AARHUS UNIVERSITY
BUSINESS AND SOCIAL SCIENCES
DEPARTMENT OF ECONOMICS AND BUSINESS

CREATES
Center for Research in Econometric Analysis of Time Series

Assessing Measures of Order Flow Toxicity via Perfect Trade Classification

Torben G. Andersen and Oleg Bondarenko

CREATES Research Paper 2013-43

Department of Economics and Business
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Assessing Measures of Order Flow Toxicity via Perfect Trade Classification[*]

Torben G. Andersen[†] Oleg Bondarenko[‡]

This version: November 2013
First version: March 2013

Abstract

The VPIN, or Volume-synchronized Probability of INformed trading, metric is introduced by Easley, López de Prado and O'Hara (ELO) as a real-time indicator of order flow toxicity. They find the measure useful in predicting return volatility and conclude it may help signal impending market turmoil. The VPIN metric involves decomposing volume into active buys and sells. We use the best-bid-offer (BBO) files from the CME Group to construct (near) perfect trade classification measures for the E-mini S&P 500 futures contract. We investigate the accuracy of the ELO Bulk Volume Classification (BVC) scheme and find it inferior to a standard tick rule based on individual transactions. Moreover, when VPIN is constructed from accurate classification, it behaves in a diametrically opposite way to BVC-VPIN. We also find the latter to have forecast power for short-term volatility *solely* because it generates systematic classification errors that are correlated with trading volume and return volatility. When controlling for trading intensity and volatility, the BVC-VPIN measure has no incremental predictive power for future volatility. We conclude that VPIN is not suitable for capturing order flow toxicity.

JEL Classification: G01, G12, G14, G17, and C58

Keywords: VPIN, Accuracy of Trade Classification, Order Flow Toxicity, Order Imbalance, Volatility Forecasting

[*] We are grateful to the CME Group for providing access to data from the CME DataMine system. Andersen also acknowledges support from CREATES funded by the Danish National Research Foundation. We thank seminar participants at Spot Trading, the Kellogg School, the Second Annual Meeting of DAEiNA, SAC Capital, Duke University, the High-Frequency Data and High-Frequency Trading Conference at the University of Chicago, and the Fields Institute at the University of Toronto for helpful comments.

[†] Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208; NBER, and CREATES; e-mail: t-andersen@kellogg.northwestern.edu

[‡] Department of Finance (MC 168), University of Illinois at Chicago, 601 S. Morgan St., Chicago, IL 60607; e-mail: olegb@uic.edu

1 Introduction

The trading environment for securities listed on financial exchanges worldwide have undergone dramatic changes over the last decade. Assets are now traded almost exclusively on electronic platforms and high-frequency trading firms have largely taken over the basic market making function. These developments have brought impressive gains in market quality, as measured in terms of the average bid-ask spread or trading costs associated with small transactions. On the other hand, concerns have arisen regarding market fragility. One example is the occurrence of "flash crashes," where market prices move extremely rapidly – typically downward – for a short period of time, only to then reverse trend and almost as quickly return to a level near the original price point. While most of these incidents are self-contained and have a limited impact on other securities or markets, they can spill over into related venues and ultimately disrupt the broader financial market infrastructure. This occurred, e.g., during the flash crash on May 6, 2010, which widely is believed to have originated in the E-mini S&P 500 futures contract at the Chicago Mercantile Exchange (CME), cf. the CFTC-SEC Report (2010). Such incidents may call into question the role of financial markets in providing credible signals for capital allocation and raise issues about the fairness and efficacy of the trading process itself. In particular, the events of May 6, 2010, spurred a major debate about the origin of these idiosyncratic and highly erratic market dynamics and alternative measures which may prevent them from propagating into systemic disruptions to the financial system.

Against this backdrop, the VPIN metric, developed in a sequence of papers by Easley, Lopez de Prado and O'Hara, henceforth ELO, (2011a, 2011b, 2011c, 2012a), has generated widespread attention among academics, practitioners, regulators and exchanges. It is designed to provide a real-time estimate of the toxic order flow in a financial market, i.e., the extent to which market makers (high-frequency traders) are being adversely selected by agents with private information. When toxicity is high, liquidity is likely provided at a loss and may be withdrawn in short order. Thus, a high VPIN reading should signal an elevated probability that liquidity may vanish, causing a market disturbance. Clearly, such a "warning system" would be valuable for traders, exchanges, and regulators alike. In the words of Marcos López de Prado:

> The measure would have been able to anticipate two hours in advance there was a high probability of a liquidity-induced event on May 6.
>
> This would be much more effective than a circuit breaker [which] stops the infection after the infection is already widespread. (*Bloomberg*, October, 29, 2010).[1]

In addition, ELO (2012b, 2012c, 2012d, 2013) build on or relate to the VPIN metric in important ways. Meanwhile, different implementations of the metric are explored in a variety of studies, e.g., Abad and Yague (2012), Bethel et al. (2012), Chakrabarty, Pascual and Shkilko (2012), Menkveld and Yueshen (2013), Yildiz, Van Ness and Van Ness (2013), Wei, Gerace and Frino (2013) and Wu et al. (2013).[2]

Nonetheless, VPIN is not without controversy. One, the metric cannot be replicated with precision without matching the entire transaction record, starting with the initial trade. Any

[1] VPIN has been featured in many other prominent news outlets, including the Wall Street Journal, it has been the topic of many key-note presentations at leading academic conferences, it is featured on a variety of web videos displaying real-time events where VPIN is interpreted as signaling extreme toxicity. Moreover, three separate VPIN related patent applications have been filed.

[2] Likewise, commentators on the general debate surrounding high-frequency trading routinely reference the ELO findings; for recent examples, see, Corcoran (2013) and MacIntosh (2013).

mismatch in the *exact* starting point or any subsequent discrepancy in the recording of a trade, say, on the first day of the sample will imply that *all* subsequent VPIN measures differ, even if the transaction record is otherwise identical. That is, the VPIN metric obtained across a given trading day hinges on the full history of the tick data.[3] Two, even more critically, the appropriate metric for gauging performance is not evident and, at least for some important predictive comparisons, it fares poorly relative to traditional forecast variables, as detailed in Andersen and Bondarenko, henceforth AB, (2014). Three, from a broader perspective, the ELO (2011a, 2012a) findings challenge the notion that market efficiency provides a sensible first order approximation to the high-frequency behavior of asset returns. If a simple metric, constructed from the real-time transaction record, embodies important information regarding the distribution of future returns and this information largely eludes the market participants and is also not adequately reflected in option prices, then (high frequency) traders systematically fail to extract critical information from the observable dynamics of trades and prices. In particular, high-frequency trading firms are known to scrutinize the trade and order book dynamics in real time, and it is surprising if they do not identify and respond to developments that can be extracted via a simple algorithm from market data.

AB (2014) address some of the issues of the current paper, but emphasize the original VPIN metric developed in ELO (2011a, 2011b, 2011c). ELO (2012a) advocate a different implementation, arguing that the Bulk Volume Classification (BVC) procedure, outlined in ELO (2012b), improves the accuracy of trade classification relative to the modified tick rule used in ELO (2011a). This opens the door to continuing refinements, reflecting the development of ever more precise order imbalance measures. On the other hand, it is not clear that the BVC approach provides a superior real-time classification of the (active) buy and sell volume relative to a more traditional tick rule procedure. The latter has recently been disputed by Chakrabarty, Pascual and Shkilko (2012), based on evidence from individual stock trades.

Consequently, the accuracy of the underlying order imbalance measure is an ever-present source of ambiguity for VPIN. We confront the issue head on by exploiting best-bid-offer (BBO) data for the E-mini S&P 500 futures from the CME Group to construct order imbalance measures with practically 100% accuracy. This enables us to monitor the precision of alternative classification techniques over time, so we can directly relate the performance of a specific VPIN implementation to its classification accuracy. Moreover, we can observe how trade classification interacts with other features to generate specific properties of the resulting metric. In particular, we document that alternative VPIN measures display a radically different degree of correlation with market activity variables such as trading volume and return volatility. We explore whether these discrepancies stem from differences in classification accuracy or other features of the VPIN construction. In short, we get "under the hood" of the VPIN algorithm and explore why different implementations produce widely diverging measures and often generate diametrically opposite conclusions.

Our results are striking. Most notably, we find that the ELO (2012a) BVC-VPIN metric forecasts return volatility *only* because it generates trade classification errors that are strongly correlated with innovations to trading volume and volatility. In contrast, the VPIN metric based on the actual order imbalance, or even the order imbalances obtained via the tick rule at the transaction level, is *negatively* correlated with future return volatility. This result was first highlighted by AB (2014), but is also consistent with observations in ELO (2011a). The latter cite this feature as an indication that trade classification based on individual transactions is misleading. However, this conjecture is counterfactual. We document that order imbalance

[3] We reproduce the algorithms for generating alternative VPIN measures, as stated by ELO (2011c), ELO (2012a), and ELO (2012b), in our Web Appendix.

measures derived from tick data strictly dominate those obtained via time or volume bar aggregation and, in fact, provide qualitatively reliable results regarding the true relationship between VPIN measures and market activity variables. We further establish that the VPIN metric, once we control for volume and volatility, has no residual correlation with future volatility. In particular, the ELO (2012a) VPIN measure provides no incremental explanatory power beyond what is already captured by traditional volatility forecast variables. Consequently, we conclude that VPIN is not a useful empirical measure of order imbalances or flow toxicity.

Our analysis goes far beyond the few existing studies exploring the properties of VPIN due to our construction of order imbalances and associated metrics based on perfect trade classification. First, this is critical for removing any ambiguity regarding the underlying source of discrepancy between the alternative VPIN metrics. It also helps in determining whether the trade classification errors are mechanically correlated with the variables that VPIN portends to forecast. In particular, we can formally test the hypothesis that BVC-VPIN, by construction, is highly correlated with concurrent realized volatility. Second, it enables us to directly assess the accuracy of the novel "bulk volume" classification strategy introduced in ELO (2012b). Third, we can directly compare the traditional cumulative signed order imbalance measure constructed via the regular tick rule to the corresponding imbalances obtained from perfect classification. This helps us assess whether the former provides economically meaningful information during stressful market conditions such as the flash crash. Fourth, it allows us to gauge the accuracy of alternative classification schemes by exploiting the diurnal patterns in volatility and volume, as these features should induce systematic, and artificial, patterns in the corresponding order imbalance measures, if the latter, indeed, are mechanically correlated with the activity variables. Fifth, we exploit a longer sample for the E-mini S&P 500 futures than used in the ELO studies of VPIN, facilitating the identification of extreme events. However, the long sample period also forces us to invoke new normalization techniques to ensure that the VPIN metric is not distorted by non-stationarity in the volume series. Taken together, the findings based on perfect trade classification speak generally to the potential of *any* VPIN metric to enhance the information content of traditional order imbalance measures. In the absence of the true order imbalances, it is obviously impossible to rule out that future VPIN metrics, exploiting alleged superior trade classification, can be successful.

The remainder of the article is organized as follows. Section 2 describes how we obtain near perfect trade classification by combining real-time trade and order book information. Section 3 introduces the VPIN metric. Section 4 develops notation to distinguish the many variants of VPIN that we explore. Section 5 shows how the trend in trading volume distorts the VPIN metric, and motivates our data-dependent detrending procedure. Section 6 analyzes classification accuracy, both in terms of average (unconditional) performance and via the correlation of errors with the activity variables. Section 8 provides predictive regressions for future return volatility. We find that the VPIN metrics have no incremental forecast power for volatility, as they are wholly subsumed by standard realized volatility and volume measures. Section 9 explores the factors behind the surprising empirical finding that VPIN, based on the actual order imbalances, is strongly negatively correlated with future return volatility. Finally, Section 10 concludes.

2 Data

We exploit best bid-offer (BBO) files for the E-mini S&P 500 futures contract from the CME Group. Among other variables, these "top-of-the book" files provide a complete record for the best bid, bid depth, best ask, ask depth, trade prices, and trade sizes. Quotes and trades are time stamped to the second. Moreover, the files contain a sequence indicator that identifies the order in which quotes and trades arrive to the exchange. Thus, we know the actual sequence of order arrivals within each second. The files are obtained directly from CME DataMine.

Our sample covers the period from February 10, 2006, to March 22, 2011. Hence, our analysis is based on more than five years of tick-by-tick data. The E-mini S&P 500 futures contract (commodity ticker symbol ES) trades exclusively on the CME GLOBEX electronic platform. The ES futures contract expires quarterly, on the March expiration cycle. The notional value of one contract is $50 times the value of the S&P 500 stock index. The ES futures contract has a tick size of 0.25 index points, or $12.50. The contract trades essentially 24 hours a day, five days a week. Specifically, from Monday to Thursday, the trading is from 15:30 to 15:15 of the following day, with a half-hour maintenance shutdown from 16:30 to 17:00. On Sunday, the trading is from 17:00 to 15:15 of the following day.

2.1 Summary Statistics

The ES futures market is among the most liquid worldwide. Table 1 provides summary statistics.[4] There were more than 136,000 trades involving 1,782,000 contracts on average per day. That implies a mean transaction size of about 13 contracts. During regular trading hours, there were around 4.8 trades per second. Although the numbers are much lower outside regular hours, the activity is still impressive, with a trade consummated about every three seconds.

2.2 Trade Classification from the Quote and Transaction Record

The BBO files record the changes in the best bid or offer. Importantly, the quote updates arrive in pairs, one for the bid and one for the ask, synchronized by the sequence variable. The following table illustrates the information content of the BBO files.

Initially at 17:02:58, the limit book is characterized by a bid of 1289.50 with an associated depth of 125 contracts and an ask of 1289.75 with a depth of 98 contracts. Within the same second, a new limit sell order of one contract arrives, raising the ask depth to 99 contracts. The new sequence number (5780) records the updated state of the limit order book. Note that, although the bid price and depth are unchanged, they are repeated in the BBO records when there is a change on the other side of the limit order book. At 17:02:59, five contracts are traded at the ask price of 1289.75. As a result, the ask depth drops by 5 contracts and the new state of the limit book is recorded (sequence number 5800). In this example, we may unambiguously classify the trade of the five contracts as buyer initiated.

Overall, the BBO files provides a snapshot of the limit order book whenever there is a trade or a change in either the quote or depth for the best bid or ask. By comparing the trade price with the preceding bid and ask, we are almost always able to identify the aggressor of the trade. The main exception occurs at the initiation of trading, at 15:30 and at 17:00 on Monday through Thursday and at 17:00 on Sunday. At these times, there is an auction for which all trade orders that cross are executed simultaneously at the identical clearing price. In this scenario, trade is initiated by both sides, and we assign half of the auctioned volume to the buy side and the other half to the sell side, thus ensuring they cancel out in the computation of order imbalances. These auction trades make up about 0.024% of the total volume in our

[4] We remove two trading days from our sample. May 9, 2008, has missing quote data, while January 13, 2010, has a highly unusual trading pattern where, within one second, almost 200,000 contracts were traded – representing about 17% of the typical daily volume for the preceding month. Seemingly, two parties coordinated to execute a block trade through the electronic platform. The event did not occur during a stressful trading period. All qualitative conclusions are robust to the inclusion of the latter day.

Table 1: **Descriptive Trading Statistics for the E-mini S&P 500 Futures Contract**

	Regular	Overnight	Holiday
# Days	1285	1285	30
Volume (1 min)	3973	208	69
# Trades (1 min)	285	23	9
# Order Book Changes (1 min)	1730	175	60
# BBO Changes (1 min)	26	8	4
Notional Value, $Mln (1 min)	235	12	4
Trade Size	13.9	8.9	7.3
Order Book Changes per Trade	6.1	7.5	6.4
Trades per BBO Changes	10.9	3.0	2.4

Order Size: Average Daily Percentiles

	Min	10%	50%	75%	90%	99%	99.9%	Max
All	1.0	1.0	2.1	7.5	30.5	233.4	643.5	1683.5

Notes: This table reports summary statistics for the trading in the E-mini S&P 500 futures contract over the period February 10, 2006 - March 22, 2011. The data are reported separately for Regular Trading Hours (Regular, 8:30-15:15), Overnight Trading Hours (Overnight, 15:30-8:30), and Holiday Trading Hours (Holiday, exchange holidays).

Time	Sequence	BidPrice	BidSize	AskPrice	AskSize	TradePrice	TradeSize
17:02:58	5770	1289.50	125	1289.75	98		
17:02:58	5780	1289.50	125	1289.75	99		
17:02:59	5790					1289.75	5
17:02:59	5800	1289.50	125	1289.75	94		

sample. In addition, during the regular course of trading, there are a small number of cases (about 0.048% of volume) for which trades are consummated strictly between the last observed bid and ask price. This tends to happen when the spread exceeds one tick and a bid or ask quote is submitted within the existing spread and then immediately (within a small fraction of a second) hit by a market order. Again, we neutralize the impact of such trades by splitting them equally between buys and sells. In total, we positively identify the trade direction for over 99.95% of the trades consummated within the usual electronic trading environment.[5]

The setting is exemplary for testing the accuracy of alternative trade classification procedures. All trades and quotes are recorded in a single electronic system and trades can only be executed at prevailing bid or ask quotes. Finally, the sequence indicator provides a comprehensive record of quotes and trades in real time. This eliminates problems stemming from having to merge separate files with quote and trade information, from dealing with reporting delays, and from having to integrate activities from several trading venues. Given the extreme trad-

[5] The number of contracts not uniquely classified is sufficiently small that our subsequent results are robust to any buy-sell assignment scheme one may apply to them.

ing intensities in today's major financial markets, accurate sequencing of events is infeasible without this type of integrated system from a single electronic platform.

We know of only one prior study that explores the accuracy of the classification techniques associated with the VPIN metrics in ELO (2011a, 2012a) and focuses on the E-mini S&P 500 futures contract, namely ELO (2012b). They exploit the detailed DataMine Market Depth database over a single year. This database provides all the raw messages required to construct the order book. However, the data structure is complex and extensive data handling capabilities are required to separate the messages, disentangle the contracts referring to separate expiry times, sequence the transactions in time (stamped up to the millisecond), and even separate fictitious trades (arising from the exchange algorithmic testing procedures) from actual ones. In contrast, the BBO data are cleaned and sequenced by the exchange itself, using in-house expertise and knowledge of the trading system. Consequently, we analyze a significantly longer five year sample without encountering pitfalls associated with inadequate handling of the data cleaning task or the order book construction, and we can readily construct the trade direction indicator for all transactions throughout our more than five-year long sample.[6]

2.3 Cumulative Order Flow Imbalance during the Flash Crash

To illustrate the relationship between signed trades and price changes we depict, in Figure 1, the cumulative signed order imbalance and the e-mini S&P 500 futures price for the regular trading hours on May 6, 2010; the day of the "flash crash." The signed order imbalance is initiated at zero at the start of regular trading on the prior day, i.e., 15:30 on May 5.

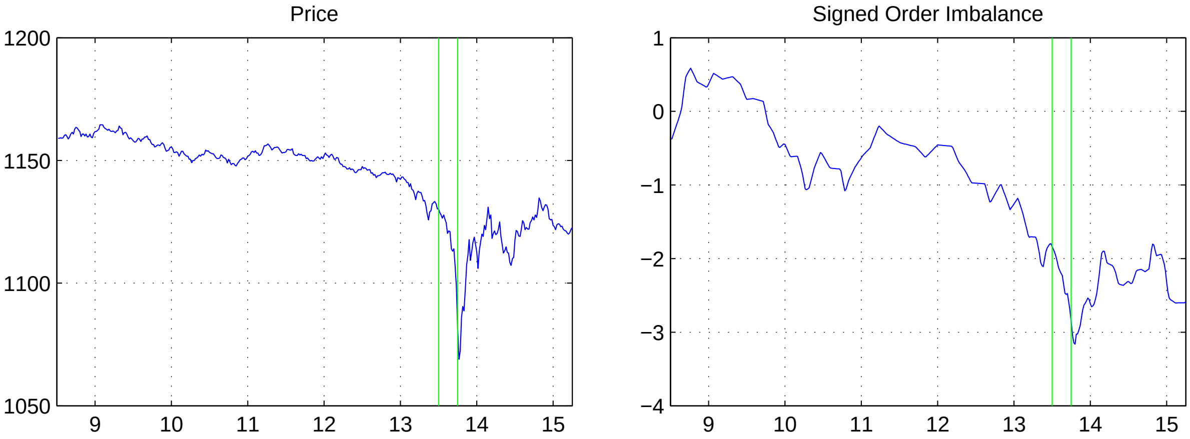


Figure 1: This figure plots the price and cumulative signed order imbalance for May 6, 2010. The solid vertical lines indicate the official CTFC report timing of the "flash crash."

Not surprisingly, there is a strong coherence between the cumulative signed order imbalance and the corresponding price movement. Close inspection of Figure 1 reveals that every distinct spike in the price path may be linked to a corresponding innovation in the signed order imbalance measure. In particular, increasing selling pressure is evident prior to and during the flash crash, while the sharp recovery is characterized by sustained buying pressure. The order imbalances appear to have a near instantaneous impact on prices, even if the relationship is far from perfect. For the full sample, the correlation between the S&P 500 log returns and the

[6] It should be noted that our database also is very large and necessitated the development of specialized storage and retrieval procedures along with the acquisition of a great deal of working memory.

signed order imbalance measure is 0.53.[7] In summary, the standard cumulative order imbalance measure is clearly economically meaningful. At the same time, it is evident that market conditions more broadly impact the sensitivity of the transaction prices to sustained shifts in the order imbalance. Some times, selling pressure causes only minor drops in price while, at other times, prices react vigorously to negative innovations in order flow.

3 The VPIN Metric

At its core, the VPIN metric is simply a rolling average of estimated absolute order imbalance measures. As such, a key ingredient is the procedure for estimating the order imbalance over a given period. However, even given a specific order imbalance measure, the construction of the moving average requires a number of choices regarding measurement units and lag length. Although many of the details are well known, see, e.g., ELO (2011c, 2012a) and AB (2014), we review the procedure in order to establish notation and identify the critical variables. The first subsection explains the calibration of the rolling average parameters, while the second focuses on alternative procedures for estimating the order imbalances.

3.1 Calibrating the Rolling Average of Order Imbalances.

This section reviews the construction of the VPIN metric for a fixed interval of calendar time, $[0, T]$, *given* a specific order imbalance measure. We represent each traded contract by the pair (t_i, p_i), $i = 1, \ldots, I$, where I is the total number of contracts traded, while t_i and p_i indicate the time and price associated with contract i. Of course, many transactions involve multiple contracts, but these may equivalently be viewed as simultaneous trades of a single contract at the same price. Hence, for a trade involving $v > 1$ contracts, there are v replications of the pair (t_i, p_i) in the (unit) trade sequence. Transaction times are measured in seconds and form a non-decreasing sequence $0 \leq t_1 \leq t_2 \ldots t_I \leq T$. While many trades occur within the same second, we know the order in which they were executed via the associated event indicator. As such, we have a complete transactions history over the sample period.

We now define the volume bucket, V, as a fixed increment to the cumulative trading volume. We set V equal to $(1/50)^{\text{th}}$ of the volume on a regular trading day. Specifically, let T_ℓ indicate the time at which bucket number ℓ has just been filled, $\ell = 1, \ldots, \mathcal{L}$, where $\mathcal{L}$ denotes the total number of buckets in $[0, T]$. The VPIN measure is updated whenever we reach a new bucket, i.e., at times $T_1, \ldots, T_{\mathcal{L}}$, so it evolves in event time, governed by the (random) intensity of trading. The trading record is characterized by the pairs, (T_ℓ, P_ℓ), where P_ℓ denotes the last transaction price in bucket ℓ. By definition, each bucket encompasses V traded contracts.

VPIN is constructed from the underlying order imbalances, defined over the volume buckets. Alternative ways of computing the order imbalances are discussed below but, as stated above, they are, for now, taken as given. Hence, each volume bucket, ℓ, has an associated order imbalance measure, denoted OI_ℓ. The VPIN metric is defined as a moving average of the preceding order imbalance measures,

$$VPIN_\ell \;=\; \frac{1}{L} \sum_{j=0}^{L-1} \, OI_{\ell-j} \, , \tag{1}$$

[7] This correlation is measured across predetermined quantities of trading volume, corresponding to volume buckets of around $(1/50)^{\text{th}}$ of the trading day, as defined in the following section.

where L indicates the length of the moving average. In accordance with ELO, we fix $L = 50$, so VPIN reflects the order imbalances across an average trading day. However, when the trading volume is elevated, the buckets will fill quickly and, thus, the VPIN measure may reflect OI measures covering substantially less (more) than one day during fast (slow) markets.

The order imbalance measures are derived from a given trade classification scheme, determining the number of contracts classified as (active) buys and sells over a given volume bucket, V_ℓ^B and V_ℓ^S, so that $V = V_\ell^B + V_\ell^S$. It is convenient to define the proportion of buy volume, $b_\ell = V_\ell^B / V$, where $0 \leq b_\ell \leq 1$, and the *signed order imbalance*, $\gamma_\ell = \left(V_\ell^B - V_\ell^S\right) / V = 2\, b_\ell - 1$, so that $-1 \leq \gamma_\ell \leq 1$. The order imbalance measure is then given by the *absolute proportional imbalance between buying and selling* over the bucket,

$$OI_\ell \;=\; \frac{\left|\, V_\ell^B \;-\; V_\ell^S \,\right|}{V} \;=\; |\, \gamma_\ell \,| \;=\; |\, 2\, b_\ell \;-\; 1 \,| \,, \tag{2}$$

Once we obtain a measure for V_ℓ^B, or equivalently b_ℓ, then γ_ℓ and OI_ℓ are also uniquely determined. Clearly, $0 \leq OI_\ell \leq 1$, with zero reflecting a perfectly balanced market and unity indicating maximal imbalance, i.e., the bucket is populated wholly by buys or wholly by sells.

3.2 Estimating the Order Imbalance.

We now present alternative classification schemes that may be used in defining a specific VPIN metric. The different schemes share a few common features. First, every bucket, ℓ, is divided into Q_ℓ smaller units, or (time or volume) "bars." The actual trade classification is performed for these bars and then aggregated to the bucket level. We denote the trading volume within bar (q, ℓ) by $V_{q,\ell}$, $q = 1, \ldots, Q_\ell$, implying $V = V_{1,\ell} + \ldots + V_{Q_\ell,\ell}$. The classification rule splits $V_{q,\ell}$ into buys and sells, $V_{q,\ell}^B$ and $V_{q,\ell}^S$, i.e., $V_{q,\ell} = V_{q,\ell}^B + V_{q,\ell}^S$. We also define the proportional buying volume, the signed order imbalance, and the volume weight for bar (q, ℓ) as,

$$b_{q,\ell} = \frac{V_{q,\ell}^B}{V_{q,\ell}}, \qquad \gamma_{q,\ell} = \frac{V_{q,\ell}^B - V_{q,\ell}^S}{V_{q,\ell}} \;=\; 2\, b_{q,\ell} - 1, \qquad \nu_{q,\ell} = \frac{V_{q,\ell}}{V}.$$

The order imbalance for the bucket is then given by,

$$OI_\ell \;=\; \frac{\left| \sum_{q=1}^{Q_\ell} \;\left(V_{q,\ell}^B \;-\; V_{q,\ell}^S \right) \right|}{V} \;=\; \left| \sum_{q=1}^{Q_\ell} \gamma_{q,\ell} \cdot \nu_{q,\ell} \right| . \tag{3}$$

Thus, the order imbalance measure is obtained from the volume-weighted average of the signed order imbalances for the bars across the bucket.

Second, trade classification is performed over bars containing fixed increments of calendar time, known as *time bars*, or trading volume, denoted *volume bars.* For the former, we let δ indicate the length of the calendar time interval in seconds, while for the latter ν denotes the size of the increment relative to the volume bucket, so the volume bar consists of $\nu \cdot V$ traded contracts, $0 < \nu \leq 1$. Of course, the finest level of granularity is achieved by classification based on individual transactions, that is, $\nu = 1/\,V$, so that each bar consists of a single traded contract. Here, the key is to determine whether the active party in the trade was the buyer or seller. For a market with a limit order book, this typically boils down to assessing whether the trade took place at the prevailing bid or ask quote. However, ELO (2012a) argue forcefully that trade classification of individual transactions in a high-frequency environment is fraught with difficulties and, inevitably, will induce significant errors. Instead, they favor classification using larger bars that cumulate individual trades into "bulks" of aggregate order flow.

Third, the length of the bars determines the degree of aggregation applied to individual trades prior to classification. For volume bars, we pick ν such that both the size of the bar, $\nu \cdot V$, and the number of bars, $Q_\ell = Q = 1/\nu$, are positive (fixed) integers. As mentioned above, choosing $\nu = 1/V$ and thus $Q = V$, we obtain a bar size of unity – or *contract-by-contract* trade classification. At the other end of the spectrum, for $\nu = 1$, we have only one bar, $Q = 1$, with V traded contracts, corresponding to *bucket-level* classification. Any intermediate choice, $1 < Q < V$, implies that we have multiple bars within each bucket, each representing an aggregated order flow. However, independent of the choice of Q, the volume bar approach ensures that OI_ℓ in equation (3) is obtained from a simple average of the signed order imbalance measure, $|\gamma_{1,\ell} + \ldots + \gamma_{Q,\ell}| \, / \, Q$. Below, we consider $\nu = 0.02$ and 0.10, producing fifty and ten bars within each bucket, respectively. In contrast, for time bars, Q_ℓ varies inversely with the trading intensity. The shortest possible bar size is one second, but we also explore larger time bars, including $\delta = 60$ seconds, which constitutes the leading case for ELO (2011a, 2012a). For this scenario, there are periods in which an entire volume bucket is filled in less than one minute, so that $Q_\ell = 1$, while at times of subdued trading it can take several hours to fill the bucket, and Q_ℓ may be hundred-fold larger.

Fourth, alternative trade classification schemes may be applied to the bars. The initial VPIN metric, developed in ELO (2011a, 2011b, 2011c), exploits a *tick rule*. To formalize this procedure, let $\Delta P_{q,\ell} = P_{q,\ell} - P_{q-1,\ell}$ be the price change over the bar.[8]

The tick rule stipulates that bar (q, ℓ) is a *buy*, i.e., $b_{q,\ell} = 1$, and $\gamma_{q,\ell} = 1$, if:

$$\Delta P_{q,\ell} > 0 \quad \text{(the price change is positive)}, \quad \text{or}$$

$$\Delta P_{q,\ell} = 0 \quad \text{and} \quad b_{q-1,\ell} = 1 \quad \text{(no price change, but the previous transaction is a buy)}.$$

Otherwise, unit (q, ℓ) is a *sell*, i.e., $b_{q,\ell} = 0$ and $\gamma_{q,\ell} = -1$.

This implies that the overall order imbalance for the bucket in equation (3) will be determined by the number of buy versus sell bars, weighted by the relative volume.

ELO (2012a) instead invoke a Bulk Volume Classification (BVC) strategy. It applies to aggregated transactions volume (time or volume bars) only, and it assigns the proportion of buy volume as a function of the price change. Letting $Z(\cdot)$ denote the cumulative distribution function of a standard normal variate, the procedure takes the form,

$$b_{q,\ell} \;=\; Z\left(\frac{\Delta P_{q,\ell}}{\sigma_{\Delta P}}\right) \qquad \text{and} \qquad \gamma_{q,\ell} \;=\; 2 \cdot Z\left(\frac{\Delta P_{q,\ell}}{\sigma_{\Delta P}}\right) \;-\; 1\,, \tag{4}$$

where $\sigma_{\Delta P_\ell}$ is the sample standard deviation of the transaction price change between adjacent bars. Thus, the BVC approach interprets no price change as balanced trading, while a large positive (negative) price change is translated into a proportionally large (small) buy volume.

A qualitatively different approach, hitherto not explored for VPIN, is to rely on contemporaneous bid and ask quotes and classify buys and sells using the relation between the transaction price and prevailing quotes. This procedure is only feasible if reliable and timely order book information is available. This condition is satisfied for the E-mini S&P 500 futures contract. In fact, we exemplified the approach in Section 2.2. This enables us to achieve (near) perfect categorization, which we label "true" classification. It allows us to construct "true" order imbalance and VPIN measures by suitable aggregation of individual transactions. In turn, these measures serve as natural benchmarks for the procedures derived solely from transactions data.

[8] For $q = 1$, i.e., the first bar within bucket ℓ, we set $P_{q-1,\ell} = P_{Q_{\ell-1},\ell-1}$, so the lagged price is the transaction price of the last traded contract in the preceding volume bucket.

3.3 The Features Behind our Alternative VPIN Implementations.

Our subsequent empirical implementation of VPIN follows ELO by fixing the number of volume buckets exploited in equation (2) to $L = 50$, and by calibrating the size of the bucket to encompass around 2% of the daily trading volume.[9] Given these conventions, we turn toward an examination of alternative VPIN metrics. The differences stem from the choice of transactions versus volume or time bars as the basic unit for trade classification, as well as the aggregation level adopted for the bars (length of time interval, number of traded contracts) and, finally, the choice of the tick rule versus the BVC strategy. In short, what type of bar, what level of aggregation, and what trade classification rule. That is, the VPIN variants arise from different choices associated with points one through four in Section 3.2.

4 Notation

We need to distinguish between several variations of VPIN and adopt the following naming convention. The first letter of the index denotes the type of data aggregation:

R) Individual trades based on tRansaction level data;

T) "Time bars" based on fixed increments to calendar time;

V) "Volume bars" based on fixed increments to trading volume;

The second letter refers to a specific trade classification rule:

A) "true" trade classification, based on whether the trade happens at the bid or ask;

B) tick rule classification, as used in ELO (2011a, 2011b, 2011c);

C) Bulk Volume Classification (BVC), as used in ELO (2012a), where the proportion of buys and sells is determined by the size of the absolute price changes relative to the average volatility over the sample (an unconditional estimate of σ);

D) same as C, but now volatility is time-varying and estimated using a rolling window of about one week (a conditional estimate of σ);

E) uninformative or "random" trade classification, based on the L_2 norm, as developed in AB (2014) (labeled U2-VPIN in their exposition).

Note that not all trade classification rules apply to all types of aggregation. In particular, Rule A can be used on transaction level data only. Also, even though Rule E could be applied to volume bars, the resulting index is constant and thus uninteresting.

Finally, if needed, we use a third symbol to indicate the degree of aggregation. For transaction level data, this is not applicable, since individual trades always are used. Hence, we only operate with the classification schemes RA and RB for this scenario. For the time bar aggregation, we consider four separate frequencies, corresponding to $\delta = 1$, 10, 60, and 300 seconds, and we denote those case 1 through 4. For example, we calculate the VPIN measures TB1, TB2, TB3, and TB4. Similarly, for volume bar aggregation, we consider $\nu = 0.02$ and

[9] The interpretation of a bucket, as representing an average trading day, is compromised if volume displays a pronounced time trend. In that case, the fixed bucket size will not correspond to the "normal" proportion of daily trading volume in the early or late part of the sample. We address this issue in Section 5 below.

0.10, i.e., $Q = 1/\nu = 50$ and 10 volume bars per bucket, and we denote these case 1 and 2, respectively. For instance, we compute the VPIN measures VC1 and VC2.[10] Table 2 summarizes the adopted notation for various VPIN indexes.

Table 2: **Notation for VPIN Measures**

Data aggregation	**Trade Classification**				
	True	Tick-rule	BVC unconditional σ	BVC conditional σ	Random
	A	B	C	D	E
Transactions, R	RA	RB	–	–	–
Time bars, T	–	TB	TC1-TC4	TD1-TD4	TE1-TE4
Volume bars, V	–	VB	VC1-VC2	VD1-VD2	–

5 Regularizing the VPIN Metric

Two features of the volume bucketing procedure render the regular VPIN implementation unsuitable for our lengthy sample. The first problem is the sensitivity of VPIN to the starting point. To replicate the computation of a single VPIN observation, one must obtain the complete tick record for the full sample. The second issue stems from the constancy of the bucket size, even as volume displays a pronounced trend. The volume trend distorts the time-bar based metrics used by, e.g., ELO (2011a, 2012a). Given our long sample, we modify the VPIN implementation to alleviate the impact of these features, while still retaining the essential character of the metric.

5.1 Enhancing the Ability to Replicate the VPIN Metric

AB (2014) establish that the VPIN metric is sensitive to initial conditions. The specific point at which the cumulation of trading volume is initiated determines the location of the buckets throughout the sample. If the starting point changes, the bars are split differently across the buckets. This feature hampers replication, verification, and comparison of results across studies. Specifically, to obtain the same results for a given trading day, two studies must exploit the identical tick-by-tick transactions and initiate the volume cumulation at the exact same time point. If there are any deviations in the reported transactions, discrepancies in the filtering of erroneous entries, or differences in the handling of the transactions consummated during the auction preceding regular trading *at any point throughout the sample*, the VPIN metrics will deviate to varying degrees.

To mitigate such effects and facilitate replication, we instead restart the bucketing process at the beginning of each trading day (normally, at 15:30 of the previous calendar day). At the end of a given day, the last bucket will usually contain less than V contracts and we simply

[10] In this terminology, the empirical work of ELO (2011a,b,c) focuses primarily on TB3, while ELO (2012a) relies on TC3. Moreover, the study of AB (2014), focusing on a shorter sample than here, analyze the following VPIN series: RB, TBj, TCj, TEj, VBk, for $j = 2, 3, 4$ and $k = 1, 2$.

discard it when computing the VPIN measures.[11] As a result, one may verify the VPIN computations for a given trading day by having access to the identical transactions record for that specific day (and the fifty preceding volume buckets necessary for the initial VPIN computation) as well as knowledge of the size of the volume bucket only.

5.2 Inducing Stationarity in the Volume Series

ELO (2011a, 2012a) set the volume bucket, V, equal to $(1/50)^{\text{th}}$ of daily average trading volume and fix it throughout the sample. In our case, this would imply setting V to about 36,500 contracts.[12] We refer to this approach as "constant V". This may be sensible for shorter samples bereft of persistent variation in the daily volume, but it is problematic if there is a pronounced trend in volume over the sample. Since its introduction in 1997, the trading volume in the E-mini contract has risen steadily. The average daily volume was about 17 thousand contracts in 1998 (the first full year of trading), while it was about 970,000 in 2006 and more than 2 million in 2010. We also note that the theory motivating the VPIN construction in ELO (2012a) implicitly assumes that the market environment is stationary.

It is evident from Figure 2 that the volume displays high-frequency spikes that are strongly correlated with the VIX index during turbulent market conditions. When the trading intensity rises, there is a smaller number of time bars per bucket and, as documented in AB (2014), this tends to inflate the time-bar based VPIN metrics. In our sample period, the volume bucket of size $V = 36{,}500$ on average gets filled within 22 minutes of regular trading in the month with the lowest volume, but in just 4.5 minutes during the month with the highest volume. ELO (2012a) scale the bucket size to the average trading intensity for different futures contracts to enhance the compatibility of the corresponding VPIN series. The argument for a similar adjustment for time variation in the trading volume *within* each series is even more compelling. Only if the metric is based on a similar number of bars across the sample can we meaningfully compare the relative VPIN values at different times. In particular, if a secular increase in trading doubles the average daily volume, then the bucket size should be scaled correspondingly, i.e., halved, to render the early and late stages of the sample comparable.

Hence, to accommodate the non-stationarity, we set V equal to $(1/50)^{\text{th}}$ of the daily volume *over the preceding month.*[13] The choice of a one-month moving average is motivated by a desire to also smooth out seasonal fluctuations in the volume, e.g., declines around major holidays, which are evident from the third panel of Figure 2. We report results based on this second approach, which we refer to as "detrended V" below. The bottom panel of Figure 2 contrasts the one-month moving average of daily volume, used in the detrended V approach, with the dashed line, representing the daily sample mean, used by the constant V approach.

To illustrate the impact of our volume regularization, we contrast the daily maxima of the TC3-VPIN series obtained from the constant V and detrended V approach, respectively, in the first and third panels of Figure 3.[14]

We identify a distinct level effect in the TC3 series in the top panel of Figure 3. For the

[11] Alternatively, we could add the volume of the last incomplete bucket to the previous one and weigh the contribution of this bucket accordingly in the VPIN computation. We have confirmed that the discrepancy between these two approaches is negligible.

[12] This compares to $V = 39{,}351$ contracts in ELO or $V = 40{,}000$ in AB (2014) for more recent and shorter samples. The average volume in the early part of our sample is considerably lower than in later years.

[13] On any given day, we round V to the closest multiple of 100. This insures that when the volume bar aggregation is used, the bars we consider always contain a whole number of contracts.

[14] We use TC3 for the illustration as it is most closely aligned with the VPIN metric adopted in ELO (2012a), but similar results are obtained for alternative VPIN measures based on time bars.

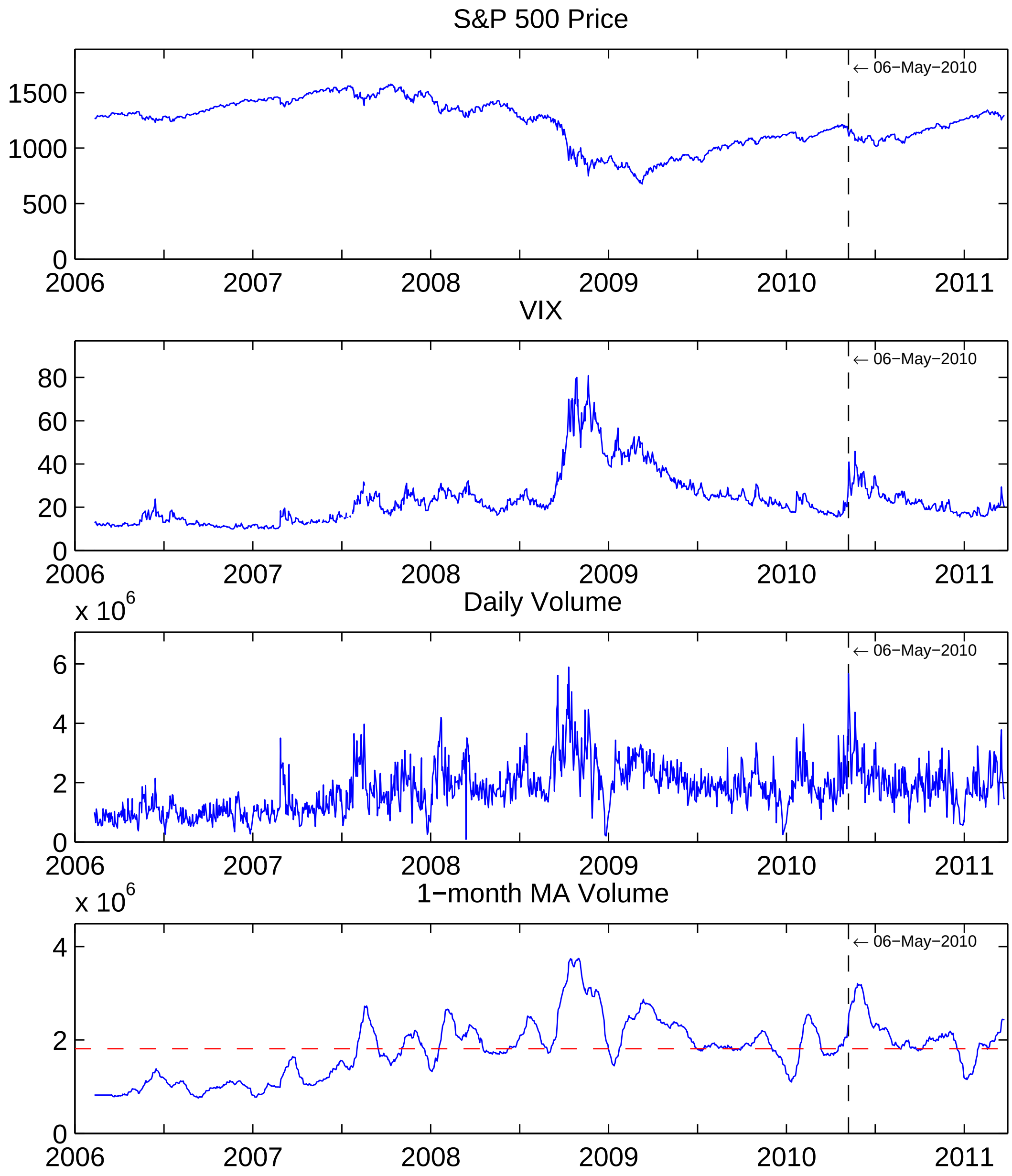
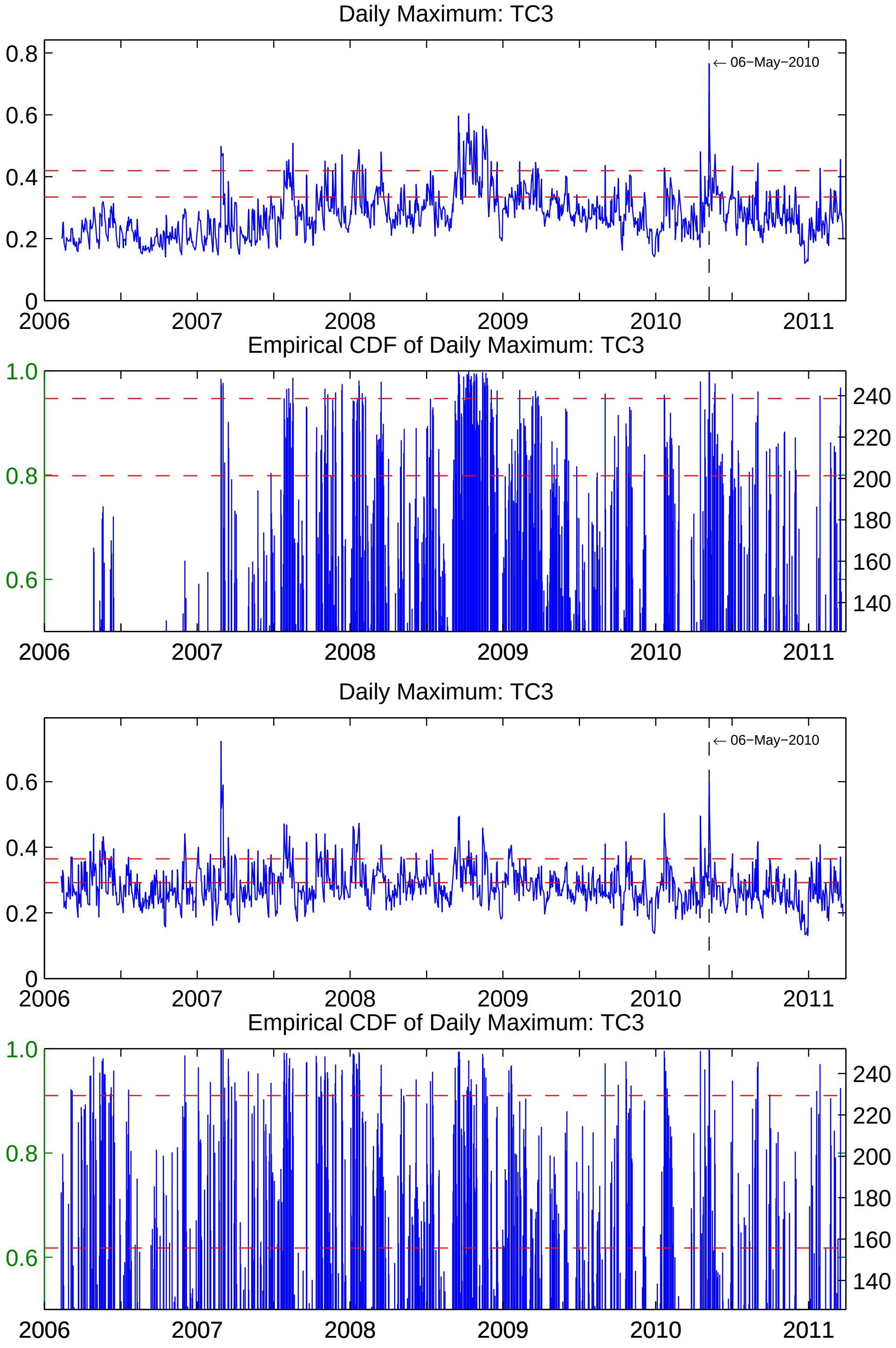


Figure 2: This figure depicts daily values of the S&P 500 index, VIX, trading volume, and a 21-day moving average of the daily volume over the entire sample, February 10, 2006 - March 22, 2011. The dashed line in the bottom panel indicates the average daily volume over the full sample.

first year of the sample, the values are noticeably lower than for the remainder of the sample. Moreover, the variation of the TC3-VPIN series is also lower in this early period, even relative to later periods when similarly low levels of "toxicity" are observed. In addition, the TC3 series only attains extreme values when volume and volatility are elevated, as can be confirmed by referring back to Figure 2. While this may reflect the fact that toxicity is high during turbulent market conditions, the pronounced pattern is troublesome. The theory underlying VPIN is developed for a stationary environment, and the figure suggests that even exceptional days during the early parts of the sample will not stand out due to the apparent dampening of


Daily Maximum: TC3

Empirical CDF of Daily Maximum: TC3

Daily Maximum: TC3

Empirical CDF of Daily Maximum: TC3

Figure 3: The two top panels concern the VPIN measure under the constant V approach. Panel one depicts the daily maximum value of TC3-VPIN. Panel two depicts the CDF (left scale) and corresponding number of days per year with a (maximum) VPIN level below that for the given day (right scale). The horizontal dashed lines indicate the level of VPIN on the Flash Crash day at 12:30 (the lower one) and 13:30 (the upper one). The two bottom panels provide the identical series for TC3-VPIN, but under the detrended V approach.

the metric when (trend) volume is low. In contrast, the detrended V series for TC3 in panel three appears stationary. In particular, the turbulent markets during the financial crisis no longer stand out as a prolonged period of elevated toxicity. This is consistent with the market fluctuations at that time being driven primarily by observable shocks to economic policy and fundamentals rather than by private information filtering into the market through trading. The extreme maximum TC3-VPIN observation for the flash crash, May 6, 2010, is evident from both TC3 series. However, for the detrended series in panel three it is no longer the largest outlier. This distinction belongs instead to February 27, 2007, when the S&P 500 index tumbled 50 points in its largest drop in about four years. Moreover, two days later, on March 1, 2007, the VPIN value almost matches that of May 6, 2010, as well. In short, the VPIN series is drastically transformed by the standardization of the volume series.

ELO (2012a) argue that the appropriate gauge for extreme VPIN values is the CDF. The empirical CDF for the maximum TC3-VPIN values are indicated on the left vertical axis in the second and fourth panels of Figure 3. On the right axis, we indicate the number of days within an average year that realizes a maximum value below the indicated VPIN. For example, the second panel shows that the TC3-VPIN value achieved one hour prior the flash crash (the lower dashed line) was exceeded on about 20% of the days in the sample, or about 50 times a year. Likewise, the TC3-VPIN value at the onset of the crash (the upper dashed line) was topped on average 13 times a year, or about once a month. More strikingly, all these extreme values were observed after 2006, and there is a massive clustering of extreme events during the financial crisis. In contrast, the third panel demonstrates that extreme VPIN events are much more uniformly distributed once we normalize for the volume effect. This appears more consistent with a stationary environment where toxicity rears its head randomly throughout time. In conclusion, for samples with a strong volume trend, we should adjust the size of the volume bucket as trading volume rises, and this is the procedure we adopt in the sequel.

6 Misclassification Measures for Order Flow Imbalance

We now explore the accuracy of alternative trade classification schemes for the E-mini S&P 500 futures contract. We rely on the strategy in Section 2.2 to obtain the "true" RA classification, and each procedure is assessed relative to this benchmark.

6.1 Order Flow Aggregation and Trade Classification

Order imbalances are computed over volume buckets and then averaged to obtain the VPIN measure. Thus, our primary interest is to determine the classification accuracy at the bucket level. We let the index c denote a specific classification scheme. Following the notation of Section 4, $c \in \{\text{RB}, \text{TB}k, \text{TC}k, \text{TD}k, \text{TE}k, \text{VB}m, \text{VC}m, \text{VD}m\}$, for $k = 1, 2, 3, 4$ and $m = 1, 2$.

We first focus on the tick-rule procedure, RB, based on the individual transactions. It uses the full trade record up through time T_ℓ to determine the buy volume for the first ℓ buckets of the sample. Since the approach generates a buy or sell indicator for each transaction we can readily assess accuracy relative to the RA classification at a contract-by-contract basis.

The above procedure is similar in spirit to the approach of ELO (2012b) who examine the tick rule accuracy relative to that of BVC classification. However, it is critical to realize that the accuracy for the tick rule is much better at the bucket level than the trade-by-trade level. Consider the following illustrative example. Assume the bucket covers 2 minutes and there are 4 trades of one contract within each minute. The actual trade sequence is BBBSSSSB, which we represent via a buy indicator sequence and compare to a candidate classification rule C,,

Actual Rule (A): $(1,1,1,0,0,0,0,1)$
Candidate Rule (C): $(1,0,1,0,1,0,1,0)$

In terms of individual contracts, or volume bars for $\nu = 1/8$, rule C misclassifies four out of eight, or 50%. However, if we focus on the accuracy for one-minute bars (or volume bars with $\nu = 1/2$), then rule C misclassifies only two contracts ($|2-3|+|2-1|$), so the inaccuracy rate drops to 25%. Finally, if we classify accuracy at the bucket (two-minute bar) level, then the candidate classification is perfect ($|4-4| = 0$).

The example highlights an important property of contract-by-contract classification: *Aggregation improves accuracy.* This result applies for bar-based classification as well. Formally, let V_q denote the volume of bar q within a given volume bucket, and V_q^B and $\hat{V}_q^B$ represent the corresponding actual and estimated buy volume, respectively, for $q = 1, \ldots, Q$, then

$$\sum_{q=1}^{Q} \left| \hat{V}_q^B - V_q^B \right| \quad \leq \quad \left| \sum_{q=1}^{Q} \left(\hat{V}_q^B - V_q^B \right) \right| . \tag{5}$$

This reflects a *diversification effect*: as volume is aggregated, positive and negative errors cancel, just as positive and negative returns of individual assets cancel within portfolios. The initial effect is typically quite substantial, but levels off as the degree of aggregation rises.

These observations have a couple of implications. Most importantly, meaningful comparison of accuracy must be performed at the identical aggregation level as, otherwise, the rule exploiting the more aggregated order flow is unduly advantaged. Second, depending on the objective, different degrees of aggregation are appropriate. Market microstructure issues often hinge on trade-by-trade classification and there are few alternatives to tick rules for this purpose. For VPIN, however, it is unequivocally the bucket level accuracy that matters. Below, we develop metrics for both applications.

Apart from RB, all the classification rules, based solely on the transaction record, exploit bar-level aggregation. Clearly, aggregation involves a loss of granularity. This does not necessarily imply that bar level classification is inferior to contract-by-contract classification. In fact, ELO (2012a, 2012b) argue forcefully that the probabilistic "bulk" volume approach is superior to tick-by-tick identification due to the noise associated with classification of individual trades in a high-frequency environment. On the other hand, the diversification effect favors the tick-by-tick approach, or the use of small sized bars, as such schemes benefit the most from diversification of errors within the buckets. Ultimately, the relative precision of alternative schemes is an empirical question which we address in Section 6.3.

6.2 Defining Misclassification Measures

This section introduces accuracy measures for the alternative classification schemes by formalizing the "misclassification" measure we, implicitly, adopted during the discussion in Section 6.1. We first define the measure at the volume bucket level and then provide a natural generalization that characterizes classification accuracy at the individual contract level.

6.2.1 The Misclassification Measure for Volume Buckets

We initially focus on the accuracy for a given bucket, ℓ, comprising V contracts. As before, the traded contracts within bucket ℓ are allocated to Q separate bars. The trading volume within bar q is denoted $V_{q,\ell}$, $q = 1, \ldots, Q$, and the (true) proportion of buys is labeled $b_{q,\ell}$.

The corresponding estimated number of buys and proportion of buys within bar q are denoted, respectively, $\hat{V}_{q,\ell}$ and $\hat{b}_{q,\ell}$.

Upon defining $\nu_{q,\ell} = V_{q,\ell}/\,V$ and aggregating by suitable volume-weighting, we arrive at a bucket-wide misclassification measure, MVB$_\ell$, for a given classification scheme,

$$MVB_\ell \;=\; \frac{\sum_{q=1}^{Q} |\hat{V}^B_{q,\ell} - V^B_{q,\ell}|}{V} \;=\; \sum_{q=1}^{Q} |\hat{b}_{q,\ell} - b_{q,\ell}| \cdot \nu_{q,\ell}\,.$$

The measure simplifies further for volume bars, where $\nu_{q,\ell} = 1/\,Q$,

$$MVB_\ell \;=\; \frac{1}{Q} \sum_{q=1}^{Q} |\hat{b}_{q,\ell} - b_{q,\ell}|\,.$$

Finally, we note that MVB$_\ell$ equals $(1 - Ar_\ell)$, where Ar_ℓ is the accuracy measure of ELO (2012b) for bucket ℓ. Thus, the two concepts are equivalent, except that the former indicates the degree of error and the latter the degree of accuracy.

The MVB$_\ell$ measure refers to a single volume bucket. Our sample contains thousands of buckets and we construct the overall measure by averaging across all buckets,

$$MVB \;=\; \frac{1}{\mathcal{L}} \sum_{\ell=1}^{\mathcal{L}} MVB_\ell\,.$$

6.2.2 Misclassification at the Individual Contract Level

Trade-by-trade classification is critical for a host of microstructure applications, so we adapt our metric to also capture the performance of alternative classification schemes along that dimension. This is somewhat controversial, however, because bar-based rules are not designed for this purpose. Nonetheless, if one decides to intrapolate the classification of individual contracts from aggregate bar-based measures, the natural procedure is self-evident. The schemes operating on aggregate order flow determine the proportion of buy versus sell volume for a given bar. We can thus assign a probabilistic weight to each contract: the chance it represents a buy is simply equated to the overall proportion of buys over the bar.

We illustrate this approach for the bar-based tick-rule and bulk volume classification strategies. Specifically, consider the "true" trade sequence (A) from above, spanning two one-minute bars with four trades within each bar, $(1,1,1,0,0,0,0,1)$. Assuming the price rises over the first bar and drops over the second, the tick rule will classify all contracts in the first bar as buys and those in the second as sales. If the price increase over the first bar exceeds the price drop over the second, the BVC approach may assign a buy volume ratio of 0.80 to the first bar and 0.25 to the second. We thus have,

Actual: $(1,1,1,0,0,0,0,1)$.

Tick Rule: $(1,1,1,1,0,0,0,0)$.

Bulk Rule: $(0.80, 0.80, 0.80, 0.80, 0.25, 0.25, 0.25, 0.25)$.

The (one-minute bar) tick rule produces a misclassification rate of 25% (2/ 8) for individual trades. Note, however, that the inferred buy proportion perfectly matches the overall proportion of the buy volume, so relying on the full two-minute bar yields a misclassification rate of

0%. For BVC, the misclassification rate for individual contracts inferred from the one-minute bars is 36.25%.[15] In contrast, the BVC rate is only 2.5% when assessed on the basis of the total buy volume within the one-minute bars.[16] Both cases illustrate the effect of interpolating proportional order flow measures from bars to individual contracts – effectively, it produces a reverse diversification effect. Most importantly, we must avoid comparing, say, the tick rule misclassification rate for individual contracts (25%) to the aggregate one-minute bar BVC rate (2.5%), and instead compare to the BVC rate for contract-to-contract classification (36.25%).[17]

Formally, for I total trades, the misclassification measure at the contract level is,

$$MCC = \frac{1}{I} \sum_{i=1}^{I} |\hat{b}_i - b_i|, \tag{6}$$

where $\hat{b}_i$ and b_i denote the estimated and true buy indicator for contract i.

6.3 Overall Classification Accuracy

This section compiles accuracy measures for the various classification techniques across our full sample for the E-mini S&P 500 futures. While we contrast our results to the only existing evidence concerning this contract, namely ELO (2012b, 2012c), the findings complement a large literature on empirical trade classification, originating with Lee and Ready (1991).[18]

Table 3 tabulates the average inaccuracy of the alternative procedures. The top panel refers to contract-by-contract measures and the bottom panel to bucket measures.

6.3.1 Empirical Misclassification Rates at the Transaction Level

The top panel of Table 3 reveals that the tick rule applied contract-by-contract classifies 11.6% of the transactions incorrectly over our five year sample (entry RB). This is roughly consistent with the finding in ELO (2012b), where a 13.6% (1-0.864) failure rate is reported. Our interpretation is entirely different, however. We stress the striking improvement of the contract-by-contract procedure relative to the tick rule applied on aggregated order flow. At the one-second level, the failure rate already reaches 24.8% (entry TB1), and the trend continues with further aggregation. Applying the tick rule to order flow at the 60- and 300-second levels produce failure rates of 38.7% and 43.1% (entries TB3 and TB4), barely improving on random classification (50%)! Finally, we note that the best time-bar based bulk volume procedure is TC3, which has a failure rate of 23.2% at the transaction level. From the perspective of classifying the trade direction of individual trades the approaches in ELO (2011a) and ELO (2012a) is an order of magnitude worse than the traditional tick rule applied to actual trades.

Ultimately, for the current application, we care about the market dynamics over longer time intervals. Thus, we now explore whether the results carry over to order flow measures based on more aggregate bulks of volume.

[15] The rate is computed as $(3 \cdot |1-0.80| + |0-0.80| + 3 \cdot |0-0.25| + |1-0.25|)/8$.

[16] This rate is computed as $(|0.80-0.75| + |0.25-0.25|)/2$.

[17] Alternatively, perform the comparison explicitly at the one-minute or two-minute bar level for all classification schemes, as outlined in Section 6.2.1.

[18] The two ELO citations refer to sequential versions of the same working paper. We reference both, because the first deals with the BVC approach as implemented in ELO (2012a), while the second explores additional features that are critical for our discussion, but relies on a different BVC procedure. The voluminous literature on trade classification includes, e.g., Aitken and Prino (1996), Asquith, Oman and Safaya (2010), Boehmer, Grammig and Thiessen (2007), Chakrabarty, Li, Nguyen and Van Ness (2007), Chakrabarty, Moulton and Shkilko (2012), Ellis, Michaely and O'Hara (2000), Finucane (2000), Odders-White (2000).

Table 3: **Error Rates for Alternative Trade Classification Rules**

Panel A: MCC

Rule	R	T1	T2	T3	T4	V1	V2
A	0						
B	0.116	0.248	0.330	0.387	0.431	0.339	0.388
C		0.283	0.242	0.232	0.245	0.233	0.170
D		0.282	0.242	0.240	0.261	0.228	0.169

Panel B: MVB

Rule	R	T1	T2	T3	T4	V1	V2
A	0						
B	0.023	0.038	0.072	0.141	0.264	0.065	0.121
C		0.040	0.045	0.083	0.161	0.043	0.047
D		0.040	0.045	0.086	0.172	0.041	0.043

Notes: This table reports misclassification rates at the contract-by-contract level (MC) and for volume buckets (MVB). Columns represent different type of data aggregation: R (transactions), T1-T4 (time bars with $\delta = 1$, 10, 60, and 300 seconds), and V1-V2 (volume bars with $\nu = 0.02$ and 0.10). Rows represent alternative trade classifications: A (true), B (tick rule), C (BVC with unconditional σ), and D (BVC with conditional σ). For example, the procedure applied in ELO (2012a), TC3 (BVC with unconditional σ on 1-minute bars) is shown in the entry of the fourth column (T3) and third row (C).

6.3.2 Empirical Misclassification Rates for Volume Buckets

Moving to the bottom panel of Table 3, we find that the bucket level error rates are sharply lower than those at the individual contract level. This implies that the diversification effect is operative and highly effective in all cases. Moreover, we also find the relative ranking of the different procedures to remain largely intact. That is, the tick rule based on individual trades outperforms the tick rule or BVC approach applied to aggregated order flows. In fact, for each procedure the error rates increase monotonically as we move to a higher level of aggregation. If accuracy of the trade classification is the objective, one should *never* aggregate the order flow prior to classification. It is always better to use tick data relative to one-minute data or one-second data relative to ten-second data. Finally, it is clear that the performance of the tick rule deteriorates more quickly with aggregation than is the case for the BVC approach. This is not surprising given the extreme allocation implied by the tick rule for aggregate order flow. All transactions are classified identically, so we have either all buys or all sells within a given bar. As the bar size grows, this becomes increasingly counterfactual. At the highest frequency, sequences of buys and sells alternate in rapid order, so an extreme and uniform classification will inevitably misclassify a very large proportion of the trades. From this perspective, the superiority of the contract-by-contract classification is a near tautology. The BVC strategy underperforms the tick rule for the small bucket sizes but, eventually, the less extreme assignment of trade direction will, almost trivially, reflect the underlying proportion of actual buys and sells better over long intervals.

Notably, the above findings are at odds with the only existing study of the identical tick rule and BVC classification strategies for the E-mini futures contract, namely ELO (2012b).

They reach the conclusion that: " ... using either time bars or volume bars is more accurate than standard tick-rule classification schemes based on individual trade data."[19] ELO (2012b) obtain this result by comparing the accuracy of the contract-to-contract tick rule to the BVC approach, *assessed at the bar level.* For a time bar of 120 seconds, ELO report a 12.4% failure rate, and for volume bars of 8,000 contracts the rate is 10.0% (relative to 13.6% for the transaction level tick rule). In our tabulations, the corresponding "improvement" is reflected in the misclassification rate of 8.3% at the bucket level for 60-second time bars (TC3 under the MVB panel) and 4.7% failure rate for $\nu = 0.1$ volume bars (VC2 under the MVB panel).

The problem with the ELO (2012b) analysis is that their bar-level error rates are *incompatible* with the tick rule accuracy at the transaction level. We have already established that precision is enhanced, realization-by-realization, when success rates are assessed via more highly aggregated order flow.[20] This effect is conveyed dramatically by the contract-by-contract tick rule improving from an 11.6% failure rate for transactions to only 2.3% at the bucket level (RB under the MVB panel). When the transaction tick rule is compared to BVC at identical levels of order flow aggregation, the tick rule performs, by far, the best. Hence, assessing accuracy within categories that are compatible, the conclusions of ELO are turned upside down.[21]

6.4 Systematic Time-Variation in Misclassification Rates

We now explore whether there is any systematic variation in the accuracy of the various trade classification techniques over our five year long sample. For example, it is unclear how the BVC classification performs in times of slow versus active trading. This is important as significant fluctuations in classification accuracy – correlated with either trading activity or return volatility – may distort the VPIN metric during turbulent market events, when the measure is of particular interest.

Figure 4 presents evidence for a representative selection of classification procedures. The two top panels display misclassification rates for the BVC procedures TC1 and TC3, with the upper panel representing transactions and the second panel volume buckets. The diversification effect clearly improves the performance of TC1 greatly. In contrast, the improvement for TC3, constructed along the lines of ELO (2012a), is more modest. At the bucket level, TC1 dominates uniformly although the discrepancies are especially striking during high volatility periods (compare with Figure 2) while they are minor when volatility is low, e.g., in the first part of the sample, the beginning of 2010 as well as early 2011. Finally, the RB series is included to facilitate comparisons to transaction level tick rule classification. In either scenario, the standard tick rule uniformly dominates the BVC procedures. At the contract level, the precision of the RB approach is near constant, and it is hard to discern much of a relationship between RB and the TC series. At the bucket level, we start seeing TC1 mimicking the properties of RB and the accuracy of the two series now appears to improve as volatility increases, inducing a distinct negative correlation with TC3. It is evident that TC3 performs poorly throughout the sample and is particularly error prone when the market is volatile.

The bottom two panels portray misclassification rates for our BVC procedures exploiting volume bars. Clearly, the use of volume bars improve accuracy. Moreover, in analogy to our prior findings, VC1 – based on the smaller bar size – inherits some features of the RB series at

[19] See ELO (2012b), page 3. Likewise, in ELO (2012c): " ... the BVC algorithm beats the tick rule for time bars of 30 seconds and longer and for volume bars of 2500 shares or longer."

[20] This point is also documented carefully by Chakrabarty, Pascual and Shkilko (2012).

[21] While ELO (2012a, 2012b) fail to acknowledge the importance of comparing accuracy at the identical level of order flow aggregation, the issue is raised subsequently in ELO (2012c). Unfortunately, this paper invokes a new implementation of VPIN, so the results are no longer compatible with the procedure in ELO (2012a).

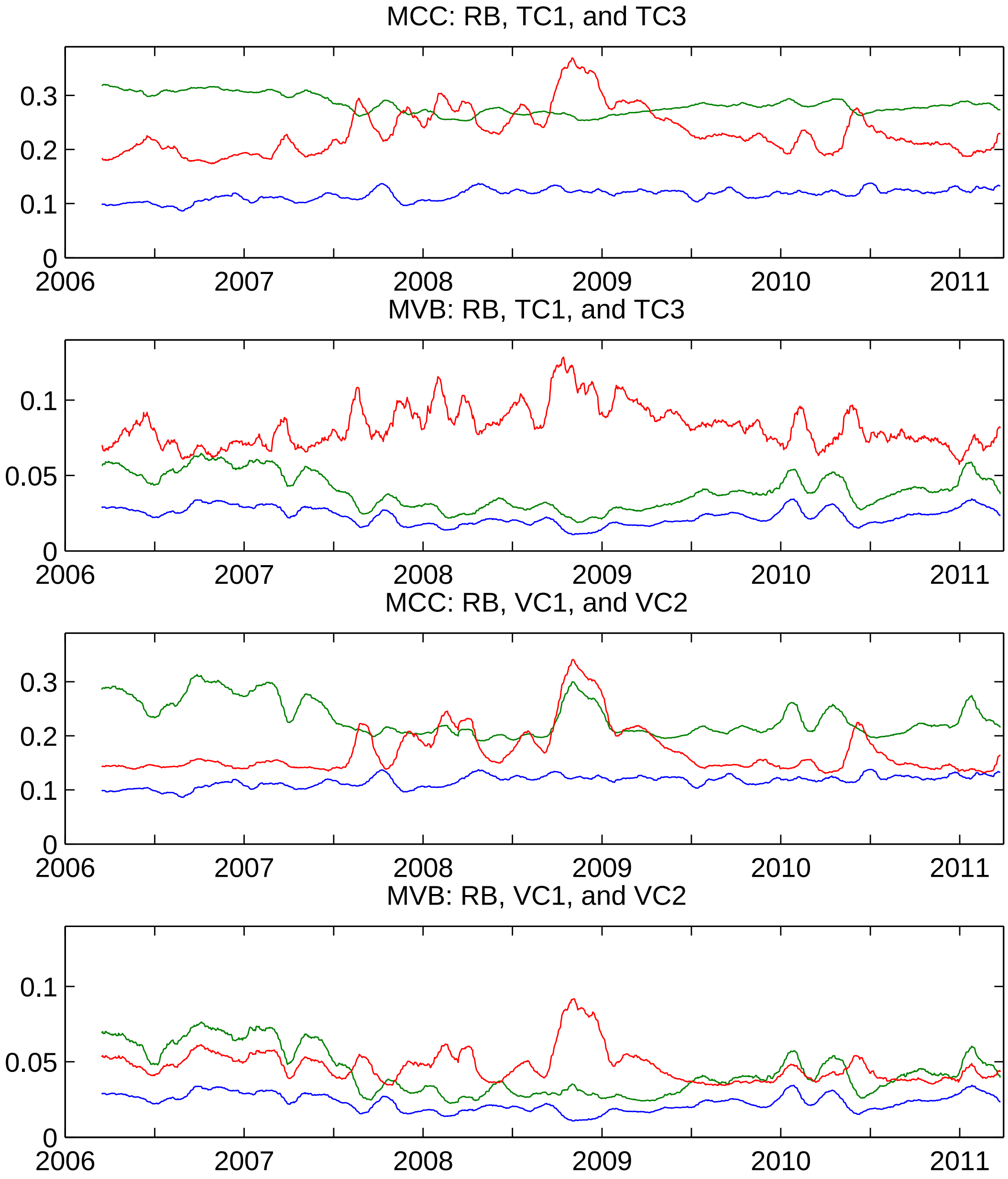


Figure 4: This figure depicts 21-day moving averages of the misclassification rates for different trade classification procedures. The top two panels depict MCC and MVB for classifications RB, TC1, and TC3, while the bottom two panels show MCC and MVB for classifications RB, VC1, and VC2. MCC stands for misclassification at the contract-by-contract level and MVB denotes misclassification at the volume bucket level. In each panel, the order of the plots is blue (first), green (second), and red (third).

the bucket level. Furthermore, VC2 retains the qualitative features observed in TC3, but the fluctuations are dampened, pointing toward a stabilizing effect of using volume rather than time bars. Overall, the main conclusion stands: in terms of bucket level accuracy, the standard tick rule uniformly dominates the BVC procedures. Overall, the most unstable approach, in terms of having its classification accuracy compromised during volatile market episodes, is TC3. This is also the series most closely aligned with the procedure of ELO (2012a).

7 The Empirical Behavior of Alternative VPIN Metrics

Any of the trade classification schemes discussed above may be applied to compute an associated VPIN metric, simply by aggregating the order imbalances across the volume bucket, as indicated in equation (2), and then constructing VPIN as the moving average in equation (1). Since we previously settled on the size of the bucket size, V, and the length of the moving average, L, the calculation of these alternative VPIN measures is unambiguous.

7.1 The Evolution of Alternative VPIN Series over a Five-Year Span

Figure 5 depicts a set of alternative VPIN series for our full sample. The RA metric in the top panel serves as our benchmark, as it is based on the correct classification. As such, it represents the ideal VPIN metric, unaffected by concerns regarding the accuracy of the underlying order imbalance measures. The VPIN series computed from RB is plotted alongside RA. The two series display a strikingly similar evolution, although RB almost invariably lies above RA. Evidently, RB generates measures that are slightly more unbalanced than they should be. Fortunately, the magnitude of this error is small and homogeneous so that, qualitatively, the series depict a near identical variation in VPIN over time. For example, both attain their *minimum* during the outbreak of the financial crisis in 2008.

Inspecting the lower three panels, we find, quite naturally, that VPIN constructed from TC1 and TD1 come closest in terms of mimicking the features of RA. The main difference is that the low frequency variation in TC1 and TD1 is notably smaller than for the transaction-based measures. In comparison, the TC3 and TD3 series display much more short term variation, but less persistent level shifts. We also note that TC3 is prone to erratic movements in either direction. Overall, there is little coherence between the TC3 and TD3 series and the VPIN measures derived from contract-by-contract classification. Finally, the bottom panel shows that there are dramatic differences in the volume-bar BVC-VPIN series, depending on whether the price changes are normalized by an unconditional or conditional volatility factor. VD2 is essentially flat, apart from some high frequency oscillations, while VC2 is erratic, attaining extremely high values during the 2008 crisis, but also hitting extreme lows in 2006 and 2010. Visually, VC2, if anything, appears inversely related to our benchmark RA series. These observations are confirmed by Table 6, which documents a small positive correlation of VD2 and a strong negative correlation of VC2 with the benchmark RA series.

7.2 The Association of VPIN with Trading Volume and Return Volatility

Section 7.1 shows that the VPIN metric is sensitive to the choice of design variables, such as whether time or volume bars are used, the length of the bars, and whether the price changes are normalized by an unconditional or conditional volatility factor. Section 6 documents that there are substantial discrepancies in the accuracy and stability of the different trade classification schemes. We now explore how the alternative VPIN measures are associated with key market activity variables, namely trading volume and return volatility. We use the "model-free" implied volatility index, VIX, and a realized volatility measure, RV, as proxies for the latter.

7.2.1 "Ideal" VPIN versus Return Volatility and Trading Volume

We start by reporting a key finding. Table 4 reveals a remarkably strong *negative* association of the ideal (RA) and RB-based VPIN measures with both trading volume and return volatility. This is eye-opening. It implies that VPIN computed from perfect trade classification has a

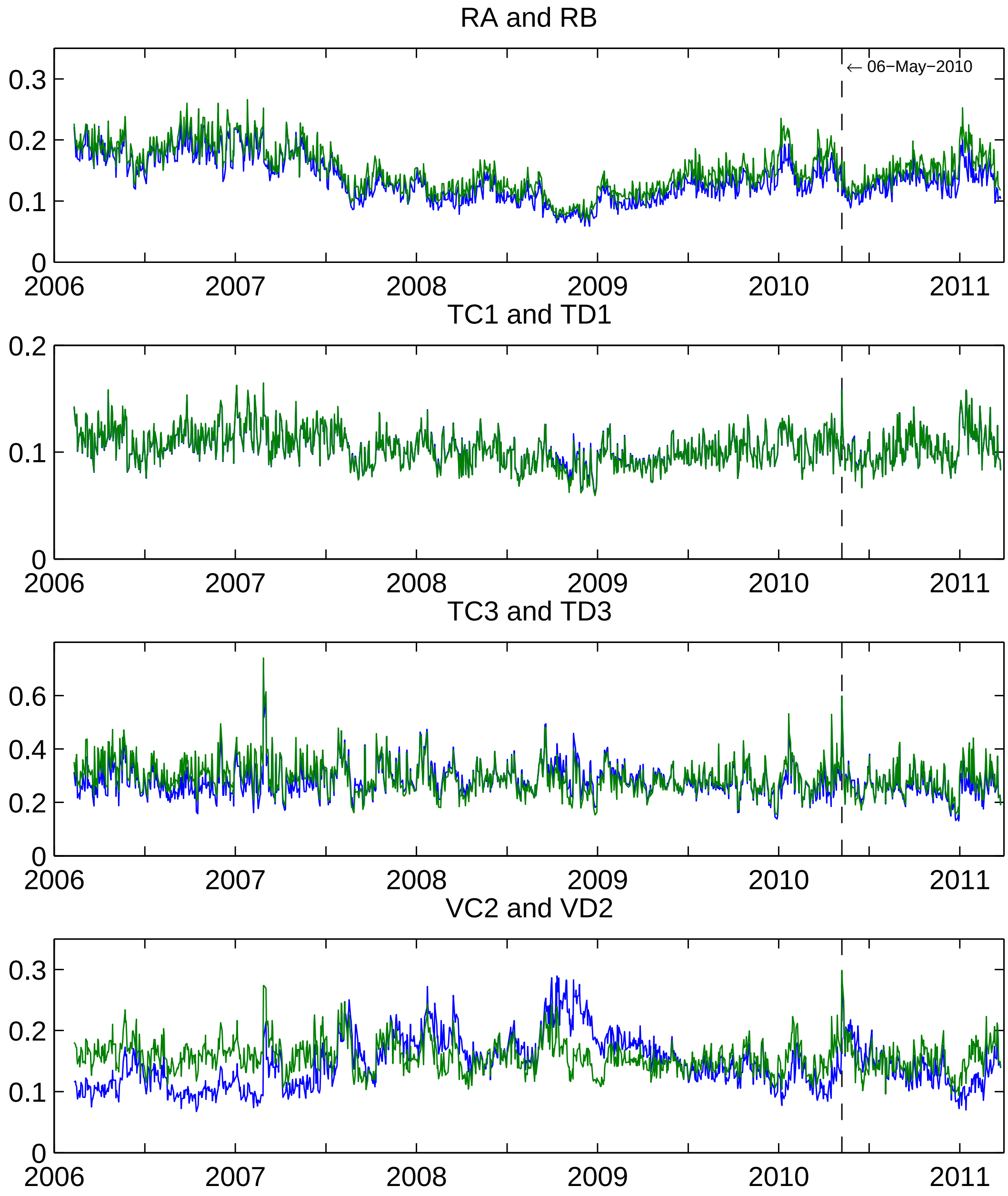


Figure 5: This figure plots the daily maximum values of various VPIN measures, February 10, 2006 - March 22, 2011. In each panel, the order of the plots is blue (first) and green (second).

fundamentally different relation to volatility and volume than the BVC-VPIN of ELO (2012a). Moreover, RB-VPIN mimics the main features of RA-VPIN very well, suggesting more generally that the former can proxy the latter and that, in this role, RB-VPIN will deliver results that are qualitatively identical to those derived from the ideal VPIN measure.

This surprising finding raises two separate questions. First, why do the VPIN measures obtained from trade classification using aggregate order flow produce diametrically opposite results to the one exploiting accurate classification? Second, why is there such a pronounced negative association between (ideal) VPIN and market activity variables in the first place? The remainder of this section explores the first issue by systematically exploring the properties of alternative VPIN implementations relative to volume, volatility and RA-VPIN. A full answer to the second question will take us beyond the current work, but we provide a more detailed look at the empirical factors that are responsible for the phenomenon in Section 9 below.

Table 4: **Correlations for VPIN Measures using Transaction-based Classification**

	RA	RB	Volume	VIX	RV
RA	1.00		-0.53	-0.72	-0.62
RB	0.96	1.00	-0.53	-0.73	-0.64

7.2.2 Time-Bar VPIN versus Return Volatility and Trading Volume

Table 5 reports corresponding correlations for VPIN measures based on time bars. TB has much lower correlation with volatility than established in AB (2014). This stems from the elimination of the trend in volume. Once we control for persistent shifts in volume, TB-VPIN is – particularly for small time bars – *negatively* related to volatility, as is the case for the transaction based measures. Second, the correlation of these indices with RA-VPIN declines monotonically with the aggregation level, suggesting that they deviate more strongly from the ideal metric when the classification accuracy deteriorates, as documented in Section 6. Third, this effect is also evident in increasing correlations between TB-VPIN and the activity variables (volume, VIX and RV) as a function of bar size. This is consistent with the assertion in AB (2014) that time-variation in volume (around trend, i.e., persistent yet mean-reverting fluctuations) severely distorts VPIN measures based on time bars. This effect stems from the induced variation in the number of bars within the volume buckets. Less bars lead to lower diversification across the bars within the bucket and a larger inferred order imbalance.

For additional evidence, it is useful to turn to the TE-VPIN series, constructed as in AB (2014). TE is compiled from the identical volume series and defines the bars in the same way as the other time-bar VPIN measures, but the trade classification is randomized to constitute an i.i.d. series with an equal chance that each bar is a buy or sell. Hence, *ceteris paribus*, it serves a control for the impact of the trade classification on VPIN. If the dramatic shift in the properties of TB-VPIN across different time bars, indeed, is driven primarily by the impact of time variation in the trading intensity, and not the trade classification per se, then the TE-VPIN series should display similar features to TB-VPIN. In fact, TE-VPIN is qualitatively identical to TB-VPIN in these respects, corroborating the above hypothesis. TE-VPIN has nearly the same correlation with RA and RB, and it displays the identical pattern in correlations with all the volume, volatility, TC- and TD-VPIN series as TB-VPIN across the different bar sizes.

Turning to TC- and TD-VPIN, we should encounter the same volume effect, but the role of the price changes in the BVC scheme introduces an additional factor, as discussed in the following section. In fact, Table 5 confirms that the 60- and 300-second bar TC measures are even more highly *positively* correlated with volume than TB. Moreover, as expected, TD is much less correlated with volume due to our normalization of the price changes by a realized volatility measure which dampens the inferred order imbalances during periods of (expected) high market activity and volatility. That is, controlling for predictable variation in volatility alters the properties of the TC-VPIN series substantially. Moreover, quite remarkably, the correlation between TE and RA *exceeds* the corresponding correlations of TC and TD with RA for every aggregation level. This, again, highlights the pivotal role of the volume pattern in shaping these VPIN measures. The speed of trading, and thus the number of bars per bucket, is clearly the key determinant for these time-bar VPIN measures. It suggests that the construction of order imbalances from the size of price changes, as for TC and TD, actively weakens the association with the ideal VPIN metric – random trade classification does better.

Table 5: **Correlations for Time Bar VPIN Measures**

$\delta = 1$

	TB	TC	TD	TE	RA	RB	Volume	VIX	RV
TB	1.00				0.90	0.92	-0.42	-0.66	-0.55
TC	0.71	1.00			0.63	0.62	-0.08	-0.39	-0.24
TD	0.74	1.00	1.00		0.65	0.65	-0.12	-0.44	-0.29
TE	0.85	0.58	0.61	1.00	0.83	0.87	-0.48	-0.72	-0.64

$\delta = 10$

	TB	TC	TD	TE	RA	RB	Volume	VIX	RV
TB	1.00				0.76	0.75	-0.16	-0.49	-0.34
TC	0.71	1.00			0.35	0.30	0.30	-0.04	0.15
TD	0.84	0.95	1.00		0.58	0.54	0.08	-0.29	-0.08
TE	0.84	0.59	0.76	1.00	0.76	0.76	-0.16	-0.53	-0.38

$\delta = 60$

	TB	TC	TD	TE	RA	RB	Volume	VIX	RV
TB	1.00				0.61	0.60	0.04	-0.34	-0.19
TC	0.65	1.00			0.06	0.01	0.55	0.21	0.38
TD	0.86	0.90	1.00		0.44	0.40	0.27	-0.15	0.05
TE	0.84	0.69	0.88	1.00	0.55	0.52	0.16	-0.28	-0.10

$\delta = 300$

	TB	TC	TD	TE	RA	RB	Volume	VIX	RV
TB	1.00				0.46	0.44	0.21	-0.20	-0.04
TC	0.55	1.00			-0.20	-0.26	0.71	0.42	0.55
TD	0.83	0.80	1.00		0.34	0.30	0.38	-0.07	0.14
TE	0.85	0.63	0.88	1.00	0.40	0.37	0.31	-0.14	0.05

In summary, our explicit control for the trend in volume and the impact of trade classification, via the TE-VPIN metric, enables us to document a close association between the properties of TB-VPIN for small time bars and the VPIN metric based on perfect classification. Moreover, the fundamentally different properties displayed by TB-VPIN for larger time bars are consistent with the time-variation in market activity increasingly distorting the measures as the order flow aggregation increases. We observe similar forces impacting the TC- and TD-VPIN metrics, but the use of the absolute price changes in classifying trade imbalances introduces a new element into the analysis. We pursue the latter point further below.

7.2.3 Volume-Bar VPIN versus Return Volatility and Trading Volume

We now focus on the VPIN metrics constructed from volume bars. The bottom panel of Figure 7.1 suggests that the VC measures are highly correlated with volatility, while VD is more stable throughout. The correlations assembled in Table 6 corroborate this point. Both VC metrics have correlations with VIX and RV exceeding 0.74, while VD is decidedly less correlated with the volatility variables – especially VIX – and the volume correlation shrinks rapidly with bar size. The stronger alignment of VD with RV relative to VIX and the overall higher correlations for smaller bars suggest that VD shares some commonality with the high-frequency variation in volatility, while the measures are largely unrelated over longer horizons. Finally, the VB measures display the same negative association to volatility and strong positive association with RA-VPIN as documented for the TB-VPIN series with small time bars. As such, they behave entirely differently than VC- and VD-VPIN, and the correlation patterns with return volatility and volume can hardly be more disparate.

Table 6: **Correlations for Volume Bar VPIN Measures**

$\nu = 0.02$

	VB	VC	VD	RA	RB	Volume	VIX	RV
VB	1.00			0.84	0.86	-0.45	-0.63	-0.53
VC	-0.59	1.00		-0.66	-0.71	0.69	0.76	0.80
VD	-0.17	0.71	1.00	-0.19	-0.24	0.55	0.28	0.46

$\nu = 0.10$

	VB	VC	VD	RA	RB	Volume	VIX	RV
VB	1.00			0.65	0.66	-0.28	-0.44	-0.35
VC	-0.32	1.00		-0.65	-0.69	0.68	0.74	0.78
VD	0.24	0.50	1.00	0.12	0.07	0.37	0.04	0.26

How do we rationalize the striking disparity in the behavior of VB-VPIN relative to VC- and VD-VPIN? By construction, the VB series annihilates the impact of time-variation in the trading volume. Thus, the dominant source of distortion in the metric has been eliminated, and for small bar sizes it should be reasonably compatible with the ideal VPIN measures. In fact, for $\nu = 0.02$ the correlation of VB-VPIN with RA- and RB-VPIN is high, and it displays remarkably similar correlations with VC- and VD-VPIN as well as the volume and volatility series as the RA-VPIN measure. This reaffirms the critical effect that time-variation in volume exerts on the behavior of the TB-VPIN metric.

In contrast, for VC- and VD-VPIN, in accordance with the BVC strategy, the size of the price changes over short intervals (volume bars) is the driving force behind the measures. As such, it is intuitively clear that the measure may share important features with regular realized volatility measures. For concreteness, our discussion below focuses on VC-VPIN which is the series most closely associated with RV.

In general, abstracting from the trade classification scheme, equations (1) and (3), plus the fact that $\nu_{q,\ell} = 1/Q$ for volume bars, imply,

$$\text{VC-VPIN}_\ell \;=\; \frac{1}{L} \sum_{\ell=1}^{L} \left| \frac{1}{Q} \sum_{q=1}^{Q} \gamma_{q,\ell} \right| . \tag{7}$$

Suppose now that the signed order imbalance measure, $\gamma_{q,\ell}$, is a monotone and increasing function $f(x)$ of the return over the bar and $f(0) = 0$, i.e., $\gamma_{q,\ell} = f(r_{q,\ell})$, where

$$r_{q,\ell} \;=\; \frac{P_{q,\ell} - P_{q-1,\ell}}{P_{q-1,\ell}} \;=\; \frac{\Delta P_{q,\ell}}{P_{q-1,\ell}}$$

is the simple return over bar q. We note that the signed order imbalance for VC is an example of a measure that satisfies the above characterization,

$$\gamma_{q,\ell} \;=\; 2 \cdot Z\left(\frac{P_{q-1,\ell}}{\sigma_{\Delta P}} \cdot r_{q,\ell}\right) - 1 \;=\; f\left(\frac{P_{q-1,\ell}}{\sigma_{\Delta P}} \cdot r_{q,\ell}\right),$$

with $f(x)$ monotonically increasing and $f(0) = 0$.

Of course, a similar characterization applies for realized volatility measures. Specifically, if $f(x) = x$ or $f(x) = \ln(1 + x)$, then $\gamma_{q,\ell}$ equals the simple or log returns, respectively.[22]

Since $\gamma_{q,\ell}$ always has the same sign as the return $r_{q,\ell}$, the following quantity,

$$\frac{1}{L}\sum_{\ell=1}^{L} |\,\gamma_\ell\,| \;=\; \frac{1}{L}\sum_{\ell=1}^{L}\left|\frac{1}{Q}\sum_{q=1}^{Q}\gamma_{q,\ell}\right| \;=\; \frac{1}{L}\sum_{\ell=1}^{L}\left|\frac{1}{Q}\sum_{q=1}^{Q} f(r_{q,\ell})\right| \tag{8}$$

proxies for realized volatility. To explore this hypothesis, we measure the coherence between the log-return, $f(x) = \ln(1 + x)$, and the bucket-wide signed order imbalance, γ_ℓ, for different VPIN measures.[23] For the BVC schemes VC1 and VC2, the correlations of γ_ℓ with $\ln(1 + r_\ell)$ are 0.86 and 0.84, respectively. In contrast, for the true classification RA, the correlation is an order of magnitude lower at 0.53. Hence, the design of the BVC measures mechanically boosts the correlation with return volatility, but renders them inaccurate indicators of actual order flow imbalances, as documented in Section 6.

Finally, we recall that VD normalizes the price changes with a factor reflecting recent realized volatility. Hence, VD controls for *expected* volatility and should be largely immune to persistent shifts in volatility, but will remain sensitive to volatility *innovations*. Therefore, we expect the association with volatility to be weaker for VD than VC. Moreover, VD should be more strongly correlated with RV than VIX, as the former captures volatility realizations more directly. These predictions are consistent with the correlations reported in Table 6.

In short, our results suggest that the VC- and VD-VPIN series serve as (suboptimal) realized volatility proxies. Moreover, given the identical dependence on price changes in the construction of the (time bar) TC- and TD-VPIN measures, these series will display identical qualitative features, albeit in a weaker form, as they, in addition, are subject to direct volume distortions arising from the use of time bars.

7.2.4 Intraday Order Flow Imbalances

We have argued that a number of empirical VPIN measures, almost mechanically, will be correlated with volume and volatility. If this explanation is correct, it must originate from an interaction of the activity variables with the order imbalance measures underlying the VPIN metrics. We now bring additional evidence to bear. It is well known that both volume and volatility display a pronounced intraday pattern. At the same time, we do not expect the actual (event time) order imbalances, or flow toxicity, to vary predictably and largely deterministically across the trading day. In other words, order imbalance measures, that are mechanically correlated with the activity variables, may inherit their intraday pattern. In

[22] The scaling factor $\frac{P_{q-1,\ell}}{\sigma_{\Delta P}}$ inside the transformation function for VC reflects the use of price changes rather than returns in the BVC classification as well as the normalization with the sample standard deviation of the price change. The latter feature constitutes another source of non-stationarity in VC-VPIN, as price changes grow larger when the asset price increases – returns rather than price levels are stationary. The use of a time-varying standard deviation, as in the VD measure, eliminates this concern. However, since the scaling factor changes little across a given bucket, this point may be ignored for the purposes of the current discussion.

[23] Note, γ_ℓ equals the log return over the entire bucket, i.e., $\ln(1 + r_\ell) = \ln\frac{P_{Q,\ell}}{P_{0,\ell}}$.

contrast, the actual order imbalances, as derived from ideal trade classification, should display a more diffuse variation across the trading day.

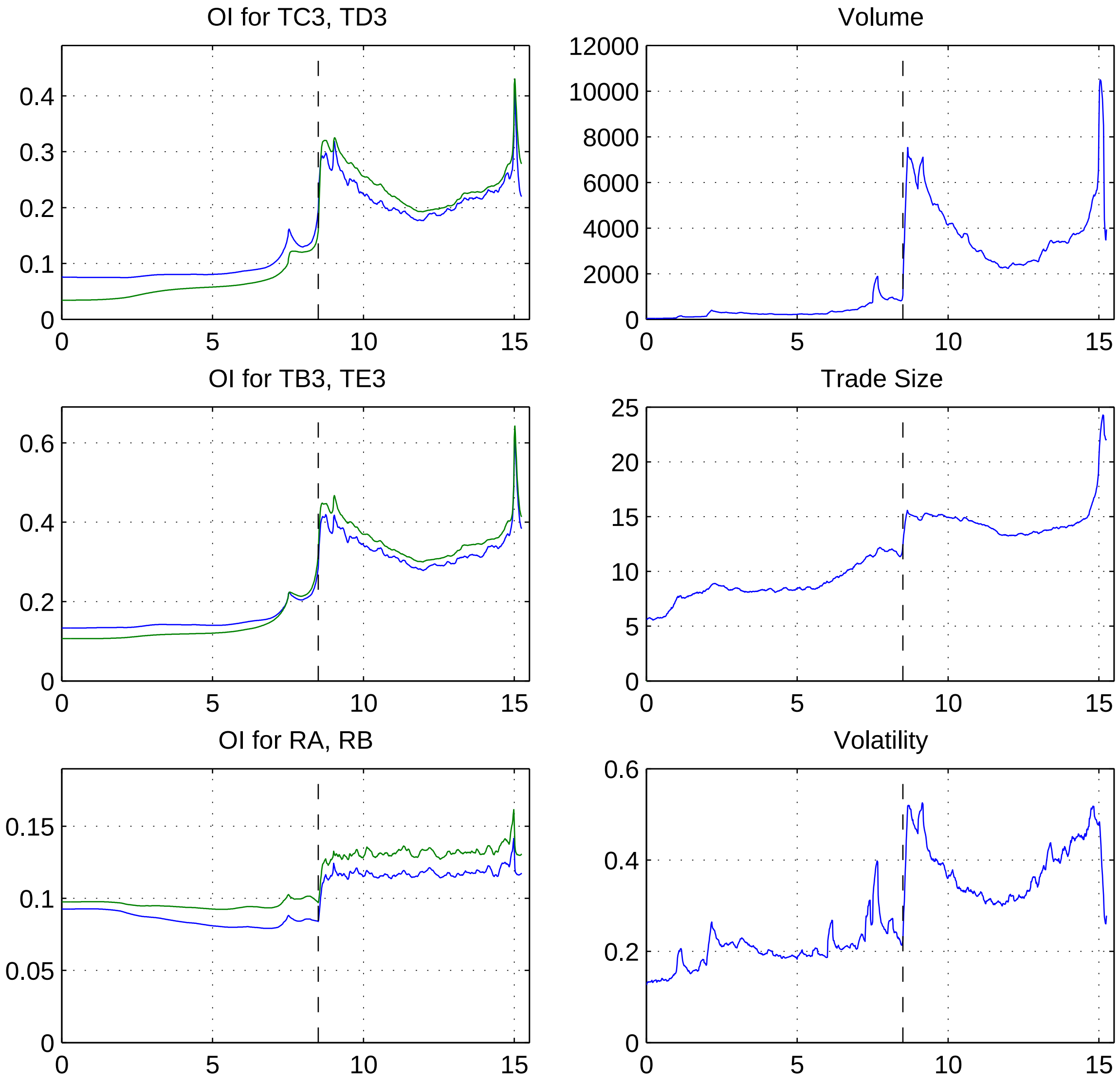


Figure 6: This figure depicts various intraday statistics from 0:00 to 15:15, averaged across all trading days from February 10, 2006 to March 22, 2011. The vertical dashed line represents the start of the regular trading hours. Left panels: Order Imbalance (OI) for trade classification schemes TC3, TD3, TB3, TE3, RA, and RB, where in each panel, the order of the plots is blue (first) and green (second). Right panels: trading volume, trade size, and volatility (square root of average return squared), where for each panel, the statistics are computed at the 1-minute frequency and then smoothed over a 10-minute window.

Figure 6 depicts the average intraday evolution of a set of activity measures, including alternative order imbalance measures as well as trading volume, trade size and return volatility. All the measures based on one-minute time bars in the two top panels of the left column display a pronounced U-shaped pattern across the regular trading hours. These curves resemble a mixture of the corresponding intraday variations in trading volume and trade size in the two top panels of the right column. The close association between TB and TE corroborates the findings of AB (2014) – the random trade classification underlying TE-OI induces the identical intraday behavior as observed for TB-OI. More strikingly, we observe the identical type of pattern for TC and TD as well. That is, the actual order imbalance does not matter. A purely randomized trade classification generates the identical pattern, corroborating the hypothesis that variation in overall market activity is the primary determinant of the VPIN measures.

Moving to the RA-OI series in the bottom panel of the left column, we find that the actual intraday evolution of trade imbalances is radically different from the depictions above. They are relatively low and stable prior to the opening of regular trading. Then, there is an initial jump at the market open, followed by a stable level throughout the regular trading hours. Just prior to the close of the equity market at 15:00, there is a spike in RA-OI, presumably due to pressure for market participants to reach their desired positions before the close of trading. The OI measure falls back to the prevailing level for the regular trading hours during the 15:00-15:15 post trading in the futures market. Finally, we note that the RB-OI series again provides a reliable proxy for the corresponding RA series.

In summary, using accurate trade classification, we find no association between the intraday patterns of order imbalances and trading intensity. In contrast, the order imbalances constructed from time bars display a pronounced pattern that mimics the diurnal features in the trading volume and trade size series. This "unconditional," or cross-sectional, finding complements the time series evidence in identifying the variation in trading intensity and volatility as major factors in driving – and distorting – the evolution of time bar VPIN.

8 The Incremental Predictive Power of VPIN

We have demonstrated that VPIN, constructed from actual order imbalances, is strongly *negatively* correlated with return volatility. At the same time, the VPIN metrics proposed by ELO (2011a, 2012a) display a pronounced *positive* correlation with volatility. Our analysis suggests that the discrepancy arises from systematic misclassification of the order imbalance by the VPIN implementations that rely on aggregate order flow for trade classification. These errors are correlated with market activity, e.g., trading volume and return volatility, thus inducing the positive correlation between VPIN and volatility for large time or volume bars.

These observations render the interpretation of existing empirical evidence ambiguous. The ability of VPIN to predict future return volatility and serve as a crash indicator – if indeed factual – is clearly not linked to the true underlying order imbalances. Instead, a potential explanation is that the nonlinear VPIN filtering procedure extracts useful information regarding toxic order flow – not regular (non-toxic) order flow – from the observed trade and price patterns. Thus, the true test regarding the usefulness of VPIN, as developed in ELO (2011a, 2012a), is the extent to which the measure provides auxiliary information beyond realized volatility and volume measures in forecasting future volatility and market disruptions.

In this section, we pursue a few different strategies. We test whether the VPIN, in general, has incremental predictive content. We also employ a dummy variable to isolate the predictive power during periods when the VPIN metric is extremely elevated. This amounts to a test of whether extreme VPIN values serve as useful signals regarding future market turmoil. Finally, in the spirit of an event study, we revisit the flash crash and investigate whether the VPIN measure potentially could have helped predict the impending market breakdown.

Table 7: **Forecast Regressions for Average Absolute Return**

Panel A: 5-Minute Forecast

Reg	Const	RA	RB	TB3	TC3	TD3	TE3	Vol	VIX	RV	$\bar{R}^2$
(1)	0.10	-0.50									19.71
	(25.50)	(-17.90)									
(2)	0.11		-0.52								22.35
	(25.43)		(-18.47)								
(3)	0.06			-0.07							1.32
	(14.90)			(-6.25)							
(4)	-0.02				0.22						9.91
	(-4.03)				(13.04)						
(5)	0.03					0.05					0.49
	(8.12)					(3.50)					
(6)	0.05						-0.04				0.27
	(9.66)						(-2.71)				
(7)	-0.01							0.26			33.42
	(-4.27)							(19.66)			
(8)	-0.01								0.22		48.42
	(-8.31)								(30.66)		
(9)	0.00									0.13	61.35
	(3.17)									(98.96)	
(10)	0.00				0.00					0.13	61.35
	(1.81)				(0.54)					(92.26)	

Panel B: 1-Day Forecast

Reg	Const	RA	RB	TB3	TC3	TD3	TE3	Vol	VIX	RV	$\bar{R}^2$
(1)	0.10	-0.50									33.30
	(26.14)	(-18.34)									
(2)	0.11		-0.52								37.40
	(25.98)		(-18.84)								
(3)	0.07			-0.08							3.36
	(16.27)			(-7.82)							
(4)	-0.01				0.18						11.87
	(-1.94)				(12.14)						
(5)	0.04					0.02					0.09
	(11.04)					(1.28)					
(6)	0.06						-0.06				1.15
	(11.48)						(-4.49)				
(7)	-0.01							0.24			48.02
	(-2.66)							(18.97)			
(8)	-0.01								0.21		77.85
	(-8.33)								(32.87)		
(9)	0.00									0.18	82.30
	(0.56)									(26.33)	
(10)	0.00				0.00					0.18	82.30
	(0.06)				(0.46)					(24.60)	

Notes: This table reports OLS forecast regressions for the average absolute 1-minute return (AAR), defined as $AAR(t, t+T) = 1/T \sum_{i=1}^{T} |r_{t+i}|$ and $r_t = 100 \ln(P_t/P_{t-1})$ is 1-minute return. Forecasts are formed every 5 minutes during regular trading hours. The forecast horizon T is 5 minutes or 1 day (defined as 405 1-minute intervals of regular trading hours), respectively. "Vol" is the one-day backward trading volume, "RV" is the realized volatility over the previous 1 hour (Panel A) or 1 day (Panel B) . Shown in parentheses are t-statistics based on HAC-standard errors with 81 lags.

8.1 Predictive Regressions

This section explores volatility forecast regressions to further assess the properties of the alternative VPIN measures. ELO (2012a) use the positive association between TC3-VPIN and future volatility to argue that VPIN may serve as a proxy for order flow toxicity.

Table 7 provides predictive regressions for return volatility over five minutes and one day ahead. The majority of the regressions are univariate so that we can identify the forecast performance of the individual variables. The first two regressions involve RA and RB. The transactions-based VPIN metrics are robustly and negatively related to future volatility. This finding is consistent with the evidence presented above. AB (2014) reach the identical conclusion and conjecture this arises from a thinning of the order book during volatile periods. We explore this feature in further depth in Section 9 below.

The next forecast variable, TB3, is closely related to the measure studied by ELO (2011a, 2011b, 2011c). AB (2014) document that this metric is correlated with volatility, but the information content regarding future return variation is subsumed by the (observable) VIX measure. Moreover, AB found a simple "uninformative" metric to dominate TB in terms of forecasting future volatility although it, by construction, has no relation to order imbalance. Instead, this association arose as the measure (TE) inherits the mechanical correlation with trading volume induced by the time bars. Here, we annihilate the majority of this correlation via volume detrending. The implication is evident in Table 7. Neither TB nor TE are positively associated with the future absolute returns and their explanatory power is essentially zero. The last VPIN metrics investigated are TC3 and TD3. Not surprisingly, TC3 has significant predictive power for future volatility – it inherits some of the realized volatility features associated with the VC3 series. Nonetheless, the explanatory power is limited with R^2 values of about 10% and 12% for the two horizons.[24] Moreover, normalizing the price changes by a recent volatility measure annihilates the predictive power, as evidenced by the regressions for TD3. This suggests that TC3 has predictive power due to its correlation with realized volatility which stems from the use of price changes to infer order imbalance. This is consistent with the vastly superior performance of either trading volume, VIX, or RV. For example, volume provides a three-fold and RV a six-fold improvement in explanatory power relative to TC3 at the five-minute horizon, and the discrepancy is even larger at the daily level.

Finally, we include an encompassing regression where TC3 and RV are joint explanatory variables for future volatility. The coefficient on TC3 is now insignificant and literally zero to three decimals. It adds nothing to the predictive power of RV. Thus, in general, TC3-VPIN is irrelevant for predicting future volatility: traditional real-time indicators like realized volatility and trading volume are vastly superior and subsume the information content of VPIN.

8.2 Extreme VPIN Realizations

The prior section documents that the information content of VPIN is subsumed by realized volatility in terms of predicting future volatility. Since VPIN is a sluggishly moving variable, constructed as a rolling moving average over 50 volume buckets, this result has even stronger implications. It suggests that realized volatility can predict the future VPIN and, furthermore, the predicted component, VPIN^P, is the one with forecast power for future volatility, while the residual component, VPIN^R, should possess little, if any, predictive power.

To test the above conjecture, we first construct a predictive regression for VPIN, based

[24] Ironically, the forecast power associated with the *negative* relation between RB and future volatility is substantially higher that implied by the positive correlation between TC3 and future volatility.

solely on RV observations obtained $\Delta = 5$ minutes earlier,

$$\text{VPIN}_t^P \;=\; a_0 \,+\, a_1 \cdot \text{RV}_{t-\Delta} \,+\, \text{VPIN}_t^R. \tag{9}$$

Next, we generalize the predictive regression from Table 7 involving TC3-VPIN by including the predictive and residual component of TC3-VPIN jointly in the regression. The findings are reported in the first row of Table 8. We find that the anticipated component of VPIN is highly significant, while the residual component is insignificant. That is, the only relevant information in VPIN is what may, a priori, be distilled from past realized volatility, while the independent variation in VPIN plays no role. We further note that RV explains only a small part of the overall variation in VPIN, with an R^2 of about 14%, while the explanatory power of equation (9) is almost identical to that of RV itself (R^2 of 82%) in Table 7. Hence, the variation in RV gets encoded almost one-for-one in TC3-VPIN, but VPIN is primarily driven by separate forces unrelated to future RV. In sum, the variation in VPIN not directly associated with past RV only serves to obfuscate the relevant information. Moreover, the latter is a major obstacle for predicting RV from VPIN, as VPIN itself only provides an R^2 of around 12%, as indicated in Table 7.

Table 8: **Two-Stage Predictive Regressions for one-day-ahead Realized Volatility**

Reg	Const	VPIN^P	VPIN^R	$\text{VPIN}^P\cdot$D	$\text{VPIN}^R\cdot$D	$\bar{R}^2$
(1)	-1.41	6.55	-0.04			81.97
	(-25.94)	(29.38)	(-1.33)			
(2)	-1.41	6.57	-0.04	-0.21	0.30	82.00
	(-25.68)	(29.07)	(-1.30)	(-1.62)	(1.77)	

Notes: The table reports the results of a two-stage predictive regression for one-day-ahead realized volatility (RV). In the first stage, VPIN is projected onto 5-minute lagged RV to obtain the "projected" and "residual" components of VPIN, i.e., VPIN = VPIN^P + VPIN^R. Specifically, we obtain $\text{VPIN}^P = 0.22 + 0.14\,\text{RV}$, with $\bar{R}^2 = 14.02\%$. In the second stage, VPIN^P and VPIN^R are used to forecast future RV. The forecast horizon is one day and forecasts are formed every 5 minutes during regular trading hours. For VPIN, we use the TC3 metric. In specification (2), we also introduce the dummy variable D, which equals one when VPIN exceeds the $99th$ percentile, and zero otherwise. t-statistics based on HAC-standard errors with 81 lags are provided in parentheses.

8.3 Extreme Realizations of VPIN

So far, we have studied general properties of VPIN. One specific goal of ELO (2011a, 2012a) is to develop a warning signal for impending market turmoil. As such, the association between extreme readings of VPIN and subsequent market conditions is of particular interest.

8.3.1 The General Statistical Evidence

To explore whether extreme VPIN readings are particularly useful as signals for future market turmoil, we include a dummy variable in the two-stage predictive regression, indicating times at which VPIN takes values in the top 99% percentile of the distribution. The second row of Table 8 shows that these observations carry no additional predictive power for future return volatility over-and-above what is already encoded in the "predictive" component associated with the past variation in RV. Likewise, the increase in R^2 is negligible. Hence, in general, we find no incremental information associated with truly extreme VPIN realizations.

8.3.2 An Event Study: The Flash Crash

ELO (2011a) find that VPIN attains a historical high during the day of the flash crash. Figure 3 reveals that, for our TC3-VPIN series, this distinction instead belongs to February 27, 2007. Furthermore, two days later, on March 1, 2007, we encounter another extreme VPIN value.[25] Nonetheless, to facilitate comparison with ELO, we focus on the flash crash. Qualitatively similar results are obtained for the other dates with extreme VPIN realizations.

The panels in the top rows of Figure 7 depict alternative VPIN series, while the panels in the bottom rows refer to activity variables during the regular trading hours on May 6, 2010. The figure exploits the same timing convention for VPIN as in Figure 1. The price series, replicated from Figure 1, serves as a reference for the other measures.

First, we note that the VIX and price series are near mirror images of each other up through the crash. Thus, VIX responds instantaneously to the price development, but does not provide auxiliary information beyond the price path.[26] The normalized volatility depicts the standard deviation of one-minute price changes obtained over a (centered) 10-minute window divided by $\bar{\sigma}$, the unconditional volatility used to construct BV-VPIN. The normalized volume series refers to trading volume per one-minute bar divided by the size of the volume bucket. When this value exceeds unity, trading is so vigorous that buckets are filled in less than one minute, and there is only one or two (broken) time bars within a given bucket.

Before discussing the time bar VPIN metrics, we focus on RA-VPIN, obtained from correct trade classification. This ideal VPIN does not display any unusual pattern. In fact, consistent with earlier findings, it drops off during the actual crash. Thus, the ideal VPIN metric does *not* signal a rise in order flow toxicity – even if the (true) signed order imbalance (SOI) in Figure 1, generated from the same classification procedure, indicates increasing selling pressure from around 12:00. Moreover, we observe the qualitatively identical development for RB-VPIN, which is constructed from trade data alone and does not require order book data. In short, the VPIN transformation of the SOI measure *destroys* whatever information regarding developing order imbalances that is conveyed by SOI. It follows that other VPIN metrics can generate a positive association with market turbulence only by systematically misclassifying trades in a manner that is correlated with volatility and volume. Indeed, this is what happens for the schemes relying on larger sized time bars, as documented in Section 6.

Inspecting the evolution of the remaining VPIN measures in Figure 7, we corroborate these predictions. As the time bars grow larger, the TB, TE and TC metrics begin mimicking the volatility and volume patterns more closely. In addition, there is little sign of any predictive content regarding an upcoming crash. The dramatic increases in VPIN occur strictly *after* the initiation of the crash at 13:32, when both volume and realized volatility also spike. This is almost a tautology, as VPIN is generated by a moving average of past order imbalance measures. When large volume and volatility innovations occur, they impact the current order imbalance measures, but they receive only $(1/50)^{\text{th}}$ of the weight in the VPIN computation. Large increases in VPIN are only feasible if there is a persistent upward shift in the level of OI,

[25] This "dethroning" of the flash crash is contingent on our "detrended V" approach. Under the ELO (2012a) "constant V" procedure, Figure 3 shows that the maximal VPIN value is, by far, during the flash crash.

[26] The subsequent extreme oscillations in the VIX measure arise from rapid fluctuations in the liquidity of the S&P 500 option market in the aftermath of the crash; see Andersen, Bondarenko and Gonzalez-Perez (2011).

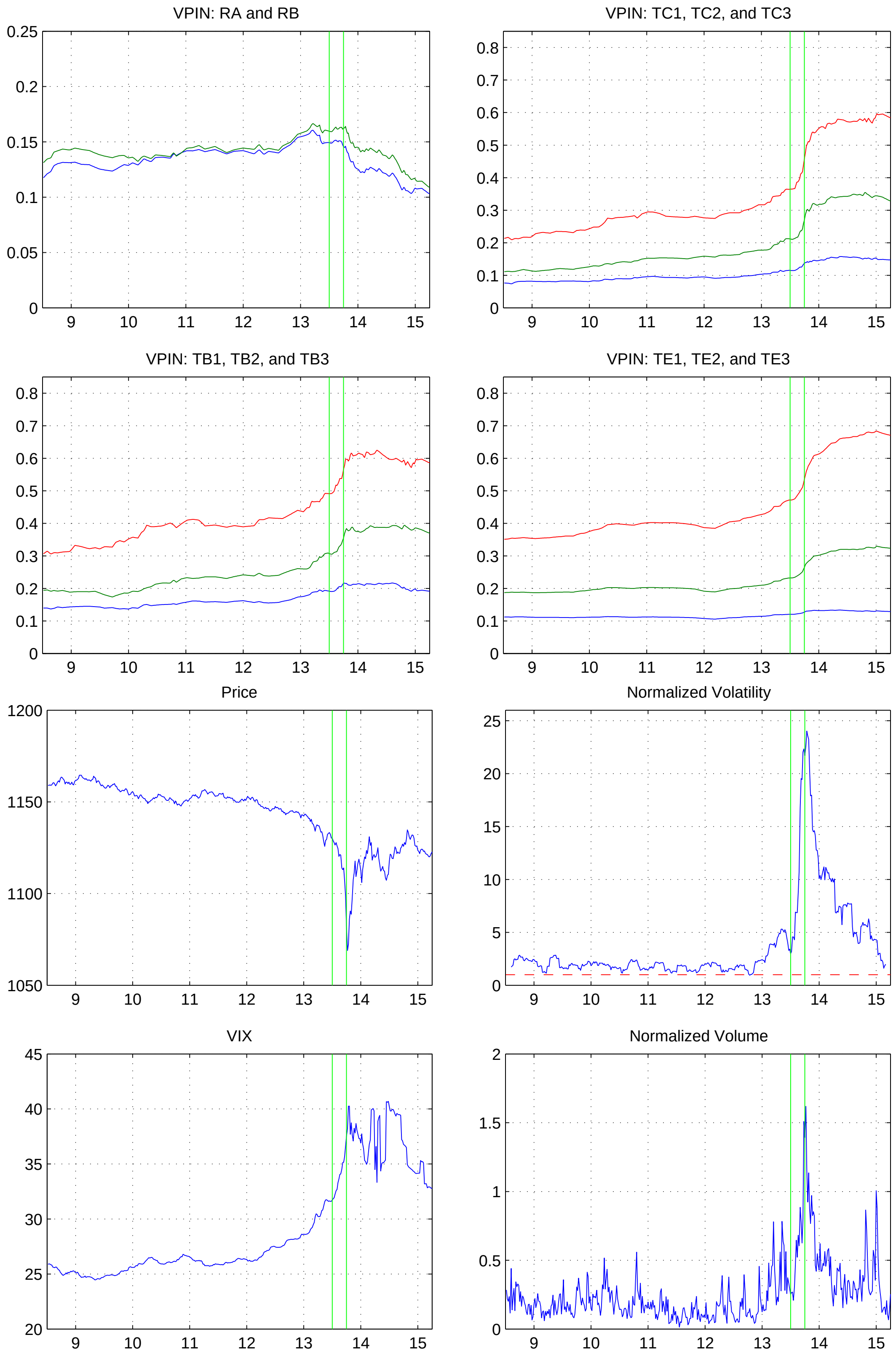


Figure 7: May 6, 2010. The solid vertical lines indicate the timing of the flash crash. In panels with multiple plots, the order of the plots is blue (first), green (second), and red (third). The normalized volatility is the ratio of the standard deviation of one-minute price changes for a (centered) 10-minute window over the unconditional volatility $\bar{\sigma}$. The normalized volume is the trading volume per one-minute bar divided by the size of the volume bucket.

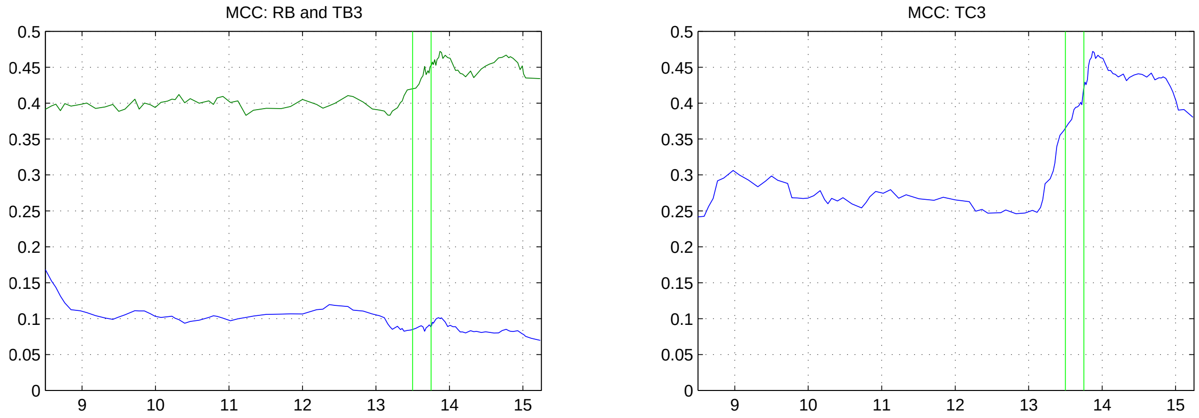


Figure 8: May 6, 2010. MCC is computed as the moving average over 10 volume buckets. The solid vertical lines indicate the timing of the flash crash. MCC stands for misclassification at the contract-by-contract level. In the first panel, the order of the plots is blue (first) and green (second).

and when this happens, it will be long-lived due to the smoothing: VPIN is, by construction, slowly moving. If (time bar) VPIN is to predict a crash, the order imbalance measures must rise well in advance of the event. Consequently, bursts in market activity precede sharp increases in (time bar) VPIN. As such, these activity variables should be superior predictors of impending volatility which is, of course, what the predictive regressions in Section 8.1 and 8.2 show. Specifically, our TC3 metric is approximately 0.292 at 12:30, about an hour prior to the crash, and 0.365 at 13:30. These values fall in the percentiles 62 and 91 of the empirical distribution of daily maximum values of TC3, as displayed in Figure 3, so we would tend to observe such values eight and two times *every month*, respectively. Only the post-crash realizations are exceptional, but these values reflect the crash and do not predict it. At the same time, RA-VPIN computed from accurate order imbalance measures drops to values well below the sample average. The extreme movement in TC3-VPIN is a testament to the volume- and volatility-induced distortions in the time-bar based order imbalance measures.

To further illustrate this effect, Figure 8 depicts the contract-by-contract misclassification rates for RB and TB3 in the left panel and TC3 in the right panel. While the TB3 classification is poor throughout, there is an extraordinarily sharp deterioration in precision for TC3 just prior to the crash. This coincides with the explosive growth in the normalized volatility and trading volume in Figure 7. As before, the time bar BVC order imbalance measure responds primarily to concurrent volatility and volume innovations and becomes increasingly distorted as the market activity rises.

9 Why Are "Ideal" VPIN and Volatility Inversely Correlated?

Our empirical analysis has documented a surprising fact: VPIN computed from actual order imbalances is systematically and significantly *negatively* correlated with return volatility. This finding must stem from market microstructure features that are ignored by the theory and implementation strategy behind the VPIN measures in ELO (2011a, 2012a). The finding effectively invalidates the VPIN approach. VPIN relies on the extraction of toxicity related signals from the underlying order flow. Given the negative correlation with the actual order imbalance, the only reason some VPIN measures can generate a positive correlation with future volatility, and market turmoil in general, is through systematically biased trade classification.

This still leaves the fundamental question of why the ideal VPIN metric is inversely related to return volatility unanswered. A thorough analysis of this phenomenon will take us beyond the scope of the current study. Nonetheless, we can identify the empirical regularities in the tick-by-tick data that induce the negative association, thus facilitating future work on rationalizing these robust features of the data.

9.1 Trade Size and Clustering in the Trade Direction

The primary determinant behind the true VPIN metric is, tautologically, the actual order flow imbalance across the volume buckets. Thus, we need to understand the clustering of active buy and sell transactions. At the micro level, a basic driver of clustering is the number of contracts exchanged per transaction. When more than one contract is bought or sold, it constitutes a sequence of unidirectional trades for the individual contracts. Likewise, it matters whether there is a tendency towards continuation rather than reversal of the trade direction, i.e., whether a buy transaction is more likely to be followed by another buy rather than a sell. These two variables, the (average) size of transactions and the probability of a continuation are, jointly, the key determinants of the buy-sell diversification. The larger the transactions and the more clustered the trade direction, the fewer effective buy and sell sequences we have over the bucket, and the larger the order flow imbalances tend to be.

We explore the trade sequence dynamics empirically, using the true trade classification. For a given volume bucket, we define the Average Trade Size (ATS), the Average Trade Run (ATR), i.e., the average number of consecutive trades in the same direction, and the Average Volume Run (AVR), i.e., the average number of consecutive contracts traded in the same direction. These measures are closely related as AVR = ATR · ATS. A large value for AVR is tantamount to having fewer, but longer, buy and sell sequences within a bucket. This can occur because of a high average trade size, ATS, and a strong trade continuation, i.e., large values for ATR.[27] Table 9 and Figure 9 summarize our findings.

Table 9: **Correlations for ATS and ATR**

Panel A: The whole sample

	ATS	ATR	Volume	VIX	RV
ATS	1.00	-0.36	-0.51	-0.79	-0.66
ATR	-0.36	1.00	0.17	0.16	-0.01

Panel B: The second half of the sample

	ATS	ATR	Volume	VIX	RV
ATS	1.00	0.48	-0.28	-0.84	-0.72
ATR	0.48	1.00	-0.45	-0.61	-0.66

Table 9 documents a pronounced negative association between the average trade size and return volatility. In contrast, the trade runs are only mildly correlated with volatility, so the net effect is a strong negative correlation between volatility and AVR, and thus also RA-VPIN and volatility. The primary explanation is that when volatility increases, the trade size drops and we have more alternation between buy and sell sequences within each bucket.

[27] The probability of a trade continuation (PC) is directly related to ATR as PC = 1 - 1/ATR.

The negatively correlation between ATS and ATR will tend to stabilize AVR and the order imbalance measure. However, Panel B documents that, in the second part of our sample, ATR is instead *positively* correlated with ATS, so they reinforce each other in generating variation in AVR and VPIN. Moreover, since both variables now are strongly inversely related to volatility, the negative association between volatility and RA-VPIN only strengthens. The pronounced negative correlation between trade size and volatility is also evident from Figure 9.

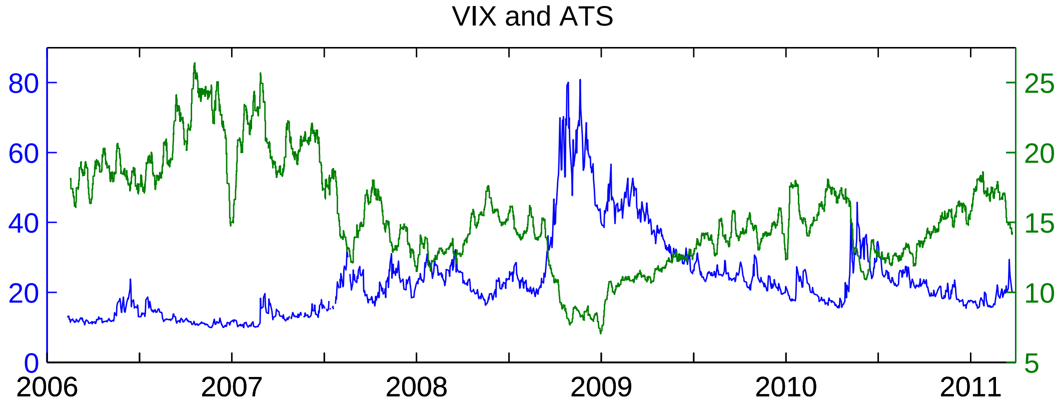


Figure 9: This figure plots VIX (the left scale, blue) and ATS (the right scale, green). ATS is computed as a 5-day moving average.

Consequently, beyond rationalizing and quantifying the extreme inverse relation between volatility and trade size, a challenge for future research is to understand the forces behind the fluctuations in ATR. One potential explanation for the pronounced change in the correlation pattern is the emergence of high-frequency trading (HFT). However, without information linking trading activity to individual accounts, it remains purely speculative to associate these developments with the increasing role of HFT in the trading process.

9.2 Evidence from the Flash Crash

ELO (2011a) stress that historically high VPIN readings could have served as a warning signal for the flash crash. As such, our finding of a negative association between "true" VPIN and return volatility in Section 8.3.2 is particularly striking. To understand whether the preceding analysis also applies to the market dynamics during this uniquely stressful episode, we depict the evolution of the relevant measures for the day of the flash crash in Figure 10.

The figure reveals that the factors identified above are out in full force. Prior to the crash, from 11:00-13:00, the average trade size, ATS, is above 15 contracts, while the average trade run, ATR, is just below four, generating an average volume run, AVR, of more than 60. Hence, we have alternating sequence of buying and selling of about 60 contracts on average. In the run-up, during, and in the aftermath of the crash, this number drops precipitously and ends up below 30, i.e., the number of alternating buy and sell sequences within a bucket more than doubles, generating a lower order imbalance measure. Of course, in reality the crash is characterized by selling pressure, as illustrated by Figure 1. However, this manifests itself through a rapid escalation in volume and only a small imbalance per bucket. The cumulative effect generated by a small, but persistent, imbalance is missed by VPIN, as the offset generated by the drop in the trade size and trade run dominates.

In summary, the VPIN metrics compiled by ELO (2011a, 2012a) attain inflated values on the day of the flash crash for the wrong reasons. The true VPIN measure, constructed from

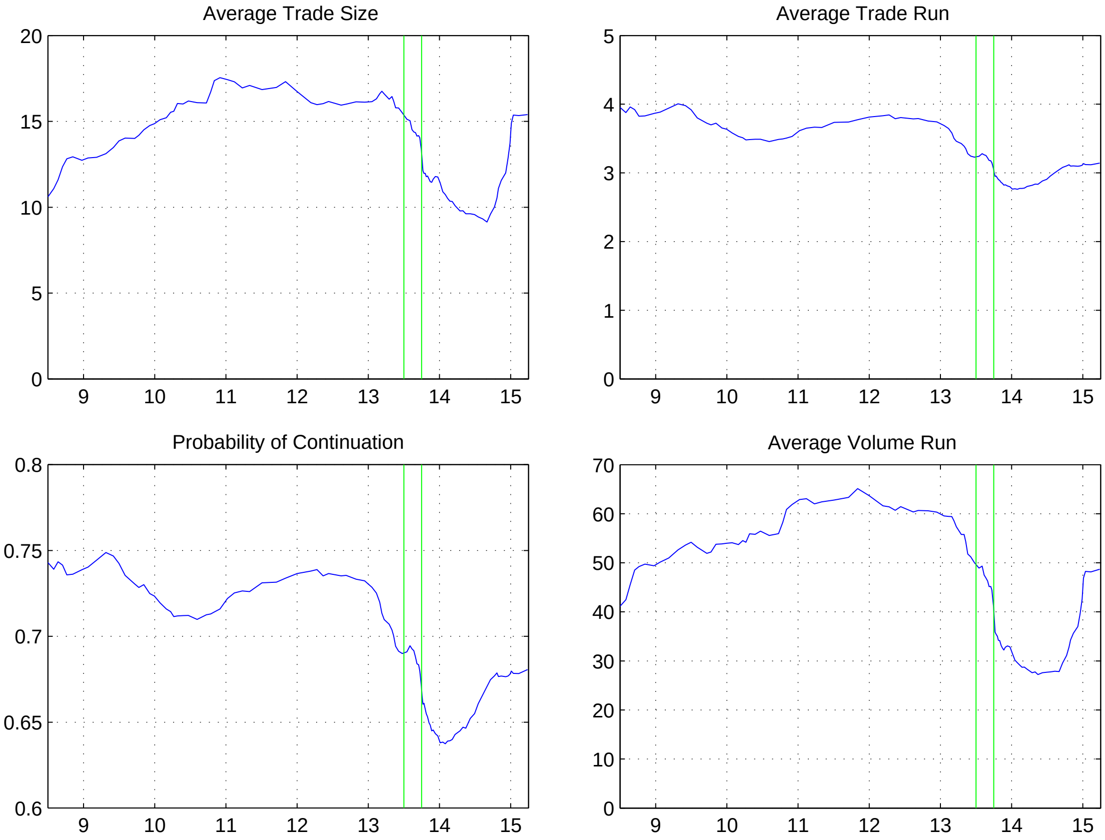


Figure 10: This figure plots ATS, ATR, PC, and AVR on May 6, 2010. The measures are computed as the moving average over 10 volume buckets. The solid vertical lines indicate the timing of the flash crash.

the actual order imbalances, *dives rather than soars*. The VPIN metrics generated from trade classification based on large bar sizes display the exact opposite correlation with volatility only because they systematically misclassify the order flow. The trade classification is instead driven by innovations to volume and volatility. This also explains why these VPIN metrics contain zero incremental predictive power for future volatility beyond a simple realized volatility measure. The problem is with the VPIN metric itself: if implemented correctly, it provides no insight into the evolving order imbalances – a standard cumulative order imbalance measure based on the tick rule is vastly superior.

10 Conclusion

This paper seeks to settle a brewing controversy regarding the usefulness of the VPIN metric as a real-time indicator of order flow toxicity and, by extension, predictor of impending market stress. We follow the basic construction of VPIN in ELO (2012a), but provide a few modifications to avoid excessive distortions in the metric due to the pronounced volume trend over our sample. The main innovation is to construct VPIN from a (near) perfect trade classification scheme. This metric constitutes an ideal VPIN which can be used as a benchmark to assess the implications of using alternative trade classification techniques. We document a systematic deterioration in the classification accuracy, as the schemes deviate from transaction-based identification. Moreover, we demonstrate how these classification errors induce a pronounced correlation with market activity variables such as trading volume and return volatility.

Our findings imply that the VPIN metric has no useful association with order flow toxicity in the S&P 500 futures market. Existing empirical results reaching the opposite conclusion are based on distorted VPIN metrics that induce a correlation with volume and volatility, by construction. The predictive content of these measures for future volatility are subsumed by traditional real-time indicators, such as realized volatility. In fact, VPIN constructed from actual order imbalances displays a pronounced negative association with return volatility. The bottom line is that the signed cumulative order flow, based on perfect trade classification, has an economically meaningful relation to concurrent price movements. Prices rise (fall) when there is excess active buying (selling) in the market. However, applying the nonlinear VPIN transformation to the (true) signed order flow generates a metric that is strongly negatively correlated with future volatility and has no sensible association with market stress.

We briefly explore the forces generating the pronounced negative correlation between transaction-based VPIN and volatility. It is natural to conjecture that it arises from a thinning of the order book during volatile periods, leading to smaller transaction sizes and more bid-ask bouncing. This suggests it may be promising to construct warning signals for impending disruptions directly from the state of the order book. The combination of diminishing depth at the top layers of the book and increasing spreads is likely to precede any major (predictable) disruption. Typically, the signed cumulative order imbalance and the short-term price trend will also be indicative of emerging tensions in the market. Future research may find such variables useful in constructing real-time warning signals.

References

Abad, D., Yague, J., 2012. "From PIN to VPIN: An introduction to order flow toxicity," *The Spanish Review of Financial Economics* 10, pp. 74–83.

Aitken, M., and A. Frino, 1996, "The Accuracy of the Tick Test: Evidence from the Australian Stock Exchange," *Journal of Banking and Finance*, 20, pp. 1715-1729.

Andersen, T.G., 1996, "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility," *Journal of Finance*, 51, pp. 169-204.

Andersen, T.G., and T. Bollerslev, 1998, "Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts," *International Economic Review*, 39, pp. 885-905.

Andersen, T.G., and O. Bondarenko, 2007, "Construction and Interpretation of Model-Free Implied Volatility," *Volatility as an Asset Class*, London: Risk Books, I. Nelken (editor), pp. 141-181.

Andersen, T.G., and O. Bondarenko, 2014, "VPIN and the Flash Crash," *Journal of Financial Markets*, forthcoming.

Andersen, T.G., O. Bondarenko, and M.T. Gonzalez-Perez, 2011, "A Corridor Fix for VIX: Constructing a Coherent Model-Free Option Implied Volatility Measure," Working Paper.

Asquith, P., R. Oman and C. Safaya (2010), "Short Sales and Trade Classification Algorithms," *Journal of Financial Markets*, 13, pp. 157-173.

Boehmer, K., J. Grammig, and E. Thiessen, 2007, Estimating the Probability of Informed Trading – Does Trade Misclassification Matter?" *Journal of Financial Markets*, 10, pp. 26-47.

Bethel, E. W., Leinweber. D., Rubel, O., and K. Wu, 2012. "Federal Market Information Technology in the Post-Flash Crash Era: Roles for Supercomputing," *Journal of Trading.*

CFTC-SEC, 2010, "Preliminary Findings Regarding the Market Events of May 6, 2010," *Joint Commodity Futures Trading Commission (CFTC) and the Securities and Exchange Commission (SEC) Advisory Committee on Emerging Regulatory Issues*, May 18, 2010.

Chakrabarty, B., R. Li, V. Nguyen, and R. Van Ness, 2007, "Trade Classification Algorithms for Electronic Communications Network Trades," *Journal of Banking and Finance*, 31, pp. 3806-3821.

Chakrabarty, B., P.C. Moulton and A. Shkilko (2012), "Short Sales, Long Sales, and the Lee-Ready Trade Classification Algorithm Revisited," *Journal of Financial Markets*, 15, pp. 467-491.

Chakrabarty, B., R. Pascual and A. Shkilko, 2012, "Trade Classification Algorithms: A Horse Race between the Bulk-based and the Tick-based Rules," Working Paper, SSRN, December 2012

Clark, P.K., 1973, "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," *Econometrica*, 41, pp. 135-155.

Corcoran, C., 2013. "Systemic Liquidity Risk and Bipolar Markets: Wealth Management in Today's Macro Risk On / Risk Off Financial Environment." Wiley.

Easley, D., N.M. Kiefer, M. O'Hara, and J.B. Paperman, 1996,"Liquidity, Information, and Infrequently Traded Stocks," *Journal of Finance*, 51, pp. 1405-1436.

Easley, D., M. López de Prado, and M. O'Hara, 2011a,"The Microstructure of the "Flash Crash": Flow Toxicity, Liquidity Crashes, and the Probability of Informed Trading," *Journal of Portfolio Management*, 37 (2), pp. 118-128.

Easley, D., M. López de Prado, and M. O'Hara, 2011b, "The Exchange of Flow Toxicity," *Journal of Trading* 6 (2), pp. 8-13.

Easley, D., M. López de Prado, and M. O'Hara, 2011c, "Measuring Flow Toxicity in a High-Frequency World," Working Paper, SSRN, February, 2011.

Easley, D., M. López de Prado, and M. O'Hara, 2012a, "Flow Toxicity and Liquidity in a High-Frequency World," *Review of Financial Studies* 25, pp. 1457-1493.

Easley, D., M. López de Prado, and M. O'Hara, 2012b, "Bulk Classification of Trading Activity," Working Paper, SSRN, March 2012.

Easley, D., M. López de Prado, and M. O'Hara, 2012c, "Bulk Classification of Trading Activity," Working Paper, SSRN, August 2012.

Easley, D., M. López de Prado, and M. O'Hara, 2012d, "The Volume Clock: Insights into the High Frequency Paradigm." *Journal of Portfolio Management*, 39 (1), pp. 19-29.

Easley, D., M. López de Prado, and M. O'Hara, 2013, "Optimal Execution Horizon," *Mathematical Finance*; forthcoming.

Ellis, K., R. Michaely, and M. O'Hara, 2000, "The Accuracy of Trade Classification Rules: Evidence from NASDAQ," *Journal of Financial and Quantitative Analysis*, 35, pp. 529-551.

Epps, T.W., and M.L. Epps, 1976, "The Stochastic Dependence of Security Price Changes and Transactions Volumes: Implications for the Mixture of Distributions Hypothesis," *Econometrica*, 44, pp. 305-321.

Finucane, T.J., 2000, "A Direct Test of Methods for Inferring Trade Direction from Intraday Data," *Journal of Financial and Quantitative Analysis*, 36, pp. 553-576.

Kirilenko, A., A.S. Kyle, M. Samadi and T. Tuzun, 2011,"The Flash Crash: The Impact of High Frequency Trading on an Electronic Market," Working Paper, SSRN, May 2011.

Lee, C. and M. Ready, 1991,"Inferring Trade Direction from Intraday Data," *Journal of Finance*, 46, pp. 733-746.

MacIntosh, J.G., 2013. "High Frequency Traders: Angels or Devils?" Commentary 391, C.D. Howe Institute; Toronto, Ontario, Canada; October 2013.

Menkveld, A.J., Yueshen, B.Z., 2013. "Anatomy of the Flash Crash," Working Paper. SSRN, April 2013.

Odders-White, E. 2000, "On the Occurrence and Consequences of Inaccurate Trade Classification," *Journal of Financial Markets*, 3, pp. 259-286.

Tauchen, G.E., and M. Pitts, 1983, "The Price Variability-Volume Relationship on Speculative Markets," *Econometrica*, 51, pp. 485-505.

Yildiz, S., R.A. Van Ness, and B.F. Van Ness, 2013, "Analysis Determinants of VPIN, HFTs' Order Flow Toxicity and Impact on Stock Price Variance," Working Paper, University of Mississippi; September 2013.

Wei, W. C., Gerace, D., and Frino, A., 2013. "Informed Trading, Flow Toxicity and the Impact on Intraday Trading Factors," *Australian Accounting Business and Finance Journal* 7, pp. 3–24.

Wu, K., Bethel, W., Gu, M., Leinweber, D. and Ruebel, O., 2013, A Big Data Approach to Analyzing Market Volatility," June 5, 2013, SSRN: http://ssrn.com/abstract=2274991.

Web Appendix

A TB-VPIN Algorithm

This appendix reproduces algorithm for VPIN computation based on the Tick-Rule trade classification (TB in our classification) as implemented in ELO (2011c) and used in ELO (2011a, 2011b).

1. ALGORITHM FOR COMPUTING THE-VPIN METRIC

In this appendix, we describe the procedure to calculate *Volume-Synchronized Probability of Informed Trading*, a measure we called the *VPIN* informed trading metric. Similar results can be reached with more efficient algorithms, such as re-using data from previous iterations, performing fewer steps or in a different order, etc. But we believe the algorithm described below is illustrative of the general idea.

One feature of this algorithm to note is that we classify all trades within each one minute time bar as either buys or sells using a tick test. We do not have data which directly identifies trades as buyer-initiated or seller-initiated so some classification procedure is necessary. One could classify each trade separately or one could classify trades in groups of an alternative size (based on either time or volume). Different schemes will lead to different levels of VPIN. We have used a variety of schemes and, as one would expect, cutting the data more finely leads to reduced levels of VPIN– measured trade becomes more balanced when groups of trades in say a one-minute time bar may be classified differently. However, our focus is on how rare a particular VPIN is relative to the distribution of VPINs derived from any classification scheme, and this is unaffected by the classification schemes we have examined (including trade-by-trade as well as groups based on one-tenth of a bucket). We focus on one minute time bars as this data is less noisy, more widely available and easier to work with.

1.1. INPUTS

1. Time series of transactions of a particular instrument (T_i, P_i, V_i) :
   a. T_i: Time of the trade.
   b. P_i: Price at which securities were exchanged.
   c. V_i: Volume exchanged.
2. V: Volume size (determined by the user of the formula).
3. n: Sample of volume buckets used in the estimation.

P_i, V_i, V, n are all integer values. T_i is any time translation, in integer or double format, sequentially increasing as chronological time passes.

1.2. PREPARE EQUAL VOLUME BUCKETS

1. Sort transactions by time ascending: $T_{i+1} \geq T_i$, $\forall i$.
2. Expand the number of observations by repeating each observation P_i as many times as V_i. This generates a total of $I = \sum_i V_i$ observations P_i.
3. Re-index P_i observations, $i = 1, \ldots, I$.
4. Initiate counter: $\tau = 0$.
5. Add one unit to τ.
6. If $I < \tau V$, jump to step 10 (there are insufficient observations).
7. $\forall i \in [(\tau - 1)V + 1, \tau V]$, classify each transaction as *buy* or *sell initiated*:

   a. A transaction i is a *buy* if either:

      i. $P_i > P_{i-1}$, or

      ii. $P_i = P_{i-1}$ and the transaction $i-1$ was also a *buy*.

   b. Otherwise, the transaction is a *sell*.

8. Assign to variable V_τ^B the number of observations classified as *buy* in step 7, and the variable V_τ^S the number of observations classified as *sell*. Note that $V = V_\tau^B + V_\tau^S$.

9. Loop to step 6.

10. Set $L = \tau - 1$ (last bucket is always incomplete or empty, thus it will not be used).

1.3. APPLY VPINs FORMULA

If $L \geq n$, there is enough information to compute $VPIN_L = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{\sum_{\tau=L-n+1}^{L} (V_\tau^S + V_\tau^S)} = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{nV}$.

B TC-VPIN Algorithm

This appendix reproduces algorithm for VPIN computation based on the Bulk-Volume Classification (BVC) (TC in our classification) as implemented in ELO (2012c, on-line Appendix) and ELO (2012b).

1. ALGORITHM FOR COMPUTING THE-VPIN METRIC

In this appendix, we describe the procedure to calculate *Volume-Synchronized Probability of Informed Trading*, a measure we called the *VPIN* flow toxicity metric. Similar results can be reached with more efficient algorithms, such as re-using data from previous iterations, performing fewer steps or in a different order, etc. But we believe the algorithm described below is illustrative of the general idea.

One feature of this algorithm is that we apply a probabilistic approach to classify the volume exchanged within each 1-minute time bars. We cannot expect users to have data which unequivocally identifies trades as buyer-initiated or seller-initiated so some classification procedure is necessary. One could classify each trade separately or one could classify trades in aggregates of an alternative size (based on either time, number of trades or volume bars). Different schemes will lead to different levels of VPIN. We have used a variety of schemes and conclude that data aggregation leads to better flow toxicity estimates than working on raw transaction data.[28] Data granularity has an impact on the magnitude of VPIN levels, however our focus is on how rare a particular VPIN is relative to the distribution of VPINs derived from any classification scheme, and this is unaffected by the classification schemes we have examined (including trade-by-trade as well as groups based on one-tenth of a bucket). We focus on 1-minute time bars as this data is less noisy, more widely available and easier to work with.

1.1. INPUTS

1. Time series of transactions of a particular instrument (T_i, P_i, V_i) :

   a. T_i: Time of the trade.

   b. P_i: Price at which securities were exchanged.

   c. V_i: Volume exchanged.

2. V: Volume size (determined by the user of the formula).

3. n: Sample of volume buckets used in the estimation.

[28] For an in-depth analysis, please refer to ELO (2012b).

P_i, V_i, V, n are all integer values. T_i is any time translation, in integer or double format, sequentially increasing as chronological time passes.

1.2. PREPARE EQUAL VOLUME BUCKETS

1. Sort transactions by time ascending: $T_{i+1} \geq T_i$, $\forall i$.
2. Compute ΔP_i, $\forall i$.
3. Expand the number of observations by repeating each observation ΔP_i as many times as V_i. This generates a total of $I = \sum_i V_i$ observations ΔP_i.
4. Re-index ΔP_i observations, $i = 1, \ldots, I$.
5. Initiate counter: $\tau = 0$.
6. Add one unit to τ.
7. If $I < \tau V$, jump to step 11 (there are insufficient observations).
8. $\forall i \in [(\tau - 1)V + 1, \tau V]$, split volume between *buy* or *sell initiated*:

   a. $V_\tau^B = \sum_{i=(\tau-1)V+1}^{\tau V} Z\left(\frac{\Delta P_i}{\sigma_{\Delta P}}\right)$

   b. $V_\tau^S = \sum_{i=(\tau-1)V+1}^{\tau V} \left[1 - Z\left(\frac{\Delta P_i}{\sigma_{\Delta P}}\right)\right] = V - V_\tau^B$

9. Assign to variable V_τ^B the number of observations classified as *buy* in step 8, and the variable V_τ^S the number of observations classified as *sell*. Note that $V = V_\tau^B + V_\tau^S$.
10. Loop to step 6.
11. Set $L = \tau - 1$ (last bucket is always incomplete or empty, thus it will not be used).

1.3. APPLY VPINs FORMULA

If $L \geq n$, there is enough information to compute $VPIN_L = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{\sum_{\tau=L-n+1}^{L} (V_\tau^S + V_\tau^S)} = \frac{\sum_{\tau=L-n+1}^{L} |V_\tau^S - V_\tau^S|}{nV}$.

C Tick-Rule and Bulk-Volume Classification

This appendix reproduces algorithms for Tick-Rule and Bulk-Volume Classification (BVC) as implemented in ELO (2012b).

1. TICK-RULE IMPLEMENTATION

Here we present a simple implementation of the Tick Rule in Python language. More efficient implementations exist, but we believe the one outlined below is the clearest.

```
queryCurs is assumed to contain the output of a SQL query such as

queryCurs.execute('SELECT Price, Volume, VolBuy FROM ' + tablename + '
ORDER BY Instrument, Time')
```

VolBuy is the field that stores the Volume from traders initiated by an aggressive buyer, as reported by the Exchange. The tick list variable will accumulate the amount matched over the entire volume. The rest of the code is self-explanatory.

```
a =queryCurs.fetchone()
flag, price, tick=1, a[0], [0,0]
while True:
   try:
      a=queryCurs.fetchone()
      # tick rule
      if a[0]>price:
         flag=1
      elif a[0]<price:
         flag=2
      if flag==1:
         tick[0]+=a[2] #correctly classified as buy
      else:
         tick[0]+=a[1]-a[2] #correctly classified as sell
      tick[1]+=a[1] #volume to be classified
      # reset price
      price=a[0]
   except:
      break
```

2. BULK VOLUME CLASSIFICATION IMPLEMENTATION

An equivalent codification of the BVC algorithm would be as follows. stDev is a real variable storing the volume weighted Standard Deviation of price changes across bars. The amount matched over the entire volume is stored in the list variable bulk.

```
a =queryCurs.fetchone()
price, bulk=a[0], [0,0]
while True:
   try:
      a=queryCurs.fetchone()
      # bulk classification
      z=float(a[0]-price)/stDev
      z=scipy.stats.norm.cdf(z)
      bulk[0]+=min(a[1]*z,a[2]) #correctly classified as buy
      bulk[0]+=min(a[1]*(1-z),a[1]-a[2]) #correctly classified as sell
      bulk[1]+=a[1] #volume to be classified
      # reset price
      price=a[0]
   except:
      break
```

Research Papers 2013

CREATES
Center for Research in Econometric Analysis of Time Series

2013-25: Nima Nonejad: Time-Consistency Problem and the Behavior of US Inflation from 1970 to 2008

2013-26: Nima Nonejad: Long Memory and Structural Breaks in Realized Volatility: An Irreversible Markov Switching Approach

2013-27: Nima Nonejad: Particle Markov Chain Monte Carlo Techniques of Unobserved Compdonent Time Series Models Using Ox

2013-28: Ulrich Hounyo, Sílvia Goncalves and Nour Meddahi: Bootstrapping pre-averaged realized volatility under market microstructure noise

2013-29: Jiti Gao, Shin Kanaya, Degui Li and Dag Tjøstheim: Uniform Consistency for Nonparametric Estimators in Null Recurrent Time Series

2013-30: Ulrich Hounyo: Bootstrapping realized volatility and realized beta under a local Gaussianity assumption

2013-31: Nektarios Aslanidis, Charlotte Christiansen and Christos S. Savva: Risk-Return Trade-Off for European Stock Markets

2013-32: Emilio Zanetti Chini: Generalizing smooth transition autoregressions

2013-33: Mark Podolskij and Nakahiro Yoshida: Edgeworth expansion for functionals of continuous diffusion processes

2013-34: Tommaso Proietti and Alessandra Luati: The Exponential Model for the Spectrum of a Time Series: Extensions and Applications

2013-35: Bent Jesper Christensen, Robinson Kruse and Philipp Sibbertsen: A unified framework for testing in the linear regression model under unknown order of fractional integration

2013-36: Niels S. Hansen and Asger Lunde: Analyzing Oil Futures with a Dynamic Nelson-Siegel Model

2013-37: Charlotte Christiansen: Classifying Returns as Extreme: European Stock and Bond Markets

2013-38: Christian Bender, Mikko S. Pakkanen and Hasanjan Sayit: Sticky continuous processes have consistent price systems

2013-39: Juan Carlos Parra-Alvarez: A comparison of numerical methods for the solution of continuous-time DSGE models

2013-40: Daniel Ventosa-Santaulària and Carlos Vladimir Rodríguez-Caballero: Polynomial Regressions and Nonsense Inference

2013-41: Diego Amaya, Peter Christoffersen, Kris Jacobs and Aurelio Vasquez: Does Realized Skewness Predict the Cross-Section of Equity Returns?

2013-42: Torben G. Andersen and Oleg Bondarenko: Reflecting on the VPN Dispute

2013-43: Torben G. Andersen and Oleg Bondarenko: Assessing Measures of Order Flow Toxicity via Perfect Trade Classification