Semi-Analytical Latency-Aware Distributional Inference for Limit Order Books

Bernardo Teixeira[†, a, b]
Renan Avila[a, b]
Oswaldo Costa[a, b]
Leandro Maciel[c]

[a]Escola Politécnica da Universidade de São Paulo
[b]BTG Alpha Lab, BTG Pactual
[c]Faculdade de Economia, Administração, Contabilidade e Atuária da Universidade de São Paulo

Abstract In modern electronic markets, execution at millisecond horizons requires fast, state-conditional probabilities for price moves and passive fills. This creates a need for models that are both computationally tractable and directly interpretable in execution settings. We study whether the semi-analytical framework of Cont, Stoikov, and Talreja (CST) can provide such signals in Brazilian equities. First, we present the first CST-based analysis of latency effects in this market, incorporating reaction delays into market-making probabilities. Second, we assess CST adherence by comparing model-implied and realized event distributions. We also introduce the BTG-OBD-A26 dataset, a reproducible message-level order-book-dynamics dataset for Brazilian equities. Our results show that the CST framework captures directional and fill probabilities reasonably well, whereas the latency extension is less accurate because the model does not capture temporal clustering in order flow.

Keywords: Financial engineering; Limit order book; Market microstructure; Fill probability; Mid-price direction; Continuous-time Markov chain; Latency; Market making.

JEL codes: G14, G17, C63, D47.

1. Introduction

Execution systems in electronic equity markets at millisecond and microsecond horizons repeatedly decide between aggressive execution, passive posting, and quote updates. These decisions depend on short-horizon conditional probabilities, especially the probability of the next mid-price direction

How to cite: Teixeira, B., Avila, R., Costa, O., & Maciel, L.. Semi-Analytical Latency-Aware Distributional Inference for Limit Order Books. Encontro Brasileiro de Finanças, 2026.
[†]bernardoteixeira@usp.br

and the probability of passive execution at the best prices. For practical use in this market, these probabilities must be state-conditional, fast to compute, and interpretable enough to support trading and execution decisions under tight latency constraints.

The market microstructure literature already offers a broad range of techniques, from stochastic models to machine-learning approaches, for mapping the local state of the order book into execution-relevant probabilities (Zaznov et al., 2022; Briola et al., 2025). Among them, the semi-analytical framework of Cont et al. (2010) is particularly attractive for electronic equity markets. In this framework, the best-price state (a,b) denotes the current sizes of the best ask and best bid queues, while quantities such as $(p_{\uparrow},p_{\text{fill}})$ summarize, respectively, the probability of the next upward mid-price move and the probability that a passive order posted at the best price is executed. The main attraction of the CST framework is operational tractability: after reducing the best-price dynamics to queue-depletion problems, the quantities of interest can be expressed through first-passage distributions and recovered from Laplace-domain representations without resorting to heavy Monte Carlo simulation (Abate and Whitt, 1992, 1995, 1999). This makes the model especially appealing when one needs repeated evaluations across many states, symbols, and intraday decision times.

The objective of this paper is to evaluate whether a transparent semi-analytical limit-order-book model can serve as a practical probabilistic layer for execution in the Brazilian equity market. To this end, we introduce message-level data from the **BTG-OBD-A26** dataset from BTG Pactual Dataservices, covering order-book dynamics in liquid B3 equities over the period from October 2025 to January 2026, calibrate the CST framework day-ahead in the one-tick-spread regime, extend it to account for latency, and compute state-conditional probabilities for the next price direction and passive execution from the best-price state. We then assess the model by comparing theoretical and realized event frequencies across assets, states, and volatility regimes, with emphasis on calibration, ranking, and operational interpretability.

Our main findings are that the CST framework retains economically meaningful predictive structure in this market, especially in its ability to rank short-horizon directional and execution outcomes. The empirical results indicate monotone agreement between model-implied and realized probabilities, while the latency extension provides a tractable way to incorporate implementation delay into execution-relevant signals. Taken together, these results suggest that even a simple structural model can remain useful when the objective is not unrestricted forecasting accuracy, but rather fast, interpretable inference.

Our contributions are twofold. First, we extend the CST framework to a latency-aware setting in which reaction delays directly affect actionable directional probabilities. Second, we provide a systematic empirical assessment of CST adherence in the Brazilian equity market by comparing model-implied and realized best-price event distributions across assets and volatility regimes. As an additional contribution, we introduce the **BTG-OBD-A26** dataset, a message-level order-book-dynamics dataset for Brazilian equities with a reproducible reconstruction and event-classification pipeline. The latency extension pursued here also follows a future-work direction suggested by Cont et al. (2010), namely to study how implementation delay distorts the actionability of CST state-conditional probabilities. The broader goal is not to claim that CST is a fully realistic market model, but to evaluate how far a transparent structural model can go as a fast probabilistic layer for execution.

The remainder of the paper is organized as follows. Section 2 reviews the basic mechanics of the limit order book and positions the paper within the structural, predictive, and simulation-oriented literatures. Section 3 introduces the semi-analytical best-price model, recalls the queueing-theoretic preliminaries, and formalizes the CST state dynamics. Section 4 derives the main probabilities of interest, namely the probability of the next upward mid-price move, the passive fill probability at the best prices, and the probability of a mid-price change during a fixed latency window. Section 5 describes the BTG-OBD-A26 dataset and the calibration procedure used to estimate the model on Brazilian equities. Section 6 presents the empirical experiments for directional prediction, best-price fill prediction, and the probability of a mid-price change during latency for market making. Section 7 discusses the results, and Section 8 concludes.

2. Order Book Background and Related Literature

2.1 Limit Order Book Background

Modern electronic equity markets are organized as continuous double auctions in which trading intentions are stored and matched through a limit order book. At each instant, the book records the visible supply and demand available at discrete price levels separated by the tick size. Buy orders populate the bid side of the book and sell orders populate the ask side. The highest bid is the best bid, the lowest ask is the best ask, and the difference between them is the bid-ask spread. The midpoint between the best bid and best ask is often used as a reference price, but actual execution occurs against the standing liquidity available on one side of the book or the other (Gould et al., 2013).

Throughout the paper, a book level means a queue identified by its price distance, in ticks, from the opposite best quote. In the one-tick-spread regime, level $i = 1$ corresponds to the best quote on the event side, level $i = 2$ to the next price level one tick further into the book, and so on. This indexing is useful because it lets us describe depth and order flow relative to the prevailing best prices rather than in absolute price units.

Three event types drive the evolution of the book. A limit order adds liquidity to a queue at a chosen price level, a market order consumes standing liquidity from the opposite side, and a cancellation removes previously posted volume. Under the standard price-time priority rule, newly arriving limit orders always join the back of the queue at their chosen price. Queue depletion, in turn, can occur in two distinct ways: through execution, which removes only the order currently at the head of the queue, or through cancellation, which can remove any resting order in the queue. Queue position is therefore economically important: two agents posting at the same price do not face the same fill probability if one stands behind the other in the queue.

For short-horizon modeling, the most informative part of the book is often the neighborhood of the best prices, that is, the best bid and best ask queues. When the spread is narrow, local price formation is tightly linked to the competition between replenishment and depletion at those two queues. New limit orders replenish depth, while market orders and cancellations erode it. If one of the best-price queues is exhausted before the other, the best quotes shift and the mid-price moves accordingly. This queue-based view is the natural entry point for structural models that describe the book as a stochastic system evolving through event arrivals (Cont et al., 2010; Cont and de Larrard, 2013). Figures 1, 2, and 3 summarize the three ingredients used throughout the paper: the local best-price geometry, the queue-depletion race that determines the next price move, and the queue-position logic that determines passive execution.

This perspective is particularly useful for the questions studied in this paper. Figure 1 fixes the basic geometry of the one-tick-spread book and the meaning of a tagged order inside a queue. Figure 2 then isolates the directional event, namely which best-price queue is depleted first. Figure 3 isolates the execution event, namely whether the bid-side volume in front of and including the tagged order is cleared before the ask queue disappears. Once execution decisions are implemented with non-negligible delay, one may then ask how latency changes the relevance of those same state-conditional probabilities. These considerations motivate our focus on the one-tick-spread regime and on the immediate dynamics of the best-price queues. In that regime, the local race between order arrivals, executions, and cancellations captures a large part

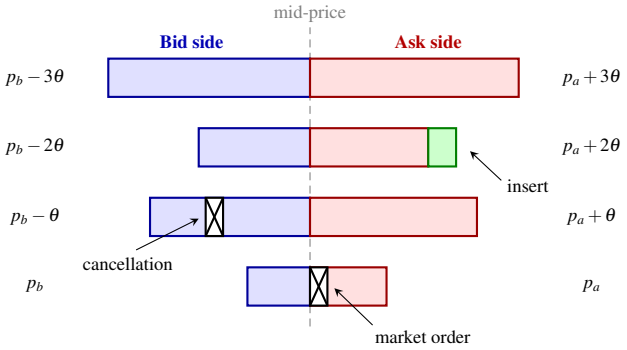

Figure 1

Local best-price geometry. This schematic represents the visible bid and ask queues around the mid-price in a limit order book. The blue bars denote bid queues and the red bars denote ask queues at successive price levels. Here p_b and p_a are the current best bid and best ask, while the labels $p_b - k\theta$ and $p_a + k\theta$ indicate deeper price levels separated by one price increment in the stylized price ladder. The green block appended to the ask queue at price $p_a + 2\theta$ illustrates a newly arriving limit order joining the back of that queue under price-time priority. The crossed white block inside the bid queue at price $p_b - \theta$ illustrates a cancellation, which may remove a resting order from any position in that queue. The crossed white block at the best ask p_a illustrates an incoming market order from the bid side, which consumes standing liquidity at the front of the ask queue.

of the short-horizon behavior that matters for execution.

2.2 Related Work

Recent surveys classify the limit order book literature into a few broad families: structural queueing and point-process models, predictive machine-learning models, and simulation-oriented frameworks for backtesting or agent training (Zaznov et al., 2022; Jain et al., 2024; Briola et al., 2025). This classification is useful because it makes clear that the field has not evolved along a single line. Rather, the literature has advanced along two complementary directions: on one side, increasingly flexible predictive models aimed at maximizing short-horizon forecasting performance; on the other, structural models that preserve economic interpretability, analytical tractability, and direct links to execution quantities such as fill probabilities, depletion races, and price impact.

The structural branch starts from early statistical and zero-intelligence descriptions of the order book as a stochastic queueing system or interacting-

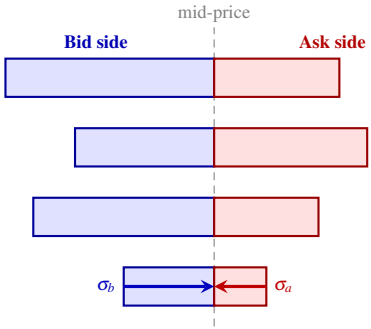

Figure 2

Upward-move probability at the best prices. This schematic shows the queue-depletion race that determines the probability of the next upward mid-price move from the best-price state (a,b), where a is the initial best-ask depth and b is the initial best-bid depth. The arrows labeled σ_a and σ_b denote the ask-queue and bid-queue depletion times, respectively. The quantity $p_\uparrow(a,b)$ is the probability that the ask queue is exhausted before the bid queue. When that occurs, the best ask is removed, the price ladder shifts upward, the next ask level becomes the new best ask, and, in the one-tick-spread setting, the previous ask price becomes the new best bid, so the mid-price increases.

particle system (Bouchaud et al., 2002; Smith et al., 2003; Luckock, 2003). Within this tradition, Cont et al. (2010) occupies a particularly important position. Their model is richer than a homogeneous Poisson benchmark because it delivers state-conditional event probabilities at the best prices, yet it remains substantially more tractable and transparent than large-scale simulators or fully nonparametric forecasting systems. Related work extended the same structural logic to queue depletion and short-horizon price dynamics (Cont and de Larrard, 2013), to order-book-event price impact (Cont et al., 2014), and to richer self-exciting order-flow specifications based on Hawkes processes (Bacry et al., 2015). For our purposes, the main advantage of the CST framework is precisely this balance between realism and usability: it yields economically interpretable, state-dependent probabilities that can be recomputed quickly and embedded directly in an execution rule. This is especially valuable when the objective is not only to predict direction, but to do so in a way that remains auditable, latency-aware, and operationally compatible with real-time trading.

In parallel, the predictive literature moved from handcrafted features and shallow classifiers toward increasingly expressive supervised-learning architectures. Early high-frequency forecasting papers used top-of-book variables, depth profiles, and support vector machines or related classifiers (Kercheval

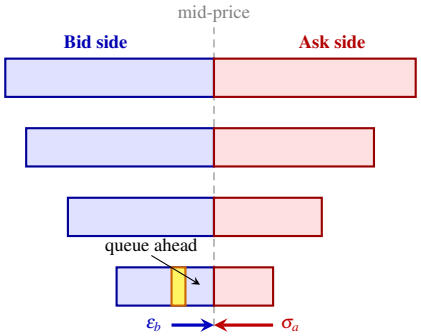

Figure 3

Fill probability for a tagged bid order. This schematic shows the execution event for a tagged order posted at the best bid from the best-price state (a,b), where a is the initial best-ask depth and b is the initial best-bid depth. The yellow rectangle marks the tagged bid order, and the label "queue ahead" indicates the resting bid-side volume that must be cleared before that order can execute. The quantity ε_b denotes the depletion time of the bid-side volume consisting of the orders ahead of the tagged order together with the tagged order itself, while σ_a denotes the depletion time of the best ask queue. The quantity $p_{\text{fill}}(a,b)$ is the probability that this bid-side volume is depleted before the ask queue is exhausted, so that the tagged order executes at the best bid before the price moves away.

and Zhang, 2015). This line of work was followed by deep architectures such as DeepLOB (Zhang et al., 2019) and by broader evidence that deep learning can extract persistent regularities from high-frequency price formation (Sirignano and Cont, 2019). More recent work emphasizes that predictive performance depends crucially on data representation, prediction horizon, and asset-specific microstructure: see, for example, Lucchese et al. (2022), Kolm et al. (2023), Prata et al. (2023), and the microstructural guide of Briola et al. (2025). These contributions make clear that modern forecasting models can be powerful, but they also reinforce a limitation that matters for this paper: high raw predictive accuracy does not automatically produce compact conditional objects that are easy to interpret, stress, recalibrate, and translate into execution decisions under tight latency constraints.

This distinction is central for positioning our contribution. We do not use the CST model as a generic black-box forecaster, nor do we claim that a low-dimensional structural model should dominate modern deep architectures in unconditional classification metrics. Instead, we exploit the comparative advantages of the Rama Cont line of work: semi-analytical tractability, ex-

plicit dependence on observable best-price-state variables, modest calibration requirements, and a natural connection between model outputs and execution-relevant events. These features make CST particularly well suited for our setting, where the goal is to map the instantaneous state (a,b) into actionable probabilities for direction and execution, and then study how those probabilities are distorted by implementation latency.

Finally, our empirical setting addresses a gap in the literature. Most public benchmarks and much of the empirical evidence behind both structural and predictive LOB models come from North American, European, or major Asian markets. By constructing the BTG-OBD-A26 dataset and evaluating CST-style probabilities on Brazilian equities, we contribute evidence on how this class of structural models transfers to a different microstructure environment. The latency extension then adds a layer that is economically important but often left implicit in forecasting studies: a probability is only actionable if it remains relevant by the time the order or quote update effectively reaches the exchange.

3. Semi-analytical Best-Price Model

In the CST model, each price level is represented as a queue that is replenished by limit-order arrivals and depleted by market orders and cancellations. Under this representation, the depletion time of a birth–death process corresponds to the time required for a queue to be exhausted and, therefore, for the best price to change. The execution of a tagged order, in turn, depends on the depletion time of a pure-death process describing the volume ahead of the order together with the order itself, since newly arriving limit orders cannot obtain priority over it. This section formalizes the CST framework by first recalling the queueing-theoretic preliminaries and then describing the best-price dynamics as a continuous-time Markov chain.

3.1 Preliminaries on Queueing Systems

Queueing theory provides the mathematical framework used in this section. We consider two distinct stochastic queues and study their evolution through continuous-time Markov dynamics. The first object is a birth–death process, used to represent a queue whose size may both increase and decrease over time. The second object is a pure-death process, used to represent a queue whose size can only decrease. These two processes form the basic queueing ingredients of the analysis developed below. Finally, we study how to compute probabilities of the form $\mathbb{P}[X < Y]$ for independent random variables X and Y through their Laplace transforms, which is the key step for computing directional probabilities and fill probabilities in the CST framework.

Hitting time of birth-death process. Consider a continuous-time Markov process on $\mathbb{N}_0$ (the natural numbers including 0) with birth rates λ and death rates μ_i, where $\mu_i > 0$ in every nonzero state and $\mu_0 = 0$. Let σ_x denote the first hitting time of state 0 starting from $x \geqslant 1$ elements in the queue. Then

$$\sigma_x = \sigma_{x,x-1} + \sigma_{x-1,x-2} + \cdots + \sigma_{1,0}, \tag{1}$$

where $\sigma_{i,i-1}$ is the first hitting time of $i-1$ starting from i. By the strong Markov property, $\{\sigma_{i,i-1}\}$ are independent. Therefore, if $f_x(t)$ is the density of σ_x and $\hat{f}_x(s)$ its Laplace transform, then

$$\hat{f}_x(s) = \prod_{i=1}^{x} \hat{f}_{i,i-1}(s), \tag{2}$$

where s is the complex variable of the Laplace transform and $\hat{f}_{i,i-1}(s)$ is the Laplace transform of the density of $\sigma_{i,i-1}$.

Abate and Whitt (1999) provide a semi-analytical expression for each factor:

$$\hat{f}_{i,i-1}(s) = -\frac{1}{\lambda} \Phi_{k=1}^{\infty} \left(\frac{-\lambda \mu_i}{\lambda + \mu_i + s} \right), \tag{3}$$

where Φ denotes the continued-fraction operator. Given sequences $\{c_n, n \geqslant 1\}$ and $\{d_n, n \geqslant 1\}$ of complex partial numerators and denominators, with $c_n \neq 0$ for all $n \geqslant 1$, define

$$w_n = t_1 \circ t_2 \circ \cdots \circ t_n(0), \qquad t_k(u) = \frac{c_k}{d_k + u}, \qquad k \geqslant 1, \tag{4}$$

where $\circ$ denotes composition. If $w \equiv \lim_{n\to\infty} w_n$ exists, the continued fraction is said to be convergent and its value is denoted by

$$w \equiv \Phi_{k=1}^{\infty} \frac{c_k}{d_k}. \tag{5}$$

In visual form,

$$\Phi_{k=1}^{\infty} \frac{c_k}{d_k} = \cfrac{c_1}{d_1 + \cfrac{c_2}{d_2 + \cfrac{c_3}{d_3 + \cdots}}}; \tag{6}$$

Combining the expressions above yields the semi-analytical representation of $\hat{f}_x(s)$:

$$\hat{f}_x(s) = \prod_{i=1}^{x} \left(-\frac{1}{\lambda} \Phi_{k=1}^{\infty} \left(\frac{-\lambda \mu_i}{\lambda + \mu_i + s} \right) \right). \tag{7}$$

Hitting time of pure-death process. Now consider a pure-death continuous-time Markov process on $\mathbb{N}_0$ with death rates μ_i, where $\mu_i > 0$ in every nonzero state and $\mu_0 = 0$. Let ε_x be the first hitting time of 0 from $x \geqslant 1$ elements in the queue. Then

$$\varepsilon_x = \varepsilon_{x,x-1} + \varepsilon_{x-1,x-2} + \cdots + \varepsilon_{1,0}, \tag{8}$$

and, by the strong Markov property, $\{\varepsilon_{i,i-1}\}$ are independent. If $g_x(t)$ is the density of ε_x and $\hat{g}_x(s)$ its Laplace transform, then

$$\hat{g_x}(s) = \prod_{i=1}^{x} \hat{g}_{i,i-1}(s). \tag{9}$$

where s is the complex variable of the Laplace transform and $\hat{g}_{i,i-1}(s)$ is the Laplace transform of the density of $\varepsilon_{i,i-1}$. In this setup, state i has a single outgoing transition, $i \to i-1$, with rate μ_i. Hence $\varepsilon_{i,i-1}$ is exponential with rate μ_i,

$$\hat{g}_{i,i-1}(s) = \int_0^\infty \mu_i e^{-(\mu_i+s)t}\, dt = \frac{\mu_i}{\mu_i + s}. \tag{10}$$

Hence,

$$\hat{g_x}(s) = \prod_{i=1}^{x} \frac{\mu_i}{\mu_i + s}; \tag{11}$$

Probabilities of independent random variables. We use two-sided Laplace transforms for random variables whose support is not restricted to $[0,\infty)$. For a density $f : \mathbb{R} \to \mathbb{R}_+$ of a random variable X, the two-sided Laplace transform, $\hat{f}_X$, is defined as

$$\hat{f}_X(s) \equiv \int_{-\infty}^{\infty} e^{-st} f_X(t)\, dt = \mathbb{E}[e^{-sX}]. \tag{12}$$

where s is a complex variable. This coincides with the usual one-sided transform for nonnegative random variables. In the rest of the paper, all references to the Laplace transform refer to the two-sided version.

If X and Y are independent random variables with densities f_X and f_Y and Laplace transforms $\hat{f}_X$ and $\hat{f}_Y$, then the two-sided Laplace transform of the density of $X - Y$ is

$$\hat{f}_{X-Y}(s) = \mathbb{E}\left[e^{-s(X-Y)}\right] = \mathbb{E}\left[e^{-sX}\right]\mathbb{E}\left[e^{sY}\right] = \hat{f}_X(s)\hat{f}_Y(-s). \tag{13}$$

Then, let F_{X-Y} be the cumulative distribution function (CDF) of $X-Y$ and $\hat{F}_{X-Y}$ its Laplace transform. Then, since $F_{X-Y}(t) = \int_{-\infty}^{t} f_{X-Y}(u)\,du$,

$$\hat{F}_{X-Y}(s) = \frac{1}{s}\hat{f}_{X-Y}(s) = \frac{1}{s}\hat{f}_X(s)\hat{f}_Y(-s). \tag{14}$$

Inverting the Laplace transform $\hat{F}_{X-Y}$ yields the CDF of $X-Y$ in the time domain,

$$F_{X-Y}(t) = \mathscr{L}^{-1}\left\{\frac{1}{s}\hat{f}_X(s)\hat{f}_Y(-s)\right\}(t). \tag{15}$$

where $\mathscr{L}^{-1}$ denotes the two-sided inverse Laplace transform operator. Finally, evaluating it at $t = 0$ yields

$$F_{X-Y}(0) = \mathbb{P}[X-Y<0] = \mathbb{P}[X<Y] = \mathscr{L}^{-1}\left\{\frac{1}{s}\hat{f}_X(s)\hat{f}_Y(-s)\right\}(0). \tag{16}$$

3.2 CST Model Description

Following Cont et al. (2010), we define the mid-price p_M and the bid-ask spread S as

$$p_M(t) \equiv \frac{p_a(t)+p_b(t)}{2}, \qquad S(t) \equiv |p_a(t)-p_b(t)|, \tag{17}$$

where p_a and p_b are the best-ask and best-bid prices, respectively, and are measured in ticks (i.e., the minimum price increment). Then, define T as the first mid-price change time:

$$T \equiv \inf\{t \geqslant 0 : p_M(t) \neq p_M(0)\}. \tag{18}$$

The main assumptions of the CST model are 1) the arrivals of limit orders, market orders, and cancellations are independent Poisson processes with constant rates, 2) the order flow at the best prices is independent of the state of the book away from the best prices, 3) the order flow at the best prices is symmetric with respect to the bid and ask sides, and 4) all orders of each type have the same size. These assumptions are not fully realistic, but they cover the main stylized facts of order-book dynamics and allow for a tractable representation of the best-price state as a two-dimensional continuous-time Markov chain (CTMC).

We are particularly interested in the $S = 1$ regime, which is the most liquid and prevalent state in many markets. We refer to the $S = 1$ regime as the state in which the spread is exactly one tick. The quantities of interest in this paper are conditional, short-horizon events given the instantaneous best-price state. By Lemma 3 of Cont et al. (2010), the best-price state can be represented on $[0,T)$ by a tractable two-dimensional CTMC with independent birth–death coordinates. This representation is the key structural reduction used throughout the remainder of the paper, since it allows the relevant event probabilities to be computed through first-passage arguments.

Concretely, letting $B(t)$ denote the best-bid queue size and $A(t)$ the best-ask queue size, we use the standard best-price dynamics. Figure 4 summarizes this two-queue representation and the associated birth–death transitions on each side.

Bid queue dynamics. Transitions at the best bid satisfy

$$B \to B+1 \qquad \text{with rate } \lambda, \tag{19}$$
$$B \to B-1 \qquad \text{with rate } \mu + B\theta, \tag{20}$$

where λ captures the arrival of a passive order at the bid, μ captures an aggressive ask order that consumes a passive bid order, and θ captures the cancellation of a passive bid order, with total cancellation rate proportional to the queue size. The same parameters govern the ask side.

Ask queue dynamics. Transitions at the best ask satisfy

$$A \to A+1 \qquad \text{with rate } \lambda, \tag{21}$$
$$A \to A-1 \qquad \text{with rate } \mu + A\theta. \tag{22}$$

These induce a two-dimensional CTMC for $X(t) = (A(t),B(t))$ in which the next mid-price move is governed by which queue hits zero first. Note that the model is symmetric with respect to the bid and ask sides, so the same parameters govern both queues. This is a simplifying assumption that may not hold in practice, but it avoids overfitting into trending regimes and allows for a more parsimonious calibration.

More explicitly, for interior states $(a,b) \in \mathbb{N}^2$, where $\mathbb{N}$ denotes the positive

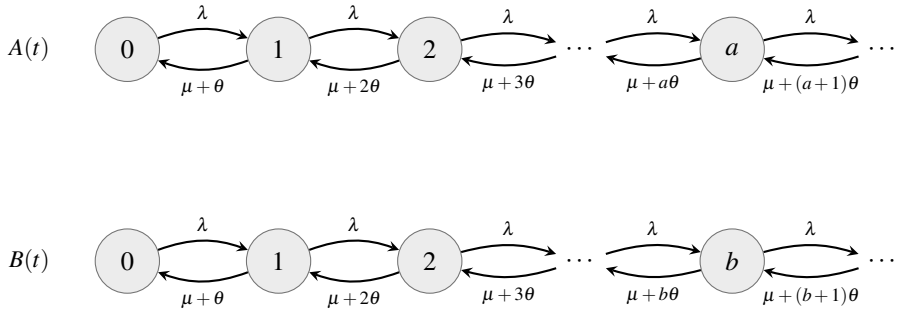

Figure 4
Two-queue CST dynamics. This figure shows the Cont-Stoikov-Talreja best-price approximation in which the best ask queue $A(t)$ and the best bid queue $B(t)$ evolve as independent birth–death processes up to the next mid-price change. On each side, transitions from state n to $n+1$ occur at arrival rate λ, representing new limit-order arrivals at the best price, while transitions from state n to $n-1$ occur at depletion rate $\mu+n\theta$ for $n \geqslant 1$, representing market-order executions and queue-size-proportional cancellations.

integers, the process $X(t)$ evolves through the transitions

$$(a,b) \to (a+1,b) \qquad \text{with rate } \lambda, \tag{23}$$
$$(a,b) \to (a-1,b) \qquad \text{with rate } \mu + a\,\theta, \tag{24}$$
$$(a,b) \to (a,b+1) \qquad \text{with rate } \lambda, \tag{25}$$
$$(a,b) \to (a,b-1) \qquad \text{with rate } \mu + b\,\theta. \tag{26}$$

That is, each coordinate behaves as a birth–death queue, and the next mid-price change occurs when either the ask queue or the bid queue reaches zero.

4. Probability Computation in the CST Model

In the CST model, the next mid-price move and the fill probability at the best prices are the easiest events to compute. However, we also provide a latency-aware extension for a practically relevant situation in market making, and the same logic can be applied to other events of interest (Cont et al., 2010).

4.1 Probability of an upward mid-price move

The probability that the next mid-price move is upward, given the initial state (a,b), can be written as

$$p_{\uparrow}(a,b) = \mathbb{P}[p_M(T) > p_M(0) \mid A(0) = a, B(0) = b] \tag{27}$$

Under the best-price coupling introduced in Section 3, if σ_a and σ_b denote the birth–death depletion times of the ask and bid queues starting from (a,b), the event above is equivalent to

$$p_{\uparrow}(a,b) = \mathbb{P}[\sigma_a < \sigma_b \mid A(0) = a, B(0) = b]\,. \tag{28}$$

Setting $X = \sigma_a$ and $Y = \sigma_b$, where both are birth–death queue-depletion times, Eq. (16) gives

$$p_{\uparrow}(a,b) = \mathscr{L}^{-1}\left\{\frac{1}{s}\hat{f}_a(s)\hat{f}_b(-s)\right\}(0), \tag{29}$$

where $\hat{f}_a$ and $\hat{f}_b$ are the Laplace transforms of the densities of the birth–death queue-depletion times σ_a and σ_b, respectively, which can be computed through the semi-analytical formula in Eq. (7).

4.2 Fill probability at the best price

The probability that an order posted at the best bid is executed before the first mid-price change at time T can be written as

$$p_{\text{fill}}(a,b) = \mathbb{P}[\varepsilon_b < \sigma_a \mid A(0) = a, B(0) = b, NC_b]\,. \tag{30}$$

Here NC_b denotes the event that no new order is inserted ahead of the considered bid order. The pure-death representation is appropriate because, under price-time priority, once the tagged order has joined the queue, later arrivals at the same price cannot move in front of it. The relevant bid-side volume therefore consists of the orders already ahead of the tagged order together with the tagged order itself, and this volume can only decrease through executions and cancellations. We therefore compare the pure-death depletion time ε_b of that bid-side volume with the birth–death depletion time σ_a of the ask queue.

Conditioning on NC_b changes the depletion rate of the observed bid queue. Using the general pure-death transform in Eq. (11), we replace

$$\mu_i \longmapsto \mu + \theta(i-1), \tag{31}$$

which yields

$$\hat{g}_b^{NC}(s) = \prod_{i=1}^{b} \frac{\mu + \theta(i-1)}{\mu + \theta(i-1) + s}. \tag{32}$$

The $(i-1)$ term appears because one order in the queue (the tagged order itself) is constrained not to cancel under NC_b. For example, in a state with four resting orders that includes the tagged order, only three are eligible for cancellation from the perspective of the tagged order, so the cancellation component is lower than in the unconstrained theoretical formula.

Setting $X = \varepsilon_b$ and $Y = \sigma_a$, where ε_b is the pure-death depletion time of the bid-side volume up to and including the tagged order and σ_a is the birth–death queue-depletion time of the ask queue, Eq. (16) gives

$$p_{\text{fill}}(a,b) = \mathscr{L}^{-1}\left\{ \frac{1}{s} \hat{g}_b^{NC}(s) \hat{f}_a(-s) \right\}(0), \tag{33}$$

where $\hat{f}_a$ and $\hat{g}_b^{NC}$ are the Laplace transforms of the densities of the birth–death depletion time σ_a and the pure-death depletion time ε_b under NC_b, with ε_b defined over the bid-side volume up to and including the tagged order, respectively.

4.3 Probability of Mid-Price Change During Latency

We now derive the probability that the mid-price changes within a fixed reaction window, extending the Laplace-based framework to incorporate a deterministic latency horizon.

Let $\delta > 0$ denote a fixed, deterministic reaction latency between observing the book state and effectively updating quotes at the exchange. The event of interest is that the first mid-price change time T, as defined in Section 3, occurs before the latency expires. Under the CST independence structure, $T = \min(\sigma_a, \sigma_b)$, and we define

$$p_{\text{chg}}^{\delta}(a,b) \equiv \mathbb{P}\left[T < \delta \mid A(0) = a, B(0) = b\right]. \tag{34}$$

Since σ_a and σ_b are independent under the CST dynamics,

$$p_{\text{chg}}^{\delta}(a,b) = F_{\sigma_a}(\delta) + F_{\sigma_b}(\delta) - F_{\sigma_a}(\delta) F_{\sigma_b}(\delta), \tag{35}$$

where $F_{\sigma_a}(\delta) \equiv \mathbb{P}[\sigma_a < \delta \mid A(0) = a]$ and $F_{\sigma_b}(\delta) \equiv \mathbb{P}[\sigma_b < \delta \mid B(0) = b]$ are the cumulative distribution functions of the birth–death depletion times

evaluated at the fixed horizon δ:

$$F_{\sigma_a}(\delta) = \mathscr{L}^{-1}\left\{\frac{1}{s}\hat{f}_a(s)\right\}(\delta), \qquad F_{\sigma_b}(\delta) = \mathscr{L}^{-1}\left\{\frac{1}{s}\hat{f}_b(s)\right\}(\delta), \quad (36)$$

with $\hat{f}_a$ and $\hat{f}_b$ given by Eq. (7).

5. Implementation

5.1 BTG-OBD-A26 Dataset

We use the BTG-OBD-A26 message-level dataset for Brazilian equities, covering October 2025 to January 2026 under an event-time reconstruction pipeline. The empirical design in this paper uses 5 liquid tickers and 9 test days, selected to span high-, low-, and median-volatility market conditions (see Tables A1 and A2); full construction details are deferred to Appendix A, and Figure A1 there provides a compact overview of the data-processing pipeline. This selection is deliberate: rather than aiming at broad market coverage, we choose a small set of liquid assets and representative days to maximize the amount of calibration and validation data in the one-tick-spread regime, which is the regime studied in this paper.

At a high level, each event is mapped to one of three actions used in model calibration: limit-order arrivals, market-order executions (trade prints), and cancellations/deletions. The full event-classification rules are provided in Appendix A.2.

Because the CST specification is calibrated in the one-tick-spread regime, we restrict rate estimation and validation to periods with $S(t) = 1$, measured over the regular session window 11:00–17:00 (local exchange time), excluding auction-related intervals. This choice aligns the empirical sample with the model assumptions and concentrates the analysis on the most liquid best-price states. More importantly, our inferential objective is explicitly restricted to the $S = 1$ regime. Outside that regime, the order-book dynamics are more complex, the original CST reduction is less directly applicable, and we do not evaluate the model in this paper.

Asset and day selection diagnostics. We justify the empirical design in Section A with simple diagnostics that make the selection rule transparent and reproducible. Table A1 reports the selected tickers, their sectors, typical price levels, and the time share with a one-tick spread. The latter is computed from 1-minute snapshots between 11:00 and 17:00 (local exchange time) to avoid auction-related periods. We select tickers with high $S = 1$ time share to ensure

that the model's assumptions are relevant for a large fraction of the trading day. Price is reported as a rounded integer (BRL) to provide a coarse notion of the typical absolute price level rather than a precise quote. Under a fixed tick size, one-tick-spread regimes are mechanically more prevalent among liquid names with lower absolute prices; highly liquid but higher-priced equities (e.g., `WEGE3` and `VALE3`) are therefore less compatible with the tight-spread ($S = 1$) modeling regime.

5.2 Calibration procedure

We estimate the model parameters directly from the BTG-OBD-A26 event stream under the $S = 1$ regime. We follow a strict day-ahead protocol: for each evaluation day d, calibration uses only the immediately previous trading day (the `calib_date`). An assumption of the CST model is that all orders of a given type have the same size. While the original model proposed a unit-size normalization based on the average size, Lee and Kim (2013) proposed a revised model for the highly liquid KOSPI 200 futures market. They found a trade-off between the stability obtained with large unit sizes (and small queues) and the sensitivity obtained with small unit sizes (and large queues).

We adopt the original unit-size normalization of Cont et al. (2010) for the Brazilian equity market, which is less liquid than the KOSPI 200 futures. We compute average sizes for limit, market, and cancellation events, denoted by S_l, S_m, and S_c, respectively. Let T^* be the total time (in minutes) used for calibration on that previous-day window.

Here the index i denotes the book level, that is, the i-th queue measured in ticks from the opposite best quote, as defined in Section 2.1. Thus $i = 1$ is the best quote on the event side, $i = 2$ is the next level one tick deeper, and so forth.

For depth-dependent arrivals, we use direct counting at the first ten distances from the opposite best quote,

$$\hat{\lambda}(i) = \frac{N_l(i)}{T^*}, \qquad i = 1, \ldots, 10, \tag{37}$$

where $N_l(i)$ is the number of limit-order arrivals observed at level i. Since the best-price model only uses the best level, we set the notation

$$\lambda \equiv \lambda(1), \qquad \theta \equiv \theta(1), \tag{38}$$

and use (λ, θ) throughout the empirical sections.

The market-order rate is estimated as

$$\hat{\mu} = \frac{N_m}{T^*} \frac{S_m}{S_l}, \tag{39}$$

where N_m is the number of trade events mapped to market-order executions and S_m/S_l is the size normalization factor that converts the observed average market order size into the number of model units.

For cancellations, let Q_i denote the average standing depth at level i, that is, the average queue size resting at the i-th book level, averaged across bid and ask sides. Then,

$$\hat{\theta}(i) = \frac{N_c(i)}{T^* Q_i} \frac{S_c}{S_l}, \tag{40}$$

where $N_c(i)$ counts size reductions attributed to cancellations. Implementation details for event classification, depth construction, and day-ahead filtering are provided in Appendix A. Calibration outputs are summarized in Appendix B.

6. Experiments

We conduct four experiments to evaluate the CST model's performance in predicting events in the Brazilian equity market. The first experiment focuses on the probability of an upward mid-price move, the second on the fill probability at the best price, the third on the probability of a mid-price change during latency for market making, and the fourth on the sensitivity of the model to the intensity of events. Computing the probabilities of interest involves two main numerical challenges: (i) the numerical computation of the continued fractions presented in Eq. (6) and (ii) the numerical inversion of Laplace transforms to recover CDF values. For the former, we use a stable recursion based on the modified Lentz method (Thompson, 1986). For the latter, we use the COS method (Fang and Oosterlee, 2009), which provided more stable results than Euler-based methods in our implementation.

The first three experiments use a truncated state space. In this setup, we only need to precompute 400 probabilities for each day and symbol. This introduces some approximation error in larger queues, but these are less common in our data. In practice, the model probabilities are precomputed only on a 20×20 grid of best-price states (a,b). To compare these grid-based predictions with realized outcomes in calendar time, we sample the reconstructed order book every 100 ms, read the observed best-price state and assign the corresponding precomputed theoretical probability from the grid. We then group these snapshots into bins of theoretical probabilities and compute the empirical frequency within each bin, weighted by the number of snapshots in that bin.

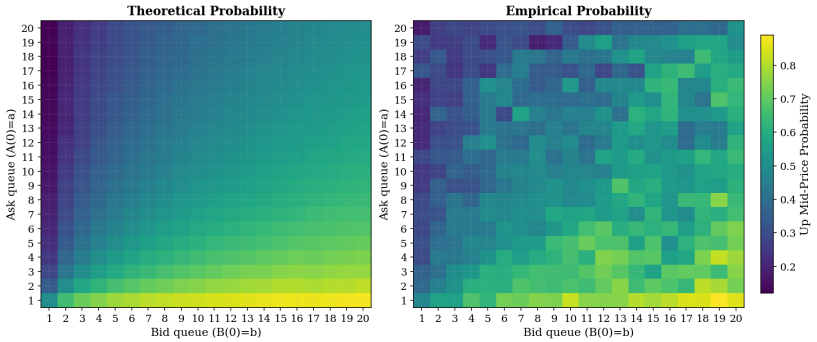

Figure 5

Directional-probability heatmaps. This figure compares CST-implied and empirical probabilities of an upward mid-price move on the truncated 20×20 best-price grid of states (a,b), where a is best-ask depth and b is best-bid depth. Each cell corresponds to one queue state, and color encodes the probability assigned to that state. The comparison highlights the expected monotone structure: states with relatively thinner ask queues and thicker bid queues are associated with larger upward-move probabilities.

6.1 Upward Mid-Price Move Prediction

The first experiment evaluates the probability that the next mid-price move is upward, given the initial best-price state. This is a fundamental quantity in many applications, including directional trading and market making.

Figure 5 provides the state-by-state comparison between theoretical and empirical probabilities across the truncated grid, while Figure 6 summarizes the same experiment through reliability bins in the spirit of Lee and Kim (2013). The reliability plot shows a clear monotone relation between model-implied bins and realized frequencies. The proximity of the empirical heatmap to the theoretical one in Figure 5 shows that the model captures the main structural features of the state-by-state probability surface.

Figure 6 reports a reliability-style summary for the same experiment. Each panel shows the aggregate or per-ticker relation between binned model-implied directional probabilities (horizontal axis) and the corresponding weighted empirical upward-move frequencies (vertical axis). The dashed line indicates perfect calibration. The background gray histogram on the secondary axis displays the number of snapshots that fall into each probability bin, providing a visual indication of where statistical support is concentrated. In the aggregate panel, the curve is monotonically increasing and lies slightly above the

diagonal over most of the support. The per-ticker panels preserve the same monotone shape with varying degrees of dispersion around the diagonal.

The main visual anomaly appears in CSAN3, where the last bin exhibits an abrupt drop in the empirical upward-move frequency. The background histogram for CSAN3 shows that this highest-probability bin contains very few orders relative to the other bins, so the drop is attributable to insufficient sample size rather than to a systematic calibration failure.

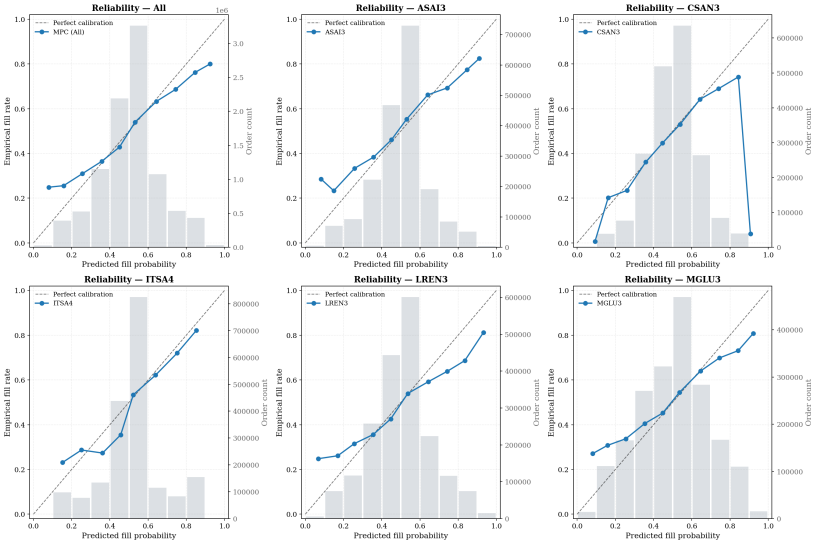

Figure 6

Directional-probability reliability plots. This figure summarizes the upward mid-price experiment by plotting, in each panel, the empirical frequency of upward moves against binned CST-implied directional probabilities. The dashed 45-degree line indicates perfect calibration, and the gray background histogram, read on the right axis, reports the number of snapshots in each probability bin. The top-left panel pools all selected tickers, while the remaining panels report ticker-level results.

6.2 Best-Price Fill Prediction

The second experiment evaluates the probability that a passive order posted at the best bid is executed before the first mid-price change. In contrast with synthetic insertion experiments, here we exploit the fact that the BTG-OBD-A26 dataset is effectively L3 rather than merely L2, so individual resting

orders can be tracked through time by their identifiers and queue evolution; see Appendix A for the reconstruction pipeline and the discussion of L2 versus L3 granularity. This allows us to follow real orders submitted at the best price instead of simulating hypothetical ones.

Concretely, we consider all orders inserted at the best price while the spread is one tick and track each order until the first mid-price change. We then partition these orders into three groups: (i) orders canceled before the first mid-price change, which are discarded from the evaluation sample; (ii) orders filled before the first mid-price change, which are labeled as successes; and (iii) orders not filled before the first mid-price change, including orders filled only after the mid-price change or canceled only after the mid-price change, which are labeled as failures. This yields a binary target for passive execution conditional on the order not being canceled before the first price move.

This sampling rule induces an important selection effect that must be made explicit. The empirical target is not representative of arbitrary market orders resting at the best quote, because it conditions away the orders that are canceled before the first mid-price change. Accordingly, the estimated probabilities should be interpreted as conditional fill probabilities for orders that the trader intends to keep active, precisely the same economic interpretation embodied in the event NC_b in the model. Under that interpretation, the experiment is directly aligned with the CST quantity $p_{\text{fill}}(a,b)$.

As in the directional experiment, we truncate the state space to a 20×20 grid, precompute model probabilities on that grid, and compare them with empirical conditional frequencies extracted from the tracked-order sample. Figure 7 reports the state-by-state comparison between CST-implied and realized conditional fill probabilities. Figure 8 complements this view with a reliability-style binning of the same predictions, summarizing how well the model ranks and calibrates passive execution outcomes across the observed states.

6.3 Probability of Mid-Price Change During Latency for Market Making

The third experiment evaluates the quantity $p^{\delta}_{\text{chg}}(a,b)$ introduced in Section 4: the probability that the next mid-price change occurs before a fixed reaction latency δ expires. This experiment is motivated by market making. A market maker who continuously posts at the best bid and ask earns the spread when both sides are filled, but is exposed to losses whenever the mid-price moves before quotes can be updated (Avellaneda and Stoikov, 2008). In the one-tick-spread regime, $p^{\delta}_{\text{chg}}(a,b)$ is therefore the probability that a posted

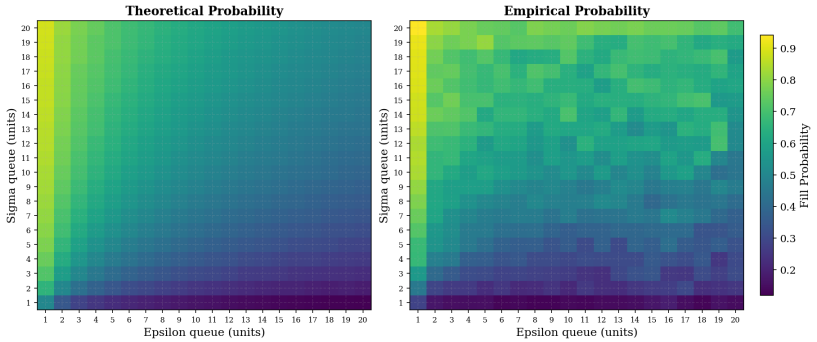

Figure 7

Best-price fill-probability heatmaps. This figure compares model-implied and empirical conditional fill probabilities on the truncated 20×20 best-price grid of states (a,b). Each cell corresponds to one best-ask/best-bid depth pair, and color encodes the conditional probability that a tagged order posted at the best bid executes before the first mid-price change. The empirical target is computed from tracked L3 orders and is conditioned on the order not being canceled before the price moves.

quote becomes stale before the quoting system can cancel.

In contrast with the first two experiments, which condition on which queue depletes first, this experiment targets the event $\{T < \delta\}$, where T is the first mid-price change time and δ is the reaction latency. The key advantage of this formulation is that T is always observed in the data. If we tried to target the depletion time of a single side, we would only observe that time for the side that depletes first, and the other side would be censored. These selection issues would complicate the evaluation and interpretation of the results, while the formulation in terms of T allows us to use all snapshots in the $S = 1$ regime without selection bias.

The theoretical object is the state-conditional probability $p^{\delta}_{\text{chg}}(a,b)$ derived in Eq. (35), namely

$$p^{\delta}_{\text{chg}}(a,b) = F_{\sigma_a}(\delta) + F_{\sigma_b}(\delta) - F_{\sigma_a}(\delta)\,F_{\sigma_b}(\delta), \tag{41}$$

with F_{σ_a} and F_{σ_b} computed from Eq. (36). We evaluate two fixed latency horizons, $\delta = 1\,\text{s}$ and $\delta = 10\,\text{s}$. For each horizon, theoretical probabilities are precomputed on the same setup used in the previous experiments. The experiment therefore compares, at the snapshot level, a model-implied theoretical staleness probability with the realized indicator of whether the next mid-price change arrived before the latency window expired. Because T is directly ob-

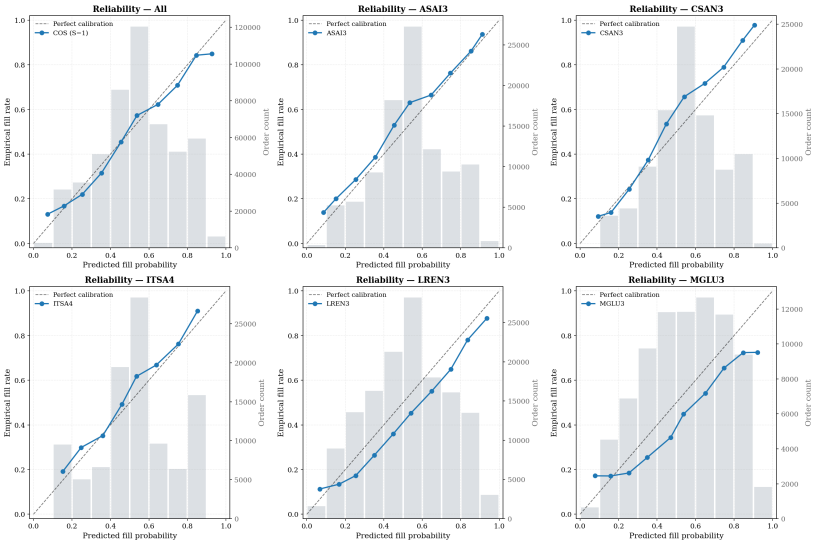

Figure 8

Best-price fill-probability reliability plots. This figure summarizes the execution experiment by plotting, in each panel, the empirical conditional fill rate of tracked L3 orders against binned CST-implied fill probabilities. The dashed 45-degree line indicates perfect calibration, and the gray background histogram, read on the right axis, reports the number of orders in each probability bin. The top-left panel pools all selected tickers, while the remaining panels report ticker-level results.

served without censoring, every snapshot in the $S = 1$ regime contributes to the evaluation regardless of which side depletes first.

Figures 9 and 10 provide the state-by-state comparison between theoretical and empirical staleness probabilities across the truncated grid for the two latency horizons, while Figures 11 and 12 summarize the same experiment through reliability bins in the pooled sample and by ticker.

Figures 11 and 12 report reliability-style summaries for the two latency horizons. Each panel shows the aggregate or per-ticker relation between binned model-implied staleness probabilities (horizontal axis) and the corresponding empirical frequency of a mid-price change before the latency window expires (vertical axis). The dashed line indicates perfect calibration. The background gray histogram on the secondary axis displays the number of snapshots that fall into each probability bin. At the shorter horizon of 1 s, the support of model-implied probabilities is compressed toward lower values, so the reliability curves are concentrated in the left portion of the plot. At the longer horizon of

10 s, the support expands and the reliability curves cover a broader range of predicted values.

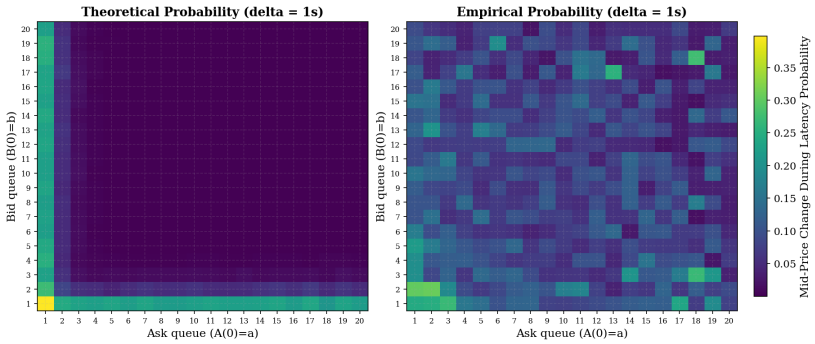

Figure 9

Latency-event heatmaps for a 1-second window. This figure compares model-implied and empirical probabilities that the next mid-price change occurs within a fixed latency horizon of $\delta = 1$ second on the truncated 20×20 best-price grid of states (a,b). Each cell corresponds to one best-price state, and color encodes the probability of the event $\{T < 1\ \mathrm{s}\}$. The theoretical surface is concentrated in thin-queue states, while the empirical heatmap is noisier because the event is relatively rare away from the lower-left corner of the grid.

6.4 Sensitivity to Short-Term LOB Activity

To understand why the latency experiment is more fragile than the state-based exercises, we run an auxiliary diagnostic on the fill-probability predictions. The goal is to test whether calibration deteriorates when the local order-book activity immediately after order insertion is unusually low or unusually high relative to the average intensities used by the CST model.

The sample contains 512,309 tracked limit orders across the same 5 Brazilian equities and 9 out-of-sample prediction dates used in the main analysis, forming a balanced panel of 45 symbol-date pairs. For each tracked order, we record the CST-implied fill probability and the realized binary outcome indicating whether the order is filled before the first mid-price change, conditional on the order remaining active as in the main fill experiment.

As a proxy for short-term local activity, for every order we count the number of non-trade LOB events arriving in the 60-second window after insertion. Denote this count by $n_{\text{events,60s}}$. In the pooled sample, this proxy has median 383, mean 628, 5th percentile 80 and 95th percentile 1,979. We

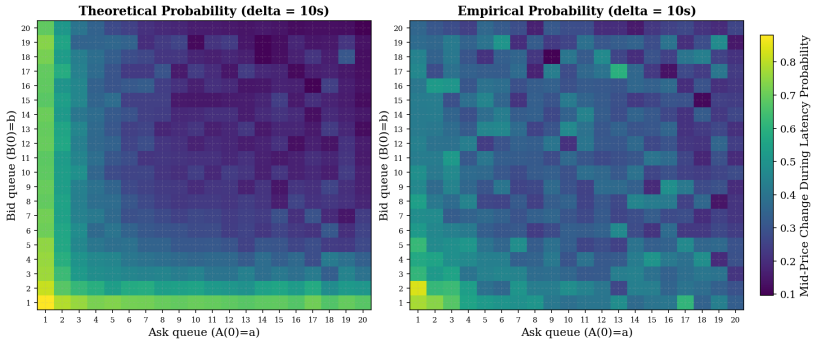

Figure 10

Latency-event heatmaps for a 10-second window. This figure compares model-implied and empirical probabilities that the next mid-price change occurs within a fixed latency horizon of $\delta = 10$ seconds on the truncated 20×20 best-price grid of states (a,b). Each cell corresponds to one best-price state, and color encodes the probability of the event $\{T < 10\,\mathrm{s}\}$. Relative to the 1-second horizon, the theoretical surface is less concentrated and the empirical heatmap shows a broader monotone gradient across the state grid.

partition the sample into six global quantile buckets using this variable, so that each bucket contains approximately 85,000 orders.

For each bucket, we report a reliability diagram together with two summary metrics. The Brier score is the mean squared error between predicted probabilities and realized binary outcomes,

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \tag{42}$$

where p_i is the model-implied fill probability and $o_i \in \{0,1\}$ is the realized outcome: filled or not filled before the first mid-price change. Lower values indicate better overall probabilistic accuracy, in the sense that the predicted probabilities are, on average, closer to the realized binary outcomes. In particular, the Brier score is a global loss function: it summarizes the total probability error, but does not isolate whether that error comes from poor calibration, weak discrimination across orders, or differences in event incidence across buckets.

We also compute the expected calibration error (ECE) using 15 equal-width

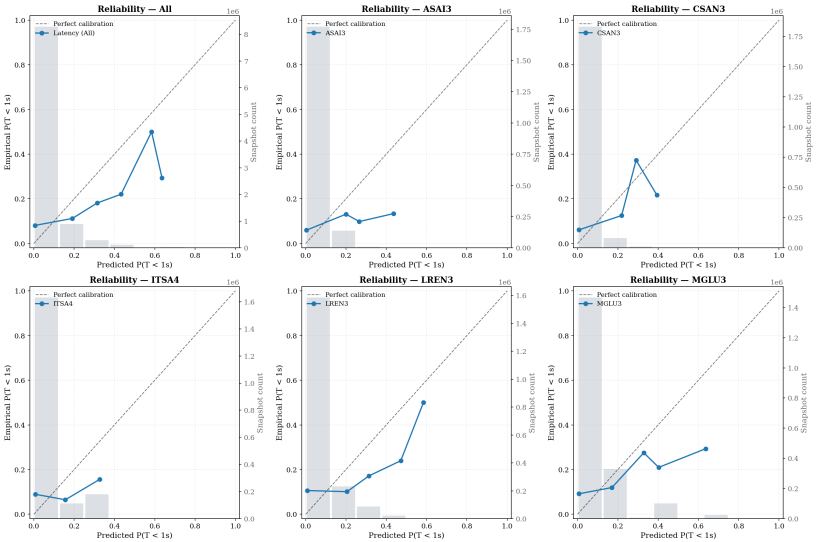

Figure 11

Latency reliability plots for a 1-second window. This figure summarizes the latency experiment at $\delta = 1$ second by comparing binned model-implied stale-quote probabilities with the corresponding empirical frequencies. In each panel, the dashed 45-degree line indicates perfect calibration, and the gray background histogram, read on the right axis, reports the number of snapshots in each bin. The short horizon produces a compressed support with substantial mass in low-probability bins, and the panels report both pooled and ticker-level results.

probability bins,

$$\mathrm{ECE} = \sum_{m=1}^{15} \frac{n_m}{N} \left| \bar{p}_m - \bar{o}_m \right|, \tag{43}$$

where n_m is the number of observations in bin m, while $\bar{p}_m$ and $\bar{o}_m$ are the mean predicted and mean realized fill probabilities in that bin. In this paper, calibration means agreement between predicted probabilities and realized frequencies: for example, among orders assigned a fill probability close to 0.7, calibration is good if roughly 70% of them are in fact filled before the first mid-price change. ECE therefore measures how far, on average across bins, predicted probability levels are from the corresponding empirical frequencies. Lower ECE indicates better calibration. Figure 13 overlays the corresponding reliability curves with the distribution of predictions inside each activity bucket.

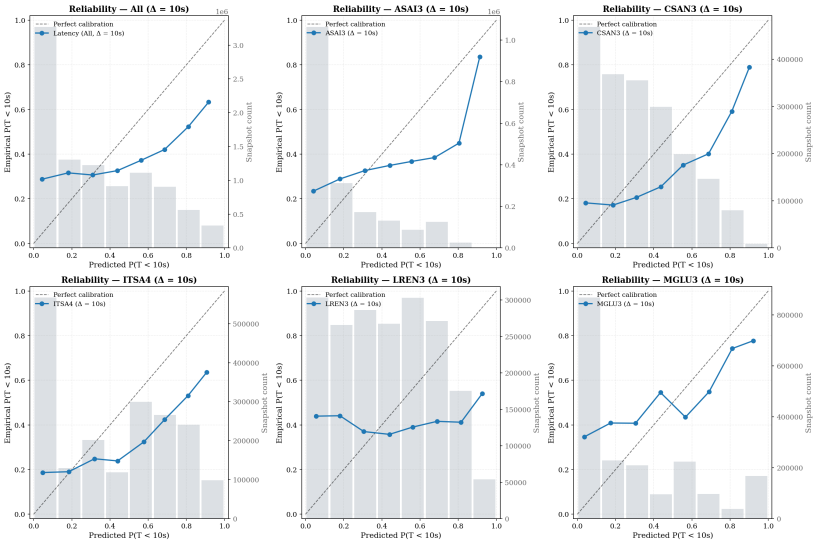

Figure 12

Latency reliability plots for a 10-second window. This figure summarizes the latency experiment at $\delta = 10$ seconds by comparing binned model-implied stale-quote probabilities with the corresponding empirical frequencies. In each panel, the dashed 45-degree line indicates perfect calibration, and the gray background histogram, read on the right axis, reports the number of snapshots in each bin. The longer horizon expands the support of model-implied probabilities and yields a broader calibration range in both the pooled sample and the ticker-level panels.

The diagnostic yields two patterns. First, the Brier score remains relatively stable across buckets, ranging from 0.1996 to 0.2237, which indicates that the model's aggregate probabilistic error remains in the same order of magnitude when short-term activity changes. Second, the ECE displays a clear U-shape: calibration is strongest in the intermediate buckets, especially $(252,383]$ with ECE 0.0289 and $(383,576]$ with ECE 0.0510, and substantially weaker in the tails, namely $(7,151]$ with ECE 0.1634 and $(961,10477]$ with ECE 0.1571.

7. Results and Discussion

The four experiments support a consistent finding: the CST framework preserves economically coherent state ranking in the Brazilian equity market. The key empirical regularity across experiments is not exact pointwise calibration

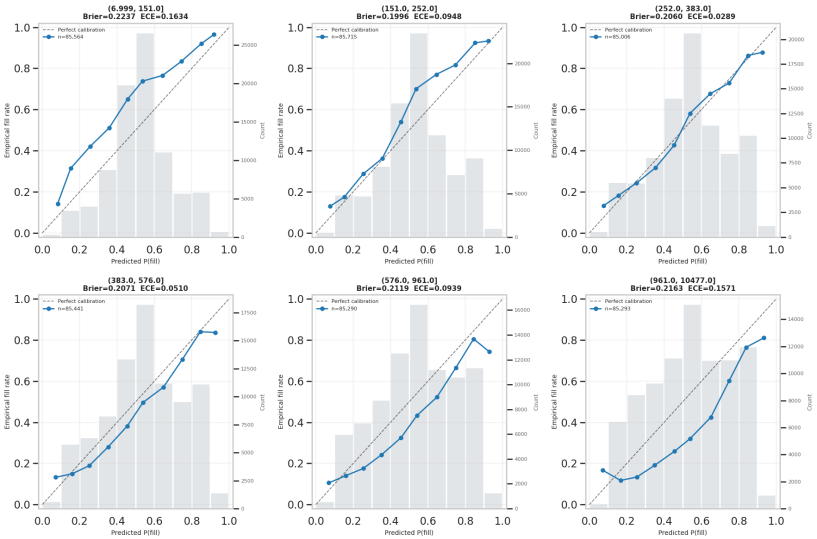

Figure 13

Fill-probability reliability by short-term LOB activity. Orders are partitioned into six quantile buckets according to the number of non-trade best-book events observed in the 60 seconds after insertion. Each panel plots the empirical fill rate against the binned CST-implied fill probability for one activity bucket. The dashed 45-degree line indicates perfect calibration, and the background histogram reports the distribution of predicted probabilities within the bucket. Calibration is best in intermediate-activity buckets and deteriorates in both tails, consistent with the view that the CST model is most reliable when local event intensities remain close to the constant-rate approximation used in calibration.

in every cell of the state grid, but robust monotone reliability patterns across tickers and regimes. This is the property that matters most for execution decision rules, which typically depend on relative state ordering and thresholding rather than on perfect probability levels.

Directional prediction and the queue-imbalance channel. In the directional experiment, the model-implied probability $p_{\uparrow}(a,b)$ exhibits the expected structural dependence on queue imbalance: states with relatively thinner ask queues and thicker bid queues are assigned larger upward-move probabilities. The reliability analysis confirms that this ordering is present in realized outcomes, with empirical frequencies increasing monotonically across theoretical bins. This result connects directly to the queue-imbalance literature. Gould

and Bonart (2016) documented empirically that the imbalance between the best bid and ask queues is a powerful one-tick-ahead directional predictor, while Gould et al. (2013) surveyed the broader role of queue dynamics in short-horizon price formation. The CST model formalizes the same economic channel through its birth–death queue-depletion race, and the monotone agreement observed in Figure 6 confirms that this structural mechanism carries over to Brazilian equities.

Execution prediction and target alignment. The fill experiment validates the queue-priority channel directly on tracked L3 orders rather than synthetic insertions. The pooled reliability curve shows a clear monotone relation between model-implied and empirical conditional fill rates, and the same qualitative pattern is preserved across assets. A relevant interpretive point is target alignment under selection: the empirical target conditions on orders that remain active until the first price change, which matches exactly the economic meaning of the model event NC_b. This improves comparability between theory and measurement and ensures that the experiment tests the object that an execution algorithm actually uses when deciding whether to keep a passive order live.

The short-term activity diagnostic in Figure 13 refines this interpretation. Here, ranking means the model's ability to order states or orders from lower to higher true fill probability, even if the numerical probability levels are not exactly correct. Under that definition, the evidence in Figure 13 suggests that ranking is less visibly affected than calibration. In other words, once local order flow moves away from the constant-intensity environment used by the CST approximation, the probability levels become unreliable even when the model may still preserve some coarse ordering of more- versus less-likely fills.

Latency experiment and the two horizons. The latency experiment benefits from a cleaner identification setup than the first two experiments. Because the target event is $\{T < \delta\}$, where T is the observed time to the next mid-price change, each $S = 1$ snapshot can be assigned a theoretical score and matched to a realized stale-quote indicator without censoring or selection. At $\delta = 1\,\mathrm{s}$ the support of model-implied probabilities is compressed toward low values, so the reliability diagnostic covers a narrow range largely driven by thin-queue states. At $\delta = 10\,\mathrm{s}$ the support expands, the reliability curves cover a broader calibration range, and the monotone agreement is clearer across the state grid. Nevertheless, the latency heatmaps reveal a qualitative difference relative to the first two experiments: the empirical probability surfaces are visibly noisier

and less in agreement with the theoretical surfaces than those observed in the directional and fill experiments. This contrast motivates a distinction between state-based and time-based metrics.

State-based versus time-based metrics. The first two experiments condition on which queue depletes first and are therefore purely state-based: they ask which of two competing processes wins the race, without reference to calendar time. The CST model's main simplification – independent Poisson arrivals – does not capture temporal clustering in real order flow, but this limitation is less consequential for state-based comparisons, because the relative ranking of queue-depletion outcomes depends primarily on the instantaneous queue geometry rather than on the fine temporal structure of arrivals. The third experiment, by contrast, explicitly targets a calendar-time threshold δ: the empirical outcome is whether the next mid-price change occurs within δ seconds. Here, temporal clustering – bursts of market orders or cancellations that arrive in rapid succession – directly affects the distribution of T in ways that independent Poisson arrivals cannot reproduce. The empirical heatmaps for the latency experiment confirm this. In short, the absence of temporal clustering appears to be a minor limitation for state-based queue comparisons but a more material one when the model is asked to predict the timing of price changes.

Figure 13 makes this mechanism visible even before one reaches the latency target. When the realized short-term order-flow intensity is close to the intermediate regime, fill-probability calibration remains reasonably good. But when activity is unusually low or unusually high, the ECE rises sharply even though the Brier score remains comparatively stable. This wedge is informative: it points more clearly to a calibration failure than to a collapse in overall probabilistic usefulness. In other words, under extreme activity regimes, the CST approximation still produces probability forecasts with similar aggregate loss, but those probability levels cease to match empirical frequencies well. Low-activity windows and high-activity bursts both move the local dynamics away from the stationary Poisson benchmark. That distortion is already visible in fill probability, which is the simpler of our execution objects. It is therefore unsurprising that the same misspecification becomes even more pronounced in the latency experiment, where the target is explicitly time-based and thus directly exposed to short-lived bursts and droughts in event intensity.

Structural trade-off of discretization. A second structural limitation concerns the unit-size normalization inherited from the CST specification. As

documented by Lee and Kim (2013) for the KOSPI 200 futures, there is an inherent trade-off between sensitivity and statistical support: a large unit size S_l maps the raw order flow onto fewer queue states with more observations per state, improving statistical stability but losing the ability to distinguish between small and large orders; a small S_l preserves granularity but increases the number of states and, more critically, introduces jumps of more than one unit per event, which violates the unit-increment Poisson assumption. In our sample, the effective unit size is of the order of 1 000 shares, whereas the minimum lot is 100 shares. A single model unit can therefore correspond to raw queue sizes ranging from roughly 100 to 1 500 shares. This coarse discretization has a direct impact on the latency experiment at $\delta = 1\,\text{s}$: events that are relevant for very short horizons – small limit orders, partial cancellations, and thin-queue depletions – are aggregated into the same unit and become invisible to the model. In this sense, larger-than-minimum units subsample the fine-grained events that dominate very short-horizon dynamics. The effect is less pronounced at $\delta = 10\,\text{s}$, where the event $\{T < 10\text{ s}\}$ is common enough that the coarser state representation still captures the main variation.

Scope of the latency horizons. Although the latency horizons used in this paper ($\delta \in \{1\text{ s}, 10\text{ s}\}$) are large by high-frequency standards, they were chosen to be commensurate with the empirical queue-depletion times of the selected Brazilian equities. In markets with shorter depletion times – typically more liquid venues with many small lots – the same framework would apply at correspondingly shorter horizons. Such environments are precisely the ones in which the latency-aware extension of the CST model would be most useful.

8. Conclusion

This paper shows that a semi-analytical CST layer can deliver useful execution-oriented probabilities in the Brazilian equity market, even under clear model simplifications. The four experiments support a consistent message: the CST framework preserves first-order microstructural economics that are useful for execution. Across the directional and fill experiments, the main empirical regularity is robust state ranking with monotone reliability patterns, which is precisely the property needed for practical decision rules. The latency extension provides an interpretable price-change probability over the full state space, with an empirical target that is directly observable and free from single-side censoring. The model does not attempt to replicate every feature of high-frequency order flow, and it should not be judged as a universal forecasting engine. Its comparative advantage is different: transparent state

conditioning, fast recomputation, and direct mapping from observable queue states to interpretable decision quantities.

From a deployment perspective, CST-like probabilities can serve as auditable baseline signals or structural priors inside hybrid execution stacks that also incorporate machine-learning forecasts. In that role, the semi-analytical model contributes robustness and interpretability while more flexible components absorb residual nonlinearities, temporal clustering, and size heterogeneity not explicitly modeled here.

Several directions for future work follow naturally from our findings. The first is operational validation: the precomputed probabilities – $p_{\uparrow}(a,b)$, $p_{\text{fill}}(a,b)$, and $p^{\delta}_{\text{chg}}(a,b)$ – should be embedded in an execution rule that decides when to wait, cancel, or cross the spread, so that the statistical evidence documented here can be translated into realized trading costs and execution quality. The second is a more systematic study of the unit-size normalization. Our results suggest a genuine trade-off between finer discretizations, which are more sensitive to small queue changes, and coarser discretizations, which provide more stable estimation and better state coverage; understanding how this trade-off varies with asset liquidity and target latency horizon would sharpen the practical scope of the CST approximation. A third direction is to convert the latency-aware probabilities into explicit economic quantities. Since the latency experiment directly produces state-conditional stale-quote probabilities, a natural next step is to map reductions in δ into reductions in expected adverse-selection losses, thereby giving a quantitative benchmark for latency-improvement decisions. More broadly, the evidence in this paper suggests that CST is most useful as a transparent structural layer rather than as a standalone forecasting engine. This makes hybrid architectures a promising extension: one can use the CST layer to provide fast, interpretable state conditioning, while more flexible machine-learning components absorb residual nonlinearities, temporal clustering, and size heterogeneity that lie beyond the constant-intensity specification studied here.

Acknowledgments The authors thank BTG Alpha Lab (BTG Pactual) for funding this research and for providing the data.

Conflict of interest The authors' affiliations with BTG Pactual are disclosed on the title page. The views expressed in this article are those of the authors and do not necessarily reflect the views, policies, or positions of BTG Pactual.

Artificial Intelligence This research utilized AI tools to assist in coding tasks. All AI-generated content was critically reviewed and validated by the authors to ensure accuracy and alignment with the scientific integrity of the study. The authors affirm that the AI tools did not compromise the originality or integrity of the work.

Data and code availability All data used in this study are available from the corresponding author upon request. All code used are available in the public repository https://github.com/btg-pactual/alphalab-LOB-inference.

References

Abate, J. and Whitt, W. (1992). The fourier-series method for inverting transforms of probability distributions, *Queueing Systems* **10**(1–2): 5–88.
URL: https://doi.org/10.1007/BF01158520

Abate, J. and Whitt, W. (1995). Numerical inversion of laplace transforms of probability distributions, *ORSA Journal on Computing* **7**(1): 36–43.
URL: https://doi.org/10.1287/ijoc.7.1.36

Abate, J. and Whitt, W. (1999). Computing laplace transforms for numerical inversion via continued fractions, *INFORMS Journal on Computing* **11**(3): 325–332.
URL: https://doi.org/10.1287/ijoc.11.3.325

Avellaneda, M. and Stoikov, S. (2008). High-frequency trading in a limit order book, *Quantitative Finance* **8**(3): 217–224.
URL: https://doi.org/10.1080/14697680701381228

Bacry, E., Mastromatteo, I. and Muzy, J.-F. (2015). Hawkes processes in finance, *Market Microstructure and Liquidity* **1**(1): 1550005.
URL: https://doi.org/10.1142/S2382626615500057

Bouchaud, J.-P., Mézard, M. and Potters, M. (2002). Statistical properties of stock order books: Empirical results and models, *Quantitative Finance* **2**(4): 251–256.
URL: https://doi.org/10.1088/1469-7688/2/4/301

Briola, A., Bartolucci, S. and Aste, T. (2025). Deep limit order book forecasting: A microstructural guide, *Quantitative Finance* **25**(7): 1101–1131.
URL: https://doi.org/10.1080/14697688.2025.2522911

Cont, R. and de Larrard, A. (2013). Price dynamics in a markovian limit order market, *SIAM Journal on Financial Mathematics* **4**(1): 1–25.
URL: https://doi.org/10.1137/110827662

Cont, R., Kukanov, A. and Stoikov, S. (2014). The price impact of order book events, *Journal of Financial Econometrics* **12**(1): 47–88.
URL: https://doi.org/10.1093/jjfinec/nbt003

Cont, R., Stoikov, S. and Talreja, R. (2010). A stochastic model for order book dynamics, *Operations Research* **58**(3): 549–563.
URL: https://doi.org/10.1287/opre.1090.0780

Fang, F. and Oosterlee, C. W. (2009). A novel pricing method for european options based on fourier-cosine series expansions, *SIAM Journal on Scientific Computing* **31**(2): 826–848.
URL: https://doi.org/10.1137/080718061

Gould, M. D. and Bonart, J. (2016). Queue imbalance as a one-tick-ahead price predictor in a limit order book, *Market Microstructure and Liquidity* **2**(2): 1650006.
URL: https://doi.org/10.1142/S2382626616500064

Gould, M. D., Porter, M. A., Williams, S. D., McDonald, M., Fenn, D. J. and Howison, S. D. (2013). Limit order books, *Quantitative Finance* **13**(11): 1709–1742.
URL: https://doi.org/10.1080/14697688.2013.803148

Huang, R. and Polak, T. (2011). Lobster: Limit order book reconstruction system. SSRN Scholarly Paper No. 1977207.
URL: https://doi.org/10.2139/ssrn.1977207

Jain, K., Firoozye, N., Kochems, J. and Treleaven, P. (2024). Limit order book simulations: A review. arXiv preprint arXiv:2402.17359; submitted to Quantitative Finance.
URL: https://arxiv.org/abs/2402.17359

Kercheval, A. N. and Zhang, Y. (2015). Modelling high-frequency limit order book dynamics with support vector machines, *Quantitative Finance* **15**(8): 1315–1329.
URL: https://doi.org/10.1080/14697688.2015.1032546

Kolm, P. N., Turiel, J. and Westray, N. (2023). Deep order flow imbalance: Extracting alpha at multiple horizons from the limit order book, *Mathematical*

Finance **33**(4): 1044–1081.
URL: https://doi.org/10.1111/mafi.12413

Lee, Y. and Kim, W. C. (2013). A stochastic model for order book dynamics: An application to korean stock index futures, *Management Science and Financial Engineering* **19**(1): 37–41.
URL: https://doi.org/10.7737/msfe.2013.19.1.037

Lucchese, L., Pakkanen, M. S. and Veraart, A. (2022). The short-term predictability of returns in order book markets: A deep learning perspective. arXiv preprint arXiv:2211.13777.
URL: https://arxiv.org/abs/2211.13777

Luckock, H. (2003). A steady-state model of the continuous double auction, *Quantitative Finance* **3**(6): 385–404.
URL: https://doi.org/10.1088/1469-7688/3/6/303

Ntakaris, A., Magris, M., Kanniainen, J., Gabbouj, M. and Iosifidis, A. (2018). Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods, *Journal of Forecasting* **37**(8): 852–866.
URL: https://doi.org/10.1002/for.2543

Prata, M., Masi, G., Berti, L., Arrigoni, V., Coletta, A., Cannistraci, I., Vyetrenko, S., Velardi, P. and Bartolini, N. (2023). Lob-based deep learning models for stock price trend prediction: A benchmark study. arXiv preprint arXiv:2308.01915.
URL: https://arxiv.org/abs/2308.01915

Sirignano, J. and Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning, *Quantitative Finance* **19**(9): 1449–1459.
URL: https://doi.org/10.1080/14697688.2019.1622295

Smith, E., Farmer, J. D., Gillemot, L. and Krishnamurthy, S. (2003). Statistical theory of the continuous double auction, *Quantitative Finance* **3**(6): 481–514.
URL: https://doi.org/10.1088/1469-7688/3/6/307

Thompson, I. J. (1986). Coulomb and bessel functions of complex arguments and order, *Journal of Computational Physics* **64**(2): 490–509.

Zaznov, I., Kunkel, J., Dufour, A. and Badii, A. (2022). Predicting stock price changes based on the limit order book: A survey, *Mathematics* **10**(8): 1234.
URL: https://doi.org/10.3390/math10081234

Zhang, Z., Zohren, S. and Roberts, S. (2019). Deeplob: Deep convolutional neural networks for limit order books, *IEEE Transactions on Signal Processing* **67**(11): 3001–3012.
URL: https://doi.org/10.1109/TSP.2019.2912336

A. BTG-OBD-A26 Dataset

This appendix describes the BTG-OBD-A26 dataset, which consists of high-frequency limit order book data from the Brazilian exchange B3, including message-level information on order arrivals, cancellations, and trades. These streams allow us to reconstruct the state of the limit order book at each timestamp and to build the state variables needed by the CST model. We detail the data structure, preprocessing steps, and the asset/day selection protocol used in our empirical evaluation.

A.1 Data Source and Structure

Dataset source. We use tick-by-tick limit order book data from the Brazilian exchange B3, disseminated through the FIX/FAST UMDF market data protocol. For each instrument and trading day, the raw feed is represented as event-based updates for the bid and offer sides together with a separate stream of trade prints.

Period and assets. The empirical experiments in this paper focus on liquid B3 instruments for which the UMDF feed provides deep and frequent updates. Our test design uses 5 tickers and 9 test dates drawn from the period October 2025–January 2026.

We choose the 5 tickers in two steps. First, we rank instruments by trading activity and retain the top $N = 50$ most traded tickers (by total traded volume over the study period). Second, among these liquid names, we select the 5 with the largest time fraction in one-tick-spread states, i.e., the highest fraction of the continuous session for which the instantaneous spread equals one tick ($S(t) = 1$), computed from 1-minute snapshots between 11:00 and 17:00 (local exchange time) to avoid auction-related periods. This criterion targets the regime where the semi-analytical best-price model is most directly applicable and where the next mid-price move is driven primarily by depletion of one of the best-price queues.

All five selected names are among the top-50 most liquid B3 equities and exhibit exceptionally high shares of tight-spread states, making them well-suited to best-price-level modeling.

Table A1 reports the selected tickers, their sectors, typical price levels, and the share of time spent in the one-tick-spread regime, which is the main criterion used to align the empirical design with the CST assumptions.

We choose the 9 test dates by stratifying trading days in October 2025–January 2026 according to a market-wide realized-volatility proxy computed

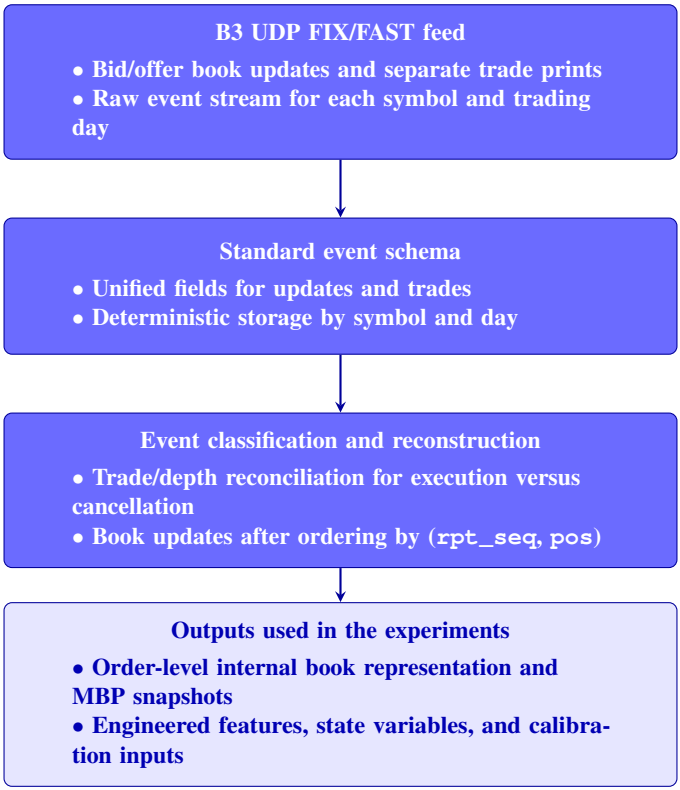

Figure A1
BTG-OBD-A26 data-processing pipeline. The workflow starts from the raw B3 UDP FIX/FAST source, containing bid and ask book updates together with separate trade prints, then normalizes these messages into the standard event schema used throughout the paper. Next, the pipeline applies the trade-reconciliation, event-classification, and order-book-reconstruction steps described in the appendix. The resulting outputs consumed by the empirical analysis include order-level book states, aggregated MBP snapshots, engineered features, and calibration inputs.

from `BOVA11`, an ETF designed to track the IBOVESPA index and thus summarize large-cap Brazilian equity conditions. Using `BOVA11` as a macro-regime proxy, we define for each day d a daily realized-variance estimator based on intraday mid-price log-returns sampled at $\Delta = 1$ minute and restricted

to 11:00–17:00 (local exchange time) to avoid auction-related periods,

$$\widehat{\sigma}_d^2 \;=\; \sum_j r_{d,j}^2, \qquad r_{d,j} = \log p_M(t_j) - \log p_M(t_j - \Delta).$$

We then select (i) the 3 highest-volatility days, (ii) the 3 lowest-volatility days, and (iii) the 3 days whose volatility is closest to the sample median, yielding a deliberately heterogeneous set of market conditions (stress, calm, and typical). The resulting 9 dates are used as the common test days for all five equities, ensuring comparability across assets and avoiding date selection based on each equity's own realized volatility.

Table A2 reports the resulting nine-date panel together with the realized-volatility values for `BOVA11` and for the five selected equities.

Table A1

Ticker selection table: generic instrument information and the time share with a one-tick spread. The $S(t) = 1$ time share is computed from 1-minute snapshots between 11:00 and 17:00 (local exchange time) to avoid auction-related periods. Tick size is BRL 0.01 for all selected tickers.

Ticker	Company name	Sector	Price (BRL, rounded)	Time share $S(t) = 1$
`CSAN3`	Cosan	Oil/gas/biofuels	7	98.24%
`ASAI3`	Assaí	Non-cyclical cons. (retail)	9	94.52%
`MGLU3`	Magazine Luiza	Cyclical cons. (retail)	11	93.94%
`ITSA4`	Itaúsa	Financials (diversified holdings)	15	91.39%
`LREN3`	Lojas Renner	Cyclical cons. (retail)	15	90.90%

Table A2

Test-date selection table: realized volatility for each of the 9 selected dates and each selected asset (5 tickers) plus `BOVA11`, computed with $\Delta = 1$ minute on the 11:00–17:00 window.

Bucket	Date	`BOVA11`	`CSAN3`	`ASAI3`	`MGLU3`	`ITSA4`	`LREN3`
high	*2025-12-05*	*3.99e-04*	*1.94e-03*	*1.46e-03*	*1.74e-03*	*6.23e-04*	*1.11e-03*
high	*2026-01-26*	*2.03e-04*	*6.46e-04*	*5.98e-04*	*6.56e-04*	*3.31e-04*	*3.79e-04*
high	*2026-01-29*	*1.91e-04*	*9.94e-04*	*5.25e-04*	*1.28e-03*	*2.97e-04*	*5.12e-04*
low	*2025-10-27*	*1.26e-05*	*3.79e-04*	*2.53e-04*	*4.76e-04*	*6.14e-05*	*1.43e-04*
low	*2025-11-10*	*1.29e-05*	*5.61e-04*	*3.52e-04*	*3.43e-04*	*5.31e-05*	*3.68e-04*
low	*2025-11-27*	*7.07e-06*	*2.40e-04*	*2.59e-04*	*2.96e-04*	*2.78e-05*	*1.31e-04*
median	*2025-10-14*	*2.89e-05*	*2.79e-04*	*3.02e-04*	*3.67e-04*	*9.98e-05*	*2.17e-04*
median	*2025-10-21*	*2.87e-05*	*7.54e-04*	*1.42e-04*	*4.11e-04*	*7.50e-05*	*2.13e-04*
median	*2025-12-30*	*2.87e-05*	*3.01e-04*	*2.35e-04*	*2.89e-04*	*6.13e-05*	*1.48e-04*
–	**mean**	**1.01e-04**	**6.77e-04**	**4.59e-04**	**6.50e-04**	**1.81e-04**	**3.58e-04**

Granularity and book representation (L2 vs. L3). An important distinction in the limit-order-book literature concerns the granularity of the data.

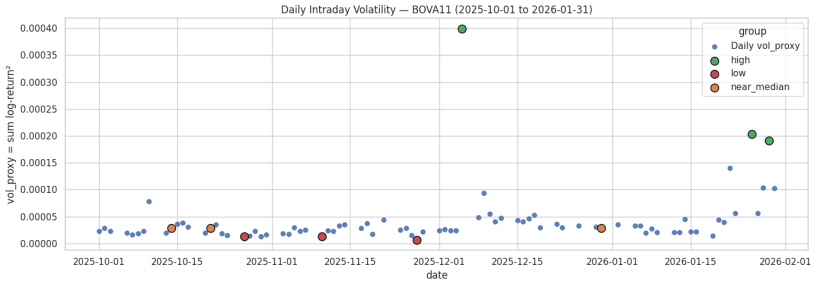

Figure A2
BOVA11 volatility-based date selection. This figure plots the daily intraday volatility of BOVA11 over the period from 01/10/2025 to 31/01/2026 and highlights the dates selected for the empirical experiments. The highlighted points identify the trading days chosen to represent high-, low-, and median-volatility market conditions in the sampling design.

Level-2 (L2) datasets report aggregated depth at each price level, but do not track the identity of individual orders within the queue. By contrast, Level-3 (L3) datasets preserve order-level information, making it possible to follow submissions, cancellations, executions, and queue reordering at the individual order level. Our dataset is effectively L3 because it tracks the `ord_id` field whenever it is available, which allows us to reconstruct the queue at the level of individual resting orders rather than only at the aggregated price-level view. This distinction matters for our application because queue-depletion dynamics, order matching, and cancellation attribution are more naturally handled in an order-level environment.

For comparison, the FI-2010 benchmark dataset of Ntakaris et al. (2018) is an L2-style dataset designed for mid-price forecasting, since it provides normalized snapshots of aggregated book depth rather than order identifiers. LOBSTER, on the other hand, is much closer to an L3 environment: it is built from Nasdaq TotalView-ITCH order messages and reconstructs event-level order-book states from order-level updates (Huang and Polak, 2011). Our BTG-OBD-A26 dataset is therefore closer in spirit to LOBSTER than to FI-2010, although our empirical pipeline ultimately produces both an order-level internal representation and aggregated MBP snapshots for downstream analysis.

Dataset schema. All raw sources are coerced into a single *standard event schema* prior to reconstruction. Each row corresponds to one UMDF update

(bid/offer) or one trade print, with the following fields:

- `dat_tim`: event timestamp with millisecond precision.
- `rpt_seq`: instrument-level report sequence used for deterministic ordering.
- `pos`: position (1-indexed) of the entry position in the queue.
- `upd_act`: update action code. For book updates, the values in $\{0,1,2,3,4,5\}$ denote new, change, delete, delete-thru, delete-from, and overlay; the value 10 denotes trades.
- `px`: price (floating-point) associated with the update.
- `size`: size/quantity associated with the update.
- `side`: `bid`, `offer` or `trade`.
- `msg_tpe`: message type, mapped to `snapshot` or `incremental`.
- `ord_id`: order identifier when available (null for trades).

More information on the UMDF feed structure can be found in the B3 documentation.

A.2 Event Classification

The CST model requires a clean partition of order flow into (i) limit orders, (ii) market orders (executions), and (iii) cancellations. We define classification rules using the schema fields above.

Relative price level. In our dataset, all stocks have a fixed tick size (i.e., the minimum price increment) of R$0.01. We let $p_b(t)$ and $p_a(t)$ be the best bid and best ask prices immediately before an event at time t. We work in tick indices $\pi(p) = p/0.01$. For a bid-side update with price p, we define the relative level

$$i = \pi(p_a(t^-)) - \pi(p), \tag{44}$$

and for an ask-side update

$$i = \pi(p) - \pi(p_b(t^-)). \tag{45}$$

Thus $i = 1$ corresponds to the best quote on the event side when the spread is one tick.

(1) Limit orders. We classify an event as a limit order arrival if:

- `msg_tpe` is `incremental`,
- `upd_act` $= 0$ (new),
- `side` $\in \{$`bid`,`offer`$\}$, and
- `px` and `size` are finite and non-zero.

The level index i is computed from the pre-event snapshot as above. In the $S = 1$ calibration regime, arrivals at $i = 1$ correspond to limit orders posted at the best quotes; arrivals at $i > 1$ correspond to deeper placements.

(2) Market orders (executions). We define an event as a market order if it is identified as a `side=trade` and has `size`> 0.

(3) Cancellations. Trade prints and book updates arrive as separate messages in the feed, so a trade print does not directly remove volume from the reconstructed book. As a result, a reduction in displayed depth observed immediately after a trade may reflect either true execution or an unrelated cancellation. To separate these two effects, we use a reconciliation procedure based on a running "pending traded volume" variable, denoted by V. Each trade print adds its reported size to V, and this pending volume is then greedily matched to subsequent size-decreasing book updates (`upd_act`$\in \{1,2,3,4,5\}$) on the bid and ask sides.

For each such update, we first compute a candidate executed quantity Δ^{cand}, that is, the amount of depth reduction that could plausibly be attributed to the most recent unmatched trade prints. When the message is a size change (`upd_act=1`) at position `pos`, Δ^{cand} is the decrease in resting size at that position. For delete-like actions, Δ^{cand} is computed from the amount of depth removed according to the current book state and the semantics of the action code (delete, delete-thru, delete-from, or overlay). We then attribute to execution only the portion of that reduction that is actually supported by pending traded volume: if the observed reduction is smaller than the remaining value of V, we treat the entire reduction as executed volume; if it is larger, we attribute execution only up to the amount still available in V. After that, we decrement V by the volume attributed to execution. If a book update reduces displayed depth when no pending traded volume is available, or if the reduction exceeds the remaining value of V, the unmatched portion is classified as a cancellation or deletion rather than as an execution.

This distinction is essential for estimating cancellation intensities θ_i without bias: partial or full executions are absorbed by the market-order process, whereas genuine non-trade removals are counted in the cancellation/deletion process.

A.3 Order Book Reconstruction

Deterministic ordering. Raw events are first normalized to the standard schema and then ordered deterministically by (`rpt_seq`, `pos`). This resolves out-of-order arrivals in wall-clock time while preserving the feed-defined sequencing required to apply position-based updates consistently.

Incremental update rules. Reconstruction maintains two side-specific vectors (bids and offers), each storing entries with fields (price, size). An incremental event with action code `upd_act` updates the selected side as follows:

- `0 (new)`: insert a new entry at position `pos`.
- `1 (change)`: update the size (and optionally price) of the entry at `pos`.
- `2 (delete)`: remove the entry at `pos`.
- `3 (delete_thru)`: delete entries from `pos` to the end of the vector.
- `4 (delete_from)`: delete entries from the front up to `pos` (or, under inverted-book regimes, from `pos` to end).
- `5 (overlay)`: update/insert if `px`> 0, otherwise delete at `pos`.

Trade events (`side=trade`) never mutate the book; they only influence execution attribution as described in Section A.2.

Snapshot materialization. After each applied book update, we materialize a snapshot consisting of: (i) an aggregated MBP view of the top K price levels on both sides, and (ii) a capped list of individual entries near the top of book. This provides a consistent state sequence that can be consumed by feature extractors (e.g., spread, midprice, imbalance) and by stochastic calibration routines.

A.4 Summary Statistics of the Dataset

This subsection reports descriptive statistics for the prediction dataset used in the empirical evaluation. Table A3 summarizes the overall size of the panel, including the number of symbols, prediction dates, events, and unique orders. Table A4 then breaks the sample down by ticker, highlighting the heterogeneity in message counts and order activity across assets. Finally, Table A5 reports the distribution of the event labels used in the preprocessing pipeline.

Table A3
Descriptive summary of the prediction dataset.

Indicator	Value
Total symbols	5
Total prediction dates	9
Total events	7,240,828
Total unique orders	2,853,006
Mean events per file	160,907.3
Median events per file	112,336.0
Minimum number of events in a file	60,377
Maximum number of events in a file	522,430

Table A4
Aggregated statistics by symbol in the prediction dataset.

Symbol	Events	Mean Events/Day	Unique Orders	New	Change	Delete	Trade	Bid	Offer
ASAI3	1,182,131	131,347.9	432,905	638,471	117,495	426,165	205,566	521,174	455,391
CSAN3	1,236,685	137,409.4	457,468	665,927	130,866	439,892	208,459	504,681	523,545
ITSA4	1,449,243	161,027.0	543,895	758,014	160,479	530,750	214,119	607,636	627,488
LREN3	2,118,158	235,350.9	887,852	1,122,201	113,517	882,440	234,349	919,125	964,684
MGLU3	1,254,611	139,401.2	530,886	682,712	54,949	516,950	151,826	551,285	551,500

Table A5
Distribution of event types in the prediction dataset.

Event Type	Total Events	Share (%)
New	3,867,325	53.41
Change	577,306	7.97
Delete	2,796,197	38.62
Trade	1,014,319	14.01

B. Calibration Details

This appendix summarizes the calibration outputs used in the empirical analysis. We report day-ahead parameter estimates by asset and date, stratified by market-regime buckets (high, low, and median realized-volatility days),

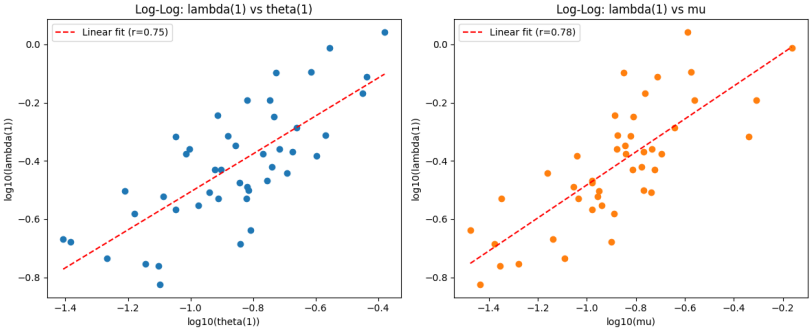

Figure B3
Log-log calibration diagnostics. This figure plots best-price-level parameter estimates in log-log scale to assess cross-sectional coherence of the day-ahead calibration procedure. The left panel compares replenishment intensity $\lambda(1)$ with cancellation intensity $\theta(1)$, and the right panel compares $\lambda(1)$ with market-order intensity μ. Each point corresponds to one asset-date calibration pair.

and we document the cross-sectional regularities used to validate the fitted intensities.

Table B6 reports the core best-price-level calibration outputs used by the CST dynamics in the $S = 1$ regime. Each row corresponds to one asset-date pair under the day-ahead protocol, where the date is the calibration date (previous trading day) and the bucket label comes from the corresponding prediction-day volatility stratum. The table is designed to make level and dispersion comparisons transparent across assets and regimes, especially for the balance between replenishment, $\lambda(1)$ and depletion forces, $\theta(1)$ and μ.

Figure B3 provides a visual consistency check for these estimates in log-log scale. The left panel compares $\lambda(1)$ and $\theta(1)$, while the right panel compares $\lambda(1)$ and μ, across all asset-date calibrations. The objective is not to impose a strict linear law, but to verify that calibrated points remain in coherent ranges and preserve stable cross-sectional scaling under the same estimation pipeline.

Table B7 summarizes the depth-profile fit for arrival rates, where $\lambda(i)$ is approximated by $ki^{-\alpha}$ for each asset-date pair. In combination with the first table, it separates best-price-level intensity information from depth-shape information, allowing us to evaluate whether day-ahead calibrations are jointly plausible both at $i = 1$ and along deeper levels.

Table B6
Compact calibration summary for best-price-level parameters $(\lambda_1, \theta_1, \mu)$ across the 9 evaluation windows and 5 selected assets. The *Date* column reports the calibration date (`calib_date`, i.e., the previous trading day used in the day-ahead protocol), not the prediction date. The *Bucket* column (high/low/median) follows the corresponding prediction-day volatility stratification.

Bucket	Calib date	Asset	λ_1	θ_1	μ
high	2025-12-04	CSAN3	0.413615	0.252597	0.091192
high	2025-12-04	ASAI3	0.334972	0.143623	0.104862
high	2025-12-04	MGLU3	0.643649	0.179169	0.491052
high	2025-12-04	ITSA4	0.774229	0.366962	0.194189
high	2025-12-04	LREN3	0.644622	0.151450	0.275474
high	2026-01-23	CSAN3	0.361072	0.203816	0.068960
high	2026-01-23	ASAI3	0.271688	0.089496	0.105330
high	2026-01-23	MGLU3	0.437529	0.099348	0.184699
high	2026-01-23	ITSA4	0.802436	0.242443	0.265930
high	2026-01-23	LREN3	0.566231	0.184487	0.154819
high	2026-01-28	CSAN3	0.420314	0.170721	0.144327
high	2026-01-28	ASAI3	0.421116	0.096693	0.202115
high	2026-01-28	MGLU3	0.569748	0.122070	0.130014
high	2026-01-28	ITSA4	1.104067	0.417930	0.258124
high	2026-01-28	LREN3	0.798725	0.188119	0.141631
low	2025-10-24	CSAN3	0.340658	0.175375	0.104884
low	2025-10-24	ASAI3	0.280198	0.105836	0.114729
low	2025-10-24	MGLU3	0.262536	0.066160	0.129646
low	2025-10-24	ITSA4	0.230072	0.155482	0.033484
low	2025-10-24	LREN3	0.436631	0.192775	0.132287
low	2025-11-07	CSAN3	0.518916	0.218138	0.228534
low	2025-11-07	ASAI3	0.372379	0.125224	0.189223
low	2025-11-07	MGLU3	0.310043	0.115041	0.182622
low	2025-11-07	ITSA4	0.488675	0.270012	0.133127
low	2025-11-07	LREN3	0.973843	0.278674	0.686499
low	2025-11-26	CSAN3	0.428101	0.211859	0.171136
low	2025-11-26	ASAI3	0.372195	0.119841	0.153660
low	2025-11-26	MGLU3	0.300387	0.081891	0.110541
low	2025-11-26	ITSA4	0.677694	0.355816	0.172856
low	2025-11-26	LREN3	0.450072	0.138997	0.143679
median	2025-10-13	CSAN3	0.173597	0.078859	0.044365
median	2025-10-13	ASAI3	0.380531	0.182217	0.166914
median	2025-10-13	MGLU3	0.210186	0.041357	0.125589
median	2025-10-13	ITSA4	0.295361	0.122455	0.092123
median	2025-10-13	LREN3	0.483959	0.131817	0.151210
median	2025-10-20	CSAN3	0.149558	0.079834	0.036689
median	2025-10-20	ASAI3	0.315236	0.152987	0.170424
median	2025-10-20	MGLU3	0.184545	0.054158	0.081121
median	2025-10-20	ITSA4	0.323511	0.151521	0.088310
median	2025-10-20	LREN3	0.483508	0.089795	0.458229
median	2025-12-29	CSAN3	0.206660	0.144333	0.041917
median	2025-12-29	ASAI3	0.176248	0.071830	0.052596
median	2025-12-29	MGLU3	0.214740	0.039170	0.072966
median	2025-12-29	ITSA4	0.296405	0.150930	0.044827
median	2025-12-29	LREN3	0.313641	0.061740	0.112308

Table B7
Compact power-law fit summary for depth-dependent arrival intensities, where $\lambda(i)$ is fitted to $y(i) = k\,i^{-\alpha}$ for each asset-date pair. The *Date* column reports the calibration date (`calib_date`, i.e., the previous trading day used in the day-ahead protocol), not the prediction date. The *Bucket* column (high/low/median) follows the corresponding prediction-day volatility stratification.

Bucket	Calib date	Asset	k	α
high	2025-12-04	CSAN3	0.450765	2.296009
high	2025-12-04	ASAI3	0.601250	2.471826
high	2025-12-04	MGLU3	1.037598	1.783385
high	2025-12-04	ITSA4	1.022944	2.481374
high	2025-12-04	LREN3	1.320363	2.278699
high	2026-01-23	CSAN3	0.719833	2.470466
high	2026-01-23	ASAI3	0.414245	2.152520
high	2026-01-23	MGLU3	0.742784	2.282147
high	2026-01-23	ITSA4	0.933546	2.146884
high	2026-01-23	LREN3	1.614750	2.557956
high	2026-01-28	CSAN3	0.754485	2.025858
high	2026-01-28	ASAI3	0.677106	2.078224
high	2026-01-28	MGLU3	0.973817	2.286642
high	2026-01-28	ITSA4	1.614631	2.292604
high	2026-01-28	LREN3	1.658275	2.324393
low	2025-10-24	CSAN3	0.484987	2.103116
low	2025-10-24	ASAI3	0.362669	2.179728
low	2025-10-24	MGLU3	0.339567	1.826729
low	2025-10-24	ITSA4	0.325459	2.217633
low	2025-10-24	LREN3	0.826640	2.267504
low	2025-11-07	CSAN3	0.635165	2.348451
low	2025-11-07	ASAI3	0.390443	2.202947
low	2025-11-07	MGLU3	0.384556	2.130368
low	2025-11-07	ITSA4	0.428740	2.205468
low	2025-11-07	LREN3	1.841430	2.559602
low	2025-11-26	CSAN3	0.393008	1.843226
low	2025-11-26	ASAI3	0.506924	2.331137
low	2025-11-26	MGLU3	0.547898	2.101441
low	2025-11-26	ITSA4	0.914326	2.617574
low	2025-11-26	LREN3	0.729487	2.087238
median	2025-10-13	CSAN3	0.277997	1.937566
median	2025-10-13	ASAI3	0.401054	2.195854
median	2025-10-13	MGLU3	0.328905	1.699973
median	2025-10-13	ITSA4	0.371479	2.489704
median	2025-10-13	LREN3	0.828645	2.273455
median	2025-10-20	CSAN3	0.299489	2.212505
median	2025-10-20	ASAI3	0.396374	2.232968
median	2025-10-20	MGLU3	0.313557	1.966451
median	2025-10-20	ITSA4	0.439241	2.203692
median	2025-10-20	LREN3	0.823906	2.044116
median	2025-12-29	CSAN3	0.345083	2.423874
median	2025-12-29	ASAI3	0.329609	2.705706
median	2025-12-29	MGLU3	0.249454	1.884639
median	2025-12-29	ITSA4	0.519438	2.601653
median	2025-12-29	LREN3	0.878531	2.487730