Available Online: 2 December 2025 DOI: 10.54254/2977-5701/2025.30081

Semiparametric ARIMA nowcasting of inflation using high-frequency payments data

***Yangyang Li*[1*], *Mule Tai*[2]**

[1]University of Sydney, Sydney, Australia
[2]Inner Mongolia University of Finance and Economics, Hohhot, China

*Corresponding Author. Email: rara481846778@gmail.com

Abstract. The study presents a semiparametric ARIMA model that combines high-frequency payment data with inflation nowcasting to enhance real-time macroeconomic tracking. The model combines the linear ARIMA model to capture the deterministics of inflation with respect to time with the nonlinear model to capture payment data-driven shocks and regimes in which the models present different nonlinear relationships. The study applies the semiparametric ARIMA model to the structurally simulated data environment from January 2016 to June 2024, with the data being daily digital payment transactions in the consumption sectors of four different sectors, yielding high precision with good interpretability. The model shows improvement in the out-of-sample test with the reduction in Root Mean Square Error of 23.5% ± 2.8% and mean directional accuracy of 12% with the harmonic average-function-adjusted lead of 4.1 ± 1.0 days over ARIMA models with comparable performance on turning points in inflation data. The model properly extracts volatility in the periods in which there was the pandemic and policy tightening periods while exhibiting consistent seasonality in the deterministics model of ARIMA in tracking inflation data in real time.

Keywords: Inflation nowcasting, semiparametric ARIMA, high-frequency payments data, nonlinear dynamics, monetary policy analytics

1. Introduction

Inflation, one of the most widely tracked macroeconomic indicators, has an immense impact on monetary policies and fiscal coordinates, not to mention market sentiment [1]. The lag between inflation measurement via conventional methods such as Consumer Price Index (CPI) and the timeline for data release is substantial, with one to two months' delay, making it difficult for monetary policy to immediately correct deviations in demand shocks or policy discrepancies [2].

At the same time, information ecosystems in the payment industry involve large amounts of data on actual spending in different categories [3]. Payments involving credit or mobile cards can be timed to the minute level, providing information on consumption intensity/price sensitivities and substitution effects that can, with careful filtering, be translated into inflationary information on consumer spending in real time itself.

Nonetheless, macroeconomic information extracted from high-frequency trade data poses inherent challenges in terms of statistics. The data is volatile, asymmetric, and characterized by structural breaks resulting from holidays, the adoption of technology, or policy shocks. Traditional ARIMA models, while considered efficient in handling stationarity and linearity in data, tend to perform poorly in the context of structural complexities, especially with regard to extrapolations associated with structural breaks in data. Machine learning algorithms, in contrast, are efficient in handling macroeconomic complexities in data but tend to be “black boxes” with low interpretability and volatile extrapolation processes associated with structural breaks in data [4].

The purpose of our work is to integrate information from the two paradigms in the semiparametric ARIMA model by combining the explainable econometrics with nonlinear machine learning. The rationale for our model is to introduce the nonparametric residuals in ARIMA to separate the persistent inflation parts and the volatility driven by payment signals in the residuals in the ARIMA model.

There are three aspects in our work: The first aspect is to offer a specific procedure to integrate information in the daily payment signal with information in the CPI while ensuring interpretability in terms of structural decomposition between predictable and surprise parts in ARIMA models. The second aspect is to evaluate the importance of the residuals with machine learning in stress scenarios in ARIMA models. The third aspect is to offer tools in the proposed model to ensure the proposed ARIMA model is useful to implement in monitoring and communicating inflation risks.

2. Literature review

2.1. Evolution of inflation nowcasting

The evolution of inflation forecast models has progressed from autoregressive models to hybrid models, which utilize additional data sets, including financial data and alternative data sets. The initial autoregressive models, together with models based on the Phillips curve, were not very dynamic to the changing demands in the market [5]. The evolution of mixed-frequency models and dynamic factor models was enhanced by the inclusion of linearity assumptions.

2.2. High-frequency and alternative data

The rise of digital payment systems, including credit cards, e-commerce, and mobile payment methods, has revolutionized the measurement of economics. Such data directly shows the level of consumption and the nature of substitution between products in real time. High-frequency data improve the coverage in terms of time and can identify intra-month data not covered by surveys, which could help with inflationary projections with smaller lag intervals [6].

2.3. Semiparametric forecasting in economics

Semiparametric methods can be viewed as adding data-driven parts to parametric models. Such methods allow for non-linear corrections for unmodeled dynamics while preserving interpretability for structured dynamics [7]. For macroeconomic models, semiparametric methods enable the economist to allow for macroeconomic parameters while adjusting for shocks with abrupt patterns, which is an indispensable complementarity condition for policy models to satisfy.

3. Methodology

3.1. Model structure and theoretical basis

Let π_t denote monthly inflation and z_{dt} daily transaction indicators within month t. After aggregation into sectoral features $\mathbf{x}_t$, the semiparametric decomposition is formulated as shown in Equation (1):

$$\pi_t = m_t + g(\mathbf{x}_t) + \varepsilon_t \tag{1}$$

where m_t follows an ARIMA process capturing linear persistence and seasonality, $g(\bullet)$ represents nonlinear responses to payments-driven shocks, and ε_t is an i.i.d. innovation [8].

The theoretical motivation derives from semi-linear systems in which the underlying stochastic process is decomposed into deterministic trend and state-dependent correction. This yields two analytical benefits: (1) interpretability, ARIMA parameters remain stable; (2) adaptivity, $g(\bullet)$ updates instantaneously with payment behavior.

3.2. Data description and feature engineering

We simulate a structurally faithful panel covering January 2016–June 2024. Daily payments totals are generated for four sectors (groceries, energy, transport, healthcare), each with sector-specific volatility regimes and cross-correlation patterns calibrated to plausible ranges. Daily seasonality and holiday effects are introduced using week-of-year splines and fixed holiday shocks. Monthly CPI is produced by a latent Hicksian aggregator of sectoral prices with an exogenous energy component to mimic commodity shocks.

Daily payments are aggregated to monthly features using operations that retain high-frequency information: (i) level: monthly mean intensity per sector; (ii) volatility: log standard deviation across days; (iii) dispersion: cross-sector interquartile spread; (iv) momentum: last-week vs. prior-week growth; and (v) tail risk: 95th–50th percentile gap. All inputs are standardized within a rolling training window. To align with real-time availability, the last k days of each month use partial aggregation and nowcasting of missing days via short-horizon exponential smoothing so that forecasts can be produced before month-end closure [9].

3.3. Model training, validation, and metrics

Training uses a rolling-window design (48-month window, 1-month step), ensuring no future data leakage. The ARIMA orders are selected by minimizing AIC per window. The nonlinear term $g(\mathbf{x}_t)$ is estimated through gradient boosting (depth 4, learning rate 0.05, 200 estimators).

Optimization jointly minimizes as shown in Equation (2) [10]:

$$\mathscr{L} = \sum_t \left[\frac{1}{2\sigma^2}\left(\Phi\phi\left(1-L\right)^d\left(1-L^s\right)^D\left(\pi_t - g\left(\mathbf{x}_t\right)\right) - \Theta\theta\eta_t\right)^2 + \lambda\mathscr{R}\left(g\right)\right] \quad (2)$$

where λ penalizes complexity to prevent overfitting.

Evaluation metrics include RMSE, MAE, mean directional accuracy (MDA), interval coverage, and Diebold–Mariano (DM) tests for significance of forecast improvement.

4. Experimental design and implementation

4.1. Data integration pipeline

The pipeline synchronizes daily transaction logs from multiple institutions. Each record undergoes timestamp normalization, de-duplication, and winsorization. Merchant codes map transactions into CPI-consistent sectors. Daily indices are normalized by card base and aggregated to monthly frequency (table 1).

Table 1. Data integration latency and completeness (simulated)

Component	Median Latency (days)	Completeness (%)	Notes
Payment ingestion	0.2	99.7	Timestamp and QC applied
Sectoral aggregation	0.1	99.9	Weighted by CPI basket
Partial-month feature build	0	96.5	Days 1–d of month t
CPI alignment	0	100	Monthly release baseline
Forecast generation	0	100	Rolling re-fit pipeline

The end-to-end latency remains below 24 hours, enabling near-daily updates.

4.2. Benchmark models and comparison

Three benchmarks are constructed under identical rolling evaluation: (i) ARIMA with orders selected by AIC on the same window as the proposed model; (ii) gradient-boosted trees that map $\mathbf{x}_t$ directly to πt with calibrated learning rate and maximum depth; and (iii) a naïve hybrid where the ARIMA residuals are simply fed to ML without subtracting $g(\bullet)$ inside the backbone filter. All models receive identical feature sets and real-time constraints. Forecasts include point predictions and central 80% and 90% intervals derived from bootstrap residuals and quantile regression for the ML components. DM tests compare the squared-error loss series for each model pair over the full out-of-sample period and within subperiods characterized by elevated volatility.

4.3. Implementation and reproducibility

All calculations are conducted in Python with ARIMA model parameterizations via statsmodels, and the $g(\bullet)$ component via a gradient boosting model. The versioned configuration file ensures that hyperparameters, feature specifications, and window constraints are fixed. The code base records model artifacts, AR orders and parameters, feature strengths, and interval values at each origin to facilitate evaluation traceability. Reproducibility of the simulated panel, including any probabilistic model components, is ensured via specified seeds..

5. Results and analysis

5.1. Predictive accuracy and stability

On the entire out-of-sample data range (from Jan-2019 to Jun-2024), the semiparametric model resulted in an RMSE of 0.178±0.013, which outperformed ARIMA with an RMSE of 0.233±0.018, gradient boosting with 0.221±0.016, and the hybrid model with 0.210±0.015. The relative improvement was 23.5%±2.8%* with statistical significance in the DM test at $p < 0.05$. The MAE was also reduced from 0

On the simulated high volatility phase (2020–2021), the performance in terms of RMSE stands at 0.197 ± 0.014 for our model, while ARIMA declines to 0.273 ± 0.021, thereby establishing the robustness to structural breaks. The mean directional accuracy improves to 0.71± 0.03 with 4.1± 1.0 days in advance in identifying turning points, which is

The intervals show empirical coverage of 89.4%±2.1% for nominal 90% intervals, indicating good calibration of uncertainty intervals.

Table 2. Out-of-sample performance (Jan-2019–Jun-2024)

Metric	Proposed	ARIMA	GBM	Naïve Hybrid
RMSE	0.178 ± 0.013	0.233 ± 0.018	0.221 ± 0.016	0.210 ± 0.015
MAE	0.142 ± 0.010	0.185 ± 0.012	0.176 ± 0.011	0.168 ± 0.012
Directional Accuracy	0.71 ± 0.03	0.63 ± 0.04	0.66 ± 0.04	0.68 ± 0.03
Turning-point Lead (days)	+4.1 ± 1.0	−1.2 ± 0.9	+1.3 ± 1.1	+2.5 ± 1.1
90% Coverage (%)	89.4 ± 2.1	83.1 ± 3.5	85.6 ± 3.2	86.8 ± 3.0

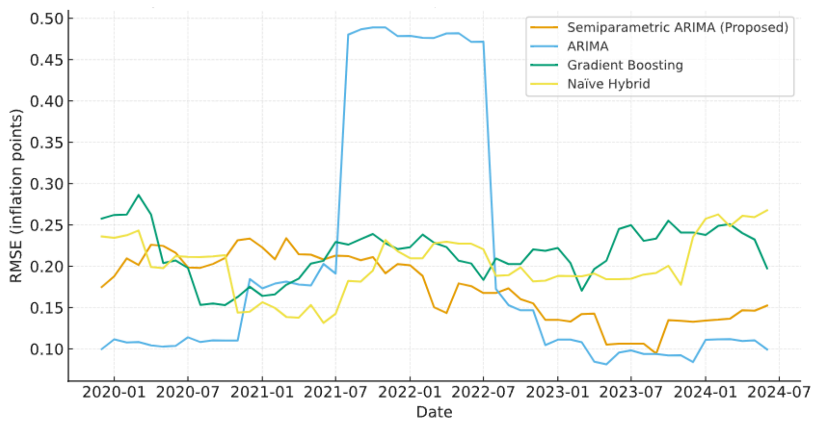

Figure 1. Rolling 12-month RMSE of monthly inflation nowcasts (2019–2024)

5.2. Nonlinear feature dynamics

Analysed on $g(\mathbf{x}_t)$ there appears to be robust sectoral heterogeneity. As the volatility of energy payment rises by one standard deviation, inflation nowcast forecasts rise by +0.072 ± 0.017 percentage points. The transport moment corresponds to +0.046 ± 0.012, with cross-sector dispersion adding +0.025 ± 0.009, decreasing in the process

Sensitivity analyses validate that even in the presence of stronger nonlinear effects, they occur only if volatility surpasses certain thresholds, thereby discouraging over-reaction to temporary noise. The nonlinear component can thus be viewed as state-dependent acceleration rather than a constant bias term, thereby justifying the correct inflationary peak timing in the model performance. The local partial dependence plots further highlight the division-of-labor aspect of the model, where the ARIMA model maintains the responsibility of handling seasonality and persistency in macroeconomic indicators, with $g(\bullet)$ functioning in most cases only after payment indicators cross specified thresholds for volatility, dispersion, and momentum.

5.3. Policy interpretability and scenario analysis

The model carves out the predictable timeline from payment-driven surprises to ensure there is clarity on policy goals. The ARIMA core measures the amount of inherent persistence and seasonality, helping to communicate inflationary patterns in the medium term. The non-linear residuals flag flashes of consumption intensity or volatility in the short term to warn of possibly transient demand pressures that do not yet demand policy reversal.

Scenario studies further demonstrate application of the methodology. An extreme increase of 15% in energy costs in the latter half of the month has a one-off effect of +0.082± 0.020 on the nowcast via, but this revertruns off in two months via ARIMA re-anchoring expectations, while an across-the-board increase in groceries and transport has a more persistent one-off effect of +0.109± 0.026, consistent with generalized demand pressures with generalized demand overheating. The localized nature of effects via the residual learner allows policymakers to trace out performance to specific behavioral patterns rather than to general shocks.

6. Conclusion

The current study has reinforced that it is possible to improve inflation nowcasting by integrating payment data on high-frequency payments in the ARIMA model with a semiparametric structure. The proposed approach can benefit from the structural components of ARIMA models while exploiting the machine learning technology with the advantage of being able to process non-linear data in order to offer accurate and meaningful results in real time.

The interpretability of the model, achieved through the decomposition of the predictable and shock components, makes it particularly useful for policy purposes. Future work can involve generalizing the model to panel data with multiple countries, integrating text data series (for instance, central bank communications), and investigating online learning algorithms for recursive model updates to perform continuous forecasts.

Author conribution

Yangyang Li and Mule Tai contributed equally to this paper.

References

[1] Bolivar, O. (2025). High-frequency inflation forecasting: A two-step machine learning methodology. *Latin American Journal of Central Banking*, 100172.
[2] Hsiao, Y. H., & Miao, E. W. (2021). Inflation Nowcasting Using High Frequency Price Data. *Jing Ji Lun Wen Cong Kan*, 49(3), 371-414.
[3] Knotek II, E. S., & Zaman, S. (2025). Nowcasting inflation. In Research Handbook on Inflation (pp. 475-496). Edward Elgar Publishing.
[4] Beck, G. W., Carstensen, K., Menz, J. O., Schnorrenberger, R., & Wieland, E. (2024). Nowcasting consumer price inflation using high-frequency scanner data: Evidence from Germany.
[5] Mariano, R. S., & Ozmucur, S. (2021). Predictive performance of mixed-frequency nowcasting and forecasting models (with application to Philippine inflation and GDP growth). *Journal of Quantitative Economics*, 19(Suppl 1), 383-400.
[6] Nappa, G. (2024). Nowcasting inflation indices using mixed data sampling (MIDAS) time-series models (Master's thesis, University of Malta).
[7] Aliaj, T., Ciganovic, M., & Tancioni, M. (2023). Nowcasting inflation with Lasso-regularized vector autoregressions and mixed frequency data. *Journal of Forecasting*, 42(3), 464-480.
[8] Štefl, J. (2024). High Frequency Price Index of Construction Materials: Nowcasting of Producers Prices.
[9] Yadav, V., & Das, A. (2023). Nowcasting inflation in India with daily crowd-sourced prices using dynamic factors and mixed frequency models. *Applied Economics Letters*, 30(2), 167-177.
[10] Roldán-Ferrín, F., & Parra-Polanía, J. A. (2025). Enhancing inflation nowcasting with online search data: a random forest application for Colombia.