How Efficient Is Brazil's FX Futures Market? Evidence from a Large-Scale Order-Book Dataset

Paulo Ramos[†] Alex Ferreira [‡] Wilson Freitas [§]

March 27, 2026

Abstract

Abstract: We investigate the informational efficiency of Brazil's FX futures market using a large-scale order-book dataset spanning 888 trading days and over 8 billion tick-by-tick observations from 2016 to 2019. We develop an algorithm to reconstruct the limited order book environment and examine efficiency along two dimensions: liquidity provision and price discovery dynamics, with a particular focus on PTAX windows. Deploying econometric and spectral methods, our results suggest stable spreads and a structural decline in price impact after 2017. We also show that the volatility increases observed in 2018 and 2019 mainly reflect microstructure noise and the first three daily PTAX windows are associated with higher volatility. Across price discovery methods, we find frequent shifts between quoting sides, suggesting that informed and uninformed traders intermittently shape price formation.

Keywords: Price Discovery; Wavelet Analysis; Foreign Exchange Market.

JEL classification: C14, G14, G15

Área: Finanças

*This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The authors declare no conflict of interest.

[†]Insper and University of Sao Paulo, paulohrb@al.insper.edu.br

[‡]University of São Paulo, alexferreira@usp.br

[§]Itaú, wilson.freitas@gmail.com

1 Introduction

How efficient is Brazil's FX futures market? This question is particularly relevant given the scale and institutional structure of USD/BRL trading. The global foreign exchange market is the largest financial market in the world, with average daily OTC turnover reaching USD 9.6 trillion according to the most recent BIS Triennial Central Bank Survey, conducted in April 2025 and released in September 2025. Within this landscape, the Brazilian real (BRL) accounts for USD 90 billion in average daily turnover, of which USD 15 billion correspond to spot transactions (BIS Triennial Survey 2025, Table 3). Prior evidence indicates that, in Brazil, the futures segment plays a leading role in price discovery for the USD/BRL exchange rate (e.g., (Garcia et al., 2015)), distinguishing it from the predominantly decentralized OTC structure observed in major currency pairs. The economic magnitude of this centralized market is substantial: in April 2025, the most liquid USD futures contract at B3 (WDO K25) registered BRL 145.4 billion in trading volume in a single day—approximately USD 25 billion in notional turnover—exceeding total global BRL spot activity and representing nearly 30% of total global OTC BRL turnover. This combination of global relevance, centralized trade, and market depth makes the Brazilian FX futures market a natural setting for examining informational efficiency.

Financial markets are populated by agents with heterogeneous information sets. In the classical sense of (Fama, 1970), market efficiency refers to the extent to which prices fully and promptly reflect available information. At high frequencies, however, this notion becomes operational: efficiency depends on how quickly and symmetrically information is incorporated into transaction prices and quotes as trading unfolds. Order flow conveys private signals, liquidity providers update quotes, and prices adjust as dispersed information is aggregated.

Against this background, we ask two central questions. First, how has the efficiency of the Brazilian FX futures market evolved in recent years, particularly in terms of liquidity conditions and price discovery dynamics? Second, do institutional features specific to this market—most notably PTAX fixing windows and the interaction between spot and

futures—systematically affect short-term price formation within the limit order book? This paper provides empirical answers to these questions by jointly analyzing liquidity measures and informational leadership in the USD/BRL futures market.

Our empirical analysis is based on an unusually large high-frequency dataset from the Brazilian FX futures market, comprising more than 8 billion observations spanning from April 25$^{\text{th}}$ 2016 to December 12$^{\text{th}}$ 2019, when data availability ceased, as B3 discontinued public access to these records through its FTP. This tick-by-tick dataset allows us to reconstruct the order book and to jointly examine liquidity and price discovery dynamics over a period marked by substantial macroeconomic and political fluctuations in Brazil.

Prior studies on emerging markets document mixed evidence regarding informational leadership between spot and futures markets (e.g., (De Boyrie et al., 2012), (Kumar, 2018)). However, this literature typically relies on transaction data and parametric time-domain models such as vector error correction frameworks (VECMs), and does not directly exploit the informational content embedded in the limit order book. As a result, relatively little is known about how quoting behavior itself contributes to price discovery, or whether informational advantages are systematically concentrated on one side of the book.

More broadly, this study contributes to the literature on price discovery and liquidity measurement. Seminal contributions such as (Hasbrouck, 1995), (Baillie et al., 2002), and (de Jong, 2002) develop econometric frameworks to quantify informational leadership across trading venues and instruments. These approaches have been widely applied to equity markets ((Booth et al., 1999); (Chakravarty et al., 2004)) and commodities ((Garbade and Silber, 1983); (Figuerola-Ferretti and Gonzalo, 2010)). In foreign exchange markets, (Cabrera et al., 2009) and (Osler et al., 2011) examine whether spot or futures markets dominate in incorporating new information, often documenting spot market leadership for major developed currency pairs. We extend this literature by shifting the focus from venue-level dominance to quote-level informational dynamics within a centralized futures limit order book in an emerging market setting.

To address our research questions, we reconstruct the limit order book from raw tick-by-tick data and extract best quotes and order flow measures at high frequency. We

then assess liquidity and price discovery using both econometric and spectral approaches. Liquidity is evaluated using dynamic models of spreads, price impact, and efficient price volatility, incorporating PTAX window effects via lead–lag structures. Price discovery is evaluated using traditional VECM-based measures—Information Share and Component Share—alongside a spectral framework based on the Cross Wavelet Transform, which provides a continuous time–frequency measure of lead–lag dynamics. An appendix presents a restricted VAR simulation to assess the behavior of these metrics.

Our contribution lies in combining quote-level data with complementary parametric and spectral tools to characterize price discovery within a centralized FX futures limit order book. While recent studies incorporate wavelet methods in commodity and exchange rate settings ((Joseph et al., 2015); (Nigatu and Adjemian, 2020); (Kumar et al., 2016); (Ghosh and Chaudhuri, 2017)), none apply them to price discovery in FX futures markets. We also build on the empirical literature on the Brazilian FX market ((Sung, 2021); (Garcia et al., 2015); (Laurini et al., 2008); (Ferreira et al., 2025)), extending it through the use of high-frequency quote data and continuous measures of informational leadership.

Our results indicate that overall market efficiency remains stable throughout the sample period, accompanied by gradual improvements in liquidity conditions. Bid–ask spreads remain broadly stable, with only limited and short-lived effects associated with PTAX fixing windows, occasionally displaying temporary spikes on specific days. Price impact measures suggest a structural shift beginning in 2017, consistent with a deepening of liquidity. PTAX windows exert negligible effects on price impact, both contemporaneously and in the surrounding intervals.

The volatility decomposition reveals that a substantial portion of the elevated volatility observed in 2018 and 2019 is attributable to microstructure noise rather than to movements in the efficient price. At the same time, the first three PTAX windows of the day are associated with a statistically significant increase in efficient price volatility. Finally, both parametric and spectral measures of price discovery point to a diffuse informational structure within the limit order book, with no persistent asymmetry between the bid and ask sides, either during PTAX windows or in adjacent intervals.

These findings have implications for researchers, policymakers, and market participants. By shifting the focus from transaction-based measures to quote-level dynamics, we provide a more granular view of liquidity provision and informational leadership in FX markets, particularly in a floating exchange rate regime where short-term distortions may inform intervention strategies and risk management decisions.

The remainder of the paper is organized as follows. Section 2 presents the methodological framework to the LOB construction. Section 3 describes the econometric and spectral models deployed to the analysis. Section 4 presents the dataset and the results. Section 5 concludes with implications for liquidity and price discovery and suggestions for future research.

2 Framework

LOB We start by providing the definition of our limited order book (LOB). As in (Gould et al., 2013) we define an order as a vector $x = (p_x, \omega_x, t_x)$ where p_x is the price, ω_x is the quantity and t_x is the time the order was submitted. If x is a sell (buy) order, then the agent commits to sell (buy) up to ω_x ($-\omega_x$) units of the traded asset at a price no less (no greater) than p_x. The prices and quantities are discretized by two resolution parameters: the lot size, ς, and the tick size, ϖ.

We can now characterize an active buy order[1]. Consider:

- A given buy order, $y = (p_y, \omega_y, t_y)$.
- A set X containing all sell orders $x = (p_x, \omega_x, t_x)$ such that $t_x \leq t$ and $p_x \leq p_y$. X comprises all sell orders that could be matched with y up to t
- A set Z containing all buy orders $z = (p_z, \omega_z, t_z)$ such that $t_z < t_y$ and $\exists x \in X(p_x \leq p_z)$. Z comprises all buy orders that could be matched with at least one order in X and were submitted before y, i.e., orders positioned ahead of y in the priority queue.

We say y is an active buy order at time t if it satisfies:

[1]The same characterization applies to an active sell order by reversing the price inequality

$$\underbrace{\sum_{x \in X} |\omega_x|}_{\text{Total Consumable Sell Volume}} - \underbrace{\sum_{z \in Z} |\omega_z|}_{\text{Prioritized Buy Volume}} < |\omega_y|. \tag{1}$$

In other words, this inequality states that a buy order is active if there is no order in the sell side that crosses its price and wasn't matched yet by a previous buy order. From this definition, we can denote $\mathcal{L}(t)$ as the set containing all the active orders (LOB) at the time t and the subsets $\mathcal{A}(t)$ and $\mathcal{B}(t)$ as, respectively, the sell side and the buy side.

From this environment we derive the variables used to examine the liquidity and price discovery dynamics. The first price level from both sides of the LOB make up the bid-ask spread, $s(t) = a(t) - b(t)$, where $b(t) = \min_{x \in \mathcal{A}(t)} p_x$ stands for the best (lowest) active sell order and $b(t) = \max_{x \in \mathcal{B}(t)} p_x$ is the best (highest) active buy order at t. Also from the first level, we can drag the mid-price $p(t)$, which is the arithmetic average between the bid and ask. Finally, the order flow imbalance, $OFI(k)$, corresponds to the net difference between buyer-initiated and seller-initiated executed orders in the time interval[2] k:

$$OFI(k) = \sum_{t_k \in k} |\omega_x| \lambda_x, \tag{2}$$

$$\lambda_x = \begin{cases} 1, \omega_x < 0 \wedge p_x > b(t) \text{ i.e., buyer-initiated} \\ -1, \omega_x > 0 \wedge p_x < a(t) \text{ i.e., seller-initiated} \\ 0, (\omega_x < 0 \wedge p_x \leq b(t)) \vee (\omega_x > 0 \wedge p_x \geq a(t)) \text{ i.e., not executed.} \end{cases} \tag{3}$$

Spread Algorithm Once we defined our framework, we can proceed to build an algorithm to extract the best bid and ask quotes from the raw order data. To do so, we represent $\mathcal{L}(t)$ as a binary tree (Glur, 2025). Binary trees are data structures used to model hierarchical relations between variables and implement search and sort operations on it. Among other features, binary trees possess nodes that act as categories. In our case, the first node (root) of $\mathcal{L}(t)$ is the book side, i.e. $\mathcal{A}(t)$ and $\mathcal{B}(t)$. For each book

[2]To an alternative definition of OFI, based on consecutive orders instead of intervals, see (Wang et al., 2021)

side we have a bunch of child nodes, corresponding to the price levels throughout the day. Finally, a price level in a specific book side has leaves (nodes with no children nodes) corresponding to the order. In each leaf we store its time, status and quantity.

With this structure, we can, for each new order, recover a snapshot of the order book at that given time. We start by initializing the binary tree as an empty tree. Then we order the raw data by the order time entry and sequentially search and sort within the binary tree to put, edit or remove the orders. Due to computational costs and to the fact that our analysis covers only the first level of the book, we restrict the set of order to those that also belong to the trade book at any time. As in Castro (2017), we found some orders crossing the spread and we dealt with it imposing that these spread violating orders are executed (or partially executed conditional to the liquidity) at the time they arrive. Another practical correction was to identify the time at which the pre-market auction happens and remove the orders entered before it. To do so, we used the trade book and found out many trades between two non-agressor orders at the beginning of the day. We marked the timestamp of these trades and remove order entered before it.

Considering the status taxonomy available at the time, we group order statuses in three categories regarding its equivalent operation within the tree (input, update and remove). New orders are simply inserted into the respective book side queued up by its price level. Modified and partially executed orders have its leaves updated according to what was changed - price or quantity. Finally, canceled, executed and expired orders have its leaves removed from the tree. For two or more observations of $\mathcal{L}(t)$ at the same t (at the millisecond aggregation) we preserve the last one. In this sense, the bid, ask and bid-ask series are the $b(t), a(t)$ and $b(t) - a(t)$ for each t of this binary tree. The pseudo code below describes how the algorithm is deployed.

3 Empirical Strategy

Liquidity Indexes For the liquidity analysis, we rely on the indexes from (Evans and Lyons, 2002), (Mancini et al., 2013) and (Zhang et al., 2005). We use these metrics to estimate autoregressive models and assess the distributional aspects of its parameters.

These estimators allow us to investigate whether there is evidence of an anomalous market behaviour during PTAX fixing windows.

The first metric is the bid-ask spread and we draw insights from (Roll, 1984) in regard to its meaning. A high spread indicates the presence of informed traders and, in response to this increase in risk, there is a decrease in liquidity. Considering the cointegration dynamics between bid and ask, we can assume that $\bar{s}(t)$ is stationary. Additionally, given that the PTAX window effect may be lagged, contemporaneous or led, we need to address the detection of this moment. Therefore, we can use the following model to detect the impact of PTAX windows fixing in the spread:

$$\bar{s}_t = \gamma + \bar{s}_{t-1} + \sum_{j=-l}^{l} \sum_{i=1}^{4} \beta_{i,j} D_{i,j,t}, \tag{4}$$

$$D_{i,j,t} = \begin{cases} j < 0 \begin{cases} 1, \forall t \in \{t_0 + j, ..., t_0 + 1\} \\ 0, \text{ cc} \end{cases} \\ j = 0 \begin{cases} 1, \forall t \in \{t_0, ..., t_1\} \\ 0, \text{cc} \end{cases} \\ j > 0 \begin{cases} 1, \forall t \in \{t_1 + 1, ..., t_1 + j\} \\ 0, \text{ cc} \end{cases} \end{cases}, \tag{5}$$

where t_0 stands for the first period where the i-th windows is active and t_1 the last one. The index $j < 0$ ($j > 0$) represents the j periods before (after) the begining (end) of the windows. After estimating (5) for a grid $J = \{-l, ..., l\}$, we pick the optimal j_-^* and j_+^* as the ones with lowest Bayesian Information Criterion (BIC). Then we proceed to estimate (4), for fixed $j = \{j_-^*, 0, j_+^*\}$ in a daily subsampling.

In the same sense, microstructure models of price impact (Evans and Lyons, 2002), especially the ones designed to high-frequency or tick-by-tick data (Cont et al., 2013) can be used to examine liquidity and price discovery. In this study, we use the order flow imbalance and an interaction with PTAX fixing windows to explain the log-return of mid-price:

$$\Delta lnp_t = \gamma + \rho OFI_t + \sum_{j=-l}^{l} \sum_{i=1}^{4} \beta_{i,j} D_{i,j,t} OFI_t. \tag{6}$$

Additionally, for a robustness check, we estimate an AR(1) of OFI and use the residuals (u_t) as a proxy to unexpected order flow. Such as:

$$\Delta lnp_t = \bar{\gamma} + \bar{\rho} u_t + \sum_{j=-l}^{l} \sum_{i=1}^{4} \beta_{i,j} D_{i,j,t} u_t. \tag{7}$$

The idea behind (7) and (8) is to verify the different market reactions to predictable and unpredictable order flows. The predictable order flow account for intraday patterns, liquidity-motivated trades and eventual inventory effects. On the other hand, the unpredictable order flow focuses on the information driven trades.

The other liquidity measure is an adaptation of (Zhang et al., 2005) volatility estimator. As in (Mancini et al., 2013) we use the two-scale realized volatility (TSRV) estimator as a proxy for efficient price volatility. The TSRV works as a smoother that filters the microstructure noise and, therefore, can be employed with the traditional realized volatility (RV) estimator to decompose volatility into efficient price fluctuations (TSRV) and microstrcture related activity.

In our approach, we use TSRV to isolate the microstructure component and analyze its autoregressive dynamics. First of all, we estimate RV to a grid of frequencies and evaluate its convergence. As in (Zhang et al., 2005) the divergence of RV to different time scales indicates the presence of microstructure noise. Following, we define the two time scales used in TSRV as the highest frequency used in RV and the highest coarse aggregation that yields a stable estimator (ie, the first aggregation that reaches a plateau when taking TSRV as a fuction of coarse approximations). After so, we construct MN as follows (Zhang et al., 2005). Consider $X_t = \log P_t$. For a given time window of length W, the realized volatility at fine grid f is described as:

$$RV_t^{(f)} = \sum_{i \in W} (X_{t_1} - X_{t_1 - 1})^2. \tag{8}$$

To construct the coarse approximation, denote $K \geq 2$ as a number that separate two

consecutive observations. In the subgrid averaging, instead of taking the return between two immediate observations, each subgrid contains the returns of two observations that occured in a K time interval. Then we define the subgrid $G_k = \{t_k, t_{k+K}, t_{k+K}, ...\}$ and compute to $k = 1, ..., K$:

$$RV_t^{(k)} = \sum_{j \in W} (X_{t_k+jK} - X_{t_k+(j-1)K})^2. \tag{9}$$

Finally, we take an average in the subgrid and use K to weight this average with the fine grid RV:

$$TSRV_t = \frac{K}{K-1} \left([\frac{1}{K} \sum_{k=1}^{K} RV_t^{(k)}] - \frac{1}{K} RV_t^{(f)} \right). \tag{10}$$

Once we are able to isolate the part of price variance accounted to microstructure dynamics , we can define our model to test if there is any efficient price dynamics regarding PTAX fixing around its windows. We have:

$$TSRV_t = \gamma + \rho TSRV_{t-1} + \sum_{i=1}^{4} \beta_i D_{i,t}. \tag{11}$$

Price Discovery To investigate price discovery between quote sides, we use the main parametric measures from (Hasbrouck, 1995) and (Booth et al., 1999), respectively the information share (IS) and the component share (CS). In addition, we propose a nonparametric estimator, using spectral methods as in (Nigatu and Adjemian, 2020) and (Kumar et al., 2016), to handle eventual nonlinearities in the cointegration pattern between the quotes.

Both IS and CS derive from the error-correction representation (ECM) from (Engle and Granger, 1987a), used to model series that, despite being non-stationary, are integrated of the same order and posse a stationary relantioship as a linear combination. So we denote $p_t = (p_{a,t}, p_{b,t})'$, the vector of log-prices representing the bid ($p_{b,t}$) and ask ($p_{a,t}$) quotes. By the error-correction formulation, we have:

$$\Delta p_t = \alpha\beta' p_{t-1} + \sum_{j=1}^{q} \Gamma_j \Delta p_{t-j} + \epsilon_t, \tag{12}$$

where $\alpha = (\alpha_a, \alpha_b)$ and $\beta = (\beta_a, \beta_b)$ are, respectively, the error-correction and cointegration vectors and $\epsilon_t = (\epsilon_{a,t}, \epsilon_{b,t})$ is the innovation process. We can interpret α as a measure of speed of adjustment to the equilibrium and β as the equilibrium relationship itself - with $\beta = (1, -1)$ when the bid-ask spread is a stationary series. We also assume that ϵ is iid with variance Ω.

Given the ECM, we can rewrite it as a vector moving average (VMA), to retrieve the price vector as a sum of lagged innovations. Denote L as the lag operator, we have:

$$\Delta p_s = \sum_{s=0}^{t} \Psi_s \epsilon_{t-s} = \Psi(L)\epsilon_t, \tag{13}$$

$$IS_i = \frac{\psi_i^2 \Omega_{ii}}{\Psi\Omega\Psi'}, \tag{14}$$

$$CS_i = \frac{\psi_i}{\psi_1 + \psi_2}, \tag{15}$$

where the IS accounts for the contribution of a quote to the efficient price and the CS the contribution of a quote to the permanent component of the efficient price.

In order to construct our espectral price discovery estimator, it may be useful to provide a quick introduction to wavelet transforms. Below we briefly address the main mathematical objetcs used to develop the estimator.

A wavelet is a function that can be compressed or stretched by its scale parameter a and moved along the timeline by its location or translation parameter τ (Daubechies, 1994). For a time series x_t, we can denote its CWT as $W_x(a, \tau)$, which can be written as:

$$W_x(a, \tau) = \int_{-\infty}^{\infty} x_t \frac{1}{\sqrt{a}} \psi^* \left(\frac{t - \tau}{a} \right) dt, \tag{16}$$

where $\psi(t)$ stands for the mother wavelet and * denotes the complex conjugate. [3]

[3]

Naturally, as we are dealing with discrete time series, we are interested in the case where $t \in \{1, 2, ..., n\}$. Considering an uniform time step δt, the discretized version of (1) is (Torrence and Compo, 1998):

$$W_{n,x}(a) = \frac{\delta t}{\sqrt{s}} \sum_{t=0}^{n-1} x_t \psi^* \left(\frac{\delta t(t-n)}{a} \right). \tag{17}$$

As it can be seen in (8), the WT coefficients form a $m \times n$ matrix where m stands for the number of scales. In addition, considering that the data arrive at a unitary time interval, which yields $\delta_t = 1$, and that our smallest scale would precisely this unit, what gives us $a_0 = 1$, the j-th scale is defined as (Torrence and Compo, 1998):

$$a_j = a_0 2^{j\delta_s}, \delta_s = \log_2 \left(\frac{n\delta_t}{a_0} \right) m^{-1}. \tag{18}$$

Once our $\psi(t)$ are complex-valued, our $W_{n,x}$ may also be divided into a real part and an imaginary part. Let's denote them, respectively, as $\mathfrak{R}\{W_{n,x}(a)\}$ and $\mathfrak{I}\{W_{n,x}(a)\}$. Then, the power spectrum of $W_{n,x}(a)$ can be defined as $||W_{n,x}(a)||^2$, which shows how the energy of the signal x_t is distributed among the different frequencies and at different moments.

The Cross Wavelet Transform (XWT) can be employed to assess how two signals interact in the time-frequency domain and to evaluate the lead-lag relationship between them (Torrence and Compo, 1998). Keeping our previous notation, we can denote an XWT between the signals x_t and y_t as $W_{n,xy}$, so $W_{n,xy} = W_{n,x}W^*_{n,y}$. Although this formulation (Torrence and Compo, 1998) has been extensively used, it leads to a bias that underestimates the small-scale (high-frequency) phenomena. For that reason, our XWT, and its consequent power spectrum, is estimated based on (Veleda et al., 2012), adding a normalization by scales that yields $W_{n,xy} = (W_{n,x}W^*_{n,y})2^{j/2}$

To help us weigh each scale's information about the relation between two series, we use the wavelet coherence defined below (Torrence and Compo, 1998). This measure can be understood as a time-frequency correlation, which indicates (normalized to spans over 0 and 1) the degree of similitude of the oscillatory patterns.

To be considered a mother wavelet, a function must satisfy (Farge, 1992) two admissibility conditions : (i) it must have zero mean, $\int_{-\infty}^{\infty} \psi(t)dt = 0$; (ii) and unit energy, $\int_{-\infty}^{\infty} ||\psi(t)||^2 dt = 1$.

$$R_n(a) = \frac{|S(a^{-1}W_{n,xy}(a)|}{S(a^{-1}|W_{n,x}(a)|)^{1/2}S(a^{-1}|W_{n,y}(a)|)^{1/2}}. \tag{19}$$

Naturally, to $R_n(a)$ present some information about the signals, it must be the case that $\exists(x,y)S(x)S(y) \neq S(xy)$, otherwise $R_n(a) = 1$ for all a. To achieve it, S is a smoothing operator defined as convolution with a Gaussian in time and a rectangular window in scale (Aguiar-Conraria et al., 2008).

Finally, following (Torrence and Compo, 1998) and (Aguiar-Conraria and Soares, 2010) formulation of phase ϕ between two signals x_t and y_t to the scale a, we have:

$$\phi_{x,y}(a) = \tan^{-1}\left(\frac{\Im\{W_{n,xy}(a)\}}{\Re\{W_{n,xy}(a)\}}\right). \tag{20}$$

With $\phi_{x,y} \in [-\pi, \pi]$ and 4 possible cases: (i) if $\phi_{x,y} \in (0, \frac{\pi}{2})$, then the series move in phase, but y_t leads; (ii) if $\phi_{x,y} \in (-\frac{\pi}{2}, 0)$, then the series move in phase, but x_t leads; (iii) if $\phi_{x,y} \in (\frac{\pi}{2}, \pi)$, then they move in anti-phase and x_t leads; (iv) if $\phi_{x,y} \in (-\pi, -\frac{\pi}{2})$, then they move in anti-phase and y_t leads

Since ϕ is precisely a measure of lead-lag, we can use it in a microstructure approach to proxy the informational flow coming from the most informed venue to the least informed one. In that sense, in a bivariate problem as ours, we may count in many frequencies the x led in a given time period t:

$$\mathbb{1}_x(a,t) = \begin{cases} 1, \text{ if } \phi_{x,y}(a,t) \in [-\frac{\pi}{2}, 0) \bigcup (\frac{\pi}{2}, \pi] \\ \\ 0, \text{ cc} \end{cases}. \tag{21}$$

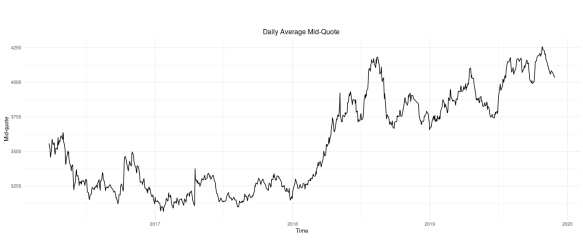
Considering a time interval K and using the wavelet coherent as a weight to translate frequency information into time information, we define our espectral price discovery estimator to the quote x as:

$$EPD_x = \mathbb{1}_x(K) = \sum_{a=a_0}^{m}\sum_{t=1}^{n} \mathbb{1}_{a,t}R(a). \tag{22}$$

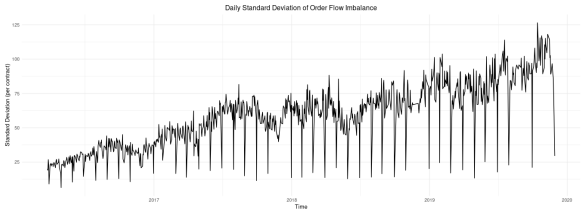
4 Data and Results

Data Our high-frequency data comprise all the bid and ask quotes from WDO (USD forward mini-contract) from April 25, 2016 to December 12, 2019, providing a total sample of 888 days. Considering the liquidity accumulated at short-term forward contracts, we fixed the maturity to the lowest one, which occurs on the first business day of each month. On the lasting trading day preceding the expiration date, the agents may rollover the contract, but these operations takes place at a different product and would not produce any significant distiortio in our analysis.

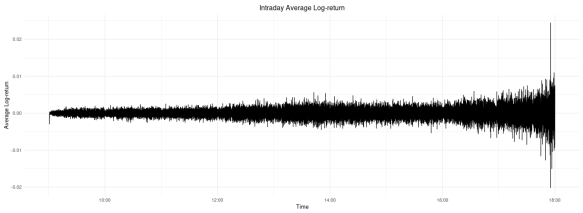
From an operational perspective, once the algorithm was executed, we had 257 million quotes observations and 18 million OFI observations. After so, we proceed to clean the pre-processed data and reach out to a dataset the encompasses 25 million observations at 1-second aggregation. The figure 1 depicts some intraday and historical series for our reconstructed LOB



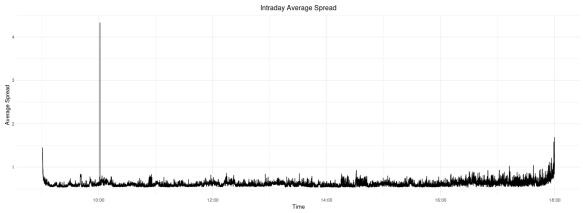
(a) Daily Average Mid-Quote



(b) Daily Standard Deviation of Order Flow Imbalance



(c) Intraday Average Log-return of Mid-Quote



(d) Intraday Average Bid-Ask Spread

Figure 1: Historical and intraday time series for mid-quote, order flow imbalance, log-return and bid-ask spread

Subfigure (a) shows the mid-quote daily average throughout the whole sample, exhibiting the non-stationary trend in the series and the respective BRL depreciation against USD. Subfigure (b) presents the OFI's daily standard deviation, highlighting the increase in variability in the liquidity provision. It can also be noted a monthly clustering in its

dispersion due to the expiration date of each contract. The intraday plots are built taking the cross-sectional average of each second across the 888 trading days - at this point, with no outlier removed. The subfigure (c) shows the log-return of mid-quote and allows us to identify an increase in price dispersion, therefore in returns, next to the closing call. The intraday dynamics of spread, in subfigure (c) also represent this narrowing in liquidity towards the end of the day. [4]

Liquidity Indexes To track if the information flows associated with PTAX windows arrive before, during or after each window, we estimate a grid of models the presents dummies representing the PTAX window being lagged or led. The figure 2 presents the trade-off between model complexity and log-likelihood for each frequency. To the spread, we can see that the addition of previous and posterior information more than compensates the inclusion of new parameters. Therefore, the 4 frequencies peaked at the most distant points from the $j = 0$.

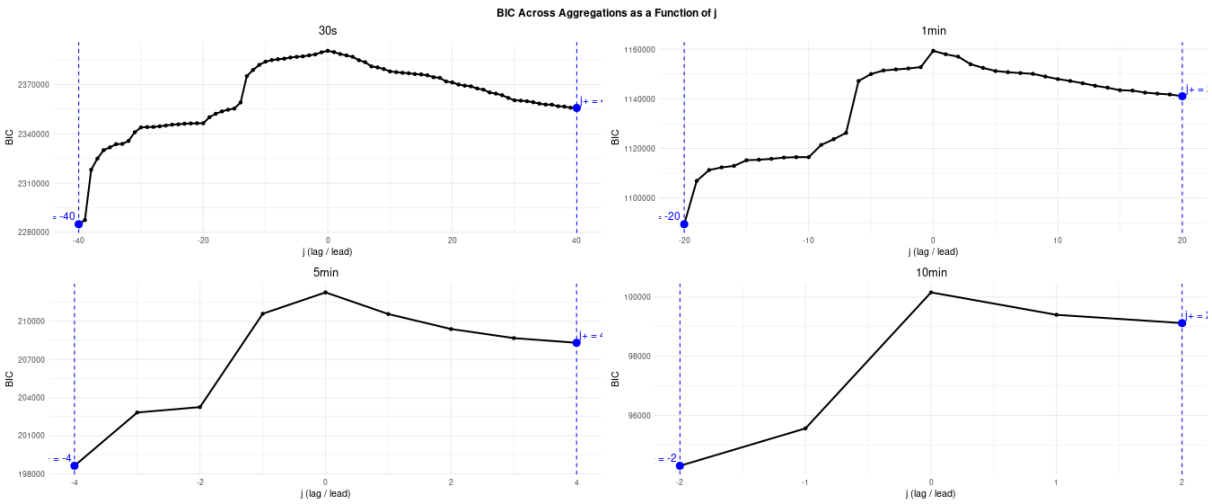


Figure 2: Bayesian Information Criterion Across Aggregations as a Function of j

Figure 3 reports the results to the autoregressive parameter of an AR(1) at the spread level using the dummies with $j = 0$. For most of the sample, the spread presented a low persistence, with low half-life and an equilibrium value close to the tick-size - consistent with a highly liquid market. Few days, though, presented statistically non-signficant estimates, probably related to less directional trading sessions.

[4]For other descriptive statistics, check table A1.

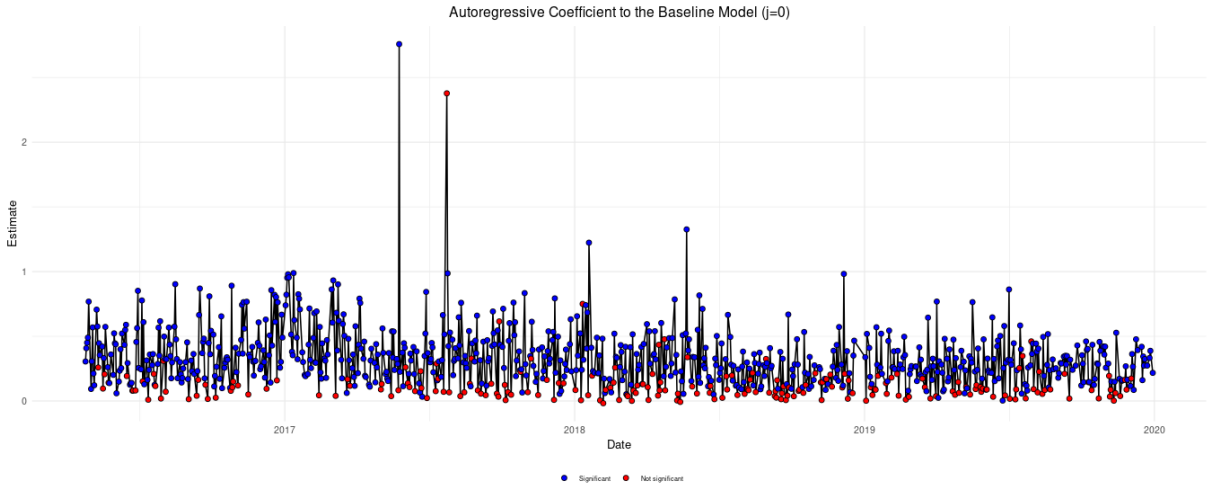


Figure 3: Autoregressive Coefficient to the Baseline AR(1) Model (j=0) to the Spread

For the sake of brevity, the further results are mainly reported at the highest aggregation (30 seconds). Figures 4, 5, 6 and 7 show the time distribution of the dummy coefficients in an augmented AR(1) of spread. The unit is the level of spread, the first graph of each figure presents the results to the dummy that encompasses 20 minutes before the beginning of the window, the second one the dummy that represents being at the window and the third one the dummy that encompasses 20 minutes after the end of the window.

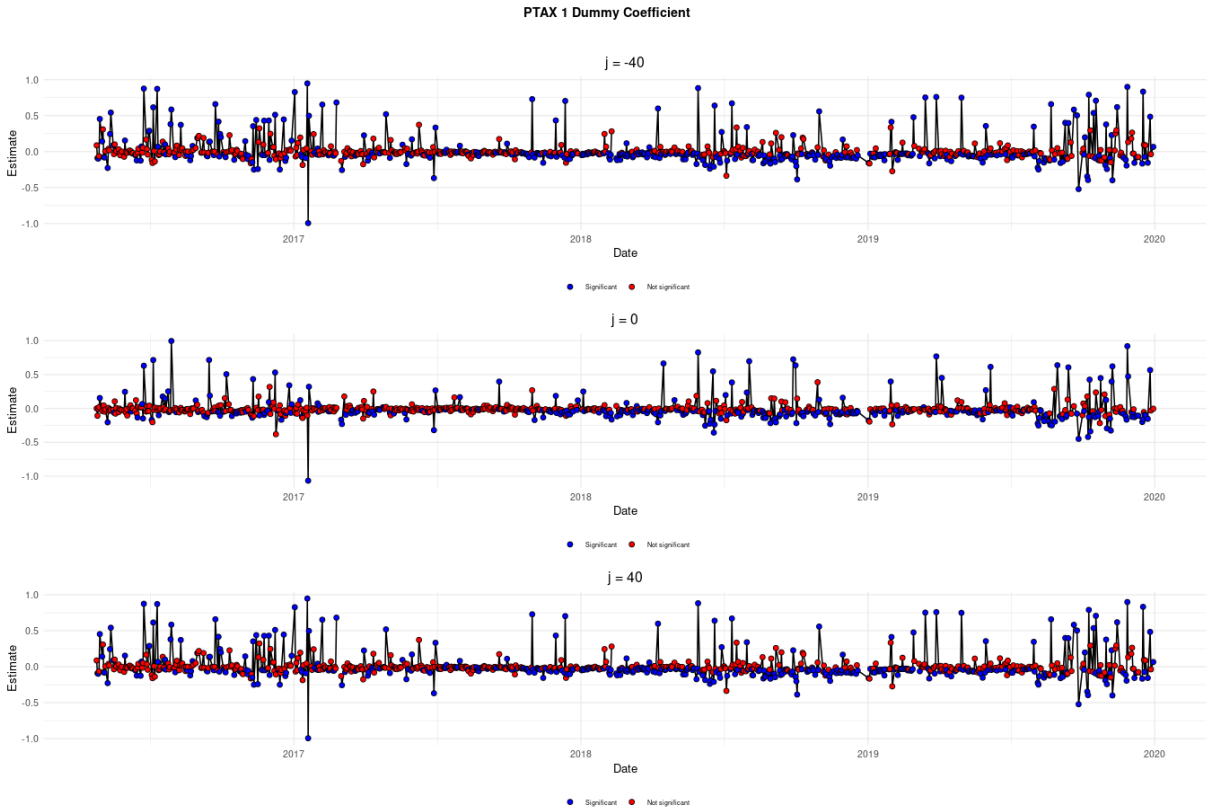


Figure 4: First PTAX Dummy Coefficient Across j

For most of the sample, the first PTAX seems to play no constant role in the widening of the spread. With mostly statistically non-significant estimates, the first window only occasionally produces a short-lived positive pressure on the liquidity.

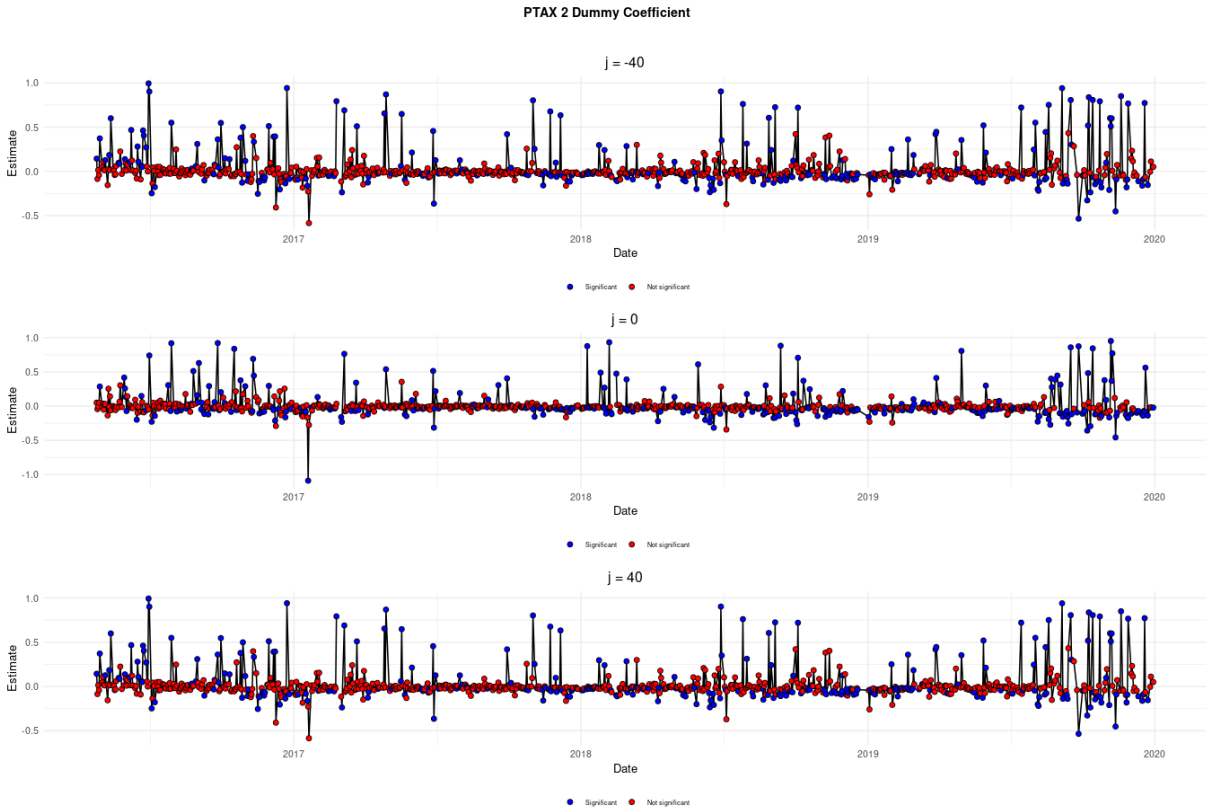


Figure 5: Second PTAX Dummy Coefficient Across j

The second PTAX show a similar pattern to the first one: mostly non-different than zero; approximately same spikes; and occasional positive pressure on liquidity. However, the impact of the second PTAX shows higher variance than the first one. A limitation worthy to be mentioned is that part of the performance that may be attributed to the first and second PTAX can, in fact, be caused by external events such as the openning of foreign markets (mainly the American one).

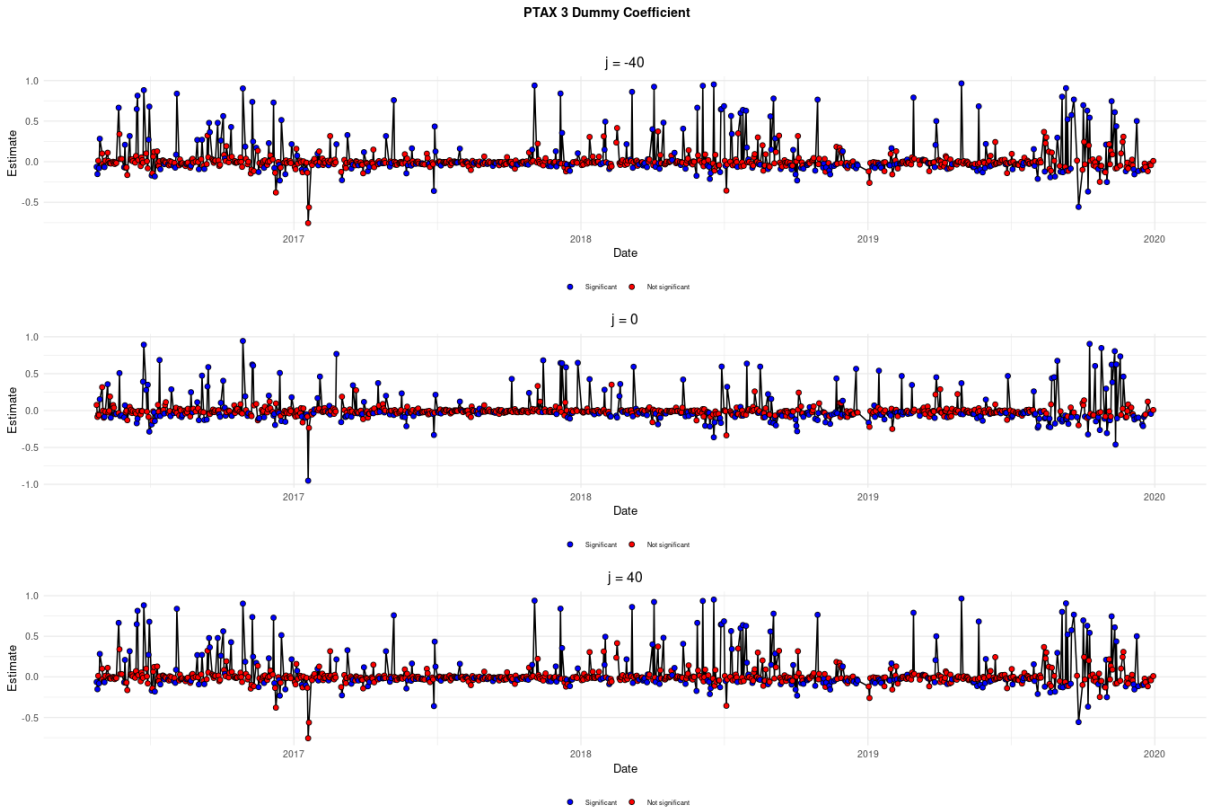


Figure 6: Third PTAX Dummy Coefficient Across j

In contrast to the previous ones, the third PTAX shows not only more persistence across different timing to the same day, but also, in some days, the liquidity pressure is higher after the end of the window.

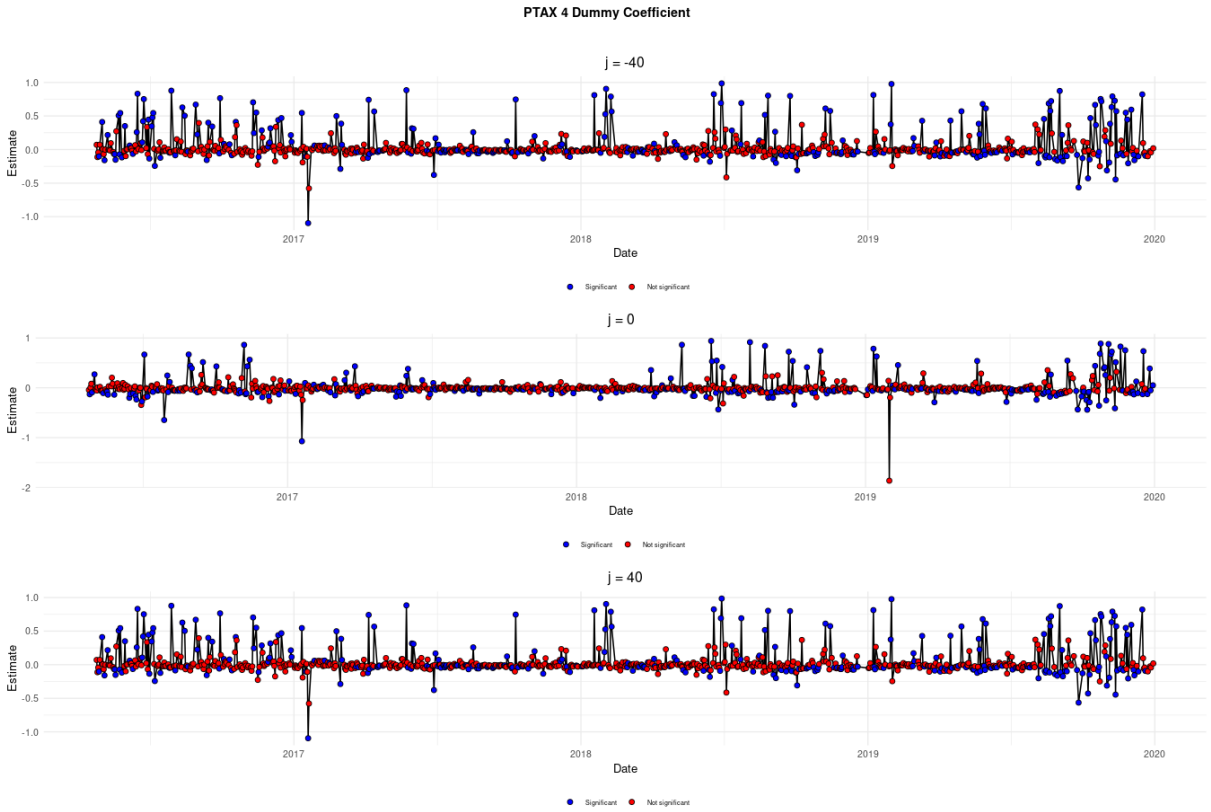


Figure 7: Fourth PTAX Dummy Coefficient Across j

The fourth PTAX displays the smallest impact amongst the other ones. Taking into account that it releases the last spot price of the day, we should expect some effect in increasing liquidity.

Figure 8 presents the BIC results to the grid of models to price impact. Differently than the spread results, there seems to be no information being accumulated before the beginning of the window or after the end of it. The frequencies diverge with the highest frequencies picking low $|j|$ while the lowest frequencies pick higher ones.

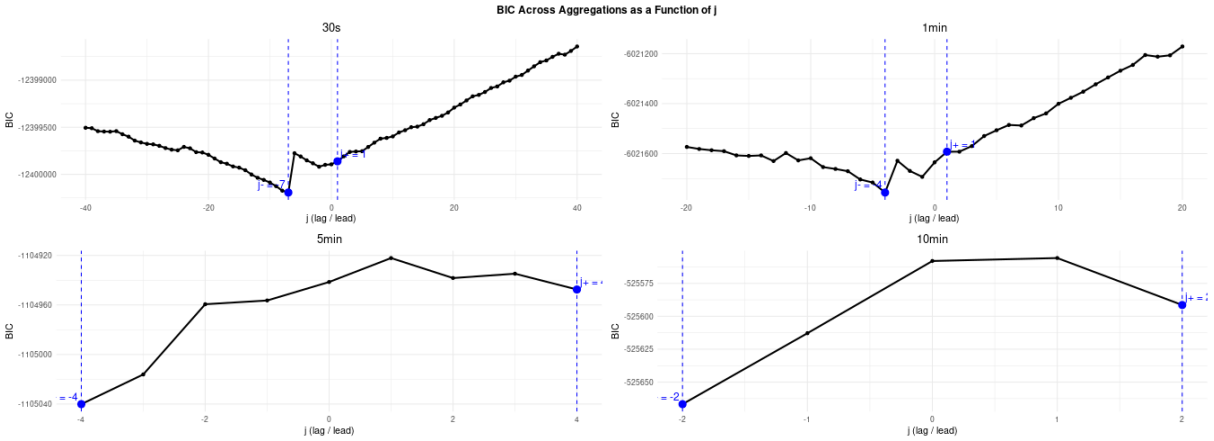


Figure 8: BIC vs j across aggregation frequencies

Figure 9 reports the results to the OFI parameter of the price impact regressions in BPS per billion. From April 2017 to the beginning of 2017 there is a downward trend that stabilizes at the beginning of 2017. The results also demonstrate a jump structure that has diminished its depth. This dynamics can be interpreted as an improvement in liquidity provision.

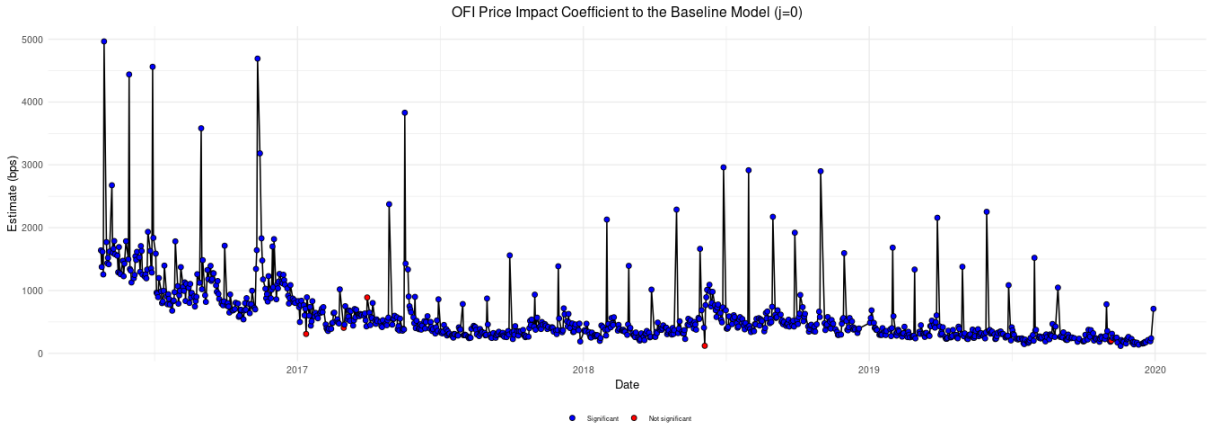


Figure 9: OFI Price Impact Coefficient to the Baseline Model (j=0)

Due to lack of dynamics and statistical significance, we omit the results from $j = 40$. As previously mentioned, the results onwards are reported in the highest aggregation frequency (30 seconds). The coefficients below correspond to the interaction between the contemporary OFI and the price variation. It can be interpreted as BPS in price per billion of flow at the given window.

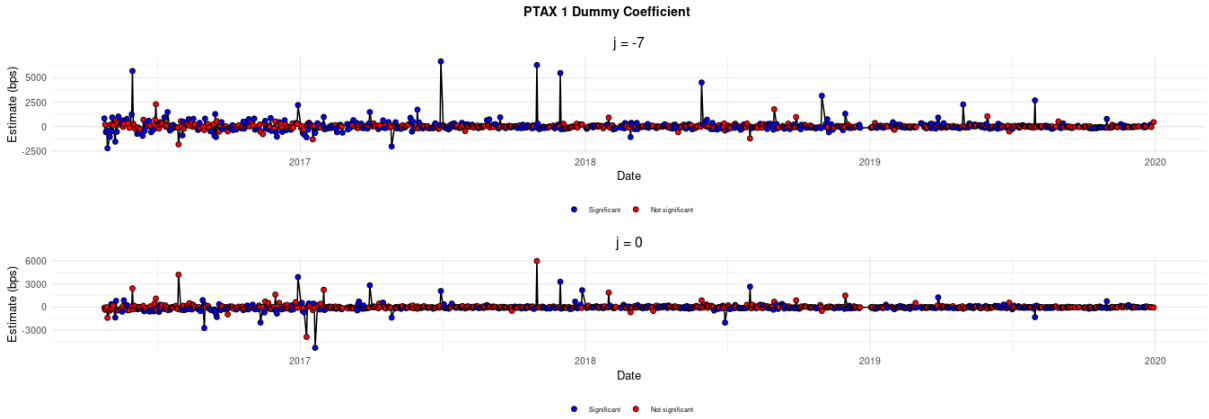


Figure 10: PTAX 1 Dummy Coefficient to Previous and Contemporaneous Effects

The PTAX 1 coefficients display no pattern regards its price impact, being mainly around zero throughout the whole sample. The few positives spikes before the window tend to vanish before the end of window.

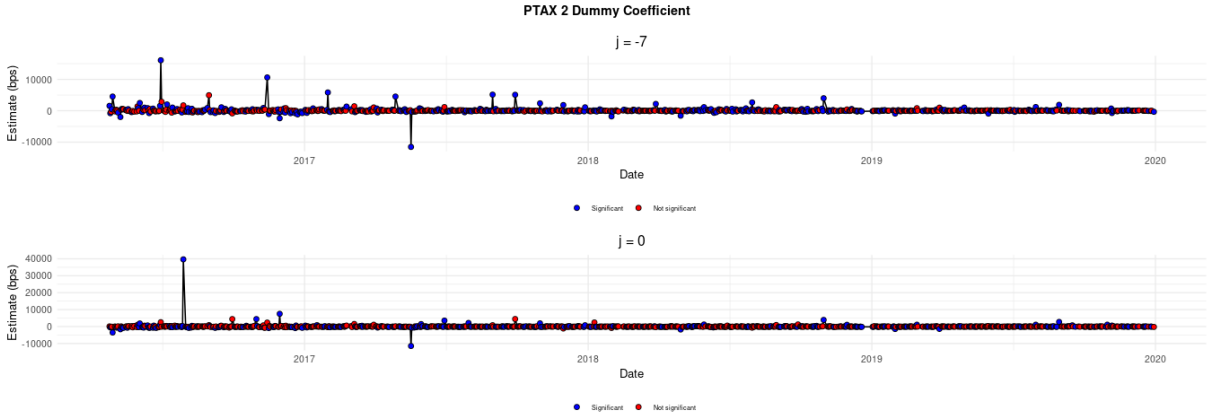


Figure 11: PTAX 2 Dummy Coefficient to Previous and Contemporaneous Effects

PTAX 2 coefficient seems to do not have any systematic impact on the price elasticity to the order flow. At that frequency, the PTAX 2 does widen the spread, but its impact on liquidity does not alter the short-term impact of order flow in prices.

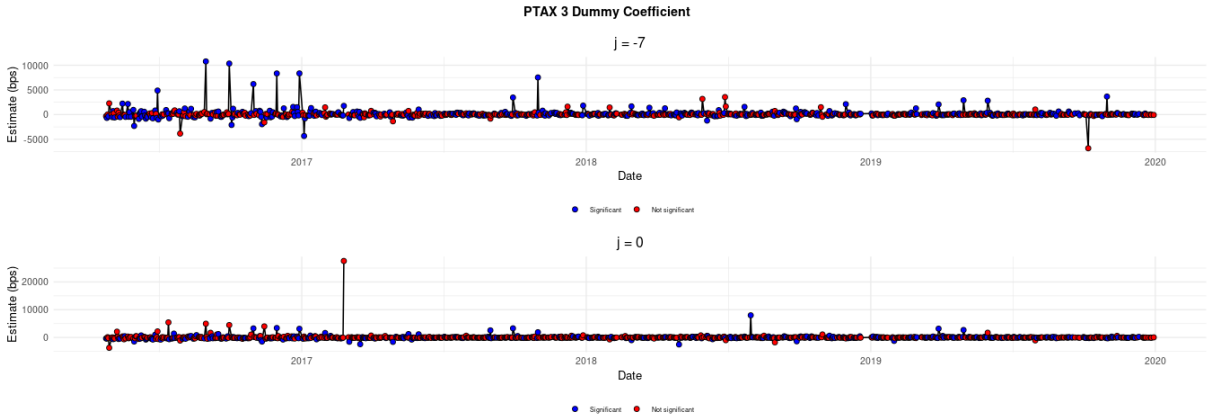


Figure 12: PTAX 3 Dummy Coefficient to Previous and Contemporaneous Effects

PTAX 3 shows some short-lived positive spikes until the beginning of 2017. From that moment, the estimates remain stable and almost equal to zero at both pre and post windows.

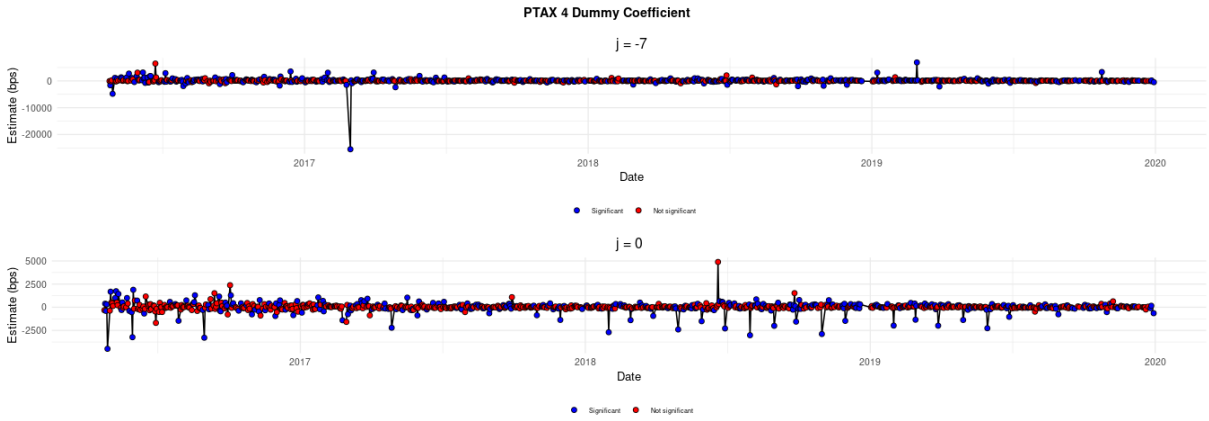


Figure 13: PTAX 4 Dummy Coefficient to Previous and Contemporaneous Effects

PTAX 4 is the only window that, in some moment, showed persistent effects in improving liquidity. From 2018 onwards, there was several trading days where being at the last PTAX windows improved liquidity.

In order to establish a common ground between the two volatility estimators used in this work, we may, first, define the coarse frequency of aggregation to the TSRV to combine with our highest one (1 second). Our benchmark to convergence was a 1% variation through 3 consecutive coarse frequencies. We estimated the TSRV to 240 coarse frequency candidates, picked the optimal coarse of each day by the previous criterion and between the daily optimal coarse frequencies, chose the most repeated. Figure 14

shows the convergence structure of both TSRV and RV. The plots allow us to see that a substantial portion of the volatility, from 2018 to the end of the sample, was due to microstructural activity and noise and not properly attributed to the fundamentals of the efficient price.

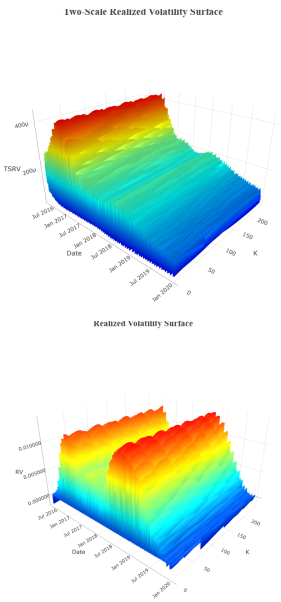


Figure 14: Volatility Estimators' Surfaces

Given the volatility series constructed by the TSRV estimator, we can then test if any of the PTAX windows are associated with higher or lower variance of the efficient price. To do so, we estimate an AR(1) to the TSRV augmented with dummies representing the contemporary effect of the i-th PTAX windows in the volatility. Table 1 reports that the first 3 PTAX windows are assoiated with higher volatility. These results also show that the effect of the PTAX window decreases monotonically during the day.

Table 1: AR(1) Coefficients to log(TSRV)

Variable	**Coefficient**	**Standard Error**	**t**	**p-value**
Intercept	-6.7347	0.1777	-37.91	$< 10^{-16}$
$\log(\text{TSRV}_{t-1})$	0.5196	0.0126	41.08	$< 10^{-16}$
PTAX 1	0.1893	0.0207	9.15	5.77×10^{-20}
PTAX 2	0.1766	0.0248	7.11	1.14×10^{-12}
PTAX 3	0.0606	0.0260	2.33	0.0196
PTAX 4	-0.0010	0.0290	-0.03	0.973

Price Discovery To analyse the price discovery dynamics, we decompose the estimations in two approaches. First we estimate, at the highest available frequency, the price discovery metrics to daily subsamples. Finally, to investigate if there is a particular asymmetry around the PTAX windows, we estimate again the metrics but to 10 minute subsamples.

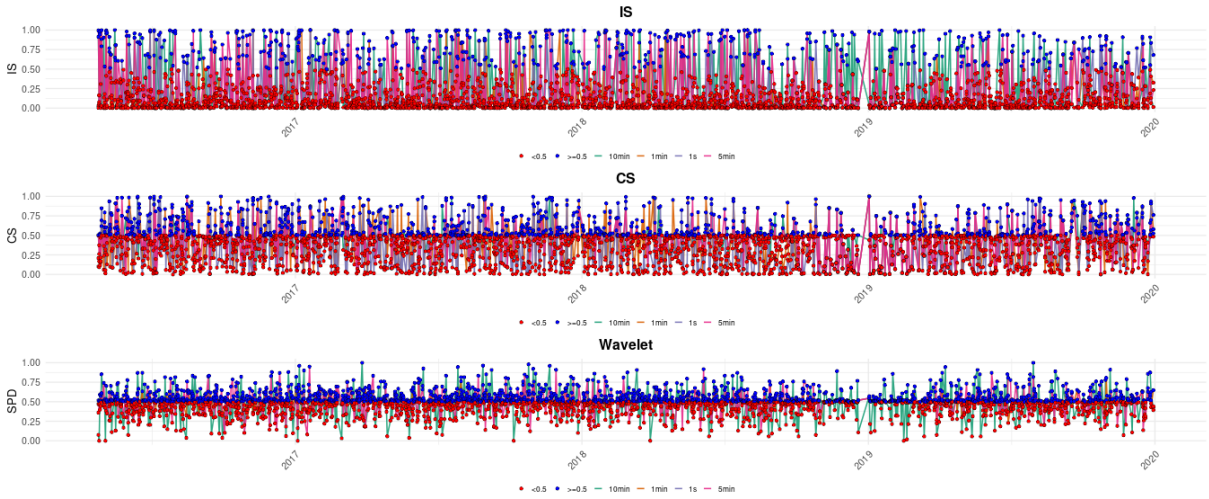


Figure 15: Daily Price Discovery Metrics to 4 Frequency Aggregations - Whole Sample

Figure 15 presents the time series of the bid contribution to the price discovery as measures by IS, CS and our spectral estimator. As it can be seen, the three estimators present short-run reversals, with no persistence dominance of any side of the book. While IS and CS present more variance and correlation between each other, the wavelet metric seems smoother and more stable. Overall, the series fluctuate around 50% all the time, suggesting that the quote leadership vanishes quickly and presents no pattern out of intraday scope.

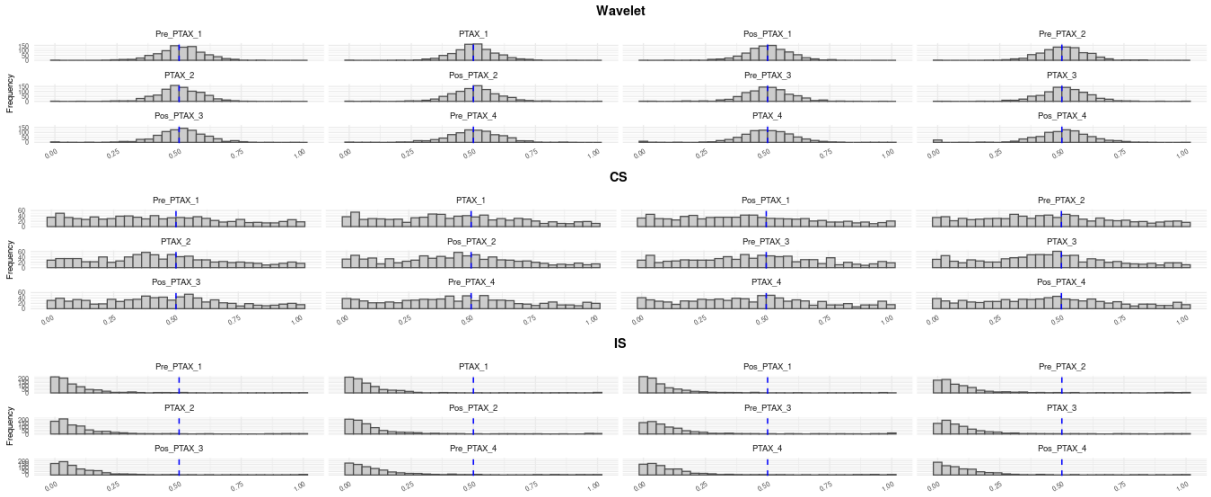


Figure 16: 10-minute Interval Histograms to PTAX Windows

Figure 16 display the histograms corresponding to the 10-minute intervals before, during and after each PTAX windows, showing the distribution (by the bid side) of the IS, CS and wavelet metrics. CS and Wavelet are both centered close to 0.5, differing only by the heaviness of the tails. However, there is a strong skewness from IS pointing to a strictly dominance of ask side through these intervals. As explored in the simulation in the appendix, the IS and CS present high variance among frequency variations and, therefore, we should not consider this as evidence of ask dominance - especially considering that the IS itself suggests no dominance in daily subsamples.

5 Final Remarks

This paper examined the efficiency of Brazil's FX futures market from the informational poinf of a view. We used a large-scale order-book dataset with over 8 billion observations and developed an algorithm to extract theoretically relevant time series from the raw data such as best quotes, bid-ask spread and order flow imbalance.Given the economic relevance and rich institutional design of this market, we addressed two central questions: how liquidity conditions evolved over the period and how price discovery dynamics performed. Additionally we took a look at the impact of PTAX windows in these channels.

Our results suggest that overall market efficiency remained stable throughout the period, with slight improvements in liquidity provision. Stable spreads, decreasing price

impact and small PTAX effects to price elasticity to OFI are evidence in this sense. However, we showed that a large portion of the increase in volatility seen in 2018 and 2019 was due to microstructure noise instead of variation in the efficient price. Moreover, despite the short-lived duration, we presented evidence that PTAX windows produces an increase in the intraday volatility - suggesting that there is still institutional improvements to be made in the market design.

Regarding the price formation leadership, both parametric and spectral measures point to a diffuse and time-varying structure of price discovery. There is no persistent dominance of either side of the book and the leadership alternates with short-run reversals either in daily subsamples or in the intervals surrounding PTAX windows.

Taken together, the evidence suggest that Brazil's FX futures market operates a highly liquid trading environment with efficient informational flow. In addition, institutional features such as PTAX windows affect volatility levels at specific moments but do not generate systematic distortions in liquidity or price discovery. We believe that future research may extend this framework with an analysis of regime shifting and delve deeper into the leadership duration of the price discovery process.

References

Agarwal, S., Singh, O., and Nagaria, D. (2017). Analysis and comparison of wavelet transforms for denoising mri images. *Biomed Pharmacol J*, 10(2):831–836.

Aguiar-Conraria, L., Azevedo, N., and Soares, M. J. (2008). Using wavelets to decompose the time–frequency effects of monetary policy. *Physica A: Statistical Mechanics and its Applications*, 387(12):2863–2878.

Aguiar-Conraria, L. and Soares, M. J. (2010). Oil and the macroeconomy: Using wavelets to analyze old issues. *Empirical Economics*, 40:645–655.

Baillie, R. T., Geoffrey Booth, G., Tse, Y., and Zabotina, T. (2002). Price discovery and common factor models. *Journal of Financial Markets*, 5(3):309–321.

Booth, G. G., So, R. W., and Tse, Y. (1999). Price discovery in the german equity index derivatives markets. *Journal of Futures Markets*, 19(6):619–643.

Cabrera, J., Wang, T., and Yang, J. (2009). Do futures lead price discovery in electronic foreign exchange markets? *Journal of Futures Markets*, 29(2):137–156.

Castro, U. C. d. (2017). Derivação de modelos de trading de alta frequência em juros utilizando aprendizado por reforço. Dissertação (mestrado profissional em economia), Fundação Getulio Vargas (FGV) – Escola de Economia de São Paulo. Accessed via Repositório Institucional Acadêmico, Fundação Getulio Vargas.

Chakravarty, S., Gulen, H., and Mayhew, S. (2004). Informed trading in stock and option markets. *The Journal of Finance*, 59(3):1235–1257.

Cont, R., Kukanov, A., and Stoikov, S. (2013). The price impact of order book events. *Journal of Financial Econometrics*, 12(1):47–88.

Daubechies, I. (1994). *Ten Lectures on Wavelets*. Springer-Verlag.

De Boyrie, M., Pavlova, I., and A.M., P. (2012). Price discovery in currency markets: Evidence from three emerging markets. *International Journal of Economics and Finance*, 4(12):61–75.

de Jong, F. (2002). Measures of contributions to price discovery: a comparison. *Journal of Financial Markets*, 5(3):323–327.

Engle, R. F. and Granger, C. W. J. (1987a). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Engle, R. F. and Granger, C. W. J. (1987b). Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276.

Evans, M. D. D. and Lyons, R. K. (2002). Order flow and exchange rate dynamics. *Journal of Political Economy*, 110(1):170–180.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417.

Farge, M. (1992). Wavelet transforms and their applications to turbulence. *Annual Review of Fluid Mechanics*, 24:395–458.

Ferreira, A., Mullen, R., Ricco, G., Viswanath-Natraj, G., and Wang, Z. (2025). Foreign exchange interventions and intermediary constraints. *Working Paper*.

Figuerola-Ferretti, I. and Gonzalo, J. (2010). Modelling and measuring price discovery in commodity markets. *Journal of Econometrics*, 158(1):95–107. Twenty Years of Cointegration.

Garbade, K. D. and Silber, W. L. (1983). Price movements and price discovery in futures and cash markets. *The Review of Economics and Statistics*, 65(2):289–297.

Garcia, M., Medeiros, M., and Santos, F. (2015). Price discovery in brazilian fx markets. *Brazilian Review of Econometrics*, 35(1):65–94.

Ghosh, I. and Chaudhuri, T. D. (2017). Fractal investigation and maximal overlap discrete wavelet transformation (modwt)-based machine learning framework for forecasting exchange rates. *Studies in Microeconomics*, 5(2):105–131.

Glur, C. (2025). *data.tree: General Purpose Hierarchical Data Structure*. R package version 1.2.0.

Gould, M. D., Porter, M. A., Williams, S., McDonald, M., Fenn, D. J., and Howison, S. D. (2013). Limit order books.

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *The Journal of Finance*, 50(4):1175–1199.

Joseph, A., Sisodia, G., and Tiwari, A. K. (2015). The inter-temporal causal nexus between indian commodity futures and spot prices: A wavelet analysis. *Theoretical Economics Letters*, 5(2):312–324.

Kumar, S. (2018). Price discovery in emerging currency markets. *Research in International Business and Finance*, 46:528–536.

Kumar, S., Pathak, R., Tiwari, A. K., and Yoon, S.-M. (2016). Are exchange rates interdependent? evidence using wavelet analysis. *Applied Economics*, 49(33):3231–3245.

Laurini, M. P., Furlani, L. G. C., and Portugal, M. S. (2008). Empirical market microstructure: An analysis of the brl/us exchange rate market. *Emerging Markets Review*, 9(4):247–265.

Li, F., Meng, G., Kageyama, K., Su, Z., and Ye, L. (2009). Optimal mother wavelet selection for lamb wave analyses. *Journal of Intelligent Material Systems and Structures*, 20(10):1147–1161.

Mancini, L., RANALDO, A., and WRAMPELMEYER, J. (2013). Liquidity in the foreign exchange market: Measurement, commonality, and risk premiums. *The Journal of Finance*, 68(5):1805–1841.

Matz, V., Smid, R., Starman, S., and Kreidl, M. (2009). Signal-to-noise ratio enhancement based on wavelet filtering in ultrasonic testing. *Ultrasonics*, 49(8):752–759.

Nigatu, G. and Adjemian, M. (2020). A wavelet analysis of price integration in major agricultural markets. *Journal of Agricultural and Applied Economics*, 52(1):117–134.

Osler, C. L., Mende, A., and Menkhoff, L. (2011). Price discovery in currency markets. *Journal of International Money and Finance*, 30(8):1696–1718.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4):1127–1139.

SCHWERT, G. W. (1989). Why does stock market volatility change over time? *The Journal of Finance*, 44(5):1115–1153.

Sung, G. (2021). The effects of the central bank intervention auctions on the foreign exchange futures market: evidence from brazil. Dissertação de mestrado, Faculdade

de Economia, Administração e Contabilidade de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, Brasil.

Torrence, C. and Compo, G. P. (1998). A practical guide to wavelet analysis. *Bulletin of the American Meteorological Society*, 79(1):61–78.

Veleda, D., Montagne, R., and Araújo, M. (2012). Cross-wavelet bias corrected by normalizing scales. *Journal of Atmospheric and Oceanic Technology*, 22(9):1401–1408.

Wang, Q., Teng, B., Hao, Q., and Shi, Y. (2021). High-frequency statistical arbitrage strategy based on stationarized order flow imbalance. *Procedia Computer Science*, 187:518–523. 2020 International Conference on Identification, Information and Knowledge in the Internet of Things, IIKI2020.

Zhang, L., Mykland, P. A., and Aït-Sahalia, Y. (2005). A tale of two time scales. *Journal of the American Statistical Association*, 100(472):1394–1411.

A Appendix

A.1 Descriptive Statistics and Tests

Table 2 presents distributional aspects of the extracted time series. The following descriptive statistics regard the cleaned version of the series at 1-second aggregation.

Table 2: Descriptive Statistics

Variable	Min	Q1	Median	Mean	Q3	Max	Variance	Skewness	Kurtosis
Best bid	3.0	3242.00	3478.50	3557.64	3860.50	4277.23	1.19e+05	0.3084	1.6471
Best ask	3047.5	3242.50	3479.00	3558.25	3861.00	4295.00	1.19e+05	0.3084	1.6467
Bid-ask Spread	0.5	0.50	0.50	0.6168	0.50	3246.00	2.94e+00	427.2363	7.04e+05
Buyer-Initiated Volume (contract)	0.0	0.00	0.00	12.8486	4.00	7325.00	3.02e+03	14.0731	466.3832
Seller-Initiated Volume (contract)	0.0	0.00	0.00	12.9040	4.00	11804.00	3.00e+03	15.0490	678.5598
Order Flow Imbalance (contract)	-9844.0	-1.00	0.00	-0.0555	1.00	6439.00	4.18e+03	-0.3178	296.4710
Mid-quote	1626.0	3242.25	3478.75	3557.95	3860.75	4285.00	1.19e+05	0.3084	1.6466

A.2 Unexpected Price Impact

Figure 17 shows the Bayesian Information Criterion to a grid of shifts in the PTAX window dummies replacing the OFI by its AR(1) residual as a proxy of unexpected order

flow. As a robustness check, it presents similar dynamics as the original specification.

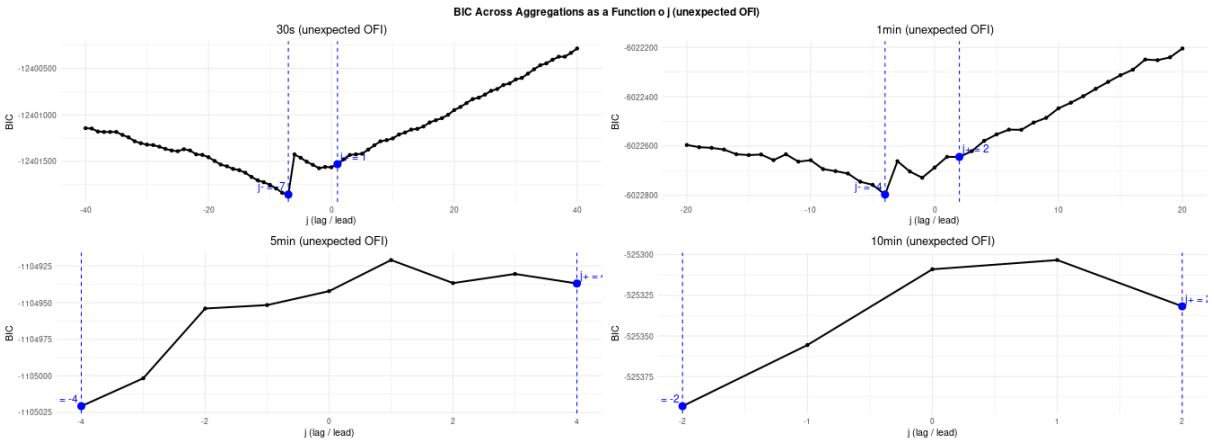


Figure 17

Figures 18 to 22 report the temporal dynamics to each parameter presented in the paper, but to the unexpected price impact functional form. As it can be seen, the parameters from the latter specification follow the same dynamics as the former one.

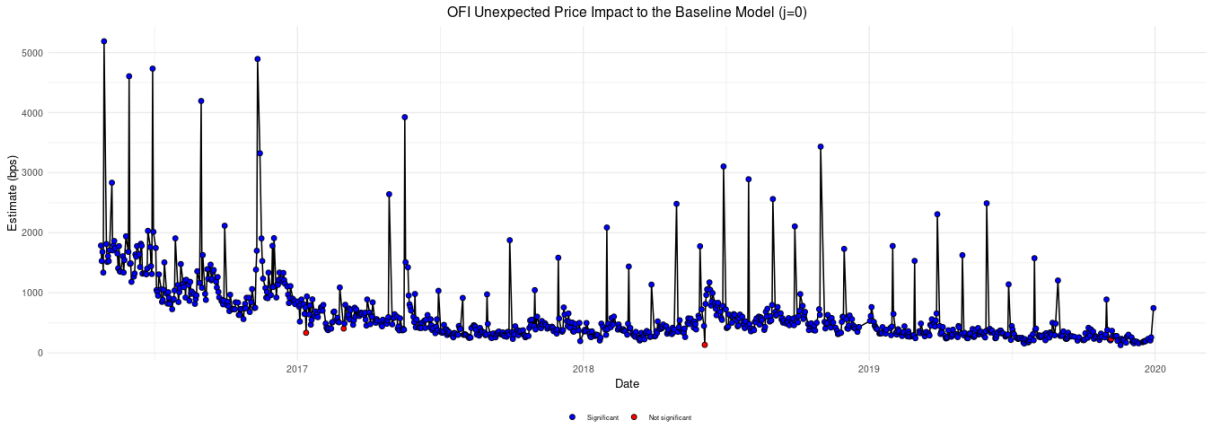


Figure 18

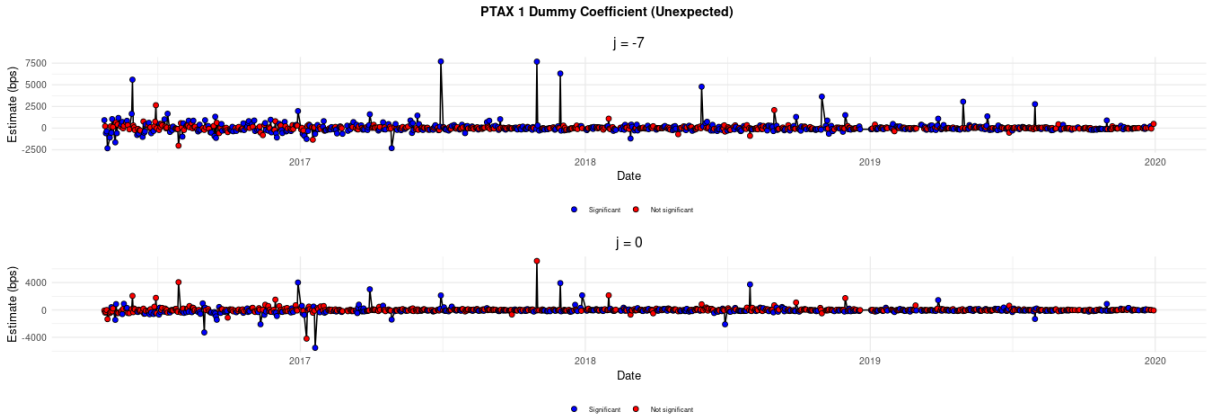


Figure 19

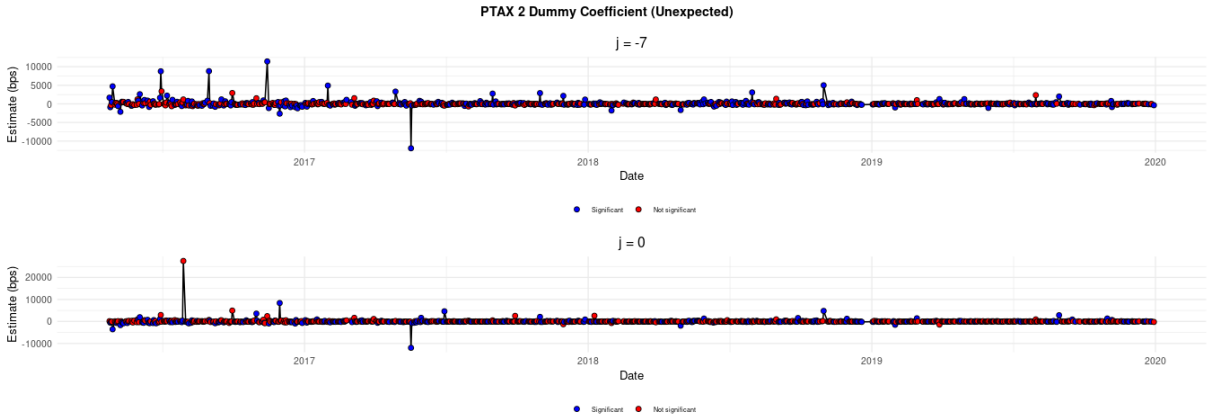


Figure 20

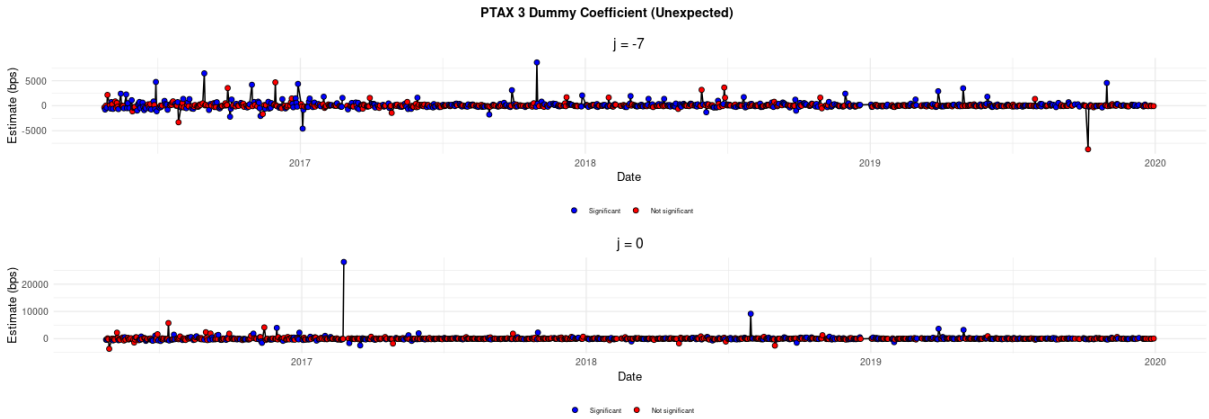


Figure 21

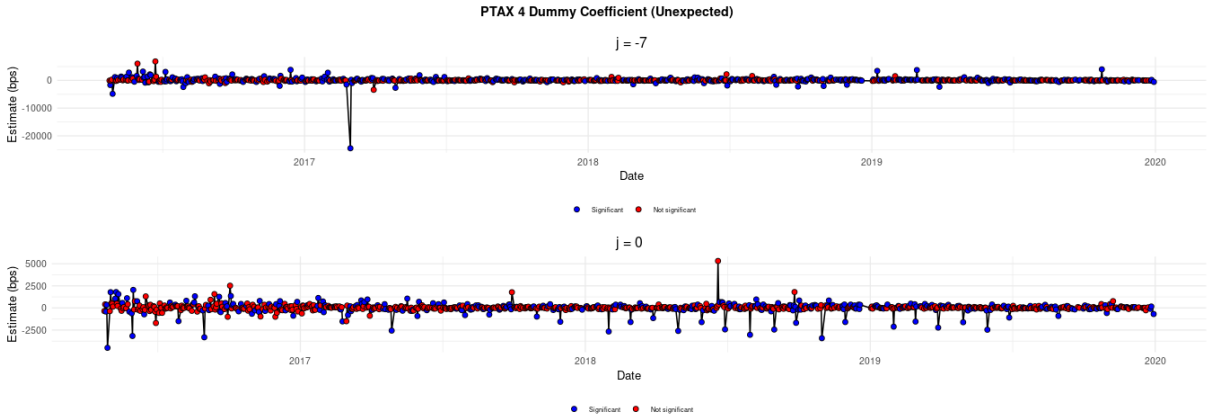


Figure 22

A.3 Wavelet Analysis

Parametrization As previously shown, a WT depends on the selected mother wavelet (wavelet function). In this sense, Li et al. (2009) the selection of the mother wavelet interferes with the accuracy and efficiency of the transformation, and this selection is case-dependent. Surveying the literature, most the applications of WTs use mother wavelets based on previous works. So far, we have not been able to find WTs' applications in economics or finance that test their parametrisation.

Some metrics can be used to assess the fit of a wavelet function to the original signal: (i) signal-to-noise ratio (SNR) Matz et al. (2009); Agarwal et al. (2017); (ii) mean squared error (MSE) Agarwal et al. (2017); (iii) Shannon entropy of wavelet coefficients (EWC) Li et al. (2009). While the first two are usually employed to analyze discrete WTs, the EWC can be used for both CWTs and DWTs and is defined as:

$$EWC(a) = \sum_{i=1}^{n} \frac{w_{a,i}^2}{(\sum_{a=1}^{m} \sum_{i=1}^{n} w_{a,i}^2)} ln \left(\frac{w_{a,i}^2}{\sum_{a=1}^{m} \sum_{i=1}^{n} w_{a,i}^2)} \right). \tag{23}$$

In this framework Li et al. (2009), the higher the energy concentrated at scale a, the lower the EWC. Therefore, for a fixed m, we select the wavelet function which achieves the minimum EWC compared to the others. For a m grid, we choose the pair $(m, \psi(t))$ that presents the global minimum EWC.

In our particular case, we define our m grid as $\{80, 81, ..., 100\}$ for $n > 80$ and

$\{0.08n, 0.081n, ..., 0.1n\}$ otherwise. We test the main complex-valued wavelet functions: Morlet, $\psi_m(t)$, and Paul, $\psi_m(t)$. The reason for excluding real-valued functions is their lack of phase content, which makes them useful mainly for detecting jumps and discontinuities but not for interpreting lead-lag relationships. Their functional forms and respective e-folding time are:

$$\psi_m(t) = \pi^{-\frac{1}{4}} e^{i\omega t} e^{-\frac{t^2}{2}}, g_m(a) = \sqrt{2a}, \tag{24}$$

$$\psi_p(t) = \frac{2^{m_0} i^{m_0} m_0! (1-it)^{-(m_0+1)}}{\sqrt{\pi(2m_0)!}}, g_p(a) = \frac{a}{\sqrt{2}}. \tag{25}$$

where ω and m_0 are nondimensional time parameters set in such a way that the admissibility conditions are sustained.

The intuition behind testing these two specific functions is also to understand what feature of the series carries the most significant information. This is accomplished considering that Morlet, with its sinusoidal formulation, gives us more frequency resolution, while Paul, considering its single peak, tends to highlight time resolution.

Simulation We are interested in understanding how a wavelet-based metric behaves in front of different scenarios and comparing it to usual price discovery metrics. Because analytical solutions are hard to obtain, we simulate series that closely resemble our data in such a way that they capture the most usual dynamics. We use assumptions that are reasonable and consistent with the observed patterns. Therefore, assume three hypotheses:

- **H1** - Bid and ask quotes follow a vector auto-regressive structure (VAR);
- **H2** - Bid and ask quotes are non-stationary;
- **H3** - There is a long-run equilibrium between these quotes, i.e they are $CI(1,1)$

H1 comes from the intuitive idea that if a good has different prices, they must relate. **H2** is a consequence of the efficient price being a martingale and, consequently, non-stationary. **H3** comes as a consequence of (Roll, 1984) definition that the bid-ask spread corresponds to a small region where the efficient price varies - thus the quotes should not drift way in the long-run.

Suppose b_t and a_t represent, respectively, the best bid and ask quotes in the FX futures market. Assume that the data generating process follows this dynamics $\forall t \in \{1, 2, ..., n\}$:

$$b_t = \rho_b b_{t-1} + \delta_b a_{t-1} + \gamma^b + \epsilon_t^b, \tag{26}$$

$$a_t = \rho_a a_{t-1} + \delta_a b_{t-1} + \gamma^a + \epsilon_t^a. \tag{27}$$

where $\epsilon_i \sim N(0, \sigma_i^2)$ are iid for $i = \{a, b\}$.

Given **H3**, from the (Engle and Granger, 1987b) representation theorem, we can write our system as a Vector Error Correction Model (VECM), where $\beta = (\beta_a, \beta_b)$ is the cointegration vector and $\alpha = (\alpha_a, \alpha_b)$ represents the speed of adjustment. In other words, quotes react to a deviation from the long-run equilibrium converging to the mid-quote:

$$\Delta b_t = \alpha_b(\beta_a a_{t-1} - \beta_b b_{t-1}) + \epsilon_t^b, \tag{28}$$

$$\Delta a_t = \alpha_a(\beta_a a_{t-1} - \beta_b b_{t-1}) + \epsilon_t^a. \tag{29}$$

Rewriting and normalizing β and α, we have $(\beta_a, \beta_b) = \left(1, \frac{(1-\rho_a)}{\delta_a}\right)$ and $(\alpha_a, \alpha_b) = \left(\delta_a, \frac{-(\delta_a \delta_b)}{(1-\rho_a)}\right)$. In order to $a_t - \beta_b b_t \sim CI(1, 1)$ it must be the case that $\rho_a > -1$ and $\delta_a \delta_b + \rho_a^2 < 1$. These restrictions follow from the root conditions of our primitive system's characteristic equation, which also imposes that one of our auto-regressive parameters must be endogenously defined based on our previous normalisation:

$$\rho_b = \frac{(1 - \rho_a) - \delta_a \delta_b}{(1 - \rho_a)}. \tag{30}$$

Given the VECM, we can map the econometric measures of price discovery from the primitive system. Denote Ω as the variance-covariance matrix of the innovative process (ϵ_t) of both quotes. Additionally, from our hypotheses of the quotes being $I(1)$, consider Ξ_i from the vector moving average (VMA) representation, $\Delta f^i = \Xi_i(L)\epsilon_t^i$, where L represents

the lag operator. Then, the information share (IS) and component share (CS) to the i-th quote can be written as:

As a way to visualize how the phase analysis captures lead-lag relations, we use our model to simulate three scenarios: ($\mathbf{b} \Rightarrow \mathbf{a}$) bid causes ask, ($\mathbf{a} \Rightarrow \mathbf{b}$) ask causes bid and ($\mathbf{b} \Leftrightarrow \mathbf{a}$) bid and ask have feedback effects. To do so, denote $\hat{\sigma}_i$ as the observed variance of the proxy for the innovation process of the i-th quote. Define the total variance of the innovation process as $\hat{\sigma} = \hat{\sigma_a} + \hat{\sigma_b}$. Therefore, as Table 2 displays, each causation scenario is based on the portion of the total variance that each quote conveys. Additionally, table 3 shows the fixed VAR parameters used to simulate the series.

Table 3: Innovation Process Parameters

Parameter/Scenario	$a \Rightarrow b$	$b \Rightarrow a$	$a \Leftrightarrow b$
σ_a	$0.75\hat{\sigma}$	$0.25\hat{\sigma}$	$0.5\hat{\sigma}$
σ_b	$0.25\hat{\sigma}$	$0.75\hat{\sigma}$	$0.5\hat{\sigma}$

Source: Own elaboration
Notes: Variance parameters of ϵ_a and ϵ_b as a function of the total variance of actual series

Table 4: Simulation Parameters

γ_a	γ_b	ρ_a	ρ_b	δ_a	δ_b
0	0	0.5	0.5	0.5	0.5

Source: Own elaboration
Notes: Simulation parameters based on equations (1) and (2)

As previously mentioned, in order to assess the causality direction by a conventional econometric approach, we use the IS (Hasbrouck, 1995) and CS (Booth et al., 1999). We follow (SCHWERT, 1989) in the lag definition as $\left\lceil 12 \times (\frac{n}{100})^{\frac{1}{4}} \right\rceil$.

The main way to extract anticipation and, therefore, to unveil price discovery within the wavelet framework is through phase content. In this sense, we built our price discovery metrics by the previous definition of lead-lag intervals in the $\phi_{xy}(a)$. Denoting our frequency aggregations by ϕ_{xy}, we count the time points where $\phi_{xy} \in [0, \pi]$ and attribute them to ask, otherwise ($\phi_{xy} \in [-\pi, 0)$) to bid. The percentage of points to each quote corresponds to its price discovery value. The rationale is to create an interval measure,

which can not only be understood in the same way as IS and CS, but also employed continuously, tracking the lead-lag dynamics at each time step.

Tables 4 and 5 present the results by minute and second aggregation, respectively.

Table 5: Price Discovery - simulated Series (by Minute)

Metrics	$a \Rightarrow b$		$b \Rightarrow a$		b $\Leftrightarrow$ a	
	Ask	Bid	Ask	Bid	Ask	Bid
Information Share	84.73%	15.26%	48.07%	51.93%	27.30%	72.70%
Component Share	57.56%	42.44%	62.68%	37.32%	40.49%	59.51%
Coherence Weighing	71.98%	28.02%	28.57%	71.43%	56.78%	43.22%
Linear discount	69.05%	30.95%	35.71%	64.29%	58.80%	41.20%
Quartile Threshold (25)	66.02%	34.98%	47.80%	52.20%	57.15%	42.85%
Quartile Threshold (50)	71.79%	28.21%	31.87%	68.13%	56.41%	43.59%
Quartile Threshold (75)	75.26%	24.73%	29.30%	70.70%	59.89%	40.11%

Source: Own elaboration
Notes: Results of price discovery metrics to simulated series by minute.

Table 6: Price Discovery - simulated Series (by Second)

Metrics	$a \Rightarrow b$		$b \Rightarrow a$		b $\Leftrightarrow$ a	
	Ask	Bid	Ask	Bid	Ask	Bid
Information Share	73.12%	26.87%	26.92%	73.08%	50.25%	49.75%
Component Share	48.71%	51.13%	51.24%	48.76%	50.41%	49.59%
Coherence Weighing	71.05%	28.95%	28.77%	71.23%	50.48%	49.52%
Linear discount	65.38%	34.62%	34.54%	65.56%	50.46%	49.54%
Quartile Threshold (25)	56.82%	43.18%	42.81%	57.19%	50.63%	49.37%
Quartile Threshold (50)	68.85%	31.15%	31.49%	68.51%	50.40%	49.60%
Quartile Threshold (75)	72.57%	27.43%	27.97%	72.03%	50.40%	49.60%

Source: Own elaboration
Notes: Results of price discovery metrics to simulated series by second.

As can be seen, IS and CS show considerable sensibility to sample size and internal inconsistency between each other, with IS overcoming CS accuracy. Analyzing the proposed frequency aggregation methods, we see that Coherence Weighing tends to be more reliable, being less sensible to sample size. The probability of falsely assuming an asymmetry in the price formation seems to diminish asymptotically for both VEC and wavelet metrics, especially for the former. In addition, one may note that the CS diverged considerably from the IS in the larger sample. Taking a look at the transition from $a \Rightarrow b$ to $a \Rightarrow b$ in Table 4, we observe, at minute aggregation, more dispersed errors between VEC and wavelet methods. The transition is smoother at the second aggregation, as shown in Table 5, and the difference between VEC and wavelet methods decreases.

Naturally, we must be conservative in extending this efficiency, given that we use only two observed moments in our simulation, and some more complex dynamics can be hidden in higher-order moments. Moreover, we must highlight that we are simulating a constant direction, i.e. no structural breaks in the innovation process.

A.4 Spread Algorithm

Algorithm 1: Reconstruction of the Bid–Ask Spread from Order and Trade Raw Data

Input: Order Data $\mathcal{O} = \{y_i\}$, Trade Data $\mathcal{N}$

Output: Time series $\{b(t), a(t), s(t)\}$

1: Identify auction time $t_{\text{auc}} = \min_{n \in \mathcal{N}} t_n$

2: Remove trades $n \in \mathcal{N}$ with $t_n = t_{\text{auc}}$

3: Remove orders $y \in \mathcal{O}$ with $t_y \leq t_{\text{auc}}$

4: Restrict $\mathcal{O}$ to orders whose Sequential Order Number (SON) appears in $\mathcal{N}$

5: Sort $\mathcal{O}$ by submission time

6: Initialize empty limit order book $\mathcal{L}(t) = \{\mathcal{A}(t), \mathcal{B}(t)\}$

7: **for** *each order* $y = (p_y, \omega_y, t_y, side_y, status_y) \in \mathcal{O}$ **do**

    if *status = New* **then**

        Insert y into $\mathcal{A}$ or $\mathcal{B}$ using price–time priority

    else if *status = Modified* **then**

        Remove previous instance and reinsert with updated attributes

    else if *status = Partial Executed* **then**

        Update remaining quantity

    else if *status* $\in$ {*Canceled, Executed, Expired*} **then**

        Remove y from the book

    Compute $b(t) = \max_{x \in \mathcal{B}(t)} p_x$ and $a(t) = \min_{x \in \mathcal{A}(t)} p_x$

    while $b(t) \geq a(t)$ **do**

        Let L_B and L_A be the queues at $b(t)$ and $a(t)$

        $q^* = \min\{\sum_{x \in L_B} |\omega_x|,\ \sum_{x \in L_A} |\omega_x|\}$

        Execute q^* using price–time priority; remove exhausted orders/levels

        Update $b(t), a(t)$

    end

    Record snapshot $(t_y, b(t_y), a(t_y), s(t_y) = a(t_y) - b(t_y))$

end

30: Deduplicate snapshots by timestamp (keep last observation)

31: Return reconstructed spread series