ISSN: 2281-1346



UNIVERSITÀ DI PAVIA
Department of Economics and Management

DEM Working Paper Series

Duration Modeling in the Presence of Zero-Duration Clusters

Massimiliano Caporin
(University of Padova)

Alessandro Morelli
(University of Milan)

Eduardo Rossi
(University of Pavia)

 # 236 (11-26)

Via San Felice, 5
I-27100 Pavia

https://economiaemanagement.dip.unipv.it/it

Duration Modeling in the Presence of Zero-Duration Clusters

Massimiliano Caporin[1], Alessandro Morelli[2], and Eduardo Rossi[3]

[1]Department of Statistical Sciences, University of Padova, Italy
[2]Department of Economics, Management and Quantitative Methods, University of Milan, Italy
[3]Department of Economics and Management, University of Pavia, Italy

November 17, 2026

Abstract

Modern electronic markets are fast, fragmented, and increasingly governed by algorithmic interaction. In this context, groups of transactions sharing the same timestamp (zero-duration clusters) should be viewed as potential signals of bursts in latent execution intensity. Using millisecond-level tick data for ten large-cap, highly liquid U.S. stocks over the last two quarters of 2024, we document three empirical regularities: durations drop immediately after a cluster and gradually revert toward their unconditional mean; larger clusters are associated with shorter subsequent durations; and clustering exhibits persistence over time, especially when clusters are large. Motivated by these findings, we propose extensions of the Autoregressive Conditional Duration (ACD) model that incorporate cluster size through a limited number of additional parameters. An extensive out-of-sample forecasting evaluation shows that specifications incorporating cluster size consistently emerge as the best-performing models. We conclude that zero-duration clusters—and their size in particular—contain relevant information regarding latent execution intensity, and that incorporating this information yields robust improvements in forecasting future durations.

Keywords: Duration, High-frequency data, Zero durations, Autoregressive Conditional Duration model, ACD, Simultaneous transactions.

JEL codes: C22, C41, C58, G14.

1 Introduction

Modern financial markets are fast, fragmented, and increasingly shaped by algorithmic trading. Order entry, matching, and execution occur at microsecond—and sometimes nanosecond—resolution (U.S. Securities and Exchange Commission, 2020), while trading activity is dispersed across venues and strategically managed by algorithms. As emphasized by O'Hara (2015), high-frequency market microstructure is not only about speed: traders optimize against market design, against other high-frequency traders, and against non-high-frequency traders, who in turn must adapt to markets populated by algorithms. In this environment, the timing of trades can carry economically relevant information.

This paper studies groups of transactions recorded with the same timestamp—hereafter *zero-duration clusters* or simply *clusters*—that generate zero durations in transaction-level data. These events are not merely technical artifacts induced by finite timestamp precision. In datasets with sufficiently fine resolution, they can reveal bursts in latent execution intensity. They may arise from the interaction of algorithmic order placement, liquidity consumption across fragmented venues, and rapid responses to local order-flow conditions. Understanding their dynamics is therefore central to trading-strategy analysis—given that both high-frequency trading (HFT) and non-HFT participants operate in environments where short-run execution intensity matters—and to market monitoring, because unusually intense execution episodes are linked to market complexity, liquidity provision, and potential instability (see U.S. Securities and Exchange Commission, 2020).

Simple point processes provide a natural framework for modeling inter-arrival times. Because events occur at distinct times with probability one, inter-arrival durations are strictly positive by construction. This feature has motivated the widespread use of duration models in finance, most notably the Autoregressive Conditional Duration (ACD) model of Engle and Russell (1998) and its many extensions (see Pacurar, 2008 for a comprehensive survey, and Bhatti, 2010, and more recently Saulo et al., 2025, among others).

In practice, however, empirical datasets record events on a discrete time grid (e.g., seconds or milliseconds), while the underlying process unfolds in continuous time with irregularly spaced arrivals. When multiple events, occurring in rapid succession, are mapped to the same grid point by rounding, identical timestamps appear and the resulting series contains runs of zero durations. This issue is especially relevant in high-frequency finance, where the resolution of transaction data is typically coarser than that at which electronic trading systems operate. Simultaneous timestamps are thus artifacts of limited temporal resolution, but they may also signal increased intensity in the latent execution process.

A common remedy, originally proposed by Engle and Russell (1998), is to merge each group of transactions sharing the same timestamp into a single event, eliminating zeros after differencing.

This practice is often justified by the "split transaction" argument: traders are assumed to break large orders into smaller pieces to facilitate execution,

so contemporaneous prints should be merged or removed. This view, however, has important limitations. In fragmented markets driven by algorithms that place, revise, and cancel orders at extremely short time scales, sequences of zero durations are more plausibly manifestations of complex order-placement and liquidity-taking strategies than mechanical fragments of a single aggressive order. Moreover, if the focus is on understanding the mechanics of modern ultra-high-frequency markets and algorithmic trading strategies, excluding them amounts to discarding economically meaningful information.

This aligns with the argument of O'Hara (2015) that, in high-frequency markets, information is no longer exclusively asset-related, but order-related as well: being informed often means reacting quickly to order flow and local liquidity conditions. Bursts of nearly simultaneous executions are therefore a natural object of interest, rather than a feature to be discarded a priori.

Transaction clustering may also reflect traders' attraction to round prices. As documented by Veredas et al. (2001), liquidity tends to concentrate at round price levels and order-flow intensity increases in their vicinity, leading to rapid sequences of execution. More recently, Lee (2026) documents how buy-sell imbalances arising from behavioral biases among traditional traders create profitable trading opportunities around round prices, which high-frequency algorithms can exploit by acting as their counterparties.

To mitigate the information loss induced by the aggregation procedure, Zhang et al. (2001) introduced an indicator variable for groups of zeros in the conditional mean equation. An alternative is to retain zeros and model them directly via zero-inflated specifications, such as Hautsch et al. (2014). Although their main focus is on regularly spaced time aggregates—where zeros reflect periods of inactivity—the approach can, in principle, be adapted to duration series. In transaction data, however, zero durations mainly arise from limited timestamp resolution that fails to capture the micro- or nanosecond granularity of electronic systems. In that setting, zero-inflated modeling is most useful when the object of interest is the conditional distribution of zero outcomes themselves, as in Blasques et al. (2024).

We contribute to the literature on zero durations by adopting a different approach. We follow the aggregation procedure of Engle and Russell (1998), but exploit the information contained in zero-duration clusters and their size, thereby maintaining the structural simplicity of a simple point process while ensuring strictly positive duration forecasts. This approach is particularly relevant in modern high-frequency settings, where a substantial share of transactions is recorded as contemporaneous: in our dataset, 50–60% of transactions share the same timestamp (see Figure 1).

We provide a comprehensive empirical investigation of the interaction between positive durations and zero-duration clusters, showing that cluster *size* is a key signal of spikes in underlying execution intensity and its persistence. We document three empirical regularities that hold systematically across ten large-cap, highly liquid U.S. stocks.

First, durations drop immediately after a cluster and gradually revert toward their unconditional mean. Larger clusters are followed by a more pronounced

drop and slower reversion, highlighting a clear size-dependent effect. This pattern is consistent with limit-order placement strategies in fragmented markets, where liquidity posted near prevailing quotes and across trading venues can be consumed rapidly, producing short-lived but persistent bursts of executions.

Second, the relationship between cluster size and the post-cluster duration is systematic and is well approximated by a quadratic polynomial in the reciprocal of cluster size. This inverse-size form is natural because the inverse cluster size proxies the average latent time per execution within a single recorded timestamp, yielding a parsimonious link between cluster size and subsequent duration.

Third, clusters exhibit persistence: the probability of observing another cluster over the next few observations decays slowly following a large cluster. This persistence indicates that clusters are not isolated events, but instead arise in sequences generated by the progressive matching of liquidity supplied through limit orders. Larger clusters are therefore more likely to trigger further clustered executions when liquidity is posted across multiple venues and book levels.

These empirical regularities motivate a parsimonious extension of existing ACD and score-driven models (Creal et al., 2013; Blasques et al., 2024) that incorporates both the occurrence and the size of zero-duration clusters into the conditional mean dynamics. We benchmark these specifications in an extensive out-of-sample comparison against standard alternatives. The forecasting exercise assesses whether bursts of execution intensity have persistent and predictable effects on subsequent trading activity, rather than being mere artifacts of timestamp aggregation.

Our results show that zero-duration clusters—and in particular their size—have a systematic impact on duration dynamics. Models that incorporate cluster size consistently dominate out of sample across all stocks, indicating that clusters contain valuable information about high-frequency duration behavior.

While zero-duration clusters could be embedded in a broader marked execution process, this paper deliberately isolates the timing dimension. The issue originates in the limited temporal resolution of transaction data and already poses nontrivial challenges for modeling the duration dynamics, as emphasized by the existing literature (Pacurar, 2008).

The remainder of the paper is organized as follows. Section 2 introduces the ultra-high-frequency transaction data, defines zero-duration clusters, and describes the diurnal adjustment procedure. Section 3 documents the empirical regularities associated with zero-duration clusters. Section 4 presents our proposed extension of the logACD model with cluster indicators and cluster-size effects, together with benchmark models and score-driven alternatives. Section 5 reports the in-sample estimation results, while Section 6 contains the forecasting exercises. Section 7 concludes.

2 Ultra-high-frequency durations

In contrast to the foundational body of econometric research on inter-trade durations, which largely relies on transaction data recorded at relatively coarse temporal resolutions, the transactions analyzed in this study are timestamped at millisecond precision and cover a sample of ten actively traded large-cap equities, traded across multiple exchanges during the last two quarters of 2024. With trading activity dispersed across fragmented and highly automated market systems operating at micro- or even nanosecond time scales, rapid execution bursts may occur over intervals shorter than the temporal resolution at which transactions are recorded, resulting in zero-duration clusters. In this section, we introduce a framework to describe the relationship between observed durations and the underlying latent execution times. We also report stock-by-stock summary statistics on zero-duration clusters, averaged across trading days, to provide a concrete picture of the empirical environment in which the analysis is conducted.

2.1 Observed versus latent durations

Consider a sequence of transactions $j = 1, 2, \ldots, N$. Let $\check{t}_0 = 0 < \check{t}_1 < \check{t}_2 < \cdots < \check{t}_N$ be the associated latent execution times expressed in the unit of resolution of the dataset. Since the unit of measurement of the dataset is typically coarser than the resolution of the latent timestamps, we allow $\check{t}_j \in \mathbb{R}_+$. For instance, executions may occur at nanosecond resolution while the dataset records timestamps in milliseconds, in which case we express all times in milliseconds.

We define a rounding function $R : \mathbb{R}_+ \to \mathbb{N}_0$ such that $R(\check{t}_j) = \lfloor \check{t}_j \rfloor + e_j$ where

$$e_j = \begin{cases} 0 & \text{if } \{\check{t}_j\} = 0 \text{ or } \{\check{t}_j\} < \varrho, \\ 1 & \text{otherwise.} \end{cases}$$

with $\varrho \in [0, 1]$. Here $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote the floor and ceiling operators, so that $\{\check{t}_j\} := \check{t}_j - \lfloor \check{t}_j \rfloor$ is the fractional part of $\check{t}_j$. We can think of R as the transformation operated by the data provider.[1]

The observed timestamps t_j and durations d_j are then given by

$$t_j := R(\check{t}_j) \in \mathbb{N}_0, \qquad d_j := t_j - t_{j-1}, \quad j = 1, 2, \ldots, N.$$

The latent durations are $\check{d}_j := \check{t}_j - \check{t}_{j-1}$. Crucially, regardless of the choice of ϱ, the difference between two rounded timestamps is always equal either to the floor or the ceiling of the latent difference:

$$d_j \in \{\lfloor \check{d}_j \rfloor, \lceil \check{d}_j \rceil\}. \tag{1}$$

A proof of this result is provided in Appendix A. Consequently, whenever the latent duration is smaller than one unit, we get $d_j \in \{0, 1\}$. A sequence of

[1] Notice that when $\varrho = 1$ this coincides with the floor function, when $\varrho = 0$ it coincides with the ceiling function, and when $\varrho = 0.5$ it corresponds to the usual round-half-up rule.

consecutive zero durations

$$d_j = d_{j+1} = \cdots = d_{j+k-1} = 0, \qquad k \geq 1,$$

arises each time $k+1$ consecutive timestamps are mapped to the same value:

$$R(\check{t}_{j-1}) = R(\check{t}_j) = \cdots = R(\check{t}_{j+k-1}).$$

2.2 Event definition

We analyze tick-by-tick transactions timestamped at millisecond precision for ten highly liquid large-cap U.S. stocks: Boeing (BA), Caterpillar (CAT), Chevron (CVX), Walt Disney (DIS), Honeywell (HON), JPMorgan Chase (JPM), McDonald's (MCD), Walmart (WMT), Johnson & Johnson (JNJ), and Amgen (AMGN). The sample covers the last two quarters of 2024, comprising 125 trading days for each stock. Each trading session runs from 9:30 to 16:00. The data are obtained from Kibot.com and consist of transaction records consolidated across multiple trading venues.

We treat each trading session independently. For each day and stock, we aggregate all transactions sharing the same timestamp into a single *event*, following Engle and Russell (1998). This yields a sequence of events for each day ℓ and stock s indexed by

$$i = 1, 2, \ldots, N_{\ell,s}.$$

Each event i is either a single transaction or a cluster of simultaneous transactions. We characterize event i by its size q_i, which equals 1 for a single trade and exceeds 1 for a merged set of trades, i.e., a cluster. Figure 1 and Table 1 provide descriptive summaries of clusters and their sizes.

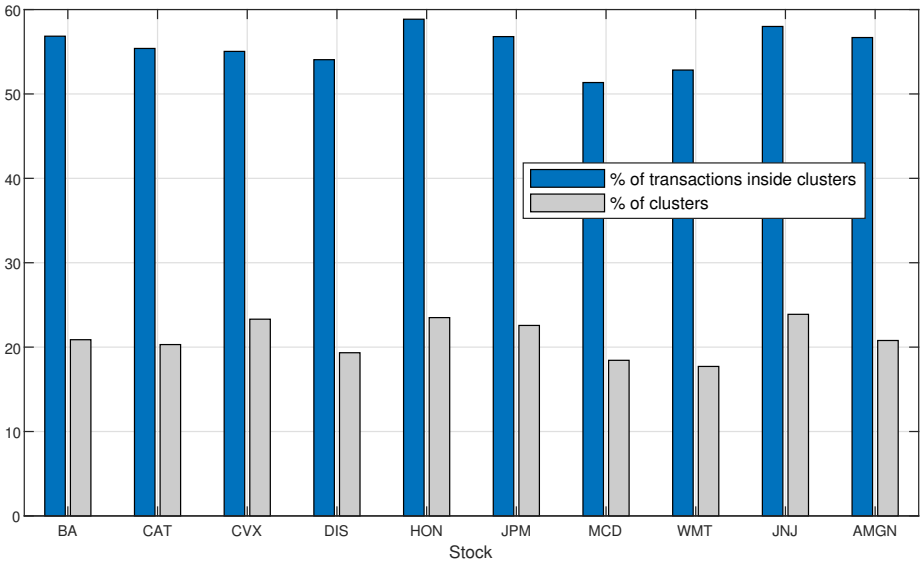
A simple comparison is informative. Based on the average daily number of events in Table 1, the implied average inter-event duration ranges from roughly 0.3 (WMT) to 1 second (AMGN), well above the millisecond resolution of the dataset. This indicates that, in our setting, zero-duration clusters are not merely a mechanical by-product of durations being close to the recording threshold.

Henceforth, for notational convenience, we suppress explicit day and stock indices, since all procedures are applied independently to each trading session.

2.3 Diurnal patterns

It is well established that high-frequency financial data exhibit pronounced diurnal patterns. Volatility and trading volume typically follow a U-shaped intraday profile, while durations display the inverse pattern: trades arrive more frequently near the market open and close, and less frequently around midday (Pacurar, 2008). Because there is no consensus on a deterministic functional form for such pattern, the standard approach is to treat it as a low-frequency component and filter it out before modeling durations. In the ACD literature, the aim is to obtain *diurnally adjusted* durations that are approximately stationary and exhibit rapidly decaying autocovariances.

Figure 1: Simultaneous timestamps



Notes The figure shows, for each stock, the proportion of events affected by simultaneous timestamps. The left bar is the percentage of transactions that occur in clusters. The right bar is the percentage of clusters relative to the total number of events.

The most common filtering method relies on cubic splines. Following Engle and Russell (1998), spline nodes are typically placed at regular intraday intervals, with additional nodes near the open and the close to capture the rapid changes in trading activity. In Engle and Russell's original implementation, the diurnal pattern is assumed to repeat identically across all days in the sample. Adjusted durations are obtained by dividing raw durations by the estimated seasonal factor. Zhang et al. (2001) note that, although joint estimation of seasonality and ACD parameters is feasible, a two-step pre-filtering procedure is asymptotically equivalent and often computationally convenient in large samples. Many subsequent studies adopt this approach: Bauwens and Giot (2000), Bauwens (2006), Fernandes and Grammig (2006), and Meitz and Teräsvirta (2006) all employ a two-step procedure, estimating the diurnal pattern separately for each stock and evaluating the average duration every 30 minutes. Similarly, Tse and Dong (2014) adopt the two-step procedure using 30-minute intervals with two additional 15-minute windows at the open and close. Meitz and Teräsvirta (2006) also emphasize that spline nodes represent conditional averages across days and that individual trading days can deviate substantially from the average profile. Alternative approaches include nonparametric smoothing methods (e.g., Zhang et al., 2001), and the time-change procedure proposed by Wu (2012), which relies on point process theory to deform the time scale based on trading intensity. Pacurar (2008) surveys additional strategies, such

Table 1: Average daily number of events and clusters by size

Stock	Events	$2 \leq q < 10$	$10 \leq q < 40$	$q \geq 40$	% $q \geq 40$
BA	62015	11545	1557	56	0.35
CAT	26341	4785	619	16	0.28
CVX	62745	13722	959	12	0.08
DIS	62017	10759	1429	31	0.23
HON	26097	5570	638	15	0.21
JPM	68003	14123	1438	25	0.15
MCD	37584	6253	751	20	0.26
WMT	77022	11965	1843	52	0.37
JNJ	54227	11892	1142	12	0.09
AMGN	23911	4451	587	26	0.44

Notes Average daily number of events and clusters by size range, reported by stock. The last column reports the average across days of the daily percentage of clusters with size $q \geq 40$ relative to the total number of clusters. Each cluster counts as a single event, regardless of its size. Each cluster counts as a single event, regardless of its size.

as employing time-of-day dummy variables to capture the diurnal pattern in discrete bins.

To control for diurnal patterns, we apply a day- and stock-specific cubic spline pre-filter. This choice reflects the substantial heterogeneity in diurnal patterns across days and stocks, and the fact that, in our ultra-high frequency dataset each interval contains many observations, allowing the precise estimation of interval means. We partition each trading session into 14 contiguous intervals: a 15-minute window at the open, twelve 30-minute windows, and a final 15-minute window at the close. For each interval, we compute the average duration and place a spline node at the interval midpoint.

We also add two boundary nodes at 09:30 and 16:00, set to the average duration in the first and last 5 minutes of the session, respectively. These additional nodes are introduced to prevent the fitted cubic spline from extrapolating negative values at the boundaries, where trading activity is extremely fast. Figure 2 illustrates fitted diurnal components for four Boeing (BA) trading days. Let $\hat{\lambda}(t_i)$ denote the fitted spline polynomial at time t_i. We define adjusted durations, x_i, as:

$$x_i = \eta \cdot \frac{d_i}{\hat{\lambda}(t_{i-1})},$$

where $\eta = 23400 / \sum_k \frac{d_k}{\hat{\lambda}(t_{k-1})}$ rescales the adjusted durations so that they sum to the length of the trading session. Adjusted timestamps are then obtained by

cumulating the adjusted durations:

$$\tau_i = \sum_{k=1}^{i} x_k, \qquad i = 1, \ldots, N,$$

with $\tau_0 = 0$, so that $x_i = \tau_i - \tau_{i-1}$.

The estimation of a separate diurnal profile for each trading day deserves clarification. Our objective is to model the dependence among consecutive transaction durations, which occur at the millisecond scale or finer, rather than the low-frequency variation in trading activity. Removing a day-specific diurnal component allows us to do so without imposing a deterministic intraday pattern that repeats identically across days, an assumption that is too restrictive for modern financial markets (see Brownlees et al., 2011 and Brownlees and Vannucci, 2013). Forecasting raw durations would instead require explicitly modeling and predicting the evolution of the diurnal component across days. Since day-specific profiles are estimated after observing the trading session, our out-of-sample analysis is not intended to assess the ability of competing specifications to forecast raw durations in real time. Rather, it aims to evaluate their accuracy in modeling high-frequency duration dynamics, and to assess the contribution of zero-duration clusters and their size. Accordingly, both forecasts and loss functions are evaluated exclusively on diurnally adjusted durations.

3 Empirical facts on zero-duration clusters

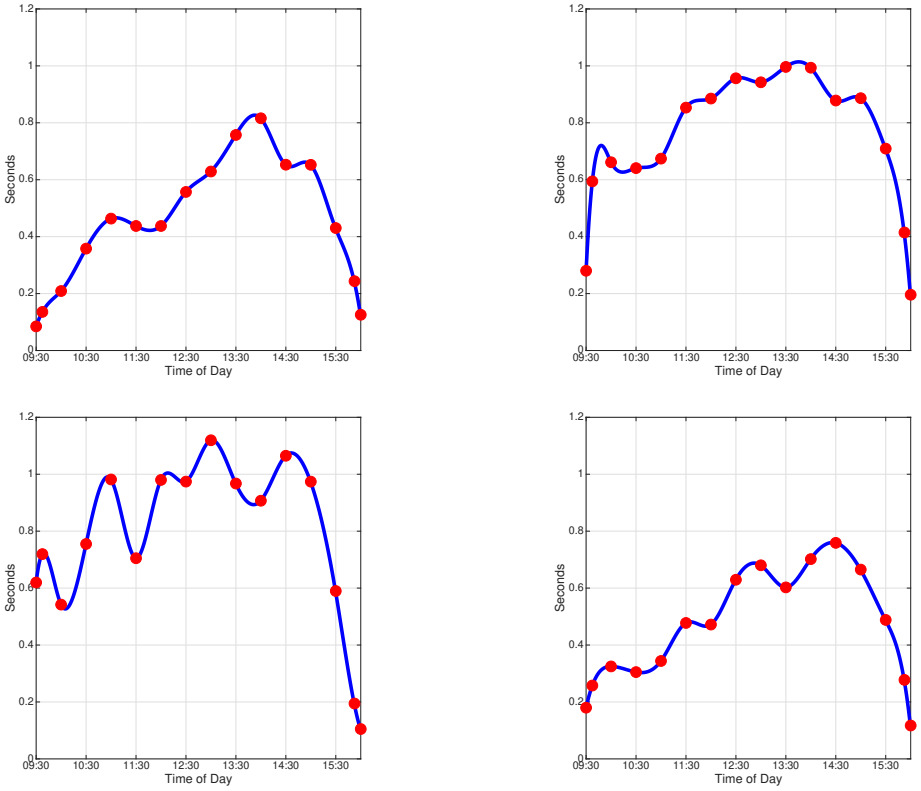
As shown in Figure 1 and Table 1, zero-duration clusters, *i.e.* groups of transactions recorded with the same timestamp, represent a sizable fraction of events included in our dataset. It is therefore important to understand how clustering affects duration dynamics. In this section, we aim to quantify the average effect of clusters on durations and examine whether this effect can be meaningfully summarized as a function of cluster size q_i. Beyond immediate impact, we also study the persistence of the cluster-induced distortion and whether it depends on the size of the initiating cluster. Overall, we document three empirical regularities in the behavior of durations around clusters, providing a detailed picture of how cluster size shapes both local temporal structure and the likelihood of further clustering.

3.1 Durations around clusters

We begin by examining whether durations in the proximity of clusters show systematically different dynamics. Specifically, we aim to measure the average duration h-steps before and after the occurrence of a cluster and to assess whether dynamics are related to the cluster size. Let q_i denote the size of event i. We classify clusters into four groups according to their size $q \in \mathbb{N}$:

$$G_2 = \{q : 2 \leq q < 10\}, \qquad G_3 = \{q : 10 \leq q < 40\}, \qquad G_4 = \{q : q \geq 40\},$$

Figure 2: Estimated intraday diurnal components for selected trading days



Notes Each panel plots the estimated diurnal component for a set of selected trading days for Boeing (BA). The dots denote the interval-specific average durations used as spline nodes, while the solid lines represent the corresponding fitted cubic splines.

and

$$G_1 = G_2 \cup G_3 \cup G_4 = \{q : q \geq 2\}.$$

The thresholds are chosen based on the empirical evidence presented in the subsequent analysis. We define the expected duration h-steps before or after the occurrence of clusters as:

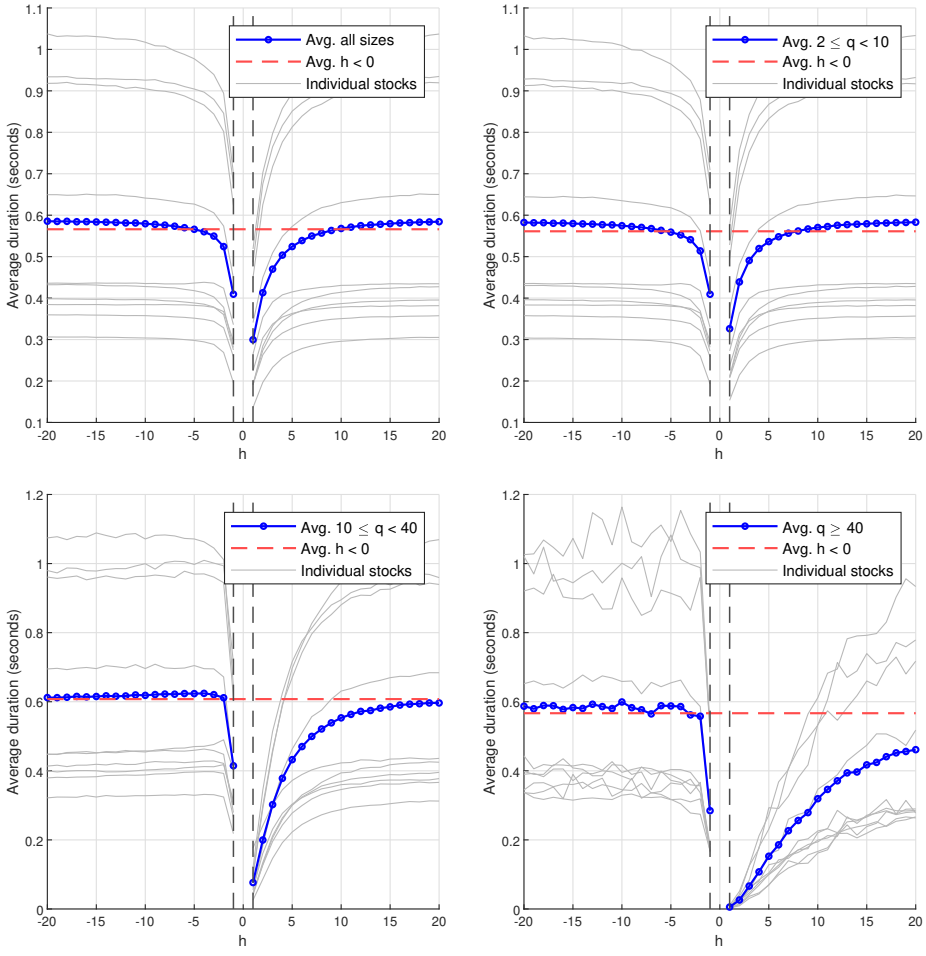
$$f(h; G_k) = \begin{cases} \mathbb{E}\left[x_i \mid q_{i-h-1} \in G_k\right], & \text{for } h = -h_{\max}, \ldots, -1, \\ \mathbb{E}\left[x_i \mid q_{i-h} \in G_k\right], & \text{for } h = 1, \ldots, h_{\max}. \end{cases} \tag{2}$$

When h is negative, we are looking $|h|$-steps before clusters, while for positive h we are looking h-steps after clusters. Given a sample of size $N \gg h_{\max}$, eq. (2)

is estimated as follows:

$$\hat{f}(h; G_k) = \begin{cases} \dfrac{\sum_{i=1}^{N+h+1} x_i \cdot \mathbb{1}\{q_{i-h-1} \in G_k\}}{\sum_{i=1}^{N+h+1} \mathbb{1}\{q_{i-h-1} \in G_k\}}, & \text{for } h = -h_{\max}, \ldots, -1, \\ \dfrac{\sum_{i=h+1}^{N} x_i \cdot \mathbb{1}\{q_{i-h} \in G_k\}}{\sum_{i=h+1}^{N} \mathbb{1}\{q_{i-h} \in G_k\}}, & \text{for } h = 1, \ldots, h_{\max}. \end{cases} \tag{3}$$

Figure 3: Conditional average duration around clusters



Notes Each panel plots the average duration across all stocks h steps before or after the occurrence of clusters ($h = 0$). The gray lines are the average durations for individual stocks. The red dashed lines represent the unconditional mean of durations computed before clusters are encountered (i.e., over $h < 0$).

Figure 3 shows the average duration h-steps around clusters, separately for different cluster size groups. In all cases, the average duration drops immediately after a cluster and only gradually reverts to its unconditional mean. This reversion is clearly size-dependent: larger clusters are followed by longer-lasting deviations from the unconditional mean. With clusters accounting for roughly 20% of events, local bursts are a pervasive feature of the execution process. The observed patterns are consistent with modern algorithmic limit-order placement strategies in fragmented markets, as discussed by O'Hara (2015). In particular, an algorithm that wants to buy (sell) may post buy (sell) limit orders across multiple venues and book levels, partly to reduce execution costs and earn rebates, partly to hide its intention, and partly to preserve the option to cancel or revise orders dynamically. Once such liquidity has been supplied, it may be consumed rapidly when trading interest arrives, generating bursts of closely spaced executions, that is, clusters. Larger clusters are consistent with larger amounts of liquidity being posted near the prevailing quotes, where execution is more likely to occur quickly. The gradual reversion of durations after a cluster then reflects the progressive nature of the same liquidity-consumption episode: the most attractive liquidity is consumed first, while liquidity posted deeper in the book or on other venues may continue to be matched over subsequent event times, but at progressively lower intensity.

3.2 Cluster size and post-cluster duration

Having established that the dynamics of average durations around clusters are size-dependent, we now measure the average relationship between cluster size and the immediately following duration. Specifically, for events with $q_{i-1} \geq 2$, we study how the mean log duration $\log x_i$ varies with the antecedent cluster size q_{i-1}, and show that this relationship is well approximated by a polynomial in inverse cluster size. We estimate

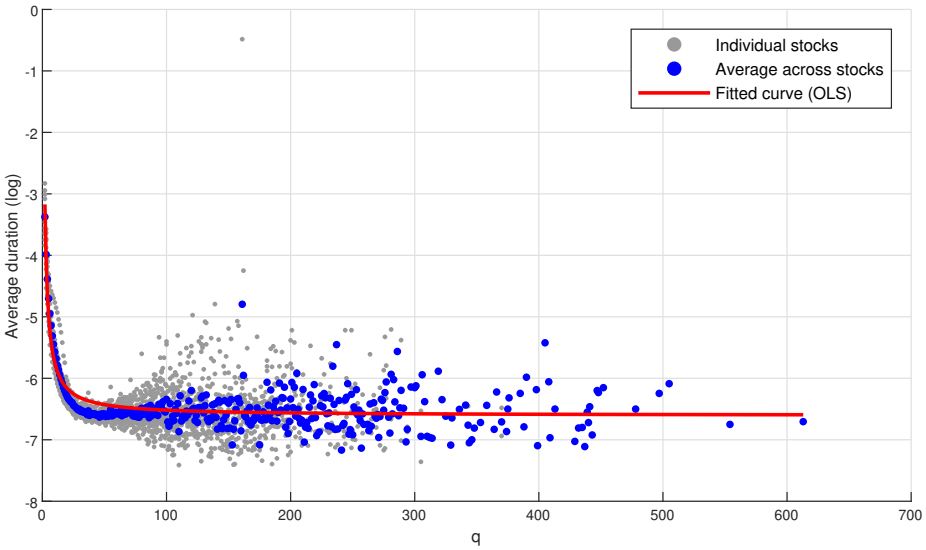
$$g(q) := \mathbb{E}\left[\log x_i \mid q_{i-1} = q\right] \tag{4}$$

with the empirical estimator

$$\hat{g}(q) = \frac{\sum_{i=2}^{N} \log x_i \cdot \mathbb{1}\{q_{i-1} = q\}}{\sum_{i=2}^{N} \mathbb{1}\{q_{i-1} = q\}}. \tag{5}$$

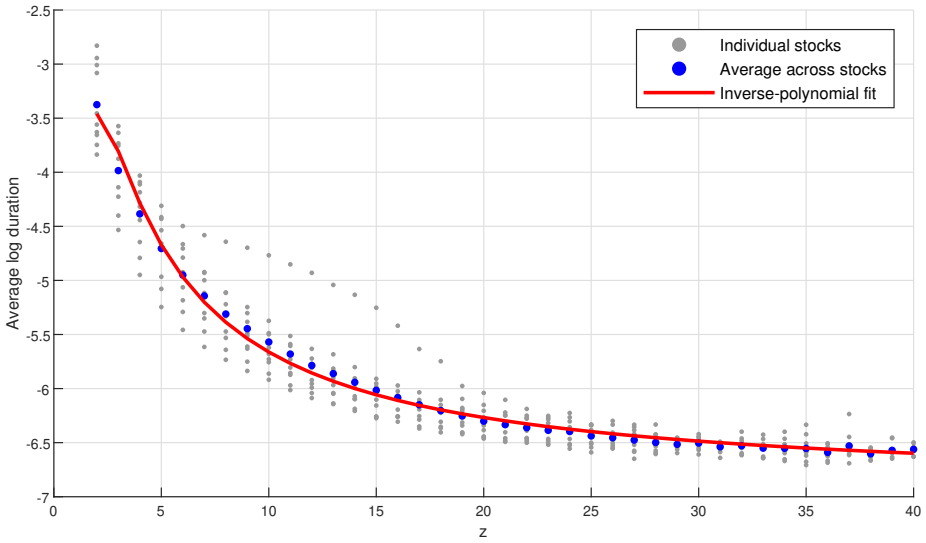
As shown in Figure 4, the variability of the estimates increases with q, since large clusters are rare in the sample (see Table 1). The average duration decays rapidly with cluster size and then stabilizes on a plateau from $q \approx 40$ onward. This plateau reflects the limited time resolution of the data: we measure the first observable duration immediately after a cluster, which is bounded below by one millisecond in the raw timestamps. After diurnal adjustment via the cubic spline, the estimated conditional means may lie slightly below this bound, but a plateau is expected nonetheless.

Figure 4: Average duration conditional on cluster size



Notes Average of log duration between τ_i and τ_{i-1} conditional on the size of the cluster observed at τ_{i-1}. Gray dots show individual stock averages, while blue dots show the average across stocks. The red curve is the OLS estimate of eq. (6) on q_i instead of z_i.

Figure 5: Average log duration and inverse-polynomial fit on thresholded cluster size



Notes: Inverse-polynomial fit of average log duration after a cluster conditional on thresholded size. Gray dots show individual stock averages, while the blue dots show the average across all stocks.

To model this relationship in a tractable way—while accounting for the rare occurrence of very large clusters—we use a thresholded size variable:

$$z_i := \min(q_i, 40),$$

so that any cluster larger than 40 is treated as having size 40. Using z_i instead of q_i substantially improves the fit, as shown in Figure 5.[2]

A natural prior is that duration increases with the inverse of the antecedent cluster size, since $1/q$ roughly corresponds to the average duration of trades within the preceding millisecond interval. We therefore adopt the following simple and interpretable specification for the conditional expectation of log duration:

$$\mathbb{E}[\log x_i \mid z_{i-1}] = b_0 + b_1 z_{i-1}^{-1} + b_2 z_{i-1}^{-2}. \tag{6}$$

Other functional forms are possible, but our aim is not to compare specifications; rather, it is to provide a parsimonious form that captures how durations depend on the size of the previous cluster, potentially in a stock-specific way.

If clusters are the visible outcome of bursts of executions generated by strategically placed limit orders across venues and book levels, then cluster size proxies for the intensity of the underlying execution episode. The inverse relationship is then intuitive: when more transactions share a timestamp, the average latent time per execution within that burst must be smaller, so inverse powers of cluster size naturally capture the effect on the subsequent duration.

3.3 Persistence of clusters

The third fundamental empirical finding concerns the temporal persistence of clustering: observing a cluster at τ_{i-1} increases the probability of observing a subsequent cluster at τ_i, as shown in Figure 6. This persistence is strongly size-dependent: the larger the initial cluster, the higher the likelihood of observing further clustered executions.

We measure the probability that event $i + h$ is a cluster (of whatever size) conditional on event i being a cluster of size G_k:

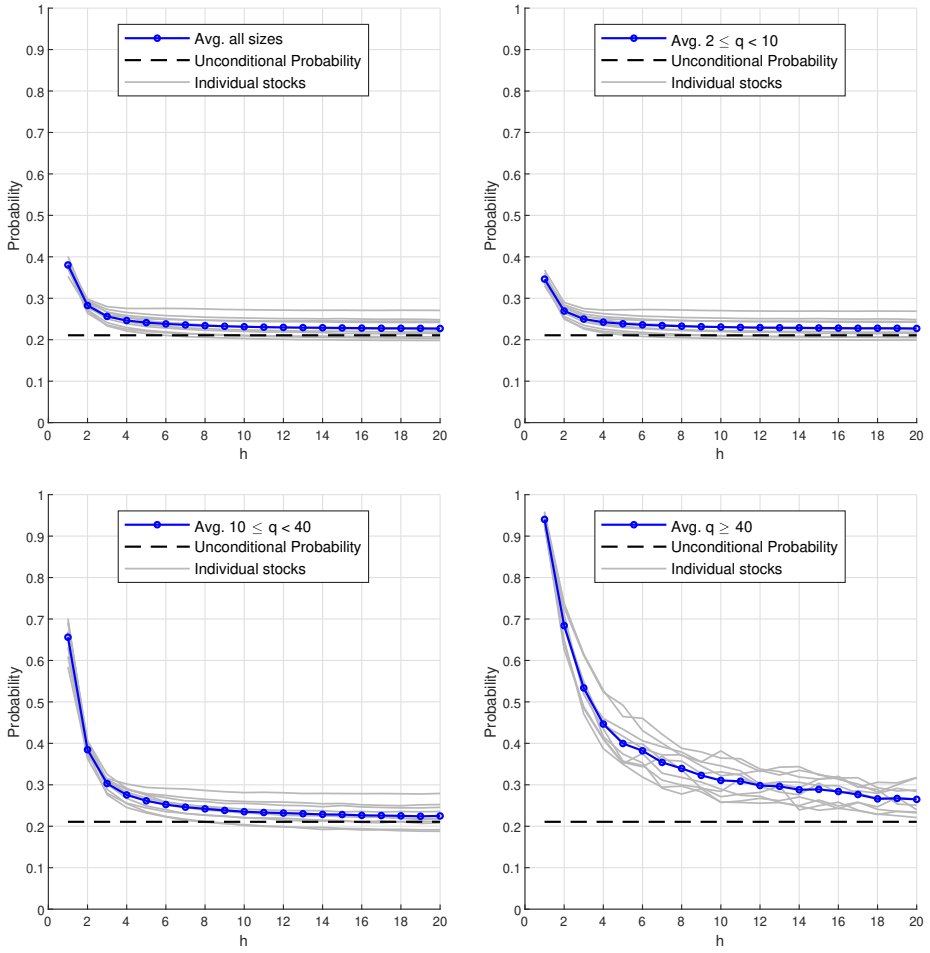
$$p(h; G_k) := \mathbb{P}\left(q_{i+h} \geq 2 \mid q_i \in G_k\right), \quad \text{for } h = 1, 2, ..., h_{\max}. \tag{7}$$

The sample counterpart is:

$$\hat{p}(h; G_k) = \frac{\sum_{i=1}^{N-h} \mathbb{1}\{q_i \in G_k,\ q_{i+h} \geq 2\}}{\sum_{i=1}^{N-h} \mathbb{1}\{q_i \in G_k\}}. \tag{8}$$

[2]Estimation results for the polynomial regressions fitted separately for each stock are reported in Appendix B. Fits obtained when capping q_i at cutoffs below 40 are worse than those obtained using z_i (cap at 40). Additional results are available upon request.

Figure 6: Persistence of clusters



Notes: Probability of observing a cluster at τ_{i+h}, given that a cluster of size $q_i \in G_k$ occurred at time τ_i.

The results show that clustering is not only persistent, but that such persistence is strongly size dependent. Larger clusters are substantially more likely to be followed by further clusters, and the decay of this effect in event time is markedly slower. This size-dependent persistence is fully coherent with the patterns observed in Figures 3 and 4: after a large cluster, we expect not only a shorter duration, but also a higher likelihood of further clusters, contributing to the slower reversion to the unconditional mean. This persistence also helps explain the pre-cluster decline in the average duration observed in Figure 3.[3] This

[3]Since clusters tend to occur in sequences, the durations preceding a cluster may already be affected by an earlier cluster. As durations drop after each cluster, this can result in the apparent pre-cluster contraction observed in the average duration.

evidence is again consistent with the functioning of modern fragmented markets, as discussed by O'Hara (2015). When liquidity is supplied through limit orders strategically posted across venues and book levels, executions may occur in runs rather than as isolated trades. A cluster then represents one observable realization of this broader execution process: part of the posted liquidity may be matched immediately at or near the best quotes, while additional liquidity may remain available at nearby prices or on other venues and be executed shortly afterward. The size of the initial cluster and the persistence of subsequent clustering therefore reflect both the unobserved scale of the underlying liquidity-provision episode and the way liquidity is strategically posted across the market. In this sense, duration dynamics and clustering persistence can be interpreted as event-time measures of liquidity consumption and replenishment.

4 Econometric specifications for the conditional mean of durations

We propose a duration model that incorporates the empirical regularities documented in the previous sections. Accordingly, our econometric specifications focus on the conditional mean of durations, using past durations and cluster information as the relevant state variables. To assess the incremental contribution of cluster size, we also consider a restricted variant in which the size component is excluded.

We additionally consider standard specifications as benchmarks. First, we include a baseline logACD model, which can equivalently be viewed as a further restriction in which all cluster-related parameters are set to zero. Second, we consider a standard ACD specification, which imposes positivity constraints on the parameters.

Finally, both our full and restricted models are augmented by including the score of the most recent observation as an additional explanatory variable in the recursion.

4.1 logACD with zero-duration cluster size

Our proposed model extends the logACD specification of Bauwens and Giot (2000), the duration analogue of the Log-GARCH model. A key motivation for working with a recursion in log durations is that it facilitates the inclusion of additional explanatory variables without imposing positivity constraints on the corresponding parameters.

Building on the empirical results in the previous sections, we augment the logACD model with an indicator for the presence of clusters. When active, this indicator adds an auxiliary ACD component that incorporates a linear parametrization of the thresholded cluster size z_{i-1}. We specify the logACD process for the adjusted duration x_i as

$$x_i = e^{\phi_i}\epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Exp}(1), \tag{9}$$

where ϕ_i evolves according to

$$\phi_i = \omega + \alpha\tilde{x}_{i-1} + \beta\phi_{i-1} + \mathbb{1}_{\{z_{i-1}\geq 2\}}\left(\gamma_1 + \gamma_2\tilde{x}_{i-1} + \gamma_3\phi_{i-1} + \boldsymbol{\delta}^\top \mathbf{z}_{i-1}\right), \quad (10)$$

with $\tilde{x}_i := \log x_i$. We refer to this specification as model M1.

In line with the evidence above, our baseline choice for the size component is a second-degree polynomial in the inverse of the thresholded cluster size:

$$\boldsymbol{\delta}^\top \mathbf{z}_{i-1} = \delta_1 z_{i-1}^{-1} + \delta_2 z_{i-1}^{-2}, \quad (11)$$

Other functional forms could be considered; however, our goal is not to select the best specification, but rather to show that cluster size is a fundamental driver of durations.

4.2 Score-driven model

In the spirit of score-driven models, we retain the same multiplicative specification as in the logACD model:

$$x_i = e^{\phi_i}\epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \text{Exp}(1). \quad (12)$$

For the baseline GAS specification, we let ϕ_i evolve according to

$$\phi_i = \omega + \alpha s_{i-1} + \beta\phi_{i-1}, \quad S_i = -1 + \frac{x_i}{e^{\phi_i}}, \quad s_i = r\tanh\left(\frac{S_i}{r}\right). \quad (13)$$

Details on the GAS recursion, the score, and the likelihood function are provided in Appendix C. The score S_i inherits the distribution of the multiplicative error ϵ_i, shifted to the support $(-1, \infty)$. Owing to the multiplicative structure, the ratio x_i/e^{ϕ_i} appears naturally; as a result, even modest underpredictions can generate large values of S_i and, in turn, overly aggressive updates.

To mitigate this issue, we apply a hyperbolic tangent transformation to S_i, so that $s_i \approx S_i$ for moderate values while the scaled score remains bounded in $(-1, r)$. This bound eliminates the possibility of explosive updates. We set $r = 6$, which progressively shrinks large scores with a noticeable effect above roughly 3 (about 5% of observations in our sample), although this percentage varies across days and stocks.

4.3 Other specifications

We also consider the following specifications, summarized in Table 2.

- **Restricted cluster model** (M2). We restrict M1 by setting $\boldsymbol{\delta} = \mathbf{0}$, so the size is omitted. This isolates the contribution of cluster size relative to the baseline indicators.
- **logACD**. By setting $\gamma_1 = \gamma_2 = \gamma_3 = 0$ and $\boldsymbol{\delta} = \mathbf{0}$ in (10) we obtain the standard logACD of Bauwens and Giot (2000).

- **ACD in levels**. The standard benchmark model where durations follow

$$x_i = \psi_i \epsilon_i, \quad \epsilon_i \sim \text{Exp}(1), \tag{14}$$

with conditional mean specified as

$$\psi_i = \omega + \alpha x_{i-1} + \beta \psi_{i-1}, \quad \omega > 0,\ \alpha \geq 0,\ \beta \geq 0. \tag{15}$$

- **Score-driven extensions** (M1s and M2s). Extensions of both M1 and M2, where recursions for ϕ_i additionally include s_{i-1}, the bounded score defined in (13), with coefficient ζ. For M1s,

$$\phi_i = \omega + \alpha \tilde{x}_{i-1} + \beta \phi_{i-1} + \zeta s_{i-1} + \mathbb{1}_{\{z_{i-1} \geq 2\}} \left(\gamma_1 + \gamma_2 \tilde{x}_{i-1} + \gamma_3 \phi_{i-1} + \boldsymbol{\delta}^\top \mathbf{z}_{i-1}\right), \tag{16}$$

and M2s is obtained by setting $\boldsymbol{\delta} = \mathbf{0}$.

Table 2: Summary of model specifications

Name	Cluster indicators	Size	Score
ACD	—	—	—
logACD	—	—	—
GAS	—	—	✓
M1	✓	✓	—
M2	✓	—	—
M1s	✓	✓	✓
M2s	✓	—	✓

Notes: This table summarizes the main features of each model.

5 In-sample estimation

The exponential log-likelihood function for a sample of N observations can be written (see Engle and Russell, 1998 and Bauwens and Giot, 2000) as:

$$\mathcal{L}_N(\theta) = -\sum_{i=1}^{N} \left[\phi_i(\theta) + \frac{x_i}{e^{\phi_i(\theta)}}\right], \tag{17}$$

where θ denotes the vector of parameters to be estimated and

$$\hat{\theta}_N = \arg\max_{\theta} \mathcal{L}_N(\theta)$$

denotes the maximum likelihood estimator of θ. Writing the log-likelihood as in (17) implicitly treats N as fixed, whereas the number of events over a fixed time span $[0, T]$, i.e. $N(T)$, is random.

Cavaliere et al. (2024, 2025) note that, when the number of observed events is random, the asymptotic theory of the maximum likelihood estimator of θ in ACD models depends on the tail behavior and the existence of unconditional moments of the ergodic and stationary durations $\{x_i\}$. Tail behavior is summarized by the tail index $\kappa > 0$ of the marginal distribution.

In particular, asymptotic normality holds only when the tail index satisfies $\kappa > 1$. When $\kappa < 2$, the Gaussian approximation can be unreliable in finite samples, and it breaks down entirely in the infinite-mean case $\kappa < 1$. Cavaliere et al. (2025, Theorem 2) show that asymptotic normality for the standard linear ACD(1,1) model can be obtained by imposing the additional condition $\alpha+\beta < 1$, which implies $\mu = E[x_i] < \infty$. A finite mean implies $N(T)/T \xrightarrow{\text{a.s.}} 1/\mu$, which is a sufficient condition for the law of large numbers for the score and the information matrix to hold in a standard way even when $N(T)$ is random.

To assess the empirical relevance of this issue for our models, which are not reconcilable to the linear ACD framework, we conduct a preliminary tail analysis using Hill diagnostics. The results are reported in Appendix D. While tail index estimation in dependent duration data should be interpreted cautiously, our estimates lie well above the critical thresholds highlighted by the theory, suggesting that the pathological cases are unlikely to be driving the in-sample results.

Table 3 and Figure 7 summarize the daily in-sample estimates across stocks for our main model (M1) and the baseline ACD. Both models are estimated independently for each stock and trading day. This is feasible because each daily series contains a large number of observations (see Table 1). The autoregressive parameters (α and β) are significant across all days and stocks, confirming stable ACD dynamics. Relative to the standard ACD in levels, the estimated baseline persistence ($\alpha + \beta$) is markedly lower (0.47 vs. 0.95), consistent with temporal dependence being partly captured by the additional regressors.

The intercept shift for the presence of clusters (γ_1) is negative and significant in 99% of days, consistent with durations shortening immediately after a cluster. The additional coefficients γ_2 and γ_3 are also significant on a substantial fraction of days (85% and 34%, respectively), although not uniformly.

Most importantly, the size–related terms (δ_1, δ_2) are consistently significant across days and stocks (98–99%), confirming that the conditional mean duration varies systematically with cluster size. Based on the average estimates in Table 3, the size component

$$c(z) = \delta_1 z^{-1} + \delta_2 z^{-2}$$

takes values $c(2) \approx 2.19$ and $c(40) \approx 0.26$. Hence, on average,

$$\frac{\mathbb{E}[x_i \mid z_{i-1} = 2]}{\mathbb{E}[x_i \mid z_{i-1} = 40]} \;=\; \exp\big(c(2) - c(40)\big) \;\approx\; 7,$$

implying that the expected duration following a cluster of size $q_{i-1} \geq 40$ is roughly seven times smaller than after a cluster of size $q_{i-1} = 2$.

Table 3: Daily in-sample estimates summary

(a) M1

Parameter	Mean	Std. Dev.	5th Pct	95th Pct	% Significant
ω	-0.1359	0.2786	-0.5445	0.3090	96.16
α	0.0828	0.0110	0.0660	0.1008	100.00
β	0.3838	0.0677	0.2747	0.4957	100.00
γ_1	-2.5871	0.7329	-3.5957	-1.0907	99.36
γ_2	-0.0377	0.0214	-0.0678	0.0062	85.04
γ_3	0.0684	0.1136	-0.1087	0.2623	34.16
δ_1	10.7400	3.2567	4.6288	15.4500	98.48
δ_2	-12.7070	4.1335	-18.7930	-5.1589	97.92

(b) ACD

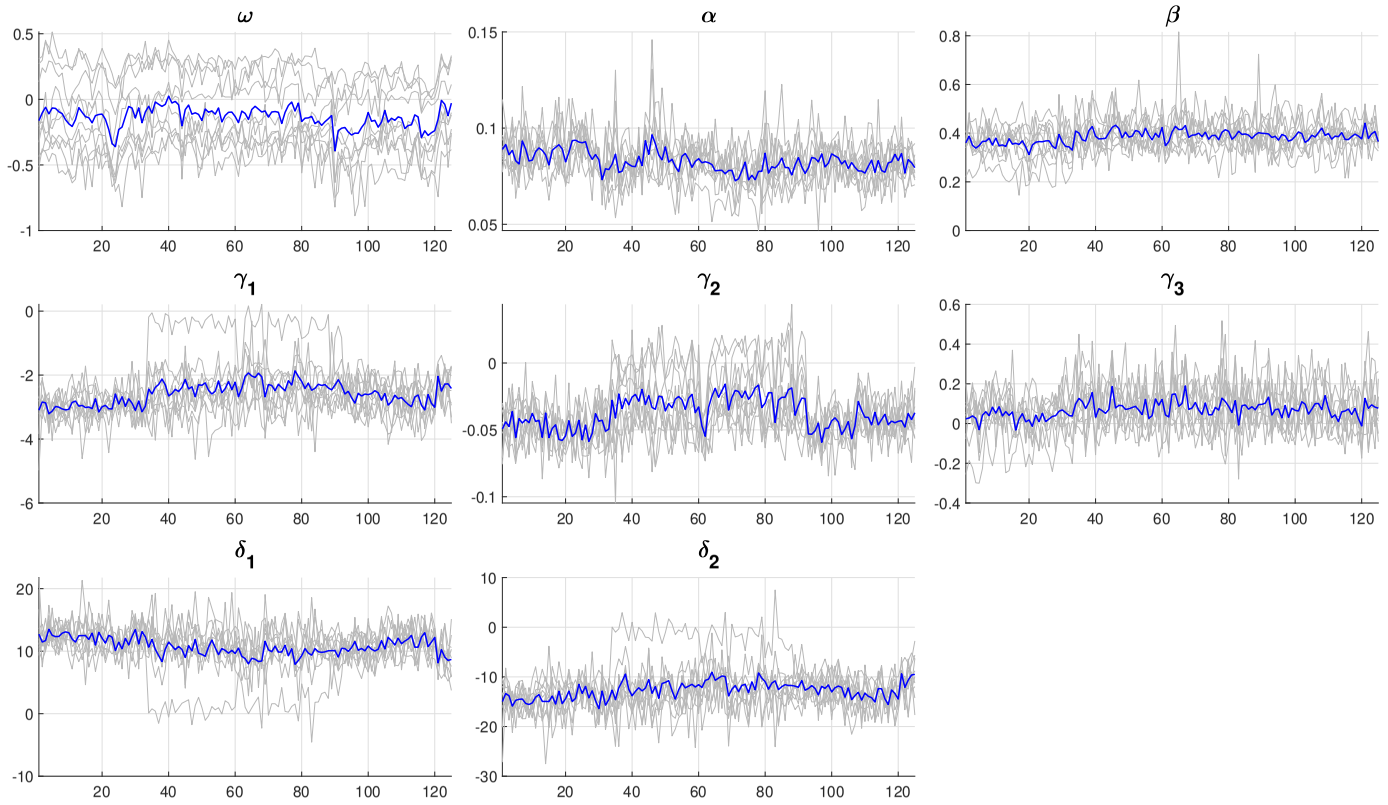
Parameter	Mean	Std. Dev.	5th Pct	95th Pct	% Significant
ω	0.0328	0.0403	0.0001	0.1137	73.28
α	0.0982	0.0505	0.0308	0.1845	99.60
β	0.8537	0.0988	0.6906	0.9688	100.00

Notes: Summary statistics of the in-sample parameter estimates. Panel (a) shows the results for the model including both cluster indicators and cluster size (M1), while Panel (b) shows the results for the standard ACD model. For each parameter, the mean, standard deviation, the 5th percentile, and 95th percentile of the estimated values are provided, aggregated across all days and all stocks. The last column shows the percentage of estimates that are statistically significant at the 5% level.

For one stock (JNJ), the estimated δ_1 and δ_2 imply a different shape for the size component than for the other stocks, although the qualitative pattern remains consistent with the results in Section 3.2. This idiosyncrasy is unsurprising given that, for JNJ, the empirical relationship between cluster size and subsequent duration already exhibits a distinctive shape and the lowest R^2 among all stocks.

Table 4 reports the results of daily likelihood-ratio (LR) tests conducted separately for each day and stock. These tests compare nested specifications in order to evaluate the incremental contribution of cluster size and the score in terms of log-likelihood gains. The unrestricted specification (M1) yields statistically significant improvements for virtually all days. Its score-driven extension (M1s) delivers further statistically significant improvements across almost the entire sample. The reported average LR statistics are extremely large, as expected given the high number of observations available per trading day (see Table 1).

Figure 7: Daily parameter estimates across stocks (M1)



Notes The figure reports a summary of the in-sample parameter estimates for the model including both cluster indicators and cluster size (M1). Gray lines represent estimates for each individual stock; the blue line shows the average value across stocks.

Table 4: Summary of likelihood-ratio tests

Unrestricted	Restricted	% Reject	Avg. stat.
M1	M2	97.4	2707.0
M1	logACD	97.7	5960.7
M1s	M1	95.5	206.1
M1s	M2s	97.6	2665.8

Notes: The table reports likelihood ratio (LR) test statistics. For each stock and trading day, the unrestricted model is tested against the restricted model under the null. The table shows the percentage of rejections at the 5% significance level and the average LR statistic computed across all days and stocks.

6 Out-of-sample evaluation

We adopt two forecasting perspectives to assess whether the information contained in zero-duration clusters and, in particular, in their size, has persistent and predictable effects on subsequent positive duration dynamics. For each stock and each trading day, parameters are estimated using the first half of the observations, and one-step-ahead forecasts are produced for the second half. This setup accommodates the variation in the distribution of durations across trading sessions, due to the natural variability in trading activity. Consistent with the discussion in section 2, we adjust diurnal seasonality separately for each session, since its shape need not be stable across days or stocks. For completeness, we also implement a *three-day rolling* scheme, in which parameters are estimated over three consecutive days and forecasts are computed over the subsequent session.

We compare predictive daily performance using the Model Confidence Set (MCS) procedure of Hansen et al. (2011) and multiple Diebold Mariano (DM) tests for selected model pairs, conducting both procedures separately for each forecast day. Forecast errors are evaluated using both MSE and QLIKE loss functions. For the MCS, we summarize results by stock using the *frequency of inclusion* in the set of superior models, that is, the fraction of days on which a model remains in the superior set. For the DM tests, we summarize results by averaging across days and stocks. We report the percentages of wins, losses, and ties together with the DM statistic averaged across all evaluations.

Results are reported in Tables 5 and 6. The MCS results show that the two specifications including the size component (M1 and M1s) are by far those most frequently retained in the set of superior models, confirming the central role of size. By contrast, whether adding the score as an additional regressor is beneficial depends on the loss function considered: under the MSE, M1 is more frequently retained, whereas under QLIKE the score-extended specification M1s attains a higher inclusion rate.

The DM tests corroborate these findings. Under MSE, M1 outperforms its score-augmented counterpart M1s. Under QLIKE, by contrast, including the

Table 5: Model Confidence Set inclusion percentages by stock (within-session forecasting)

(a) MSE loss function

Model	BA	CAT	CVX	DIS	HON	JPM	MCD	WMT	JNJ	AMGN
M1s	20.8	31.2	43.2	37.6	52.8	24.0	12.0	21.6	28.0	35.2
M1	**97.6**	**97.6**	**90.4**	**99.2**	**94.4**	**97.6**	**96.0**	**96.8**	**76.0**	**95.2**
M2s	0.0	4.8	1.6	1.6	4.8	0.8	1.6	0.8	4.0	8.8
M2	0.8	13.6	0.0	0.8	8.0	12.0	6.4	1.6	38.4	12.0
GAS	0.0	0.8	0.0	0.0	0.8	0.8	0.0	0.0	0.8	1.6
logACD	0.0	0.8	0.0	0.0	0.0	0.0	2.4	0.0	2.4	0.0
ACD	0.0	1.6	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.8

(b) QLIKE loss function

Model	BA	CAT	CVX	DIS	HON	JPM	MCD	WMT	JNJ	AMGN
M1s	**96.0**	**95.2**	**82.4**	**95.2**	**96.0**	**76.8**	54.4	**88.0**	64.0	**94.4**
M1	68.0	66.4	74.4	71.2	53.6	65.6	**85.6**	59.2	**72.8**	56.0
M2s	6.4	20.0	6.4	4.0	24.8	4.0	9.6	1.6	24.0	26.4
M2	6.4	15.2	3.2	4.0	16.8	2.4	14.4	2.4	30.4	16.8
GAS	0.0	2.4	1.6	0.0	3.2	0.8	4.8	0.0	12.0	4.0
logACD	0.8	1.6	0.8	0.8	0.8	0.0	4.8	0.0	5.6	1.6
ACD	0.0	1.6	1.6	0.0	0.0	0.8	3.2	0.0	6.4	2.4

Notes: The tables report, for each stock, the percentage of days in which a given model is included in the 95% model confidence set. For each trading session, the first half of the observations is used for parameter estimation, and the second half for forecasting. Panel (a) is based on the MSE loss function, and Panel (b) on the QLIKE loss function.

score improves accuracy for a moderately larger fraction of days, with an average DM statistic around 0.5. Across all pairwise comparisons, under both MSE and QLIKE, models that omit the size component perform markedly worse, confirming that size is a key driver of predictive performance.

The three-day rolling evaluation (reported in Appendix E) leads to the same conclusion regarding the importance of size, while the relative performance of the score-augmented specification M1s improves. Although M1 remains the most frequently included in the superior set under MSE, the gap between M1 and M1s narrows considerably. Under QLIKE, M1s—already dominant in the within-session analysis—achieves a substantially higher inclusion frequency than the model without the score. The DM tests reinforce this evidence, indicating that including the score provides a mild improvement under MSE and a clear advantage under QLIKE.

This improvement of score-augmented models under the three-day rolling scheme is not surprising. As discussed above, the distribution of durations varies across days and stocks, so estimating parameters over multiple days yields estimates that are closer to an "averaged" representation of duration dynamics.

Table 6: Diebold–Mariano tests summary (within-session forecasting)

(a) MSE loss function

Competitor	Baseline	%Win	%Loss	%Tie	Avg. stat.
M1	M2	90.0	2.4	7.6	-6.6
M1	ACD	98.9	0.6	0.5	-14.8
M1	M1s	68.7	3.4	27.8	-4.1
M2s	M1s	5.4	84.4	10.2	6.4

(b) QLIKE loss function

Competitor	Baseline	%Win	%Loss	%Tie	Avg. stat.
M1	M2	86.3	0.6	13.0	-5.5
M1	ACD	95.2	1.0	3.8	-9.4
M1	M1s	15.4	32.6	51.9	0.5
M2s	M1s	0.7	84.0	15.3	5.3

Notes: The tables report the percentage of Diebold–Mariano tests in which *competitor* outperforms (%Win), underperforms (%Loss), or ties (%Tie) *baseline* at the 5% significance level. The loss differential is evaluated as the loss of *competitor* minus the loss of *baseline*. The average DM test statistic across all days and stocks is also reported. For each trading session, the first half of the observations is used for parameter estimation, and the second half for forecasting. Panel (a) is based on the MSE loss function, and Panel (b) on the QLIKE loss function.

While the longer estimation window can improve stability, it may reduce forecast precision for the subsequent session relative to the within-session setup. In this context, the score appears to add incremental information beyond the size component.

7 Conclusion

This paper offers a new perspective on modeling intraday durations by focusing on zero-duration clusters—groups of transactions that share the same timestamp—in ultra-high-frequency data. In modern electronic markets, where trading is fragmented across venues and shaped by algorithmic interaction, such clusters should not be viewed merely as artifacts of finite timestamp precision. Instead, they may signal bursts in the latent execution intensity underlying transaction-level activity.

Using a comprehensive dataset of ten highly liquid large-cap stocks, we show that 50–60% of all transactions share the same timestamp. This finding calls for a reconsideration of how durations are modeled and interpreted in ultra-high-frequency settings. It is relevant not only for econometric modeling, but also for understanding trading behavior, since both high-frequency and non-high-frequency traders operate in environments where short-run execution intensity is driven by algorithmic interaction, liquidity consumption, and rapid responses

to order-flow conditions.

Our contribution is twofold. First, we document how zero-duration clusters shape duration dynamics and establish three empirical regularities: durations tend to decrease before a cluster and increase afterward; larger clusters are associated with shorter subsequent durations; and clustering is persistent over time, especially for large clusters.

Second, we propose a modeling approach within the Autoregressive Conditional Duration (ACD) framework that incorporates *cluster size* as a key determinant of expected durations and, in an extended specification, adds a *score* component to capture residual dynamics beyond what size explains.

The relevance of this modeling approach is confirmed through extensive out-of-sample testing. Specifications that include the size component systematically outperform alternative models across stocks, days, and evaluation settings, indicating that execution bursts have persistent and predictable effects on subsequent duration dynamics. Accordingly, zero-duration clusters are not merely a data-processing complication; they are a compact and informative proxy for short-run execution intensity. This makes them useful for trading-strategy analysis and for market monitoring, where identifying unusually intense and persistent execution episodes is key to understanding liquidity dynamics and algorithmic market activity.

A Rounding property

In this appendix, we provide a proof of result (1), which states that the observed duration is always either the floor or ceiling of the latent duration, i.e.,

$$d_j \in \{\lfloor \check{d}_j \rfloor, \lceil \check{d}_j \rceil\}.$$

Proof. Recall that the rounding function is defined as

$$t_j := R(\check{t}_j) = \lfloor \check{t}_j \rfloor + e_j, \quad e_j := \mathbb{1}\left(\{\check{t}_j\} > 0 \text{ and } \{\check{t}_j\} \geq \varrho\right), \quad \{\check{t}_j\} := \check{t}_j - \lfloor \check{t}_j \rfloor,$$

where $\varrho \in [0,1]$. The latent and the observed durations are, respectively, $\check{d}_j = \check{t}_j - \check{t}_{j-1}$ and $d_j = t_j - t_{j-1}$. Then

$$d_j - \check{d}_j = (t_j - \check{t}_j) - (t_{j-1} - \check{t}_{j-1}) = (e_j - \{\check{t}_j\}) - (e_{j-1} - \{\check{t}_{j-1}\}),$$

where the last equality follows from the fact that

$$t_j - \check{t}_j = (\lfloor \check{t}_j \rfloor + e_j) - (\lfloor \check{t}_j \rfloor + \{\check{t}_j\}).$$

Observe that for any k,

$$e_k - \{\check{t}_k\} \in \begin{cases} [0,1) & \text{if } \varrho = 0, \\ (-\varrho, 1-\varrho] & \text{if } \varrho \in (0,1]. \end{cases}$$

Hence, the difference of two such terms satisfies

$$-1 < d_j - \check{d}_j < 1,$$

independently of ϱ. Since $d_j \in \mathbb{N}_0$, the only integer values compatible with this inequality are $\lfloor \check{d}_j \rfloor$ and $\lceil \check{d}_j \rceil$. Therefore,

$$d_j \in \{\lfloor \check{d}_j \rfloor, \lceil \check{d}_j \rceil\},$$

as claimed. □

B Polynomial regression estimates

Table 7: Polynomial regression estimates

Stock	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	Adj. R^2 (%)
All Stocks	-6.95 (0.02)	14.31 (0.31)	-14.68 (0.67)	99.46
BA	-6.95 (0.01)	11.37 (0.22)	-10.10 (0.48)	99.63
CAT	-6.95 (0.02)	15.12 (0.40)	-14.81 (0.88)	99.21
CVX	-6.92 (0.02)	14.32 (0.36)	-15.37 (0.78)	99.21
DIS	-6.87 (0.01)	11.78 (0.20)	-10.79 (0.43)	99.72
HON	-6.94 (0.03)	16.07 (0.42)	-16.58 (0.91)	99.21
JPM	-6.99 (0.02)	14.26 (0.40)	-15.42 (0.87)	99.01
MCD	-7.05 (0.02)	13.91 (0.33)	-11.93 (0.71)	99.48
WMT	-6.90 (0.01)	9.94 (0.20)	-7.77 (0.44)	99.64
JNJ	-6.93 (0.09)	20.24 (1.51)	-28.01 (3.29)	90.13
AMGN	-6.98 (0.02)	16.13 (0.33)	-15.99 (0.72)	99.52

Notes: Estimates are obtained from the polynomial regression in eq. (6). Coefficients reported in the first row are obtained from the regression of the pooled average log durations across stocks.

C Log-likelihood and score functions for the logACD models

Consider the model $x_i = e^{\phi_i}\epsilon_i$, where ϵ_i's are i.i.d. and follow an Exponential distribution with mean 1 and density

$$p(\epsilon_i) = \exp(-\epsilon_i), \quad \epsilon_i > 0.$$

By a change of variable, we obtain the density of the model:

$$p(x_i \mid \phi_i) = \frac{1}{e^{\phi_i}} \exp\left(-\frac{x_i}{e^{\phi_i}}\right). \tag{18}$$

Taking logarithms yields the log-density:

$$\log p(x_i \mid \phi_i) = -\phi_i - \frac{x_i}{e^{\phi_i}}. \tag{19}$$

Since in the GAS specification the time-varying parameter of interest is ϕ_i, the score is given by:

$$S_i := \frac{\partial \log p(x_i \mid \phi_i)}{\partial \phi_i} = -1 + \frac{x_i}{e^{\phi_i}}. \tag{20}$$

The updating recursion for the time-varying parameter becomes:

$$\phi_i = \omega + \alpha S_{i-1} + \beta \phi_{i-1}. \tag{21}$$

This is the version based on the raw (untransformed) score. In the empirical analysis, S_{i-1} is replaced by the bounded score $s_{i-1} = r\tanh(S_{i-1}/r)$, which coincides with the raw score as $r \to \infty$. The log-likelihood function for a sample of N observations is given by:

$$\mathcal{L}_N(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \phi_i), \tag{22}$$

where $\theta := (\omega, \alpha, \beta)'$ denotes the vector of static parameters that can be easily estimated by maximum likelihood.

D Hill estimates of the tail index

In this appendix we implement the Hill estimator for the tail index of the adjusted durations (Hill, 1975). Let $x_1, x_2, \ldots, x_N$ denote the adjusted durations for a given stock and trading day, and let $x_{(1)} \geq x_{(2)} \geq \cdots \geq x_{(N)}$ denote the corresponding order statistics sorted in descending order. For a given number k of upper order statistics, the estimator is

$$\hat{\kappa}_H(k) = \left(\frac{1}{k} \sum_{i=1}^{k} \log\left(\frac{x_{(i)}}{x_{(k+1)}} \right) \right)^{-1}.$$

Table 8: Hill tail index estimates

Stock	1%	2%	3%	5%	Min	Max
BA	4.573	3.947	3.564	3.092	2.641	5.152
CAT	4.619	4.012	3.638	3.147	2.862	5.433
CVX	4.551	3.911	3.542	3.077	2.597	5.114
DIS	4.600	3.977	3.605	3.156	2.885	5.232
HON	4.556	3.911	3.525	3.043	2.578	5.555
JPM	4.557	3.956	3.588	3.138	2.718	5.139
MCD	4.629	4.018	3.656	3.177	2.971	5.590
WMT	4.713	4.061	3.684	3.195	2.893	5.375
JNJ	4.600	3.949	3.574	3.123	2.780	5.631
AMGN	4.586	3.952	3.582	3.092	2.424	5.508

Notes The table reports Hill estimates of the tail index for the adjusted durations. For each stock and trading day, the estimator is computed using upper tail fractions equal to 1%, 2%, 3%, and 5% of the sample size. For each stock we report averages across trading days, together with the minimum and maximum estimate observed across all days and thresholds.

Table 8 presents Hill estimates of the tail index for the adjusted durations computed under multiple upper-tail fractions, namely $k = \lceil cN \rceil$ with

$c \in \{0.01, 0.02, 0.03, 0.05\}$. Thus, for each stock and day, the estimator uses the largest 1%, 2%, 3%, and 5% of adjusted durations. These results provide a preliminary diagnostic related to the concerns raised by Cavaliere et al. (2024). Although tail index estimation with dependent duration data is inherently challenging, the estimated values of κ are comfortably above the critical thresholds, supporting the applicability of asymptotic normality and Gaussian approximations in the present setting.

E Out-of-sample tests

Table 9: Model Confidence Set inclusion percentages by stock (three-day rolling forecasting

(a) MSE loss function

Model	BA	CAT	CVX	DIS	HON	JPM	MCD	WMT	JNJ	AMGN
M1s	50.8	50.8	51.6	45.1	59.0	54.1	52.5	45.9	38.5	51.6
M1	**52.5**	**74.6**	**77.0**	**72.1**	**69.7**	**70.5**	**75.4**	**77.9**	**55.7**	**82.0**
M2s	7.4	1.6	3.3	1.6	4.9	7.4	4.1	0.8	22.1	2.5
M2	0.0	0.0	0.8	0.0	0.0	3.3	0.0	0.0	34.4	0.8
GAS	3.3	2.5	0.8	4.1	0.8	2.5	0.8	2.5	4.9	3.3
logACD	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.8	0.0
ACD	11.5	9.0	6.6	6.6	5.7	7.4	4.9	4.1	9.8	4.1

(b) QLIKE loss function

Model	BA	CAT	CVX	DIS	HON	JPM	MCD	WMT	JNJ	AMGN
M1s	**88.5**	**96.7**	**92.6**	**92.6**	**96.7**	**91.8**	**98.4**	**98.4**	**81.1**	**94.3**
M1	30.3	35.2	41.8	31.1	36.9	23.0	36.9	29.5	34.4	31.1
M2s	4.1	8.2	4.9	0.8	8.2	2.5	15.6	3.3	23.0	10.7
M2	0.8	1.6	2.5	0.0	1.6	0.0	3.3	0.0	8.2	3.3
GAS	5.7	8.2	8.2	2.5	6.6	11.5	25.4	0.8	31.1	7.4
logACD	0.0	0.0	0.0	0.0	0.0	0.8	0.0	0.0	0.8	0.0
ACD	6.6	4.1	8.2	5.7	3.3	16.4	23.0	2.5	26.2	2.5

Notes The tables report, for each stock, the percentage of days in which a given model is included in the 95% model confidence set. Parameters are estimated over a window of three consecutive days, and forecasts are computed for the following day (three-days rolling forecasting). Panel (a) is based on the MSE loss function, and Panel (b) on the QLIKE loss function.

Table 10: Diebold–Mariano tests summary (three-day rolling forecasting)

(a) MSE loss function

Competitor	Baseline	%Win	%Loss	%Tie	Avg. stat.
M1	M2	90.3	5.4	4.3	-7.1
M1	ACD	89.8	7.4	2.8	-11.0
M1	M1s	42.5	26.9	30.6	1.4
M2s	M1s	6.8	87.6	5.6	6.6

(b) QLIKE loss function

Competitor	Baseline	%Win	%Loss	%Tie	Avg. stat.
M1	M2	92	1.0	7.0	-7.2
M1	ACD	85.7	4.7	9.6	-7.8
M1	M1s	3.3	67.0	29.7	5.1
M2s	M1s	1.0	91.4	7.6	6.9

Notes: The tables report the percentage of Diebold–Mariano tests in which *competitor* outperforms (%Win), underperforms (%Loss), or ties (%Tie) *baseline* at the 5% significance level. The loss differential is evaluated as the loss of *competitor* minus the loss of *baseline*. The average DM test statistic across all days and stocks is also reported. Parameters are estimated over a window of three consecutive days, and forecasts are computed for the following day (three-day rolling forecasting). Panel (a) is based on the MSE loss function, and Panel (b) on the QLIKE loss function.

References

Bauwens, L. (2006). Econometric analysis of intra-daily trading activity on the Tokyo Stock Exchange. *Monetary and Economic Studies 24*(1), 1–24.

Bauwens, L. and P. Giot (2000). The logarithmic ACD model: An application to the bid-ask quote process of three NYSE stocks. *Annales d'Economie et de Statistique* (60), 117–149.

Bhatti, C. R. (2010). The Birnbaum–Saunders autoregressive conditional duration model. *Mathematics and Computers in Simulation 80*(10), 2062–2078.

Blasques, F., V. Holý, and P. Tomanová (2024). Zero-Inflated Autoregressive Conditional Duration model for discrete trade durations with excessive zeros. *Studies in Nonlinear Dynamics & Econometrics 28*(5), 673–702.

Brownlees, C. T., F. Cipollini, and G. M. Gallo (2011). Intra-daily volume modeling and prediction for algorithmic trading. *Journal of Financial Econometrics 9*(3), 489–518.

Brownlees, C. T. and M. Vannucci (2013). A Bayesian approach for capturing daily heterogeneity in intra-daily durations time series. *Studies in Nonlinear Dynamics & Econometrics 17*(1), 21–46.

Cavaliere, G., T. Mikosch, A. Rahbek, and F. Vilandt (2024). Tail behavior of ACD models and consequences for likelihood-based estimation. *Journal of Econometrics 238*(2), 105613.

Cavaliere, G., T. Mikosch, A. Rahbek, and F. Vilandt (2025). A comment on: "Autoregressive Conditional Duration: A new model for irregularly spaced transaction data". *Econometrica 93*(2), 719–729.

Creal, D., S. J. Koopman, and A. Lucas (2013). Generalized Autoregressive Score models with applications. *Journal of Applied Econometrics 28*(5), 777–795.

Engle, R. F. and J. R. Russell (1998). Autoregressive Conditional Duration: A new model for irregularly spaced transaction data. *Econometrica 66*(5), 1127–1162.

Fernandes, M. and J. Grammig (2006). A family of Autoregressive Conditional Duration models. *Journal of Econometrics 130*(1), 1–23.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The Model Confidence Set. *Econometrica 79*(2), 453–497.

Hautsch, N., P. Malec, and M. Schienle (2014). Capturing the zero: A new class of zero-augmented distributions and multiplicative error processes. *Journal of Financial Econometrics 12*(1), 89–121.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics 3*(5), 1163–1174.

Lee, A. J. (2026). Buy-sell imbalances on and around round numbers and high-frequency trading. *Journal of Empirical Finance 89*, 101776.

Meitz, M. and T. Teräsvirta (2006). Evaluating models of Autoregressive Conditional Duration. *Journal of Business & Economic Statistics 24*(1), 104–124.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics 116*(2), 257–270.

Pacurar, M. (2008). Autoregressive Conditional Duration models in finance: A survey of the theoretical and empirical literature. *Journal of Economic Surveys 22*(4), 711–751.

Saulo, H., S. Pal, R. Souza, R. Vila, and A. Dasilva (2025). Parametric quantile Autoregressive Conditional Duration models with application to intraday value-at-risk forecasting. *Journal of Forecasting 44*(2), 589–605.

Tse, Y.-K. and Y. Dong (2014). Intraday periodicity adjustments of transaction duration and their effects on high-frequency volatility estimation. *Journal of Empirical Finance 28*, 352–361.

U.S. Securities and Exchange Commission (2020). Staff report on algorithmic trading in U.S. capital markets. Technical report.

Veredas, D., J. M. Rodríguez-Poo, and A. Espasa (2001). On the (intradaily) seasonality and dynamics of a financial point process: A semiparametric approach. Working Paper 01-33, Universidad Carlos III de Madrid.

Wu, Z. (2012). On the intraday periodicity duration adjustment of high-frequency data. *Journal of Empirical Finance 19*(2), 282–291.

Zhang, M. Y., J. R. Russell, and R. S. Tsay (2001). A nonlinear Autoregressive Conditional Duration model with applications to financial transaction data. *Journal of Econometrics 104*(1), 179–207.