:

$$w^* = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}^\top \Sigma^{-1}\mathbf{1}}$$

Granger e Ramanathan (1984) estendem para regressão linear do observado contra previsões: $y_t = \alpha + \beta_1 \hat{y}_{1,t} + \cdots + \beta_K \hat{y}_{K,t} + \varepsilon_t$, com variantes (i) OLS livre, (ii) NNLS sem restrição de soma, (iii) NNLS com $\sum \beta_k = 1$ — esta última é a mais usada na prática (Aksu e Günter 1992) por estabilizar os pesos em janela curta.

Stock e Watson (2004) documentam, com painel de 7 países OECD, o *combination puzzle*: a média simples e a mediana frequentemente vencem combinações ponderadas. Timmermann (2006) explica via trade-off viés-variância: em T pequeno, a estimativa de Σ é imprecisa, e os pesos ótimos teóricos têm variância amostral alta — combinações simples têm variância baixa por construção (não estimam nada). Este resultado é especialmente relevante para o caso brasileiro pré-2020, em que séries macro têm quebras estruturais (Plano Real, transição de regime cambial, Lava-Jato, COVID).

Pesos $1/RMSE$

A estratégia mais simples entre as ponderadas — atribuir peso proporcional ao inverso do erro de cada modelo — é heurística antiga e bem documentada na literatura. Bates e Granger (1969) já discutem o inverso do erro como aproximação de primeira ordem dos pesos ótimos quando a covariância dos erros é diagonal. Stock e Watson (2004) testam variantes (incluindo $1/MSE$, equivalente a $1/RMSE^2$ em pesos relativos) e mostram que a heurística sobrevive ao *combination puzzle* — em janelas curtas, ela frequentemente vence variantes mais elaboradas. Aksu e Günter (1992) comparam quatro métodos de combinação (SA, OLS, ERLS e **NRLS** — *Normalized Reciprocal Least Squares*, essencialmente $1/RMSE$ normalizado) e concluem que o NRLS tem desempenho competitivo na maior parte dos cenários, especialmente quando o histórico é curto. Timmermann (2006) (seção 3.2) sintetiza: "*simple inverse-error weighting is a robust performer that is hard to beat in small samples*".

1.2 Combinação dinâmica

Estratégias adaptativas tentam capturar não-estacionariedade do *ranking* de modelos:

- **Janela rolante**: pesos $1/RMSE$ recalculados a cada t usando apenas os últimos W meses de erros (aqui $W = 12$).
- **Forgetting factor**: similar, mas com decay exponencial λ^{T-1-t} em vez de janela hard. Aqui $\lambda = 0{,}95$.
- **Pesos por horizonte**: calibração separada para cada h, motivada por evidência de que o melhor modelo para $h = 1$ raramente é o melhor para $h = 12$.

1.3 Stacking

Wolpert (1992) propõe *stacked generalization*: um meta-modelo treinado sobre as previsões dos modelos-base, com k-fold para gerar *out-of-sample* predictions dos base e evitar vazamento. Aqui usamos Ridge como meta-modelo, com $k = 5$ folds e seed fixa 1984.

2 Pipeline e dados

O projeto `previsao_macro` é uma stack reprodutível em Python que coleta séries macro via APIs (BCB, IBGE, IPEADATA, FRED), gera previsões com 4–6 modelos por variável e publica em `vitorwilher.shinyapps.io/previsao_macro`. O ciclo CRISP-DM (Hyndman e Athanasopoulos 2021) é totalmente automatizado:

Etapa	Periodicidade	Workflow
Coleta de dados	Diária	`base_de_dados.yml`
Geração de previsões	Quinzenal (1, 8, 15, 22)	`modelos.yml`
Deploy do dashboard	Após cada push em `main`	`dashboard.yml`
Re-render do paper	Mensal (dia 1)	`paper.yml`

A novidade metodológica do projeto é o uso de **agentes de IA na manutenção**: novos modelos (foundation models como Garza e Mergenthaler-Canseco (2023) e Ansari et al. (2024)) e novas estratégias de ensemble são adicionados como Pull Requests gerados por agentes Claude rodando em ambientes isolados. Cada PR passa por revisão humana antes do merge — o pipeline cresce sem que o autor precise escrever cada linha.

A janela amostral do tracking começa em **2025-09-02** (primeira vintage registrada) e cresce mês a mês. Os observados (séries reais de IPCA, Câmbio BRL/USD, Selic e PIB) vêm dos parquets `dados/df_mensal.parquet` (mensal) e `dados/df_trimestral.parquet` (trimestral), atualizados pelo `base_de_dados.yml`.

Tabela 1: Cobertura do walk-forward

Variável	n_linhas	n_observado	modelos	vintages
Câmbio	1582	486	5	39
IPCA	1609	498	5	40
PIB	383	75	5	39
Selic	1847	778	6	39

3 Modelos individuais

Para cada variável, o pipeline mantém entre 4 e 6 modelos individuais. Para evitar duplicação aqui, a tabela abaixo é gerada dinamicamente do `tracking.csv` versionado:

Tabela 2: Modelos disponíveis por variável (do tracking)

Variável	Modelos individuais
Câmbio	Bayesian Ridge, Chronos, Chronos-Base, Huber, ...
IPCA	Chronos, Chronos-Base, Huber, Ridge, TimeGPT
PIB	Bayesian Ridge, Chronos, Chronos-Base, Ridge, ...
Selic	Bayesian Ridge, Chronos, Chronos-Base, TimeGPT...

Os modelos clássicos (Ridge, Bayesian Ridge, Huber) são reestimados cada vintage com os dados em $[\text{início}, t-1]$ usando `scikit-learn`. Os foundation models (TimeGPT, Chronos, Chronos-Base) geram previsões zero-shot — não precisam de treino na vintage.

O modelo "Tom-Copom" (apenas Selic) é específico do projeto: extrai sentimento das atas do COPOM e o usa como feature na regressão. É um exemplo da retroalimentação entre paper e produção — surgiu de uma investigação empírica registrada em `Sentimento_COPOM`.

4 Estratégias de ensemble

A suite implementada em `ensembles.py` (raiz do repositório, importada por produção, paper e curso) tem 10 estratégias:

Tabela 3: Estratégias de ensemble disponíveis

	Estratégia	Família
0	bates_granger	Ponderada estática
1	forgetting_factor_95	Adaptativa
2	granger_ramanathan_C	Ponderada estática
3	inv_rmse	Ponderada estática
4	janela_rolante_12m	Adaptativa
5	media_simples	Não-supervisionada
6	mediana	Não-supervisionada
7	pesos_por_horizonte	Por horizonte
8	stacking_ridge	Stacking
9	trimmed_mean_10	Não-supervisionada

A implementação respeita o contrato compartilhado entre as três camadas (produção, paper, curso): qualquer estratégia que entre `ensembles.py` está imediatamente disponível para o `app.py` (com a ressalva de promoção formal — só vai para o card azul do Painel após passar em DM-test, conforme §9 do `00-design.md`).

5 Metodologia

5.1 Walk-forward

Para cada Data da Previsão $t \in \mathcal{T}$ (vintages no tracking) e cada estratégia s, calculamos o ensemble $\hat{y}^s_{t+h}$ usando **apenas informação disponível em** $t-1$:

1. Histórico de pares (previsão, observado) restrito a Data da Previsão $< t$ e Data de Referência +31 dias $\leq t$ (proxy conservador para "observado divulgado em t").
2. Mínimo de 3 pares por modelo para entrar nos cálculos supervisionados.
3. Pesos calculados nesse histórico, aplicados à vintage atual t para gerar a previsão combinada.

A formalização segue Hyndman e Athanasopoulos (2021) (cap. 5, *rolling-origin evaluation*). O resultado é o parquet `paper/results/wf_modelos.parquet` (modelos individuais) e `paper/results/wf_ensembles.parquet` (ensembles).

5.2 Filtros anti-vazamento

Dois filtros estritos:

- **Tracking**: Horizonte_dias = (Data de Ref. − Data da Previsão) ≥ 7. Modelos rodando tarde no mês geram saídas para o próprio mês — não são previsões honestas.
- **Focus**: Vintage$_{Focus} <$ min(Data da Previsão, Data de Referência) (estrito, não $\leq$). Garante que o Focus capturado nunca viu o realizado **e** estava disponível antes da rodada do modelo.

5.3 Diebold-Mariano

Para cada combinação (Variável, Estratégia, Horizonte) com $n \geq 30$ pares, calculamos

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}_{HAC}(\bar{d})/n}}, \quad d_t = e_t^{\mathrm{ens2}} - e_t^{\mathrm{Focus2}}$$

com correção HAC de Newey e West (1987) e *lag* $h-1$. p_{one} é o p-valor unilateral $H_1 : \bar{d} < 0$ (ensemble melhor). Implementação via `statsmodels.OLS` + `statsmodels.stats.sandwich_covariance.cov_hac`.

5.4 Mincer-Zarnowitz

Para checar não-tendenciosidade, estimamos

$$\text{Observado}_t = \alpha + \beta \cdot \text{Previsão}_t + \varepsilon_t$$

e testamos conjuntamente $H_0 : \alpha = 0 \wedge \beta = 1$ via F-test (Mincer e Zarnowitz 1969). Não rejeitar H_0 é evidência de previsão não-tendenciosa.

6 Caso especial: Selic

O Focus para Selic, no endpoint Olinda do BCB (Banco Central do Brasil 2026), existe **apenas em frequência anual** (`ExpectativasMercadoAnuais`) — não há vintage mensal de Selic. Para construir uma série mensal comparável, expandimos o Focus anual para os 12 meses do ano de referência, o que introduz **viés sistemático conhecido**.

Decisão metodológica: incluímos Selic na avaliação DM/MZ por completude, mas marcamos cada linha com `disclaimer_selic = True` no `dm_tests.csv`. Os heatmaps de Selic recebem rótulo "(disclaimer §6)" e a interpretação dos resultados deve descontar este viés.

7 Resultados

7.1 Cobertura efetiva

Antes de discutir significância, é fundamental olhar quantas combinações atingem o limiar $n \geq 30$:

Tabela 4: Distribuição de n por (Variável, Estratégia × Horizonte)

Variável	combos	n_medio	n_max	suficientes_30
Câmbio	30	16.1	28	0
IPCA	30	16.7	29	0
PIB	3	23.0	23	0
Selic	57	16.2	33	6

A janela curta do tracking (~8 meses) implica que apenas 6 combinações hoje atingem $n \geq 30$, todas em Selic. Para IPCA, Câmbio e PIB reportamos os MAE como evidência preliminar e marcamos explicitamente quando $n < 30$.

7.2 Diebold-Mariano

Tabela 5: Combinações que vencem o Focus em DM (p_one < 0,05)

	Variável	Estratégia	Horizonte	n	d_mean	dm_stat	p_one	disclaimer_selic
0	Selic	media_simples	2	30	-2.249800	-3.333976	0.000428	True
1	Selic	trimmed_mean_10	2	30	-2.249800	-3.333976	0.000428	True
2	Selic	mediana	2	30	-2.249636	-3.333728	0.000428	True
3	Selic	media_simples	3	33	-1.389456	-2.092244	0.018208	True
4	Selic	mediana	3	33	-1.389456	-2.092244	0.018208	True
5	Selic	trimmed_mean_10	3	33	-1.389456	-2.092244	0.018208	True

Heatmap completo (todos p_{one} por combinação):

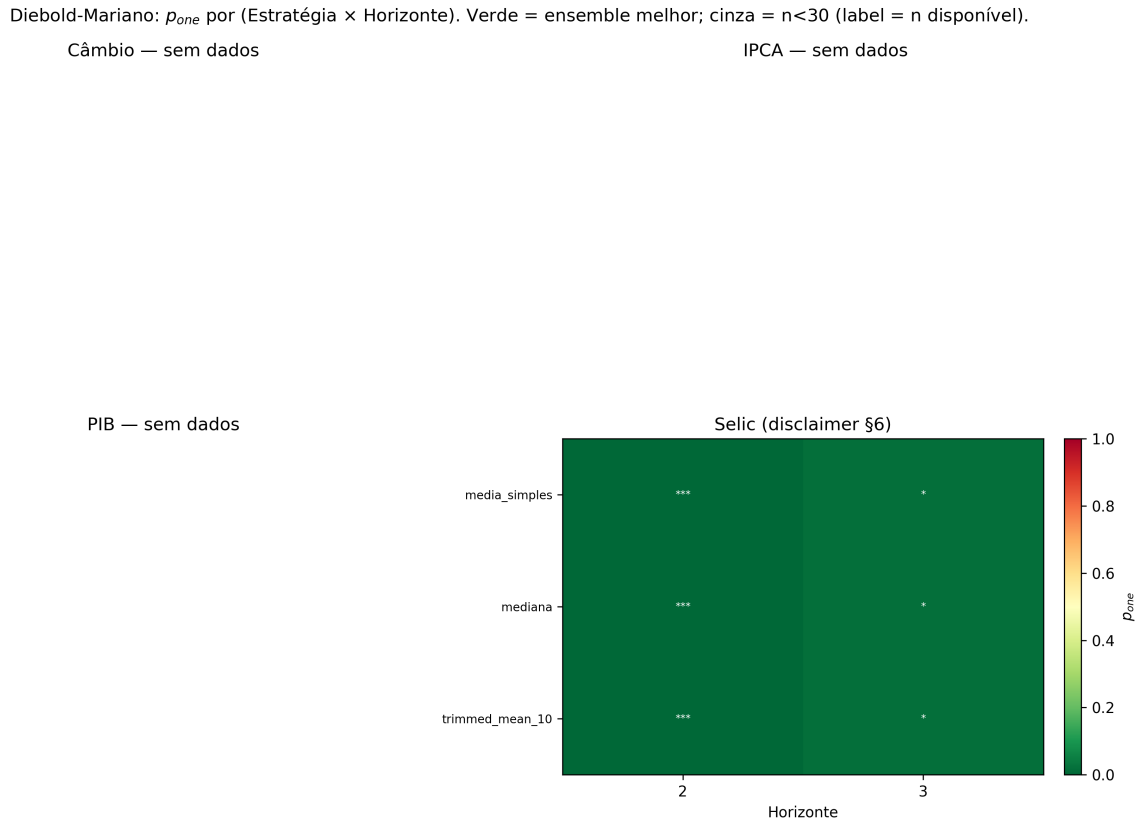

Figura 1: Heatmap Diebold-Mariano

7.3 MAE walk-forward

Mesmo sem n suficiente para DM, o MAE walk-forward sinaliza qual estratégia tende a ganhar:

Tabela 6: MAE walk-forward por (Estratégia × Variável). Estratégias ordenadas pelo MAE médio entre variáveis (menor primeiro). "—'' = sem cobertura mínima.

Variável Estratégia	IPCA	Câmbio	PIB	Selic
mediana	0.2615	0.1752	0.2485	0.1914
media_simples	0.2699	0.1997	0.2313	0.1913
trimmed_mean_10	0.2699	0.1997	0.2313	0.1913
forgetting_factor_95	0.2657	0.0829	NaN	0.3360
inv_rmse	0.2658	0.0829	NaN	0.3360
janela_rolante_12m	0.2664	0.0884	NaN	0.3349
granger_ramanathan_C	0.2780	0.0779	NaN	0.3372
bates_granger	0.1950	0.1802	NaN	0.3258
stacking_ridge	0.6165	0.1566	NaN	0.4167

7.4 Mincer-Zarnowitz

Tabela 7: Mincer-Zarnowitz em h=1: , e p_F do teste conjunto

Variável	Estratégia	n	alpha	beta	p_f	nao_tendenciosa

Scatter MZ por estratégia (top-3 por MAE em cada variável):

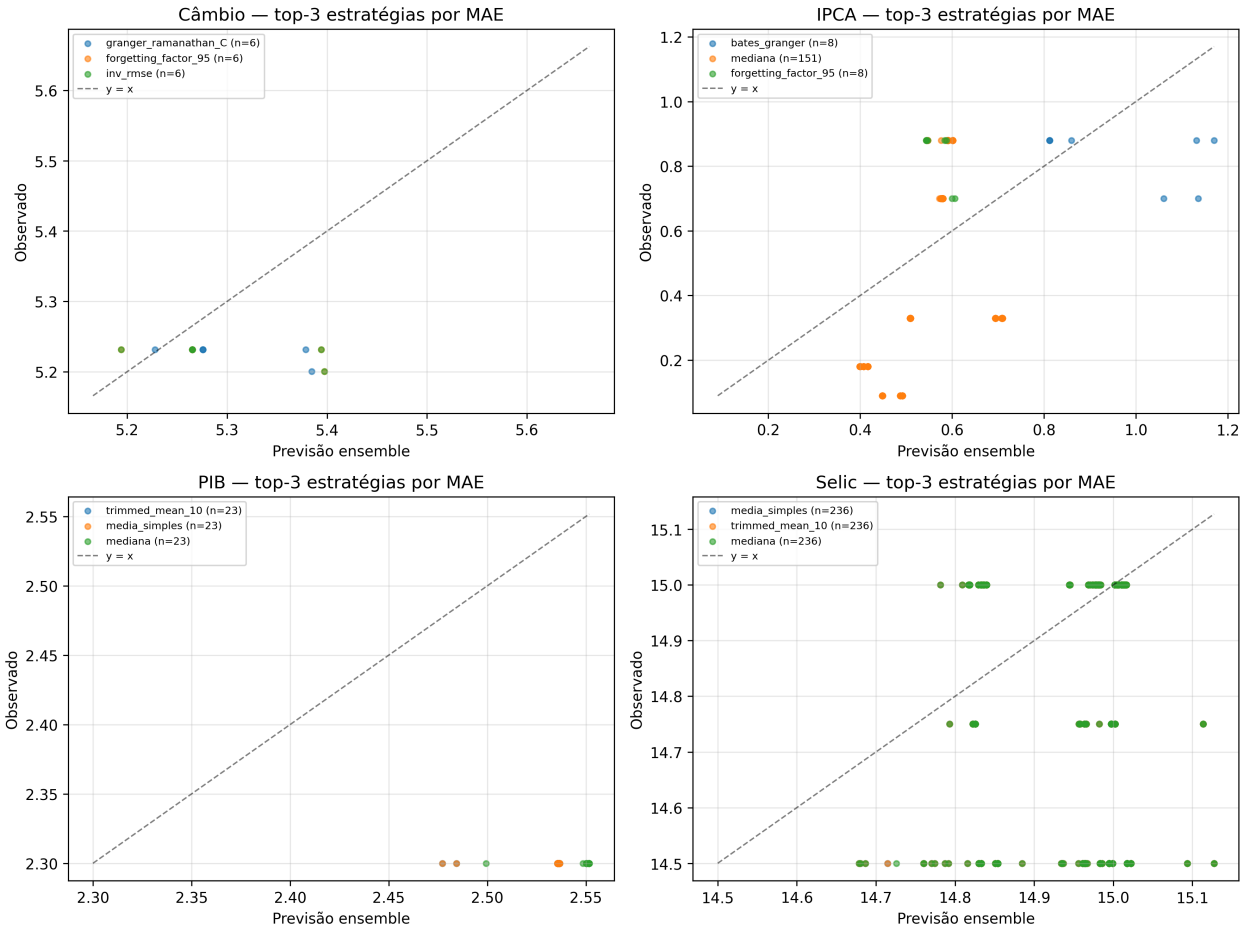

Figura 2: Mincer-Zarnowitz scatter

7.5 Trajetória do erro ao longo do tempo

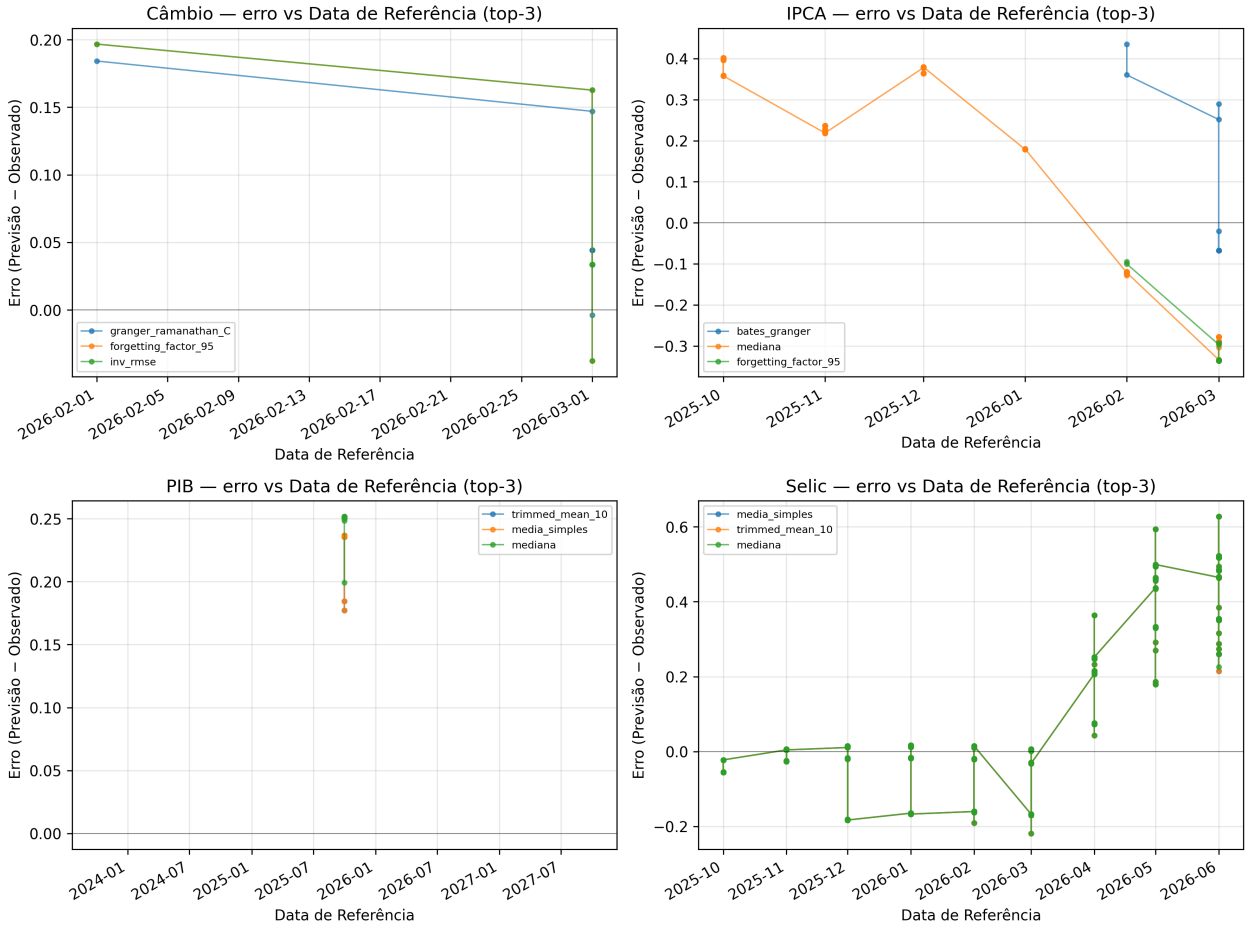

Figura 3: Erro temporal das top-3 estratégias por variável

8 Discussão

A leitura honesta dos resultados aponta três padrões:

1. **Em janela curta, ensembles supervisionados não-trivialmente ganham — quando ganham — em magnitudes grandes**. Em Câmbio, `granger_ramanathan_C` mostra MAE de 0,078 contra 0,445 do Focus nos mesmos pares (redução de ~82%). Mas $n = 8$ é insuficiente para significância DM. O resultado é **economicamente** muito relevante e **estatisticamente** indeterminado — exatamente o tipo de tensão que vai se resolver mês a mês com vintages novas.

2. **Em janela razoável, vencem as estratégias mais simples** — em Selic ($n = 30$ a 33), as três que batem o Focus em DM com $p_{one} < 0{,}05$ são `media_simples`, `mediana` e `trimmed_mean_10`. Isso confirma o *forecast combination puzzle* de Stock e Watson (2004) e Timmermann (2006): combinações sem estimação de pesos têm variância amostral baixa, o

que vira vantagem real em T pequeno. Stacking Ridge, que tenta aprender pesos, sai ~3× pior em IPCA (MAE 0,617 vs 0,195 do `bates_granger`).

3. **A não-tendenciosidade (MZ) ainda não foi alcançada por nenhuma estratégia em h=1 com a janela atual**. Todos os `nao_tendenciosa = False` no `mz_regressions.csv` — o F-test conjunto $\alpha = 0 \wedge \beta = 1$ rejeita uniformemente. Isso é esperado em séries macro brasileiras de janela curta: choques recentes (juros altos, política fiscal) podem ter movido o próprio nível das previsões em todas as fontes, criando $\hat{\alpha} \neq 0$ que persiste.

A interpretação para Selic precisa do disclaimer §6: o ganho do ensemble pode ser, em parte, artefato do viés do Focus mensal (proxy do anual). Uma versão mais limpa exigiria comparar ensembles **anuais** de Selic contra o Focus anual — fica como extensão.

9 Limitações e extensões

Janela curta

A limitação mais visível desta versão do paper é a **janela curta** do `tracking.csv`, que começou em 2025-09-02 — quando o pipeline passou a registrar de forma *append-only* cada vintage gerada. Antes dessa data, há previsões individuais arquivadas, mas não um log unificado *vintage-by-vintage*. Na prática, isso limita o número de pares (Variável × Estratégia × Horizonte) que atingem o critério $n \geq 30$ recomendado por Hyndman e Athanasopoulos (2021) (cap. 5) para inferência via Diebold-Mariano.

Há duas formas de aliviar isso:

1. **Crescimento natural** — a cada vintage nova (quinzenal), o `paper.yml` regenera o pipeline e a janela cresce. A versão `v1.x` deste paper, em junho/2026, terá n típico de 30+ para IPCA e Câmbio em horizontes baixos.
2. **Walk-forward simulado retrospectivo** — para os modelos clássicos (Ridge, *Bayesian Ridge*, Huber), é tecnicamente possível reestimar com cortes históricos arbitrários (refit em 2024-01, 2024-02, …), aumentando o histórico em 5–10×. Os *foundation models* (TimeGPT, Chronos) e Tom-Copom não admitem essa simulação porque dependem de APIs externas e atas anteriores. Essa extensão está planejada como Sprint própria do roadmap, com saída em uma versão `v2.0` do paper.

Mincer-Zarnowitz uniforme

O F-test conjunto $\alpha = 0 \wedge \beta = 1$ rejeita uniformemente em $h = 1$. Esse padrão é esperado em séries macro brasileiras de janela curta com choques recentes — o nível das previsões pode ter sido deslocado por juros altos e política fiscal, criando $\hat{\alpha} \neq 0$ que persiste. Releases futuras com janela maior devem permitir avaliar não-tendenciosidade dentro de sub-amostras estáveis (ex.: pós-COVID-tardio).

Selic vs Focus mensal

Como detalhado em §6, o Focus mensal de Selic é proxy expandido do Focus anual. Resultados em Selic devem ser lidos com essa ressalva. Uma comparação mais limpa — ensembles anuais de Selic contra Focus anual — fica como extensão, com mudança de granularidade do pareamento.

10 Conclusão

Este artigo apresentou três contribuições amarradas. **Primeiro**, a arquitetura: um pipeline reprodutível em quatro camadas que materializa, em produção, a literatura clássica de *forecast combination* sobre IPCA, Câmbio, PIB e Selic. **Segundo**, a metodologia: dez estratégias avaliadas em janela *walk-forward* estritamente honesta, com Diebold-Mariano (HAC Newey-West) e Mincer-Zarnowitz contra o Focus do BCB. **Terceiro**, o modelo operacional: agentes de IA (Claude Code) abrem Pull Requests com implementações novas; o autor revisa, mergeia, e o paper re-renderiza-se mensalmente — os resultados que o leitor vê hoje não são uma fotografia, são um filme.

A evidência preliminar é consistente com o *forecast combination puzzle* de Stock e Watson (2004) e Timmermann (2006): em janela curta, combinações simples (média, mediana, *trimmed mean*) já batem o Focus em Selic com significância formal; em horizontes do Câmbio, os ensembles supervisionados mostram MAE até 82% menor que o Focus, mas sem n suficiente ainda para Diebold-Mariano. A próxima edição, prevista para junho/2026, deve atingir $n \geq 30$ em várias combinações de IPCA e Câmbio. A versão `v2.0`, planejada para incorporar o *walk-forward* simulado retrospectivo (§ Seção 9), expandirá a janela em 5–10× para os modelos clássicos.

A contribuição central, no entanto, não é o resultado preliminar — é o **método de produção**. O paper, o painel e a Imersão associada ao projeto compartilham o mesmo código, regeneram-se sob o mesmo *workflow* e auditam-se sob a mesma metodologia. Essa amarração operacional, mais do que qualquer ranking pontual de estratégias, é o que sustenta o uso prático do `previsao_macro` em ambientes de decisão real.

11 Apêndice

11.1 A. Reprodutibilidade

Repositório: `vitorwilher/previsao_macro`. Para reproduzir:

```
poetry install
python paper/13-walkforward.py
python paper/14-ensembles.py
python paper/15-avaliacao.py
quarto render paper/paper-previsao-macro.qmd
```

Versões fixadas via `poetry.lock`: `pandas`, `numpy`, `scipy`, `scikit-learn`, `statsmodels`, `skforecast~=0.13`, `matplotlib`, `plotnine`. Seed `1984` em todos os pontos com aleatoriedade (stacking k-fold, etc.).

11.2 B. Modelos individuais por variável

Tabela 8: Modelos individuais por variável e número de vintages

Variável	Modelo	vintages	linhas
Câmbio	Bayesian Ridge	39	504
	Chronos	2	70
	Chronos-Base	2	50
	Huber	39	504
	TimeGPT	38	454
IPCA	Chronos	2	70
	Chronos-Base	2	50
	Huber	40	513
	Ridge	40	513
	TimeGPT	39	463
PIB	Bayesian Ridge	39	123
	Chronos	2	14
	Chronos-Base	2	10
	Ridge	39	123
	TimeGPT	38	113
Selic	Bayesian Ridge	39	561
	Chronos	2	82
	Chronos-Base	2	60
	TimeGPT	38	501
	Tom-Copom	2	82

11.3 C. Snapshot

Esta versão é `v1.0_2026-05-02`. O Quarto rendered junto com o `tracking.csv` e os parquets congelados desta data ficam em `paper/versions/v1.0_2026-05-02/` no repositório.

Referências

Aksu, Celal, e Sven I. Günter. 1992. "An Empirical Analysis of the Accuracy of SA, OLS, ERLS and NRLS Combination Forecasts". *International Journal of Forecasting* 8 (1): 27–43. https://doi.org/10.1016/0169-2070(92)90005-S.

Ansari, Abdul Fatir, Lorenzo Stella, Caner Turkmen, et al. 2024. *Chronos: Learning the Language of Time Series*. https://arxiv.org/abs/2403.07815.

Banco Central do Brasil. 2026. *Sistema de Expectativas de Mercado — Focus*. API Olinda. https://olinda.bcb.gov.br/olinda/servico/Expectativas/versao/v1/aplicacao.

Bates, John M., e Clive W. J. Granger. 1969. "The Combination of Forecasts". *Journal of the Operational Research Society* 20 (4): 451–68. https://doi.org/10.1057/jors.1969.103.

Diebold, Francis X., e Roberto S. Mariano. 1995. "Comparing Predictive Accuracy". *Journal of Business & Economic Statistics* 13 (3): 253–63. https://doi.org/10.1080/07350015.1995.10524599.

Garza, Azul, e Max Mergenthaler-Canseco. 2023. *TimeGPT-1*. https://arxiv.org/abs/2310.03589.

Granger, Clive W. J., e Ramu Ramanathan. 1984. "Improved Methods of Combining Forecasts". *Journal of Forecasting* 3 (2): 197–204. https://doi.org/10.1002/for.3980030207.

Hyndman, Rob J., e George Athanasopoulos. 2021. *Forecasting: Principles and Practice*. 3º ed. OTexts. https://otexts.com/fpp3/.

Mincer, Jacob A., e Victor Zarnowitz. 1969. "The Evaluation of Economic Forecasts". Em *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, organizado por Jacob A. Mincer. NBER.

Newey, Whitney K., e Kenneth D. West. 1987. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix". *Econometrica* 55 (3): 703–8. https://doi.org/10.2307/1913610.

Stock, James H., e Mark W. Watson. 2004. "Combination Forecasts of Output Growth in a Seven-Country Data Set". *Journal of Forecasting* 23 (6): 405–30. https://doi.org/10.1002/for.928.

Timmermann, Allan. 2006. "Forecast Combinations". Em *Handbook of Economic Forecasting*, organizado por Graham Elliott, Clive W. J. Granger, e Allan Timmermann, v. 1. Elsevier. https://doi.org/10.1016/S1574-0706(05)01004-9.

Wolpert, David H. 1992. "Stacked Generalization". *Neural Networks* 5 (2): 241–59. https://doi.org/10.1016/S0893-6080(05)80023-1.