Spurious Fit in Thin Markets: A Preprocessing Cautionary Tale Using Puerto Rico GO Spreads

Jorge A. Arroyo
Independent Researcher
arroyo.jorgeantonio@gmail.com

February 2026

Abstract

Post-restructuring Puerto Rico GO bond spreads achieve strong in-sample fit but fail dramatically in walk-forward evaluation. Using transaction-level yields from EMMA for the most actively traded Series 2022A maturities, matched to constant-maturity U.S. Treasury yields from FRED, I construct a monthly fiscal-stress signal from maturity-matched spreads in a thin municipal market. In a burn-in window, a one-lag predictive regression of Puerto Rico economic activity on the lagged spread signal produces a high in-sample R^2 (about 0.52) and a permutation-style p-value of 0.006. However, pseudo out-of-sample performance collapses: out-of-sample R^2 values are deeply negative (about -6.4 to -8.3), and an AR(1) benchmark dominates. A forensic preprocessing audit shows that carry-forward rules and EWMA smoothing do not materially affect conclusions, while the rolling standardization window is decisive. Short (126-day) rolling Z-scores induce an adaptive regressor that fits local regimes in the training period but does not generalize; extending the window to 252 days sharply improves walk-forward performance yet eliminates in-sample significance, indicating no stable predictive content. The results highlight that in thin markets, preprocessing choices can manufacture apparent predictability and walk-forward validation is essential.

1 Introduction

A standard premise in macro-finance is that asset prices are forward-looking and therefore may contain information about future economic activity. This premise motivates a large literature that uses high-frequency financial indicators to improve forecasts of lower-frequency macroeconomic outcomes Andreou et al. (2013); Ghysels et al. (2004). Credit spreads are a particularly salient candidate predictor: by construction, they embed market perceptions of risk and compensation for bearing that risk, and in thick corporate markets they have been shown to forecast real activity under disciplined pseudo out-of-sample evaluation schemes Faust et al. (2013).

Applying this logic to municipal bonds is tempting. Municipal debt is exposed to local fiscal conditions, and in principle municipal yields and spreads should aggregate information about future tax capacity, fiscal stress, and economic performance. At the same time, municipal markets differ sharply from corporate markets in microstructure. They are decentralized, dealer-intermediated, and characterized by substantial search frictions and infrequent trading Green et al. (2007). Sparse trading and stale pricing are documented features of fixed-income markets and are especially pronounced in municipals Choi et al. (2022); Craig et al. (2018), implying that measurement and preprocessing choices can be first-order determinants of empirical results.

I study these issues in the post-restructuring Puerto Rico general-obligation (GO) market. Puerto Rico provides a natural laboratory because its fiscal crisis and restructuring represent a sharp institutional break Medioli et al. (2022); U.S. Government Accountability Office (2025), and because prior work documents that credit risk and credit spreads were closely linked to real economic deterioration during the pre-default period Chari et al. (2017). The central question is whether a comparable relationship can be recovered in the post-restructuring regime, using newly issued recovery instruments traded in a thin municipal market.

My starting point is a simple predictive regression that maps a monthly fiscal-stress signal—constructed from maturity-matched Puerto Rico GO spreads—to the Puerto Rico Economic Activity Index (EAI). In a burn-in (training) window, this specification appears highly promising: the fiscal-spread regressor delivers a large in-sample coefficient of determination ($R^2 \approx 0.52$) and a statistically significant slope estimate, with a permutation-style p-value of 0.006. The puzzle emerges immediately in real time. When I evaluate the same mapping using a walk-forward procedure, performance collapses. Out-of-sample R^2 values become extremely negative (approximately -6 to -8 across evaluation windows), implying that the model increases mean-squared prediction error by a wide margin relative to even a constant-mean forecast. Moreover, a simple AR(1) benchmark for EAI—the natural prevailing forecast for

a highly persistent macro series—dramatically outperforms the fiscal-spread model West (2006); Hubrich and West (2010). The combination of strong in-sample fit and catastrophic walk-forward failure is the central empirical fact the paper seeks to explain.

This pattern echoes classic warnings in time-series econometrics: regressions can exhibit high in-sample fit and "significance" even when the underlying relationship is illusory Granger and Newbold (1974), and persistence can further exaggerate apparent evidence for predictability in small samples Stambaugh (1999). Yet the magnitude of the walk-forward failure in my setting is unusual. Before drawing substantive conclusions about whether post-restructuring municipal spreads contain information about future economic activity, it is essential to understand why the relationship looks strong in-sample but disintegrates out-of-sample.

I consider three broad explanations. First, the relationship could be genuinely unstable: the mapping from fiscal spreads to economic activity might differ across regimes and break in the post-restructuring period. Second, the predictor could be contaminated by thin-market measurement error: nonsynchronous trading and stale pricing can generate spurious serial dependence and false predictability Lo and MacKinlay (1989); Choi et al. (2022). Third, the relationship could be an artifact of preprocessing: when the underlying series are sparse and noisy, the transformations used to create a usable regressor can themselves manufacture apparent signal. This paper focuses on the third explanation and asks a targeted question: *which preprocessing component, if any, is responsible for producing the strong in-sample relationship and the catastrophic out-of-sample failure?*

My approach is deliberately simple. The predictive model is a one-lag regression of monthly EAI on the lagged fiscal-spread signal, evaluated against an AR(1) benchmark. The key methodological ingredient is the evaluation design: I treat walk-forward (pseudo out-of-sample) performance as the primary diagnostic, consistent with the forecast-evaluation literature emphasizing that in-sample fit can fail to translate into predictive ability, especially in nested-model comparisons West (2006); Clark and West (2007). Formal comparisons can be framed in terms of MSPE differentials Diebold and Mariano (1994), but the headline evidence in my setting is already decisive: the fiscal-spread model performs dramatically worse than the benchmark out-of-sample.

To isolate the mechanism, I conduct a one-at-a-time preprocessing sensitivity analysis under the same walk-forward design and benchmark. Starting from a baseline signal construction (daily maturity-matched spreads, composite aggregation across bonds, carry-forward treatment of non-trading days, EWMA smoothing, rolling Z-score standardization, and monthly aggregation), I vary three components: (i) the carry-forward protocol (uncapped carry-forward versus no carry-forward); (ii) EWMA smoothing (enabled versus disabled);

and (iii) the rolling standardization window (a short 126-business-day Z-score window versus a conventional 252-business-day window). This "forensic" design distinguishes between two qualitatively different sources of failure. If the predictive relationship is economically meaningful but contaminated by thin-trading noise, then modest changes to imputation and smoothing should materially alter results. If, instead, the relationship is induced by adaptive transformations that overfit local regimes, then normalization choices should be decisive.

Four findings emerge. First, the baseline in-sample relationship is not stable: the fiscal-spread model delivers a strong burn-in fit but fails catastrophically under walk-forward validation, while the AR(1) benchmark dominates. Second, carry-forward rules do not materially affect conclusions. Under the monthly aggregation procedure used to align daily spreads to monthly EAI, differences in daily carry-forward implementation are largely attenuated. Third, EWMA smoothing is not the source of spurious fit and may modestly stabilize the regressor; disabling smoothing does not repair out-of-sample performance. Fourth, and most importantly, the rolling Z-score window length is the smoking gun. Short (126-day) standardization windows make the regressor highly adaptive to local mean and volatility regimes, enabling a spurious alignment with the burn-in sample. Extending the window to 252 days dramatically reduces the severity of the walk-forward collapse but simultaneously eliminates statistical significance and does not produce a positive, stable forecasting relationship. Taken together, these results indicate that the apparent in-sample predictability is primarily preprocessing-induced overfitting rather than genuine information content.

The paper makes four contributions. First, it provides a negative result with positive value: post-restructuring Puerto Rico GO spreads, constructed from thinly traded municipal bonds, do not reliably predict monthly economic activity beyond persistence benchmarks. Second, it identifies a concrete mechanism for spurious fit in thin markets: adaptive rolling standardization windows can dominate empirical outcomes, creating the appearance of predictability in small samples even when none exists. Third, it emphasizes best practices for macro-finance forecasting in settings with sparse financial data. Walk-forward validation should be treated as essential rather than optional, and preprocessing choices—especially normalization windows—should be audited for sensitivity. This complements broader warnings about predictor mining and forecast breakdowns in macroeconomic forecasting Faust et al. (2013). Fourth, it adds nuance to the broader mixed-frequency forecasting literature. While high-frequency financial data can improve macro forecasts in liquid markets Andreou et al. (2013); Ghysels et al. (2004), thin-market microstructure and preprocessing sensitivity can overturn that promise in municipal settings.

The remainder of the paper proceeds as follows. Section 2 describes the data sources

and constructs the fiscal-stress signal from post-restructuring Puerto Rico GO spreads, including thin-trading diagnostics. Section 3 presents the baseline predictive regression results and documents the discrepancy between strong burn-in fit and catastrophic walk-forward performance relative to an AR(1) benchmark. Section 4 conducts the preprocessing forensics and isolates the rolling Z-score window as the primary driver of spurious fit. Section 5 interprets the mechanism, relates the findings to the literature, and draws practical implications. Section 6 concludes.

2 Data and Signal Construction

2.1 Data Sources

My fiscal-risk data are drawn from post-restructuring Puerto Rico General Obligation (GO) bonds in the MSRB Electronic Municipal Market Access (EMMA) system (Chirinko et al., 2018). I focus on the most actively traded maturities in the Series 2022A cohort (first available on March 15, 2022) (Medioli et al., 2022): a 2031 GO (CUSIP 74514L3J4), a 2037 GO (CUSIP 74514L3M7), and a 2046 GO (CUSIP 74514L3P0). For each CUSIP, I use transaction-level information to construct daily volume-weighted average yields (VWAY).

To compute maturity-matched spreads, I pair each GO maturity with a constant-maturity U.S. Treasury benchmark from FRED: DGS7 for the 2031, DGS10 for the 2037, and DGS20 for the 2046.[1] The macroeconomic outcome is the Puerto Rico Economic Activity Index (EAI), observed monthly (Department of Economic Development and Commerce, 2026; U.S. Government Accountability Office, 2025).

Table 1 summarizes sample spans and highlights the frequency mismatch between daily financial series and the monthly macro target. In the absence of dedicated mixed-frequency methods, I align the series through temporal aggregation, following the common practice in the mixed-frequency forecasting literature (Andreou et al., 2013; Ghysels et al., 2004).

2.2 Spread Construction and Fiscal Stress Signal

Daily maturity-matched spreads. Let $y_{j,t}^{GO}$ denote the daily VWAY for GO maturity $j \in \{2031, 2037, 2046\}$ on business day t, and let $y_{m(j),t}^{UST}$ denote the corresponding Treasury yield (DGS7/DGS10/DGS20). The raw daily spread is

$$s_{j,t} \;=\; y_{j,t}^{GO} - y_{m(j),t}^{UST}. \tag{1}$$

[1] Treasury constant-maturity series are from the Federal Reserve Economic Data (FRED) system; see the series documentation for DGS7, DGS10, and DGS20.

Table 1: Sample characteristics and data availability.

Series	Frequency	Period	N	Coverage / thin-trading statistics
PR GO 2031 (74514L3J4)	Daily (trades)	2022-03-15–2026-02-17	1026	12.28% zero-trade; median no-trade run 1.0d; max 6d
PR GO 2037 (74514L3M7)	Daily (trades)	2022-03-15–2026-02-17	1026	14.72% zero-trade; median no-trade run 1.0d; max 15d
PR GO 2046 (74514L3P0)	Daily (trades)	2022-03-15–2026-02-17	1026	13.84% zero-trade; median no-trade run 1.0d; max 23d
UST DGS7/DGS10/DGS20	Daily	2010-01-01–2026-02-12	4205	—
EAI, ΔEAI (N=316)	Monthly	1999-07-31–2025-11-30	317	—

Thin-trading protocol and composite spread. Municipal bonds trade in a decentralized, dealer-intermediated OTC market, and observed prices can be stale and unevenly updated across securities (Green et al., 2007; Craig et al., 2018; Choi et al., 2022). To produce a daily series, I apply a carry-forward rule: when a given CUSIP does not trade on day t, I set its spread to the most recently observed value.[2] I then aggregate across maturities using a daily cross-sectional median,

$$s_t \;=\; \text{median}\{s_{2031,t},\; s_{2037,t},\; s_{2046,t}\}, \tag{2}$$

which reduces sensitivity to idiosyncratic pricing noise and maturity-specific quotation irregularities.

Smoothing, standardization, and temporal aggregation. I optionally smooth the daily composite spread s_t using an EWMA filter (baseline span = 10 business days). I then standardize the (smoothed) spread using a rolling Z-score,

$$z_t \;=\; \frac{s_t - \mu_{t-1}(s)}{\sigma_{t-1}(s)}, \tag{3}$$

where $\mu_{t-1}(s)$ and $\sigma_{t-1}(s)$ are the mean and standard deviation computed over a trailing window of length W ending at $t-1$ (baseline $W = 126$ business days). The $t-1$ convention avoids look-ahead by ensuring that the normalization on day t uses only information available prior to t.

Finally, to align with the monthly EAI target, I aggregate the daily standardized signal

[2] In the baseline configuration, carry-forward is uncapped (no maximum carry duration). This choice is convenient for daily alignment and is audited in Section 4.

to month m using a flat within-month average:

$$RT_m = \frac{1}{|D(m)|} \sum_{t \in D(m)} z_t, \tag{4}$$

where $D(m)$ denotes the set of business days in month m (Andreou et al., 2013; Ghysels et al., 2004). Figure 1 plots the daily composite spread and the standardized signal and indicates the burn-in and evaluation windows used in subsequent sections.

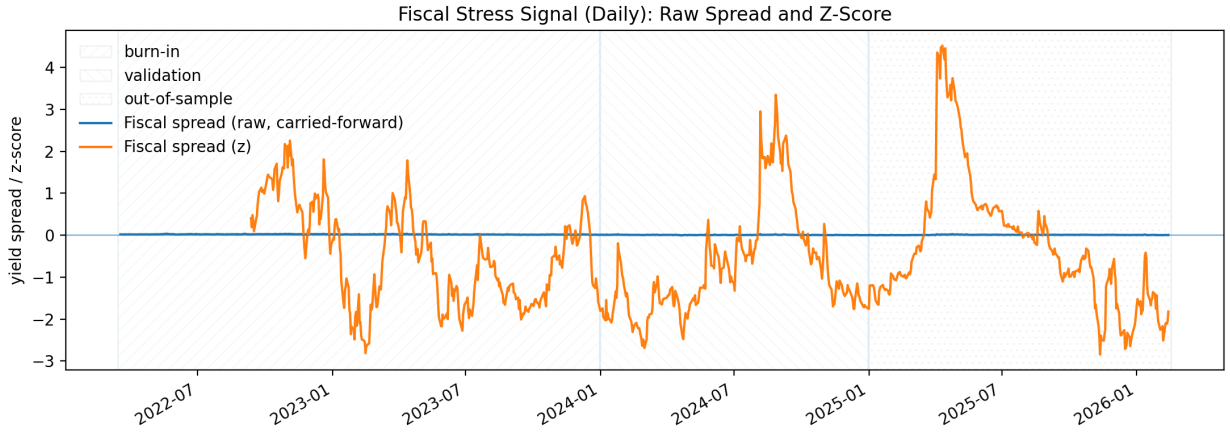

Figure 1: Fiscal stress signal (daily): raw composite spread and rolling Z-score. Shaded regions denote the burn-in, validation, and out-of-sample windows.

2.3 Thin-Trading Diagnostics

Carry-forward rules can interact with sparse trading and nonsynchronous price-setting to induce mechanical persistence and spurious serial dependence in daily series (Lo and MacKinlay, 1989; Choi et al., 2022). To quantify market thinness, I compute a daily coverage ratio: the fraction of the three CUSIPs that trade on day t. Figure 2 plots this coverage ratio over time, and Table 2 reports summary diagnostics for the composite construction.

Two features are worth emphasizing. First, although the mean coverage ratio is relatively high (0.866), there remains a nontrivial mass of days with incomplete trading across the three CUSIPs. This motivates both the composite-median aggregation and the explicit thin-trading protocol (Green et al., 2007; Craig et al., 2018; Wu, 2025). Second, because inference and forecasting are conducted on the monthly aggregate RT_m, some daily-level protocol differences (such as the fine details of carry-forward implementation) may be attenuated by temporal aggregation. This observation motivates the one-at-a-time preprocessing forensics in Section 4.

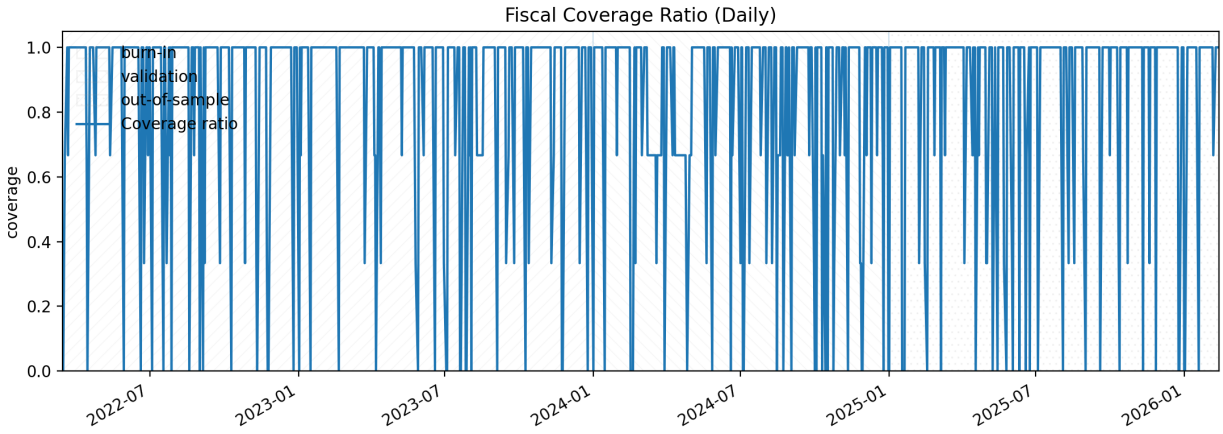

Figure 2: Fiscal coverage ratio (daily): fraction of configured GO CUSIPs trading on each business day.

Table 2: Thin-trading diagnostics for the composite fiscal spread (daily).

Diagnostic	Value	Notes
Zero-trade days (%)	8.11	No configured bonds trade
Stale days (%)	4.1	Only one bond trades (multi-bond mode)
Median carry duration (days)	1	Consecutive carry-forward length (applied)
Max consecutive carry (days)	4	Capped at MAX_FISCAL_CARRY_DAYS
Mean coverage ratio	0.866	Average fraction trading

3 Empirical Procedure and Baseline Results

This section asks whether the post-restructuring Puerto Rico GO fiscal-spread signal contains incremental predictive information for Puerto Rico economic activity beyond simple persistence. Throughout this section I use the baseline preprocessing configuration described in Section 2: (i) uncapped carry-forward (`carry=None`); (ii) EWMA smoothing enabled (`ewma=True`, span = 10 business days); and (iii) rolling standardization using a 126-business-day window ($W = 126$).

3.1 Evaluation windows and the prevailing benchmark

I partition the post-restructuring sample into an initial burn-in (training) period and two subsequent walk-forward evaluation periods (validation and out-of-sample). Table 3 reports the calendar spans used to define these regimes.

Because the Economic Activity Index (EAI) is highly persistent at the monthly frequency,

Table 3: Evaluation windows (calendar definitions).

Window	Start	End	Months	Purpose
Burn-in	2022-03-15	2023-12-31	22	Initial estimation / mapping calibration
Validation	2024-01-01	2024-12-31	12	Walk-forward evaluation I
Out-of-sample	2025-01-01	2026-02-17	14	Walk-forward evaluation II

the relevant benchmark is a univariate autoregression (Department of Economic Development and Commerce, 2026; U.S. Government Accountability Office, 2025; Hubrich and West, 2010). In the spirit of a "prevailing forecast" benchmark, I treat an AR(1) model as the standard that any fiscal-spread predictor must beat in walk-forward evaluation (Hubrich and West, 2010; West, 2006).

3.2 Models

Let y_m denote the level of EAI in month m and let RT_m denote the monthly fiscal-stress signal constructed in Section 2. I consider two one-step-ahead forecasting models.

Fiscal-spread model. The baseline predictive regression is a one-lag distributed-lag specification,

$$y_m = \alpha + \beta\, RT_{m-1} + \varepsilon_m. \tag{5}$$

AR(1) benchmark. The benchmark is the prevailing AR(1),

$$y_m = \alpha + \phi\, y_{m-1} + u_m. \tag{6}$$

3.3 Walk-forward protocol and out-of-sample performance

Forecasts are generated in a walk-forward manner. At each evaluation month, parameters are estimated using only data available up to that month, and a one-step-ahead forecast is produced (West, 2006; Clark and West, 2007; Hubrich and West, 2010). I summarize predictive accuracy using an MSPE-based out-of-sample R^2,

$$R^2_{\text{OOS}} \;=\; 1 - \frac{\sum_{m\in\mathcal{T}}(y_m - \widehat{y}_m)^2}{\sum_{m\in\mathcal{T}}(y_m - \overline{y}_{\mathcal{T}})^2}, \tag{7}$$

where $\mathcal{T}$ denotes the evaluation window and $\overline{y}_{\mathcal{T}}$ is the sample mean of y_m over $\mathcal{T}$. Negative values indicate forecasts that are worse than a constant-mean forecast over that window (West, 2006; Clark and West, 2007).

3.4 Burn-in evidence: strong fit, but not relative to persistence

I begin with burn-in (in-sample) performance. Table 4 reports static fit for the fiscal-spread model (5) and the AR(1) benchmark (6) under the baseline preprocessing (carry=None, EWMA=True, $W = 126$). The fiscal-spread regression attains a high in-sample R^2 (0.517) and a statistically significant slope estimate ($\widehat{\beta} = -1.435$, $p = 0.0026$). However, persistence dominates: the AR(1) benchmark fits the burn-in sample substantially better ($R^2 = 0.905$; RMSE 0.636 versus 1.431 for DL(1)), consistent with the high monthly persistence of EAI (Department of Economic Development and Commerce, 2026; Hubrich and West, 2010).

Figure 3 provides the visual counterpart to Table 4. The AR(1) fitted values track the level of EAI almost point-for-point over the burn-in window. The fiscal-spread model also produces a close in-sample tracking, but with noticeably larger deviations at several dates (most visibly around early 2023), which is reflected in its higher RMSE despite the seemingly strong R^2.

Table 4: Burn-in static fit (EAI level). Baseline preprocessing: carry=None, EWMA=True, $W = 126$.

Model	n	R^2	RMSE	Key coefficient	p-value
DL(1): $y_m = \alpha + \beta RT_{m-1}$	15	0.517	1.431	$\widehat{\beta} = -1.435$	0.0026
AR(1): $y_m = \alpha + \phi y_{m-1}$	15	0.905	0.636	–	< 0.001

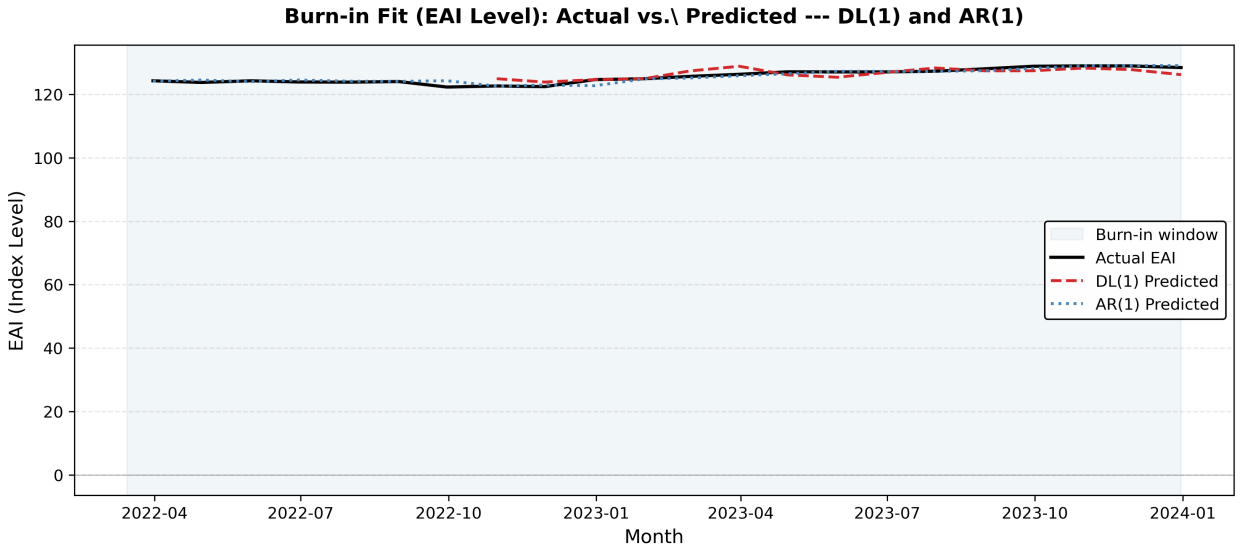

Figure 3: Burn-in fit (EAI level): actual versus fitted values under DL(1) and AR(1). The burn-in window spans 2022:04–2023:12. AR(1) tracks the highly persistent EAI series closely; DL(1) also fits in-sample but with larger deviations at several dates, consistent with its higher RMSE in Table 4.

3.5 Walk-forward evidence: large and persistent forecast failure

The central empirical fact is that the fiscal-spread model does not survive walk-forward evaluation. Table 5 reports pseudo out-of-sample performance in the validation and out-of-sample windows. The fiscal-spread model produces extremely negative out-of-sample R^2 values, indicating MSPE far worse than a constant-mean forecast over those windows. In contrast, the AR(1) benchmark performs well in validation and remains far more stable out-of-sample (West, 2006; Clark and West, 2007; Hubrich and West, 2010).

Table 5: Walk-forward performance (EAI level). Baseline preprocessing: carry=None, EWMA=True, $W = 126$.

Model	Window	n	R^2_{OOS}	RMSE
DL(1)	Validation (2024)	12	-6.442	1.766
DL(1)	Out-of-sample (2025–2026)	11	-8.270	1.332
AR(1)	Validation (2024)	12	+0.669	0.372
AR(1)	Out-of-sample (2025–2026)	11	-0.080	0.455

3.6 A Goyal–Welch diagnostic: cumulative squared error differences

To identify when the fiscal-spread model underperforms the benchmark, I follow Goyal and Welch (2008) and plot the cumulative difference in squared one-step-ahead prediction errors (West, 2006):

$$\Delta\text{SPE}(T) \;=\; \sum_{m \le T} \left(e^2_{\text{AR(1)},m} - e^2_{\text{DL},m} \right), \tag{8}$$

where $e_{\text{AR(1)},m} = y_m - \widehat{y}_m^{AR(1)}$ and $e_{\text{DL},m} = y_m - \widehat{y}_m^{DL}$ are walk-forward forecast errors. With this sign convention, upward movements indicate months in which the fiscal-spread model reduces squared error relative to AR(1), while downward movements indicate benchmark dominance.

Figure 4 shows that $\Delta\text{SPE}(T)$ is negative throughout the evaluation sample and becomes more negative in two pronounced episodes. The series begins the validation window already below zero (about -7.4 at 2024:01) and remains near -7 through mid-2024, with only small month-to-month reversals. Benchmark dominance then accelerates sharply in late 2024: the cumulative differential drops by roughly 19 units in 2024:09 (from about -11.1 to -30.2), and reaches about -35.0 by 2024:10. After a period of slower drift in early 2025, a second large deterioration occurs in 2025:05 (a drop of roughly 11.8 units, from about -38.0 to -49.8). The cumulative differential continues to decline thereafter, ending near -53.0 by 2025:11. These dynamics provide a time-localized view of the same message conveyed by the

window-level R^2_{OOS} statistics: the fiscal-spread model loses persistently and by a wide margin in real time.

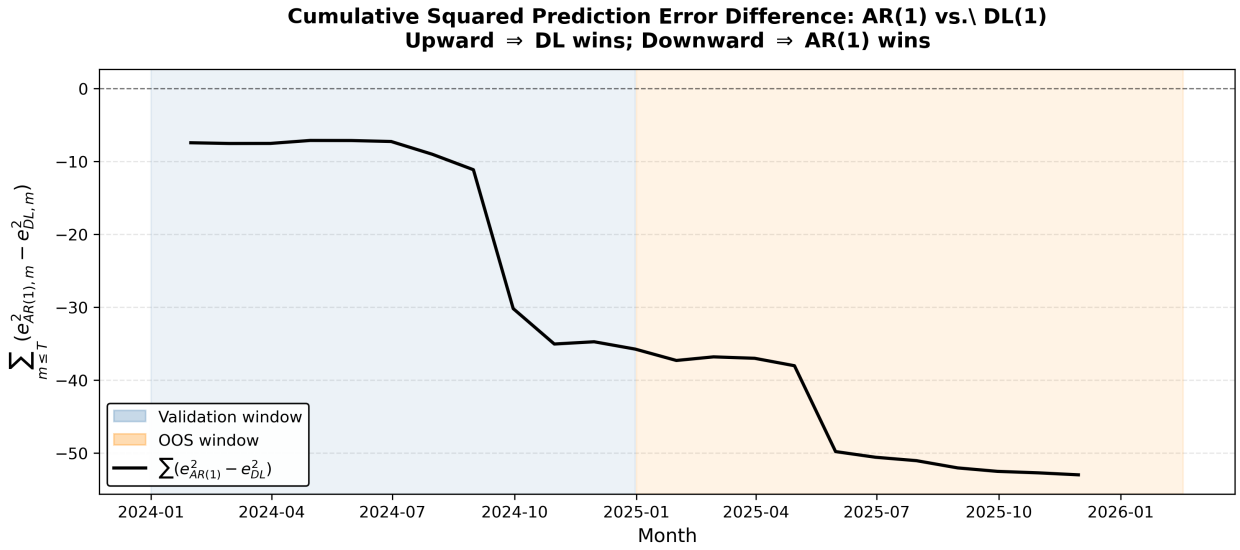

Figure 4: Cumulative squared prediction error difference between AR(1) and DL(1), $\sum_{m \le T} \left(e^2_{\text{AR}(1),m} - e^2_{\text{DL},m} \right)$. Upward movements indicate months where DL(1) improves on AR(1); downward movements indicate AR(1) dominance. Shaded regions denote the validation and out-of-sample windows.

3.7 Interpretation

The baseline results therefore sharpen the puzzle motivating the preprocessing forensics in Section 4. The fiscal-spread signal looks informative in the burn-in period, yet walk-forward performance is dominated by a simple persistence benchmark and the cumulative error differential drifts steadily downward, with discrete "cliffs" in late 2024 and mid-2025 (Figure 4). This pattern is consistent with the view that the apparent in-sample relationship does not reflect a stable mapping that generalizes out-of-sample (Granger and Newbold, 1974; Stambaugh, 1999; West, 2006; Clark and West, 2007). The remainder of the paper isolates which preprocessing component(s) generate the burn-in fit and the extreme pseudo out-of-sample instability.

4 Preprocessing Forensics

Section 3 documents the central empirical pattern: under the baseline configuration (`carry=None`, `EWMA=True`, $W = 126$), the fiscal-spread regression fits the burn-in period well yet fails

dramatically in walk-forward evaluation. This section conducts a one-at-a-time preprocessing audit to determine which transformation is responsible for the apparent in-sample relationship and the extreme out-of-sample instability (Granger and Newbold, 1974; West, 2006; Clark and West, 2007).

The design is intentionally diagnostic. I hold the forecasting regression fixed (DL(1) in (5)) and vary only a single preprocessing component, leaving all remaining steps at their baseline values. For each configuration, I report burn-in (static) fit and pseudo out-of-sample performance in the validation and out-of-sample windows. The question is not which model performs best in-sample, but which preprocessing change materially alters the out-of-sample failure mode.

4.1 Test A: Carry-forward protocol

Design. The baseline daily spread construction uses carry-forward to fill non-trading days. To assess whether this imputation rule is mechanically generating persistence and spurious fit, Test A eliminates carry-forward by setting the cap to zero (trade-days only).

Result. Test A produces essentially no change in burn-in fit or walk-forward performance (Table 7). In this application, where the forecasting exercise is conducted on the monthly aggregate RT_m, the details of daily carry-forward appear to be largely washed out by temporal aggregation.

4.2 Test B: EWMA smoothing

Design. The baseline applies EWMA smoothing (span = 10 business days) prior to rolling standardization. Test B disables smoothing.

Result. Disabling EWMA does not repair the walk-forward failure and can worsen out-of-sample performance (Table 7). The evidence therefore does not support smoothing as the driver of the baseline pattern; if anything, smoothing modestly stabilizes the constructed signal.

4.3 Test C: Rolling Z-score window length

Design. The baseline standardizes the daily composite spread using a short trailing window ($W = 126$ business days). Test C increases the window length to $W = 252$ business days, holding all other steps fixed (carry=None, EWMA=True).

Result. Table 6 shows that the normalization window materially changes the empirical conclusions. The short-window configuration produces a high-R^2 burn-in regression and a statistically significant slope estimate, but extremely negative walk-forward R^2_{OOS} values. Under the longer 252-day window, walk-forward performance improves sharply (validation $R^2_{\text{OOS}} \approx -0.01$ and out-of-sample $R^2_{\text{OOS}} \approx -1.01$). At the same time, the burn-in relationship weakens: statistical significance disappears (permutation $p = 0.08$), and the effective burn-in sample shrinks because the longer window requires a larger initial lookback.

Table 6: Z-score window sensitivity: baseline (126 days) vs. conservative (252 days).

Metric	126-day (Baseline)	252-day (Conservative)	Change
Burn-in			
Static R^2	0.517*	0.387	-0.130
RMSE	1.43	0.63	-0.80
Sample N	15	9	-6
Walk-forward validation			
R^2_{OOS}	-6.442	-0.013	+6.430
RMSE	1.77	0.60	-1.17
Walk-forward out-of-sample			
R^2_{OOS}	-8.270	-1.006	+7.264
RMSE	1.33	0.62	-0.71
Permutation test			
p-value	0.006*	0.080 (n.s.)	–

Figure 5 summarizes the sensitivity in Table 6. The improvement in R^2_{OOS} when moving from $W = 126$ to $W = 252$ indicates that short-window standardization is a primary driver of the baseline out-of-sample collapse. However, even the conservative choice does not deliver positive and stable out-of-sample performance.

Interpretation. Rolling standardization with a short window makes the transformed regressor highly adaptive to local mean and volatility regimes. In small samples, this adaptivity can generate a variable that aligns closely with the burn-in period yet does not represent a stable mapping that persists out-of-sample. Lengthening the window reduces this adaptivity, attenuating the apparent in-sample relationship and substantially reducing the severity of the walk-forward failure. This interpretation is consistent with the forecast-evaluation literature emphasizing that estimation noise and instability can inflate pseudo out-of-sample prediction error in finite samples, particularly in nested comparisons (West, 2006; Clark and West, 2007).

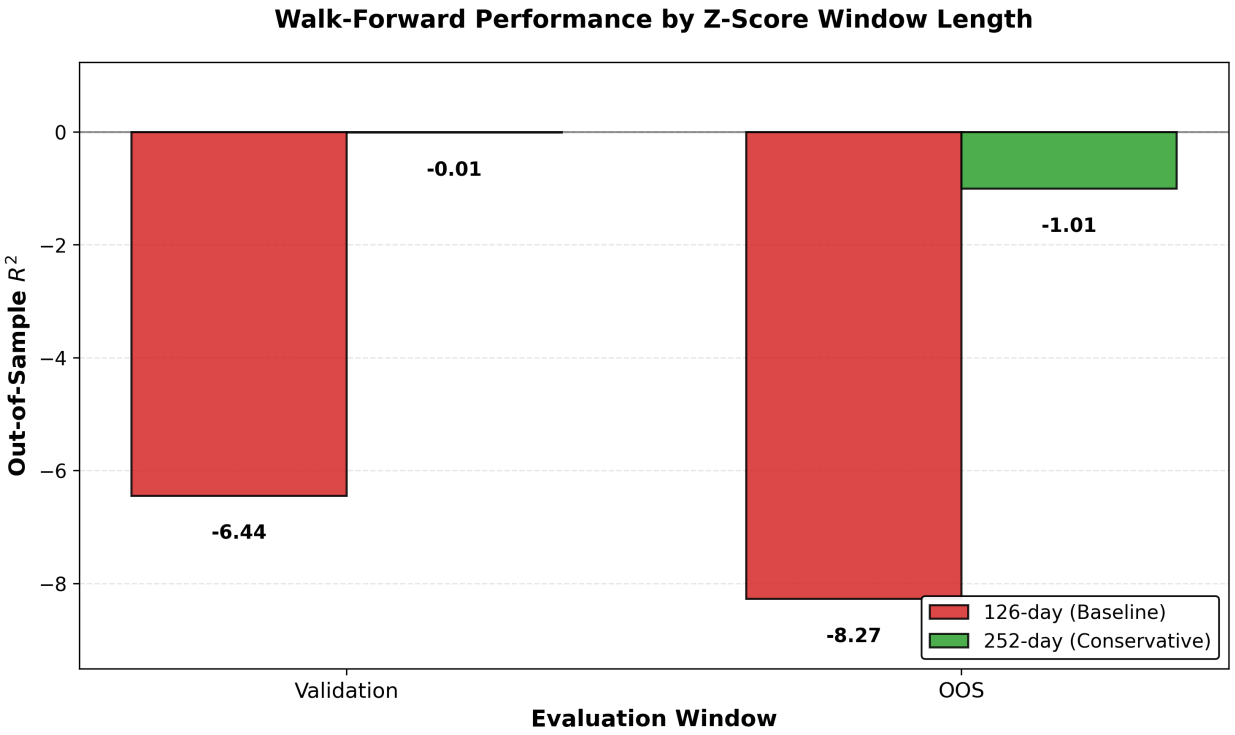

Figure 5: Walk-forward performance by rolling Z-score window length. Bars report out-of-sample R^2_{OOS} for the validation and out-of-sample windows under $W = 126$ (baseline) and $W = 252$ (conservative).

4.4 Summary

Table 7 consolidates the three tests. In this design, carry-forward details have negligible effects once the signal is aggregated to the monthly frequency. EWMA smoothing is not responsible for the baseline pattern and may modestly stabilize the constructed series. By contrast, the rolling Z-score window length is the dominant determinant of both the apparent burn-in significance and the magnitude of the walk-forward failure.

Table 7: Preprocessing robustness summary (one-at-a-time variations).

Parameter tested	Configuration	Val R^2_{OOS}	OOS R^2_{OOS}	Burn-in R^2	Sample N	Conclusion
Carry-forward cap	Baseline: Unlimited	-6.44	-8.27	0.517	15	–
	Test A: 0 days	-6.44	-8.27	0.517	15	Negligible effect
EWMA smoothing	Baseline: Enabled	-6.44	-8.27	0.517	15	–
	Test B: Disabled	-6.16	-12.25	0.546	15	Not the driver; may stabilize
Z-score window	Baseline: 126 days	-6.44	-8.27	0.517*	15	High sensitivity
	Test C: 252 days	-0.01	-1.01	0.387	9	Short window drives IS/OOS divergence

5 Discussion

Sections 3–4 deliver a consistent message. In the post-restructuring sample and at the monthly horizon, the Puerto Rico GO spread signal does not yield robust incremental forecasting power for Puerto Rico economic activity relative to a simple persistence benchmark. The apparent in-sample relationship is highly sensitive to preprocessing, with the dominant sensitivity concentrated in the rolling standardization step. This section interprets the mechanism suggested by the forensics, explains why more conservative choices do not generate stable predictability, and relates the findings to the broader macro-finance forecasting literature.

5.1 Mechanism: instability induced by adaptive normalization

The baseline results exhibit a familiar empirical pattern: strong burn-in fit and apparently significant coefficients coincide with sharp deterioration in walk-forward forecasting performance. In time-series settings, such outcomes can arise when conventional inference overstates evidence for a relationship because key assumptions are violated (Granger and Newbold, 1974). In predictive regressions, small-sample distortions can be particularly important when regressors are persistent and correlated with the regression disturbance (Stambaugh, 1999).

In my setting, the evidence points to a distinct but related mechanism. The instability is not driven primarily by the raw spread level; rather, it emerges from an adaptive transformation of the spread. Short-window rolling Z-scores make the standardized series highly sensitive to local shifts in mean and volatility. With limited burn-in data, this adaptivity can produce a transformed regressor that tracks regime-specific variation in the training period, even if no stable mapping exists between spreads and subsequent economic activity. The

resulting predictor can therefore look informative in-sample while failing to generalize.

This interpretation is consistent with standard forecast-evaluation logic for nested models. When the true incremental predictive content is weak or absent, estimating additional parameters introduces noise that raises out-of-sample MSPE for the larger model (West, 2006; Clark and West, 2007; Hubrich and West, 2010). From this perspective, strongly negative R^2_{OOS} values are not paradoxical; they can occur when an unstable predictor is introduced in a finite sample and evaluated in real time.

5.2 Why thin markets can magnify preprocessing artifacts

Municipal bond markets are decentralized, dealer-intermediated, and illiquid relative to exchange-traded assets (Green et al., 2007). Sparse trading and stale pricing are structural features of fixed-income markets and are especially pronounced in municipals (Choi et al., 2022; Craig et al., 2018). Nonsynchronous trading can induce mechanical serial dependence and cross-dependence in observed prices that is not directly tied to fundamentals (Lo and MacKinlay, 1989). These frictions motivate imputation and smoothing in data construction, but they also create conditions under which adaptive transformations can manufacture apparent "signal."

In my application, the daily fiscal spread is constructed from a small set of bonds with observable non-trading episodes (Section 2.3). The forensics indicate that carry-forward and EWMA smoothing are not the primary drivers of the baseline IS–OOS discrepancy (Table 7). Instead, the key interaction appears to be between thin-market-induced persistence/noise in the daily spread and a short normalization window that amplifies local variation into large standardized movements. This interaction makes the predictor's statistical properties highly sample-dependent and helps explain why the burn-in relationship does not survive walk-forward validation.

5.3 Why conservative Z-windows do not recover stable predictability

Increasing the Z-score window from 126 to 252 trading days sharply reduces the severity of the walk-forward collapse (Table 6 and Figure 5). This sensitivity is informative: it indicates that the most pathological out-of-sample behavior is largely induced by short-window adaptivity. At the same time, the conservative window does not produce reliably positive R^2_{OOS} values, and the burn-in relationship becomes statistically weaker.

Taken together, these results support a simple reading of the forensics. The short window is sufficiently adaptive to fit transient co-movements between the standardized spread and

EAI during the burn-in period; when conditions change, that mapping breaks down. The longer window reduces regime-tracking capacity and therefore reduces instability, but it also reveals that any underlying relationship is not strong enough to deliver robust incremental forecasting gains over persistence. Conservative preprocessing can mitigate overfitting; it cannot create signal that is not present in the data.

5.4 Relation to the literature

Two strands of evidence provide context. First, corporate credit spreads in thick markets have been shown to forecast real activity when measured carefully and evaluated using pseudo out-of-sample designs (Faust et al., 2013). This aligns with the macro-finance premise that market prices may incorporate information about future macroeconomic outcomes (Andreou et al., 2013). Second, the same literature emphasizes that apparent predictability is vulnerable to instability, data mining, and post-documentation breakdowns (Faust et al., 2013). My findings illustrate a distinct channel for such breakdowns in thin markets: preprocessing choices can materially shape inference and pseudo out-of-sample performance, and can generate the appearance of predictability in small samples.

Puerto Rico also provides a useful historical contrast. During the pre-default period, Puerto Rico credit risk and spreads were closely linked to real economic deterioration (Chari et al., 2017). In the post-restructuring period, the informational content of GO spreads may be attenuated by institutional features of the new instruments and by market microstructure, including sparse and opaque trading (Medioli et al., 2022; Green et al., 2007). More broadly, official sources emphasize that Puerto Rico's post-restructuring fiscal environment and economic outlook differ meaningfully from the pre-default regime (U.S. Government Accountability Office, 2025). In such conditions, it is plausible that a simple spread-based predictor does not outperform a persistence benchmark in real time.

5.5 Implications

Two implications follow.

For researchers. Pseudo out-of-sample validation should be treated as a central diagnostic rather than a secondary robustness check, particularly when predictors are persistent or constructed through adaptive transformations (West, 2006; Clark and West, 2007; Diebold and Mariano, 1994). In addition, preprocessing choices—especially rolling standardization windows—should be reported transparently and audited for sensitivity. In thin markets, where raw series are sparse and noisy, preprocessing can dominate empirical outcomes.

For practitioners. For practical forecasting of Puerto Rico economic activity in the post-restructuring period, GO spreads constructed from thinly traded issues do not appear to provide a reliable standalone leading indicator in my sample. Simple persistence benchmarks can dominate in real time, and strong in-sample relationships may reflect data handling rather than stable information aggregation.

5.6 Limitations and future directions

My analysis is intentionally narrow. It focuses on a short post-restructuring window with limited monthly observations and on a parsimonious mapping from a composite GO spread to EAI. Future work could consider mixed-frequency models that use daily information more efficiently than flat aggregation (Ghysels et al., 2004, 2006), incorporate broader measures of financial conditions, or explicitly model thin-trading processes. Such extensions may improve forecasting performance in principle, but the core lesson remains: without disciplined walk-forward evaluation and preprocessing audits, apparent predictability can be fragile and may be induced by transformation choices rather than underlying economic information.

6 Conclusion

This paper examines whether post-restructuring Puerto Rico GO bond spreads forecast Puerto Rico economic activity at the monthly horizon. Under the baseline signal construction, the fiscal-spread predictor appears informative in the burn-in period, delivering a high in-sample R^2 and statistically significant slope estimates. However, the relationship does not survive real-time evaluation. In walk-forward forecasting, out-of-sample R^2 values are sharply negative, and a simple AR(1) persistence benchmark dominates.

A one-at-a-time preprocessing audit clarifies the source of this discrepancy. Varying carry-forward rules and EWMA smoothing has negligible effects on pseudo out-of-sample performance. By contrast, the rolling standardization window is pivotal. Short (126-day) Z-score windows yield a highly adaptive transformed regressor that can fit local mean and volatility regimes in small samples, producing an in-sample relationship that does not generalize. Extending the window to a conventional 252-day horizon substantially reduces the severity of the walk-forward collapse, but it also weakens the burn-in evidence and does not produce positive, stable out-of-sample forecasting gains.

The empirical conclusion is therefore negative but informative. In the post-restructuring period and in my sample, Puerto Rico GO spreads constructed from thinly traded bonds do not provide robust incremental predictive content for monthly economic activity beyond

simple persistence. The broader lesson is methodological. In thin markets, preprocessing choices can materially shape inference and forecast performance, and disciplined walk-forward validation is essential even when in-sample fit and permutation-style significance tests appear compelling. These findings reinforce the general warning that time-series regressions can generate statistical mirages (Granger and Newbold, 1974) and highlight, in a modern forecasting setting, how estimation noise and instability can degrade pseudo out-of-sample accuracy in nested comparisons (West, 2006; Clark and West, 2007).

References

Andreou, E., Ghysels, E., and Kourtellos, A. (2013). Should macroeconomic forecasters use daily financial data and how? *Journal of Business & Economic Statistics*, 31(2):240–251.

Chari, A., Leary, R., and Phan, T. (2017). The costs of (sub) sovereign default risk: Evidence from puerto rico. Working Paper 24108, National Bureau of Economic Research.

Chirinko, R., Chiu, R., and Henderson, S. (2018). What went wrong?: The puerto rican debt crisis and the "treasury put". Working paper, University of Illinois at Chicago and CESifo.

Choi, J., Kronlund, M., and Oh, J. Y. J. (2022). Sitting bucks: Stale pricing in fixed income funds. Working paper, University of Illinois at Urbana-Champaign, Tulane University, and Hanyang University.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311.

Craig, L., Kim, A., and Woo, S. W. (2018). Pre-trade information in the municipal bond market. White paper, U.S. Securities and Exchange Commission, Division of Economic and Risk Analysis (DERA).

Department of Economic Development and Commerce (2026). The puerto rico economic activity index (pr-eai): October & november 2025. Technical report, Government of Puerto Rico, San Juan, Puerto Rico.

Diebold, F. X. and Mariano, R. S. (1994). Comparing predictive accuracy. Technical Working Paper 169, National Bureau of Economic Research.

Faust, J., Gilchrist, S., Wright, J. H., and Zakrajšek, E. (2013). Credit spreads as predictors of real-time economic activity: A bayesian model-averaging approach. *The Review of Economics and Statistics*, 95(5):1501–1519.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The midas touch: Mixed data sampling regression models. Scientific Series 2004s-20, CIRANO, Montréal.

Ghysels, E., Sinko, A., and Valkanov, R. (2006). Midas regressions: Further results and new directions. Technical report, University of North Carolina at Chapel Hill and UCSD. Working Paper.

Goyal, A. and Welch, I. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

Granger, C. W. and Newbold, P. (1974). Spurious regressions in econometrics. *Journal of econometrics*, 2(2):111–120.

Green, R. C., Hollifield, B., and Schürhoff, N. (2007). Dealer intermediation and price behavior in the aftermarket for new bond issues. *Journal of Financial Economics*, 86(3):643–682.

Hubrich, K. and West, K. D. (2010). Forecast evaluation of small nested model sets. *Journal of Applied Econometrics*, 25(4):574–594.

Lo, A. W. and MacKinlay, A. C. (1989). An econometric analysis of nonsynchronous trading. Working Paper 3003-89-EFA, Alfred P. Sloan School of Management, Massachusetts Institute of Technology.

Medioli, A., Hampton, T., Chea, P., and Raimes, E. (2022). Puerto rico bondholder recovery patterns echo major municipal bankruptcies. Sector comment, Moody's Investors Service.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, 54(3):375–421.

U.S. Government Accountability Office (2025). U.s. territories: Public debt and economic outlook - 2025 update. Report to Congressional Committees GAO-25-107560, U.S. Government Accountability Office.

West, K. D. (2006). Forecast evaluation. In *Handbook of Economic Forecasting*, volume 1, pages 99–134. Elsevier.

Wu, S. (2025). A comparison of transaction costs for municipal securities and other fixed-income securities. Technical report, Municipal Securities Rulemaking Board.