Received: 28 July 2016 | Accepted: 19 December 2017
DOI: 10.1002/ijfe.1608

RESEARCH ARTICLE

WILEY

Examining drivers of trading volume in European markets

Bogdan Batrinca | Christian W. Hesse | Philip C. Treleaven

Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK

Correspondence
Bogdan Batrinca, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK.
Email: bogdan.batrinca.09@ucl.ac.uk

Funding information
Engineering and Physical Sciences Research Council (EPSRC) UK

JEL Classification: C32; C52; C58; G12; G15; G17

Abstract
This study presents an in-depth exploration of market dynamics and analyses potential drivers of trading volume. The study considers established facts from the literature, such as calendar anomalies, the correlation between volume and price change, and this relation's asymmetry, while proposing a variety of time series models. The results identified some key volume predictors, such as the lagged time series volume data and historical price indicators (e.g. intraday range, intraday return, and overnight return). Moreover, the study provides empirical evidence for the price–volume relation asymmetry, finding an overall price asymmetry in over 70% of the analysed stocks, which is observed in the form of a moderate overnight asymmetry and a more salient intraday asymmetry. We conclude that volatility features, more recent data, and day-of-the-week features, with a notable negative effect on Mondays and Fridays, improve the volume prediction model.

KEYWORDS
asymmetric models, behavioural finance, European stock market, feature selection, price volume relation, trading volume

1 | INTRODUCTION

This study investigates the drivers affecting the trading volume with an in-sample analysis. We explore the interaction between truly exogenous determinants and trading volume. Several hypotheses are evaluated while looking at the previous literature, where various factors are discussed in isolation, and we propose a liquidity extraction model by placing these findings in a unified context.

Identifying the drivers of trading volume is crucial in order to anticipate and minimize market impact, by accurately sizing and executing orders. Achieving optimal order sizing relies on precise volume prediction, that is, planning trades and deciding how much to trade given the current market context and the predicted volume levels. To better illustrate the importance of trading volume, some recent facts include the total turnover value, which was \$63tn in 2011 (World Federation of Exchanges, 2012) and \$49tn in 2012 (World Federation of Exchanges, 2013). The NYSE's turnover averaged more than 100% between 2004 and 2009, with 138% in 2008 (NYSE Euronext, 2016), meaning that the entire market value has changed hands once a year, although it has decreased to significantly lower levels during the following years, averaging 72% for the 2010–2015 period.

In order to better understand the factors affecting the trading volume, it is necessary to survey and combine apparently disjoint literature concepts. We start by reviewing the relevant areas of the behavioural finance literature. Here, a large amount of research has mainly investigated the calendar effects on price returns, and there is very little emphasis on the calendar effects on trading volume. We particularly focus on the day-of-the-week effect, which, once investigated, can formulate several hypotheses to analyse other calendar effects (e.g. the effect of stock index futures expiries and cross-market holidays). We then connect the behavioural finance findings with evidence from the literature on the relation between price changes and volume (i.e. the price–volume lead–lag effect). Following this reverse path, we test the direct relation between calendar effects, represented in this study by the day-of-the-week effect, and volume.

Behavioural finance mainly consists of regression models built on a collection of indicator variables, implying a certain limitation with regard to its statistical significance. We propose a model based on lagged time series and lagged smoothed time series in order to explain observed volumes in terms of recent time series; this follows the behavioural finance paradigm and represents market dynamics on the run, while assuming stationarity and disregarding outliers. However, the financial data are a strong nonstationary and nonconstant mean time series, due to the existence of notable event dates (e.g. MSCI rebalance dates, futures expiry dates, and company earnings announcement dates). This analysis aims to bridge the gap between behavioural finance and traditional finance and explores the feasibility of a potential special event effect (e.g. futures expiries or cross-market holidays) on trading volume by starting with an analysis of the day-of-the-week effect on trading volume. The financial markets are event-driven, and their dynamics are permanently shifting. Therefore, it is important to predict the trading volume at different time horizons.

The main motivations of this study include the following: the insufficiency of literature looking at the calendar effects on trading volume (and not on returns), the inconclusive results of the price–volume relation and whether it is characterized by asymmetry, and the abundance of studies investigating certain volume determinants in complete isolation from other types of volume drivers.

Out of a total number of 55 surveyed articles, which are all cited in this study and investigate the price–volume relation and the day-of-the-week effect, only seven of them use data sets after 2000 and none of the cited papers employs market data after 2006. Moreover, only seven studies include a few European stocks or indices among their international data sets, and only two papers are based on European data sets exclusively. Given the lack of a broad European stock universe and post-2000 data sets, we employ an extensive pan-European stock universe consisting of 2,353 stocks, for which we use daily market data between 1 January 2000 and 10 May 2015, and we also test for structural breaks by comparing the results before and after the financial crisis of 2007–2008.

The aim of this study is to define a unified volume prediction model, while exploring the endogenous variables in conjunction with exogenous variables and performing feature selection. We investigate a pan-European stock universe for a sample period of over 15 years in order to test the improvement of an autoregressive volume model, by sequentially adding features, such as volatility, more recent data, and day-of-the-week, and test additional hypotheses such as the existence of an asymmetric price–volume relation. The rest of the paper is organized as follows: Section 2 provides a literature review of the main research topics addressed by this study: volume dynamics, price actions, and volume–price correlations, along with a survey of the relevant calendar effects; Section 3 introduces the sample data set, whereas Section 4 outlines the main models and the analytical approach; Section 5 describes the methodology of the trading volume study, while gradually introducing the various variables we are examining in order to better predict the volume; Section 6 exhibits the main results of the various volume prediction models and the types of price–volume asymmetry; Section 7 provides a conclusion of this study, together with a discussion on the results and potential suggestions for other researchers to further extend this study.

2 | BACKGROUND

The literature review starts by setting the context of this study, that is, why volume prediction is important, followed by a review of studies on the types of the price–volume relation and its potential asymmetry. We then switch to the behavioural finance literature by outlining the main calendar effects and elaborate on the day-of-the-week effect.

2.1 | Trading volume historical dynamics

Trading volume is extraordinarily large across developed stock exchanges, and many interesting patterns in prices and returns are closely related to the volume movement; volume is highly used in conjunction with price actions. For instance, the volume of high-priced "glamour stocks" tends to be larger than the volume of low-priced value stocks, and a stock with higher trading volume tends to have lower future returns (Hong & Stein, 2007). Trading volume is a strong indicator of economic activity.

Auctions account for a high trading volume, and there are three types of auctions: opening auctions, intraday auctions, and closing auctions. A normal day starts with pretrading auctions or opening auctions, in order to set the price after the nontrading hours during the night, when news came out, and is followed by continuous trading. In Europe, this phase can be temporarily halted by volatility interruptions, which trigger a 2- to 5-min auction, called intraday auction, in case the price is changing more than ±5%, in order to set the price correctly.

The literature has mostly examined the relation between trading volume and the following three variables: (a) bid-ask spread (i.e. negatively correlated); (b) price changes—the literature has predominantly found a positive correlation between volume and the absolute price change; (c) information—a volume increase means that the investors interpret the information either differently or identically by beginning with different priors; the market institutional design affects volume around the informational events (Karpoff, 1986).

2.2 | Volume–price relation

The relation between trading volume and prices is important in order to better understand the financial market structure. Price changes indicate the market response to new information, and the trading volume measures the level of disagreement of the information among investors (Beaver, 1968).

There is wide evidence in the literature (Harris & Raviv, 1993; Hong & Stein, 2007) for a positive correlation between trading volume and price dynamics. Volume has been found to be positively correlated either with the magnitude (i.e. absolute value) of the price change (Assogbavi & Osagie, 2006), that is, $|\Delta p|$, or with the price change per se (i.e. the raw value of the price change), that is, Δp (Karpoff, 1987), where price changes can be represented as log-price difference or percentage price change.

Moreover, Godfrey, Granger, and Morgenstern (1964) reported a modest correlation between volume and the difference between the daily high and low, that is, the intraday range, which will also be employed as a price metric in this study.

2.2.1 | Volume–price relation asymmetry

Given that there is a potential relation between trading volume and price change, we further investigate the price change representation, by taking into account the evidence on the asymmetric relation between trading volume and price changes, as previously shown in the finance literature. Having an in-depth understanding of the asymmetry in price indicators helps fit the models more accurately. More specifically, instead of having a single feature for the magnitude of a price indicator (e.g. intraday or overnight price returns), we can discriminate between positive and negative values and represent the magnitude of each sign independently, that is, one feature for the magnitude of positive values and another feature for the magnitude of negative values, which would result in two different coefficients when regressing. The following two equations summarize the symmetry and asymmetry of the volume/price relation.

$$Symmetry\text{:}\left(v_t|\Delta p_t^+\right) = \left(v_t|\Delta p_t^-\right) \tag{1}$$

$$Asymmetry\text{:}\left(v_t|\Delta p_t^+\right) > \left(v_t|\Delta p_t^-\right) \text{ or } \left(v_t|\Delta p_t^+\right) < \left(v_t|\Delta p_t^-\right) \tag{2}$$

Ying's early work (1966) was the first to draw attention to the asymmetry in the volume–price relation, finding evidence that price asymmetry exhibits greater volume when associated with a price increase than when associated with a price decrease. Furthermore, Jain and Joh (1988) used hourly volume and returns data from NYSE, concluding that the relation between volume and price returns is steeper for positive price changes. This was subsequently confirmed by other researchers, such as Epps (1975, 1977), Smirlock and Starks (1985), and Al-Deehani (2007), who also offered a possible interpretation for the price–volume relation asymmetry, namely, the fact that short sellers respond faster to information stimulating price change than long investors, causing higher volume on price upticks. Conversely, Woord, McInish, and Ord (1985) and Moosa, Silvapulle, and Silvapulle (2003) found a reverse asymmetry, where the volume/price change ratio is smaller for upticks than for downticks in stock markets. The results of these studies imply that the absolute price changes must depend on whether the price change is positive or negative, which will be an important decision in the price indicator variables across this study. Other common explanations for the existence of asymmetry include the optimistic and pessimistic investors' disparity of opinion or the higher costs of short selling compared to the costs of taking long positions (Assogbavi, Khoury, & Yourougou, 1995).

2.3 | Calendar effects

This research builds on the top of previous literature by merging disjoint findings on volume prediction, price–volume relation, and the asymmetry of this relation. However, we complement these findings by introducing some exogenous variables from the behavioural finance literature that could potentially drive trading volume. Given the abundance of isolated papers analysing the calendar effects on price returns and the papers discussing the price–volume relation, there is an auspicious context for exploring a direct relation between calendar effects and trading volume. Moreover, in this particular study, the day-of-the-week effect is investigated in a volume prediction context, along with the endogenous variables based on time-lagged volume and price-related metrics; this is opposite to most of the behavioural finance articles on calendar effects, where the authors define dummy variables only for particular effects in complete isolation from the endogenous predictors. Discovering an explicit relation between trading volume and the days of the week would allow us to subsequently investigate the effect of futures expiries and cross-market holidays, because their abnormal returns might potentially explain the day-of-the-week effect and they could impact on the trading volume. The market is typically in a steady state with a relatively constant price formation process that drives the fairly expected price and volume metrics. However, when special events occur, such as the futures expiries or cross-market holidays, the market is in a different condition during these days and calls for a state-switching model.

2.3.1 | Behavioural finance and calendar effects

The behavioural finance literature introduced several anomalies (e.g. calendar effects) that affect prices. This contradicts the traditional paradigm that markets are efficient and suggest that markets switch to different states that disturb the equilibrium.

In the 1960s, Eugene Fama introduced the efficient market hypothesis, defining an efficient market as one that efficiently processes information, that is, prices fully reflect the publicly available information at a given time (Fama, 1969). This hypothesis is shared among the finance traditionalists and was the driver of an opposing view from behaviourists, who explored various stock return patterns that violate the market efficiency. The field of behavioural finance explains the decision-making process of investors and its consequences on the market movements.

Behaviourists analysed a huge amount of samples of past market data and identified evidence of market inefficiency in the form of anomalies, which can either occur once or follow a periodic pattern. The most popular anomalies include the calendar effects, medium-term momentum, value effect, size effect, and postearnings announcement drift.

The calendar effects are market anomalies that involve a sudden change in the behaviour of stock markets at certain times of the year. These event-driven irregularities have been documented in a wide range of studies. Some of the most interesting calendar effects include the weekend effect (and more generally the day-of-the-week effect), the month-of-the-year effect, the January effect, the holiday effect (and more specifically the cross-market holiday effect), the expiry day effect, and the intramonth effect.

2.3.2 | The weekend (day-of-the-week) effect

The weekend effect (or Monday effect) consists of a lower closing price on Monday than the closing price of the previous Friday. It is a particular instance of the broader day-of-the-week effect. The literature on calendar effects focuses on the connection between these effects and returns; extremely few articles investigate the impact of calendar effects on trading volume, and hence, it is important to first understand the findings on calendar effects and price returns and then connect them with the insights on the price–volume relation, in order to infer a direct link between calendar effects and trading volume.

The weekend effect is intriguing because empirical results contradict the expectation to have higher returns on Monday, because its returns reflect three consecutive days. The average return for Monday is negative (French, 1980; Gibbons & Hess, 1981; Jaffe & Westerfield, 1985; Pettengill, 2003), as Cross (1973) first indicated that Monday returns are significantly different from Friday returns. The weekend effect has been widely documented in the literature (Dubois & Louvet, 1996; Harris, 1986; Abraham & Ikenberry, 1994). Other authors found noncyclical patterns for the day-of-the-week effect, which could be explained by other calendar effects: Rogalski (1984) found that the day-of-the-week returns are connected to the January, firm size, and turn-of-the-year anomalies, whereas Wang, Li, and Erickson (1997) found that the Monday effect occurs mainly in the last 2 weeks of the month (i.e. the fourth and fifth weeks).

Contrarily, the research of Steeley (2001) suggests that the weekend effect in the UK stock prices has disappeared after 1990, whereas Smirlock and Starks (1986) conclude that this weekend return is positive. More confusingly, Brusa, Liu, and Schulman (2000) confirmed the existence of a weekend effect for small firms, but reported the existence of a reverse weekend effect for medium- and large-sized firms, where Monday returns are positive and significantly greater than the average of the other four weekdays.

Berument and Kiymaz (2001) found a day-of-the-week effect in both returns (with highest returns on Wednesday and lowest returns on Monday) and volatility (with highest volatility on Friday and lowest on Wednesday); later, they discovered that the maximum and minimum days are different across international markets, with highest volatility occurring on Thursdays in the UK (Berument & Kiymaz, 2003). As for the relation between the day-of-the-week and the trading volume, Lakonishok and Maberly (1990) found a relative increase in the individuals' trading activity on Mondays.

Potential justifications of the strong Monday effect include general measurement-error explanations (Keim & Stambaugh, 1984), the delay between trading and stocks settlement, and in clearing checks (Lakonishok & Levi, 1982), the individual investors' trading pattern (i.e. selling pressure) on Monday (Lakonishok & Maberly, 1990), and, partially, the institutional behaviour (Flannery & Protopadakis, 1988; Sias & Starks, 1995).

Most of the literature on calendar effects consists of an ample collection of studies conducted on isolated one-off models applied to certain past samples of market data. Because the calendar effects are highly data-driven and the interdependence of economic variables is ambiguous, the calendar effects have been investigated ex post and usually the stock universe of the data samples used in the studies is too narrow in order to draw a generalized conclusion. Besides the small stock universe, most of the studies are conducted on older sample periods. However, the market structure keeps changing and what happened in the 1970s might not be valid anymore. This motivates this study to consider structural breaks around the financial crisis of 2007–2008, and we fit the models for 2000–2007 and 2008–2015 in order to explore potential structural breaks.

3 | DATA SET

The analysis is conducted on a pan-European stock universe comprising 2,353 stocks (7,197,065 daily observations) with price and volume market data for the period between 1 January 2000 and 10 May 2015. The midpoint of our data set coincides with the financial crisis of 2008/2008, whose peak consisted of the collapse of Lehman Brothers on 15 September 2008. Therefore, we investigate a potential structural break in the market dynamics before and after the crisis, by splitting the data set into two subsets: the precrisis subset (1 January 2000 to 31 December 2007) and the postcrisis subset (1 January 2008 to 10 May 2015).

3.1 | Data acquisition

The analysis market data for the extensive pan-European stock universe consist of the constituents of the most important European indices, along with a comprehensive index from Thomson Reuters. The indices' constituent list is compiled as of 10 May 2015 and does not contain historical evidence of index additions and eliminations throughout the entire duration of the study; this list is a representative stock sample for the European stock market. The final stock universe consists of the list of unique constituents of the indices included in Table 1, along with their Reuters Identification Codes (RICs).

The daily market data were retrieved from Thomson Reuters Eikon by developing a VBA script to automate the process of stock-specific data retrieval. The VBA script takes a list of desired indices as input, retrieves their constituents, and then, for each stock, it returns the daily

TABLE 1 The European indices whose constituents were part of the study data sample

RIC	Index name	RIC	Index name
.STOXX	STOXX Europe 600 EUR Price Index	.PSI20	Euronext Lisbon PSI 20 Index
.FTSE	FTSE 100 Index	.OMXS30	OMX Stockholm 30 Index
.FTMC	FTSE Mid 250 Index	.OBX	Oslo Stock Exchange Equity Index
.FTLC	FTSE 350 Index	.OMXHPI	OMX Helsinki_PI
.FTSC	FTSE Small Cap Index	.BFX	BEL 20 Index
.FTAS	FTSE All Share Index	.OMXC20	OMX Copenhagen 20 Index
.GDAXI	Deutsche Boerse DAX Index	.ATG	Athex General Composite Share Price Index
.MDAXI	MDAX Performance Index	.ISEQ	ISEQ Overall Price Index
.SDAXI	SDAX Share Index	.JTOPI	Johannesburg Stock Exchange Top 40 Tradeable Index
.FCHI	CAC 40 Index	.ATX	Austrian Traded Index
.CN20	CAC Next20 Index	.FTMIB	FTSE MIB Index
.CACMD	CAC Mid 60 Index	.MSPE	MSCI International Pan Euro Price Index
.CACS	CAC Small Index	.MCX	MICEX Composite Index
.SSMI	Swiss Market Index	.WIG20	Warsaw SE WIG-20 Single Market Index
.AEX	Amsterdam Exchanges Index	.TRXFLDEUPU	Thomson Reuters Europe Index
.IBEX	IBEX 35 Index		

Note. RIC = Reuters Identification Codes.

market data (i.e. opening, high, low, and closing prices, and trading volume) for the 15 years covered by this study (i.e. 1 January 2000 to 10 May 2015). The data set was further extended using these stocks' primary RICs and attaching their MTF (i.e. Multilateral Trading Facilities) RICs for the following MTFs: BATS, CHI-X, and Turquoise. Then, we retrieved the daily prices and trading volumes for each new MTF RIC.

The market data ranges from 1 January 2000 to 10 May 2015, comprising the daily summary of corporate actions-adjusted volumes (e.g. controlling for stock splits, stock dividends, mergers and acquisitions, spinoffs, and rights issues).

3.2 | Data preprocessing

The market data preprocessing stage starts by eliminating the instruments for which there is no available market data. There were 595 stocks without MTFs and 194 MTF RICs with no market data. Missing data points are checked in the primary exchange volume and price data (e.g. zero volumes and Thomson Reuters data retrieval errors) and in the MTF volume data only, as the MTF's market data are only used for computing the consolidated volume and the MTF prices are not of interest. The consolidated volume is then computed for each stock by summing up the primary volume and the MTF trading volume. The primary exchange volume is hereafter replaced by the consolidated volume for all stocks. The market data are further processed by discarding the stocks whose number of days of available market data is less than 100 trading days.

We include South Africa in the stock universe due to its liquidity and high level of similarity with European stocks, as it is sharing the same time zone with Eastern Europe.

Throughout the volume analysis, we will be using the logarithmic trading volume due to the high nonnormality and outliers of linear volume; from this point forward, we will refer to log-volumes only. Taking natural logarithm of the volumes and price ratios helps normalize the errors, as it reduces skew.

Given that the market data provided by Thomson Reuters does not cover the auction volumes, it is impossible to compute a highly accurate breakdown of trading volume breakdown, although these could be approximated by getting tick data and aggregating their values based on the millisecond timestamp, for example, same time and price values for the first points of the day for the opening auction, and only the ticks at 16:35 (UK time zone) for the closing auction. Therefore, we use the total daily volume as the dependent (or response) variable in our analysis. It includes all the trades executed for the day, and it disregards the overnight and off-market trades.

Furthermore, because the data for opening auction volume is unavailable, we define the overnight return as a proxy for the opening auction volume in order to quantify the improvement of more recent data. The overnight return is divided by the number of intervening nights in order to account for nontrading days (i.e. bank holidays and weekends). We explore two variants of defining the overnight return, one that applies a correction (by dividing by the number of intervening nights) and one that is not corrected, which includes an additional variable for the number of extra nights.

4 | AIMS OF STUDY AND ANALYSIS APPROACH

The objective of this study is to explore a prediction framework to understand what drives the trading volume. This study's linear regression framework tests a variety of hypotheses using various factors, which are ultimately reduced through feature selection. This is an exploration of the endogenous and exogenous factors affecting trading volume, and it is important to note that the effect size is not our main concern in this study. For each stock in our pan-European universe, different models are compared in order to accomplish the best explanation, while keeping as few predictors as possible and eventually identifying a parsimonious model.

The proposed framework conducts a stock-specific analysis, where each stock is investigated by fitting different stock-specific models, performing feature selection and model comparison. Eventually, we report the overall findings and provide a summary of the pan-European stock universe analysis, despite having a per-stock approach.

The stock-specific analyses were normalized by representing the different effects for each stock and account for idiosyncrasies; models vary for each stock independently. The normalized results were aggregated across the 2,353-stock universe.

The methodology for model comparison consists of 10-fold cross-validation (CV), where the objective function seeks to minimize the average mean squared error (MSE). We used CV even for nested models, as it is more robust (instead of an F-test, which assumes Gaussian errors and is sensitive to nonnormality) and avoids overfitting. The CV folds are defined at the beginning of the analysis, and they are constant throughout the various models that are fit for each stock. After defining the 10 CV folds, we perform stepwise regression on the various models (i.e. multiple linear regression, followed by sequential feature selection). Feature selection reduces dimensionality by producing a reduced model fit on fewer variables, while minimizing the predictive error. Whenever a feature is added to or removed from a model, feature selection performs 10-fold CV at each step in order to guarantee that the overall model error is reduced. The objective function of the sequential feature selection minimizes the average MSE across the CV folds. Therefore, features are sequentially added (for forward selection) or removed (for backward elimination) at each step, until no other features can be added or eliminated, while decreasing the criterion (i.e. MSE). Because of the unfeasibility of following an exhaustive approach and fitting all of the possible feature subsets, the sequential feature selection technique moves only in one direction, meaning that the candidate feature set is always growing (in the case of forward selection) or shrinking (in the case of backward elimination).

4.1 | Randomisation tests

When looking at the return asymmetry and the magnitude of the overnight return depending on the number of intervening nights, we aim to evaluate whether two data vectors are significantly different or whether they come from a similar distribution. The randomisation (or permutation) tests were employed mainly because they make no assumptions about the data distribution, unlike other popular parametric tests such as the student's t-test, where data points are assumed to come from a normal distribution. The randomisation test is a robust and rigorous statistical significance test, and it is appropriate for this study especially because the log-ratio returns and the volumes are not exactly Gaussian (although they are significantly normalized). Nonparametric tests, for example, the Mann–Whitney U test, could be used alternatively, but because the p-values are based on approximations and using rankings reduces the information inferred from the data (i.e. information loss), randomisation tests are considered a superior methodology (Edgington, 1964). For two vectors (X and Y), the permutation test computes the observed statistic as the absolute difference of the two vectors. The labels of the data points from vectors X and Y are randomised 1,000 times, and for each reshuffling, we compute the randomised statistics using the same equation as the initial statistic. Finally, the test assesses whether the randomised absolute differences are more extreme than the observed absolute difference, resulting in an empirical p-value; this value represents the percentage of times when the observed absolute difference is greater than the randomised absolute differences for a significance level $\alpha = 0.05$. The randomisation test rejects the null hypothesis if the empirical p-value is less than the significance level (5%).

4.2 | Model outline

In this study, we are investigating several factors that could potentially drive the trading volume. Therefore, for each stock, the analysis consists of a number of volume prediction models for hypothesis testing and effect quantification, starting from a basic volume model and expanding it subsequently. All of the models in this study include an intercept unless stated otherwise. For a given date (t_{-1}), the target (or dependent) variable is the next trading day (t_0) logarithmic volume, whereas the model is trained on past data up to the test date ($t_{-n}...t_{-1}$). The regression design matrix is computed for each target day (t_0), and then the CV partitions the target date vector accordingly. The structural breaks we investigate in connection with the financial crisis of 2007–2008 do not destroy the CV process, because the feature matrix is computed before partitioning the data, and hence, it does not interfere with the subsequent data partitioning (e.g. structural breaks or CV). When two or more predictors are linearly dependent, the linear regression sets the maximum number of coefficients to zero in order to obtain a basic solution.

In order to test the statistical improvement of the various potential endogenous and exogenous determinants of trading volume, we start by defining a basic prediction model for trading volume (i.e. the "volume model") based on time-lagged observations, both raw (i.e. autoregressive past observations) and smoothed (i.e. moving average of the last observations). We employ 10-fold CV to find the optimal orders for the time lags of the autoregressive volume and the time windows for the moving average volume. These volume features, as well as the intercept, are kept in all of the subsequent models when performing feature selection.

Next, the price features for the previous day (i.e. intraday range and intraday return) are added to the volume model, and we perform feature selection on these price features. The best model in this state is called the "State A model."

Then, we add more recent data in the form of overnight return (as a proxy for opening auction volume) to the full "State A model" (i.e. the model with the full feature set) and perform feature selection on all price features. The model with the lowest MSE is called the "State B model."

Up to this point, we use endogenous variables to fit a volume prediction model. We then switch to exogenous variables (i.e. day-of-the-week) and start from the best model up to this point, that is, the best model among the State A and State B models, which is called “the historical dynamic model.”

Two day-of-the-week models are defined. The first one (i.e. the “raw day-of-week model”) is a basic model consisting of five dummy variables for each workday, with feature selection applied to them; there are no volume or price features—this is the traditional model employed in the behavioural finance literature. The second model, called the “historical dynamic and day-of-week model,” adds five indicator variables for Monday to Friday on the top of the historical dynamic model (i.e. the best model between States A and B, based solely on price and volume features) and then performs feature selection on the day-of-week features, while forcing the historical dynamic features to remain in the model.

The feature selection consists of backward elimination for the volume and price models (i.e. State A and State B models). Considering these models have a large feature set, the runtime of backward elimination is significantly faster than the forward selection technique, although their final results (i.e. the reduced models) are almost identical. However, in the case of the day-of-the-week models, we used forward selection due to the design of the dummy variables for the day-of-the-week categorical variable. A categorical variable with n possible values is normally encoded as $n - 1$ dummy variables. In this study, the day-of-the-week dummy variables are mutually exclusive because the aim is to perform feature selection and extract the days that are the most statistically significant for volume prediction, and this is performed in a feature selection framework. Forward selection is preferred to backward elimination because matrix inversion would not be possible for determining the optimal beta when adding a collinear variable. Using backward elimination and starting from the full feature set would result in collinearity issues.

Table 2 includes a preview of the models fit for each stock and their full candidate feature sets. The historical dynamic model includes the model with the lower CV MSE among State A and State B. Depending on the chosen model, the intraday returns and overnight returns can be defined either symmetrically or asymmetrically. The aim of these models is to determine if a certain variable improves the forecast significantly. The main concern is not to evaluate the effect size (e.g. whether a certain feature improves the prediction accuracy by n shares). The models are stock-specific, and their coefficients cannot be generalized to the entire data set. Moreover, we did not standardize the variables as their interpretation would not be meaningful in terms of standard deviations, and we believe that each stock exhibits slightly different correlations.

5 | VOLUME ANALYSIS

In this section, we describe the methodology for the volume analysis study, while introducing a number of particular aspects we are investigating.

5.1 | Volume model

In the initial volume model, we explain observed volumes in terms of recent time series. Essentially, this model aims to predict today's (i.e. t_0) volume using the trading volume of the previous days. Therefore, the model contains a lagged time series part, comprising contiguous volume lags, and a lagged smoothed time series part, comprising contiguous moving average volume windows. The lagged time series part reveals the persistence or autocorrelation in the time series assuming stationarity, and the lagged smoothed time series part is a low-pass filter effect in the data; they are both

TABLE 2 Regression models with full candidate feature sets

	Volume model	Volume and price models						Historical dynamic model	Historical dynamic and day-of-week model	Raw day-of-week model
		State A models		State B models						
Intercept	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Volume lag ($1 \ldots n_1$)	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Volume window ($1 \ldots n_2$)	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Symmetric intraday return		✓		✓	✓			✓*	✓*	
Asymmetric intraday returns			✓			✓	✓			
Symmetric overnight return				✓		✓		✓*	✓*	
Asymmetric overnight returns					✓		✓			
Day-of-the-week (1...5)									✓	✓

used later to evaluate the potential importance of other features in the model. Equation 3 defines the volume model, where vol_{t_0} refers to the logarithmic trading volume of the current day and $vol_{t_{-1}}$ refers to the log-volume of the previous day; β_0 is the intercept coefficient, β_{lag_i} is the coefficient for the *lag(i)* feature (i.e. the volume lags), and β_{smooth_j} is the coefficient for the *smooth(j)* feature (i.e. the moving average volume window). Hence, the volume model is composed of three types of terms: the intercept, the autoregressive lagged predictors, and the moving average lagged smoothed predictors; the model has two parameters, that is, *recentTimeSeries(p,q)*, where *p* is the autoregressive lagged order and *q* is the moving average lagged smoothed order. The model can also be represented using the feature names that correspond to the *lag* and *smooth* terms, that is, volume lag (*volLag*) and volume window (*volWin*). The *lag* and *smooth* underlying features are contiguous. The *lag* model comprises autoregressive orders from *1* to *p*, and the *smooth* model comprises the moving average orders from *2* to *q*. For example, *lag(6)* contains all the volume lags from 1 to 6, and no in-between lag can be excluded by the subsequent feature selection.

$$vol_{t_0} = \beta_0 + \sum_{i=1}^{p} \beta_{lag_i} vol_{t_{-i}} + \sum_{j=2}^{q} \beta_{smooth_j} \frac{\sum_{k=1}^{j} vol_{t_{-k}}}{k} \tag{3}$$

In order to build an optimal volume model from time-lagged volume observations, we first need to identify the optimal *lag* and *smooth* orders. These are identified by fitting autoregressive (*lag*) and moving average (*smooth*) models up to order 15, that is, *lag(1)* ... *lag(15)* and *smooth(2)* ... *smooth(15)*. These models consist of the constant term and the autoregressive or moving average terms. Each of these models is cross-validated, and the average MSE is returned, which is then used as the criterion of comparing two nested models at a time in an incremental manner, for example, *lag(1)* against *lag(2)*; *lag(2)* and *lag(3)*; and *lag(14)* and *lag(15)*. We start with the lowest possible order and increment it by one; we compare the pair of models with consecutive autoregressive or moving average orders, and if the full model (i.e. the one with the greater order) statistically improves the reduced model (i.e. the one with the lower order), then we increment the order once again and compare the pair of models with the two largest orders at this point. We repeat this process until incrementing the order does not statistically improve the reduced model anymore. For the autoregressive model, that is, *lag(p)*, we start with $p = 1$ and compare it to the next integer value, that is, $p = 2$. If *lag(2)*, that is, the full model, improves *lag(1)*, that is, the reduced model, then we increment *p* and compare *lag(2)* to *lag(3)*; if it does not improve the reduced model, then we stop the incremental process and pick the lower *p* of the last comparison pair as the optimal *lag* order. Similarly, we determine the optimal order for *smooth*, but we start from $q = 2$, that is, *smooth(2)*, because *smooth(1)* is the same as *lag(1)*. Equations 4 and 5 show the models with consecutive orders (i.e. *p* and $p + 1$) for the comparison of statistical improvement for detecting the optimal *lag* order.

$$Reduced\ model\ lag(p){:}vol_{t_0} = \beta_0 + \sum_{i=1}^{p} \beta_{lag_i} vol_{t_{-i}} \tag{4}$$

$$Full\ model\ lag(p+1){:}vol_{t_0} = \beta_0 + \sum_{i=1}^{p} \beta_{lag_i} vol_{t_{-i}} + \beta_{lag_{p+1}} vol_{t_{-(p+1)}} \tag{5}$$

This comparison of nested models based on 10-fold CV average MSE tells whether a higher order statistically improves the goodness of fit of the model. The comparison of the CV average MSE is performed in the initial phase instead of stepwise regression in order to enforce the *lag* and *smooth* predictors to contain contiguous features (i.e. successive volume lags/windows).

5.2 | Contribution of volatility and volume–price asymmetry

After identifying the optimal lag and smooth orders and defining the volume model, we extend it by adding a couple of price metrics for the previous trading day, namely, intraday return (i.e. the difference between a trading day's closing and opening prices) and intraday range (i.e. the difference between a day's high and low prices). Each of these metrics can be represented as percentages or log-ratios. The log-ratio representation was preferred to percentages because the percentage returns cannot drop under −100%, but they can go up over 100% (due to the non-negative nature of prices), and therefore, price percentages can lie on the interval $(-100\%, +\infty)$, whereas log-ratio returns can in principle belong to $(-\infty, +\infty)$, providing a better representation for price returns that is closer to the Gaussian distribution. Therefore, we used logarithmic price ratios to compute the intraday return and intraday range price metrics. The log-ratios were calculated by taking the log of the raw price ratios (i.e. p_1/p_2). The formulae for the log-ratio returns, which were used for the features added to the volume baseline model, are presented in Equations 6 and 7. It is worth clarifying that the target variable of the model is the trading volume for t_0 based on previous information, that is, up to and including t_{-1}.

$$\textit{Intraday return log ratio} = \log \frac{close_{t_{-1}}}{open_{t_{-1}}} \quad (6)$$

$$\textit{Intraday range log ratio} = \log \frac{high_{t_{-1}}}{low_{t_{-1}}} \quad (7)$$

The intraday return and the overnight return, the latter of which will be introduced in the next model, allow for both positive and negative results. Given the literature findings on the volume–price relation asymmetry, we define each of these metrics in two ways. The first method regards the intraday and overnight returns as being symmetric in terms of magnitude, and therefore, each of them corresponds to a single feature taking the absolute value of the log-ratios (the features are generally called "abs," e.g. "abs" intraday return or "abs" overnight return). The other method is based on the fact that the magnitude of price returns is asymmetric, depending on the sign of the price return; instead of having a single feature, this method generates two features based on the price movement direction (e.g. positive or negative); this allows the positive and negative returns to be represented by two features, which will potentially result in having different coefficients when being fit into the regression model. These two features are called "absPos," representing the absolute value of positive returns only, and "absNeg," standing for the absolute value of negative returns only (e.g. "absPos" intraday return and "absNeg" intraday return). These indicators split the log-ratio returns at zero, into positive absolute values and negative absolute values.

We extend the volume model, which was constructed based on the optimal values of the *lag* and *smooth* orders, with two more price-related feature sets: symmetric intraday price features and asymmetric intraday price features. The former includes the "abs" intraday return and the intraday range log-ratios. The latter model includes the "absPos" intraday return, "absNeg" intraday return and the intraday range log-ratios.

Once these full models consisting of volume and symmetric/asymmetric intraday prices have been linearly fit using OLS regression, we perform feature selection while enforcing the volume-related features (i.e. "volLag" and "volWin") to be kept in the model, along with the constant term.

Table 3 outlines the component features for each model in the current state, called "state A." The question marks in the table represent a feature that might be selected or not after the feature selection process. The tick means that the feature is definitely present in the model and a cross notes its absence. The model with the lowest CV MSE in this state is called "the State A model."

TABLE 3 The features of State A models

	Volume and symmetric intraday prices model	Volume and asymmetric intraday prices model
intercept	✓	✓
$\mathrm{volLag_p}$	✓	✓
$\mathrm{volWin_q}$	✓	✓
intraday range	?	?
"abs" intraday return	?	✗
"absPos" intraday return	✗	?
"absNeg" intraday return	✗	?

5.3 | Contribution of overnight return

We extend the State A full models (i.e. the entire set of features) by adding the overnight return (i.e. the difference between today's open and yesterday's close). This variable is employed in order to test whether information that is more recent improves the volume prediction model significantly. The overnight return indirectly measures trading volume and can be expected to be a leading indicator of the day's volume because the opening price is associated with the opening auction. The overnight return incorporates the information accumulated during the nontrading period, when investors rebalance their portfolios. The opening auction plays a major role in the daily price discovery process and reflects the private and public information flowing while the market was closed. This theory has been argued in the price formation models formulated in established literature, as surveyed by Gerety and Mulherin (1994).

The overnight return uses the opening price of t_0. This feature incorporates the afterhours trading (i.e. market-moving events occurring overnight, between yesterday's close and today's open, such as earnings reports, preearnings announcements, or M&A activity, which drive prices) and does not reflect any of the trading activity for t_0; therefore, there is no look-ahead bias. The reason for adding this feature is to investigate whether more recent information proves to be beneficial to the prediction of the following day's trading volume. Due to licencing constraints, we did not have access to the opening auction volume, and the overnight return is implemented as a proxy for the opening auction volume (and hence for the more recent data).

As with the State A models, the models in this state, which is called "State B," are defined in two ways, either with asymmetric price ratios or with symmetric price ranges. Each of the intraday return and the overnight return has its own asymmetry, resulting in four models in State B, whose features are listed in Table 4. Feature selection is performed on the price features of these models (e.g. intraday range, intraday return abs/absPos/absNeg, and overnight return abs/absPos/absNeg), while keeping fixed the features of the volume model. The model with the minimum CV average MSE is called "the state B model."

At this stage, we performed an intermediate analysis in order to decide whether to correct the overnight return (i.e. dividing by the number of intervening nights) or not. Therefore, we define two variants for each of the four State B models outlined in Table 4. The first method provides a correction factor for the higher coefficient magnitude associated with a larger number of intervening nights, dividing by the number of intervening nights; this is the corrected overnight model type. The alternative model considers that the corrected model might have artificially small overnight returns and does not divide by the number of intervening nights (although this could potentially result in artificially high returns). Table 5 outlines the frequency table for the intervening nights up to nine nights; the most common successive trading days have one intervening night (e.g. two trading days during the same week) or three intervening nights (e.g. a Friday and a Monday, which are separated by two additional nights because of the weekend).

Because the data points with 1, 2, 3, and 4 intervening nights account for 98.93% of this study's data set, we used them as four different classes of overnight returns in order to assess whether the overnight returns are significantly different for each pair of these overnight return classes. For each stock, we performed a randomisation test with 1,000 reshufflings for each of these pairs: Class 1 and Class 2, Class 1 and Class 3, Class 1 and Class 4, Class 2 and Class 3, Class 2 and Class 4, and finally, Class 3 and Class 4. Using the empirical *p*-values, we evaluated which pairs of classes are significantly different (i.e. where the null hypothesis is rejected) and then aggregated the results across the entire stock universe by computing the percentage of tests where the null hypothesis is rejected. The results outlined in Table 6 show that each pair is predominantly found as coming from the same population, as there is no pair where the two overnight return classes are considered significantly different in more than 50% of the cases. Based on these pairwise randomisation tests, we report that the overnight return magnitude is not different based on the number of intervening nights. Therefore, the number of intervening nights is not a salient factor in determining the trading volume, and we did not include it in the list of predictors for the State B models.

In order to compare the performance of the two model types in conjunction with trading volume, we fit four

TABLE 5 Frequency table of intervening nights

Class (number of intervening nights)	Count	Percentage
1	5,590,714	77.71
2	72,852	1.01
3	1,369,188	19.03
4	84,731	1.18
5	52,805	0.73
6	13,354	0.19
7	3,867	0.05
8	1,478	0.02
9	655	0.01

TABLE 4 The features of state B models

	Volume, symmetric intraday prices, and symmetric overnight prices model	Volume, symmetric intraday prices, and asymmetric overnight prices model	Volume, asymmetric intraday prices, and symmetric overnight prices model	Volume, asymmetric intraday prices, and asymmetric overnight prices model
intercept	✓	✓	✓	✓
volLagp	✓	✓	✓	✓
volWinq	✓	✓	✓	✓
intraday range	?	?	?	?
"abs" intraday return	?	?	✗	✗
"absPos" intraday return	✗	✗	?	?
"absNeg" intraday return	✗	✗	?	?
"abs" overnight return	?	✗	?	✗
"absPos" overnight return	✗	?	✗	?
"absNeg" overnight return	✗	?	✗	?

TABLE 6 Statistical significance of the overnight returns based on the number of intervening nights

Class (number of intervening nights)	Null hypothesis rejected (%)
1-night versus 2-night overnight returns	36.36
1-night versus 3-night overnight returns	46.87
1-night versus 4-night overnight returns	34.03
2-night versus 3-night overnight returns	25.13
2-night versus 4-night overnight returns	5.86
3-night versus 4-night overnight returns	20.66

models (all combinations of symmetric and asymmetric intraday returns/overnight returns) for each of the corrected and uncorrected overnight returns, for each stock in our universe. Then we compare the corrected models with their corresponding uncorrected models, and we choose the model with the lowest 10-fold CV average MSE. There are 2,353 stocks, resulting in 9,412 fit models, which are to be compared in terms of MSE between the corrected overnight return variant and the uncorrected one. The overnight return models with corrections applied to them for the intervening nights perform better than their uncorrected alternatives in 77.54% of the cases.

Based on these results, we correct the overnight return in this study. Therefore, it is divided by the number of intervening nights in order to avoid having higher magnitude coefficients as a correction factor for the trading days following one or more nontrading days (e.g. weekends or bank holidays). Equation 8 shows the log-ratio representation of the overnight return.

$$\mathit{Overnight\ return\ log\ ratio} = \frac{\log\frac{open_{t_0}}{close_{t_{-1}}}}{\#\mathit{intervening\ nights}} \tag{8}$$

5.4 | Asymmetry randomisation analysis

Asymmetry can be either strong (with absPos or absNeg features that are explicitly picked by the stepwise regression) or weak. The weak asymmetry is determined through a procedure based on randomisation tests, which are described below.

First, a randomisation test is performed for the intraday return absPos and intraday return absNeg, excluding the zero-valued observations in order to evaluate the null hypothesis that the data in vectors absPos and absNeg come from independent random samples from distributions having equal means and variances. If the null hypothesis is rejected at the 5% significance level, then the vectors absPos and absNeg are significantly different, coming from populations with unequal means. Similarly, a second randomisation test is conducted on the same samples (i.e. intraday return absPos and intraday return absNeg) for the models where both asymmetric features are picked by the feature selection process, but this time, they are multiplied by their regression coefficients. If the randomisation test does not reject the null hypothesis, then it means that the vectors absPos and absNeg would have different magnitudes, which are eventually corrected by the regression coefficients in order to ultimately get a symmetric representation of intraday return.

For the intraday return asymmetry, we consider each stock's State A model (among symmetric and asymmetric intraday price features) where an intraday return feature is present in any way (e.g. abs, absPos, or absNeg); the stocks whose State A model has no intraday return feature kept in the model after feature selection are disregarded. Similarly, the overnight return asymmetry analysis only considers the stocks whose State B model has at least an overnight return feature in the reduced model (e.g. abs, absPos or absNeg).

There are two scenarios where we investigate whether the lack of significant difference provided by the randomisation test is consistent with the regression symmetry.

In the first scenario, the model with the best error is asymmetric. If the presence of absPos and absNeg is mutually exclusive (i.e. if absPos is present and absNeg is absent; or absNeg is present and absPos is absent) or both absPos and absNeg are present and their randomisation test is negative (meaning that the vectors absPos and absNeg come from populations with equal means, but their regression coefficients provide empirical evidence that their impact on volume is different depending on their sign), then there is return asymmetry. However, if both absPos and absNeg are present and their randomisation test is positive (i.e. the two vectors are significantly different), we need to check whether the regression coefficients might reverse their asymmetry, causing them to behave symmetrically. In this case, if the randomisation test of the two vectors multiplied by their regression coefficients is positive, then it means that the two vectors are still significantly different even after accounting for their coefficients; otherwise, if the randomisation test is negative, it means that the coefficients act as a correcting factor for the apparent asymmetry, transforming the two vectors into a symmetric vector.

In the second scenario, the model with the lowest error is symmetric. If the randomisation test of the absPos and absNeg vectors is positive, then these vectors are significantly different and the model is actually asymmetric. Otherwise, it means that they come from populations with equal means, and the model is indeed symmetric.

5.5 | Temporal context: Day-of-the-week effects

The starting point of this stage consists of the best model (i.e. having the lowest CV MSE) among the State A model and the State B model. The resulting model is called "the historical dynamic model." In order to investigate the day-of-the-week effect, we define a couple of models: one that extends the historical dynamic model with additional dummy variables for each working day (i.e. Monday to Friday) and an elementary one with features for each of the five workdays only (without any volume or price features), due to the broad implementation of this model in the behavioural finance literature on the day-of-the-week effect and weekend effect.

In order to be able to assess which day-of-the-week features are the most salient, either in improving the historical dynamic model or simply in the raw day-of-the-week model, we performed feature selection on the two models. The intercept and the other features of the historical dynamic model are kept fixed in the model, and we attempt to reduce only the day-of-week feature set. Table 7 summarizes the day-of-the-week models and their potential features, along with the historical dynamic model, which acts as a benchmark.

In theory, if a model has a categorical variable with k possible values, one should assign $k-1$ dummy variables if the model has an intercept and k dummy variables if there is no intercept; although such a model would result in identifiability issues due to assigning dummy variables to all of the 5 days (and not to only four of them), we use the five dummy variables in order to define the full model that is to be reduced using feature selection. This approach is necessary for the interpretability of each day-of-the-week. Feature selection is applied to this model, having fixed the volume and price features (i.e. the features of the historical dynamic model), in order to find the most significant days of the week.

TABLE 7 The features of the day-of-week models

	Historical dynamic model	Historical dynamic and day-of-week model	Raw day-of-week model
Intercept	✓	✓	✓
Volume features $volLag_p$ $volWin_q$	✓	✓	✗
Price features Intraday range Intraday return Overnight return	?	?	✗
Monday	✗	?	?
Tuesday	✗	?	?
Wednesday	✗	?	?
Thursday	✗	?	?
Friday	✗	?	?

6 | RESULTS

This section first introduces the results on the volume–price relation and then switches to the day-of-the-week findings. We start by investigating the effects for each stock and then provide a summary of the model performance across the entire stock universe. Tenfold CV is performed on all of the models in States A and B and the day-of-the-week models in order to get the average MSE. We provide goodness-of-fit illustrative examples while introducing each driver of trading volume (i.e. volatility, overnight return, and day-of-the-week features).

6.1 | Contribution of volatility and asymmetry

Based on the aggregated results for the entire data set in Table 8, we report that the volatility features (i.e. the previous day's price changes, i.e. intraday range and intraday return) generally improve the autoregressive volume model, where 60% of the models are trained on asymmetric intraday returns. These results are computed by selecting the State A model for each stock in our 2,353 stock universe and examining the feature presence of each model. A representative model is exhibited in Figure 1 for Fenerbahce Futbol AS, which also contains a zoomed plot of the time series spanning the last 6 months for better visualization. The improvement over the volume model for this stock is 5%. Figure 2 depicts the different distributions of the vectors "intraday return absPos" and "intraday return absNeg" for PCC Rokita SA.

There are no significant structural changes around the financial crisis of 2007–2008. However, the overnight return asymmetry is salient in the precrisis subset, but becomes rather neutral in the postcrisis data set, with a performance that is similar to the symmetric overnight return.

By using the State B model for each stock, similar results are reported for the contribution of more recent information, in the form of overnight returns, which act as a proxy indicator for the opening auction volume in this study. The opening auction volume represents the most recent piece of information that could be publicly available at the opening of trading, because the continuous trading phase begins based on the conclusion of the opening

TABLE 8 Volatility findings

Hypothesis	Observation percentage		
	Entire data set	Structural break subsets	
	2000–2015 (%)	2000–2007 (%)	2008–2015 (%)
Volatility improves the volume model	86.95	82.08	84.62
Asymmetry of volume—Intraday return relation (State A models)	60.16	59.64	62.52
More recent price information improves the model	87.21	81.57	86.87
Asymmetry of volume—Overnight return relation (State B models)	54.09	60.27	47.11
Historical dynamic model: Intraday return asymmetry	56.53	58.61	57.81
Historical dynamic model: Overnight return asymmetry	54.07	60.26	47.11
Historical dynamic model: Total asymmetry	73.59	75.70	70.02

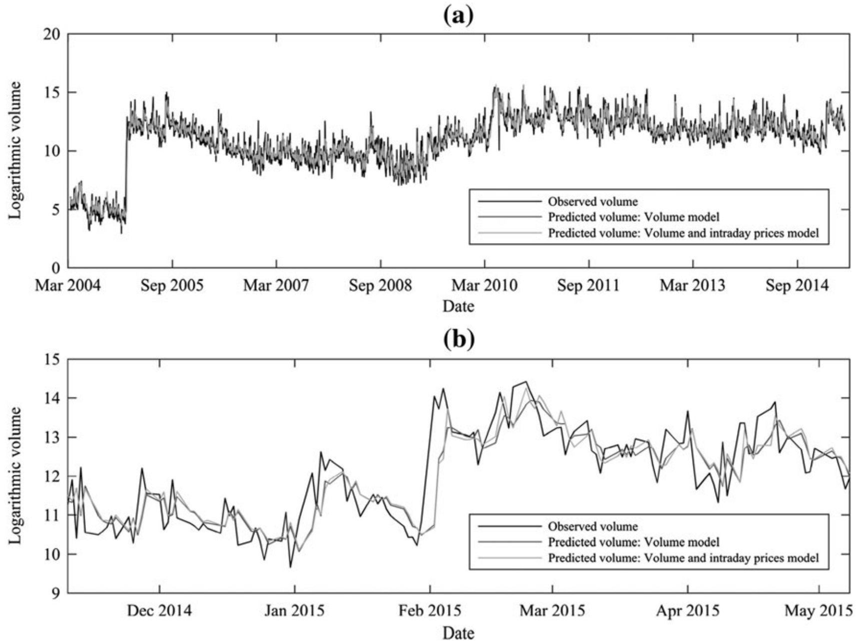

FIGURE 1 Improvement of intraday prices (intraday return and intraday range) over volume by 4.72%. Both panels illustrate the improvement for Fenerbahce Futbol AS (FENER.IS). Panel (a) shows the entire time series between 12 March 2004 and 8 May 2015, whereas Panel (b) provides a zoomed time series for the last 6 months of data

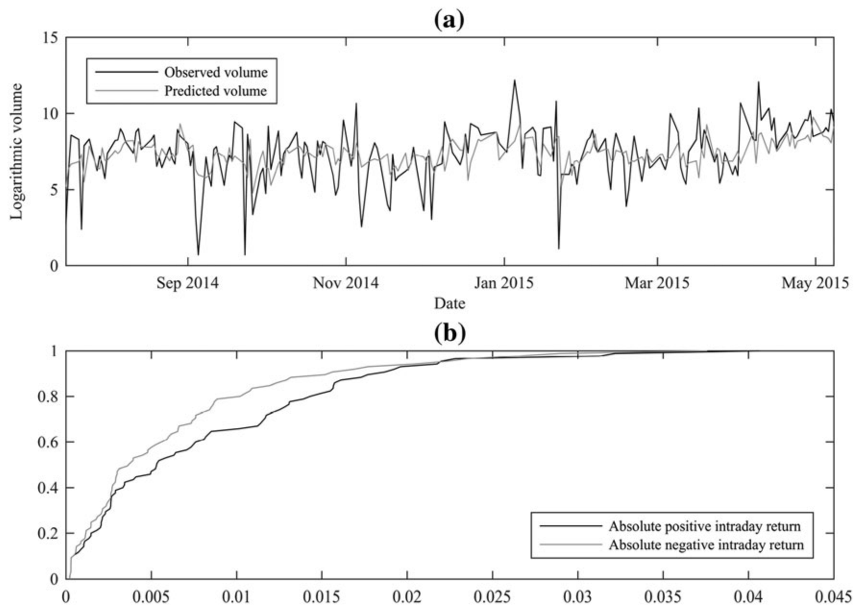

FIGURE 2 Intraday return asymmetric distribution. Panel (a) illustrates the observed volume against the predicted volume using the asymmetric intraday return model for PCC Rokita SA (PCR.WA) from 16 July 2014 to 8 May 2015 (201 trading days). Panel (b) illustrates the cumulative distribution breakdown asymmetric intraday return for this volume prediction model

auction. The overnight return improves the volume model predominantly, with a slightly lower asymmetry for the overnight return of 54%. Figure 3 shows the recent data improvement (16%) over the State A model for H & M Hennes & Mauritz AB; Panel (b) is a magnified view of the last 6 months of the same time series for easier visualization. The distribution of the asymmetric overnight return vectors is shown in Figure 4, along with the predicted and observed volume time series for Aeffe SpA.

Considering that the intraday return asymmetry and overnight return asymmetry provide better performance in more than 50% of the stocks, we argue that the volume–price relation should be modelled with asymmetry.

Further analysis on the asymmetry is computed on the historical dynamic model (i.e. the model with the lowest CV MSE throughout States A and B, fit using volume and price features). Here, we use the historical dynamic model across both types of asymmetry. The total asymmetry for each stock is the logical disjunction (logical or) between the two types of asymmetry for that stock. When examining the historical dynamic model, we report moderate intraday return asymmetry (56%) and overnight return asymmetry (54%). However, the majority of models (74%) exhibit some type of asymmetry, be it in the form of intraday returns or overnight returns.

We explored the transition of intraday price features across States A and B and investigated whether having

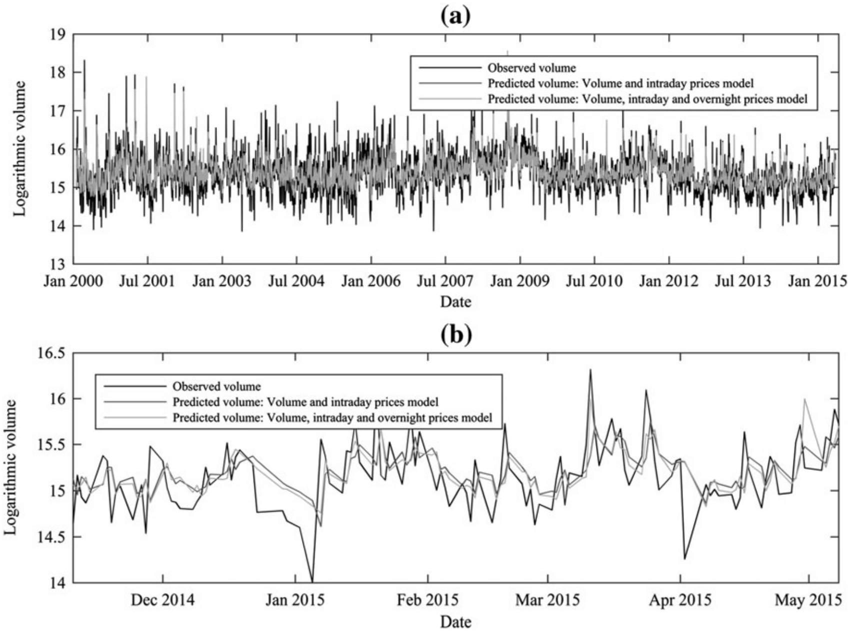

FIGURE 3 Overnight return improvement. Both panels show the overnight prices improvement over volume and intraday prices (16.26%) for H & M Hennes & Mauritz AB (HMb.ST). Plot (a) contains the entire time series between 25 January 2000 and 8 May 2015 (3,837 observations), and Plot (b) provides a magnified view of the last 6 months

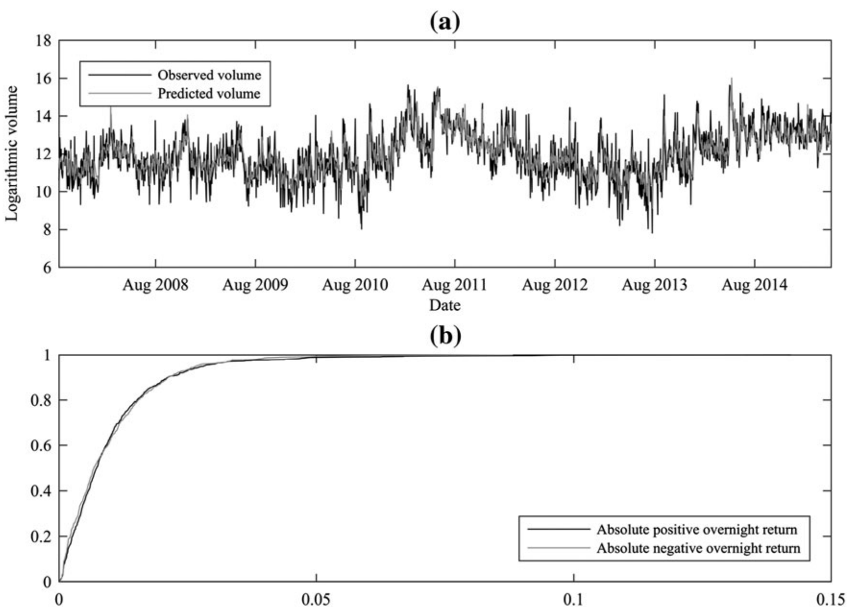

FIGURE 4 Overnight return asymmetric distribution. Panel (a) shows the observed volume against the predicted volume using the asymmetric overnight return model for Aeffe SpA (AEF.MI) from 14 August 2007 to 8 May 2015 (1,949 trading days). Panel (b) illustrates the cumulative distribution breakdown of the asymmetric overnight for this volume prediction model

the overnight return feature selected in State B could potentially cancel any intraday price features previously selected in State A. Table 9 shows the occurrence frequency of each intraday price feature across the two states for the entire data set, whereas Table 10 includes the feature presence for the two structural break data subsets (i.e. the precrisis and postcrisis data). The results do not exhibit any significant frequency changes for any volatility feature, apart from a general increase in the number of models with the intraday range predictor selected, which is also consistent for our structural break.

6.2 | Contribution of day-of-the-week effects

The study further investigates the temporal context of the volume time series, analysing the day-of-the-week effect. We compared the historical dynamic model (resulting from States A and B) to the day-of-week model that is traditionally employed in the calendar effect literature (i.e. the raw day-of-week model). Based on the results outlined in Table 11, we find that the historical dynamic model fit with volume and price features clearly dominates the traditional raw day-of-the-week model in terms of performance (in almost 100% of the analysed stocks). Moreover, we augmented the historical dynamic model with day-of-week features, which, after performing feature selection, improved the historical dynamic model with at least 1 day-of-week feature in approximately 91% of the models—an illustrative day-of-week improvement (7%) is shown in Figure 5 for E.ON SE.

Table 12 outlines the day-of-the-week feature selection process for the two models, namely, the raw day-of-week model (i.e. a model consisting only of five dummy variables for each workday) and the historical dynamic and day-of-week model (i.e. the model extending the historical dynamic model, which consists of endogenous variables, namely, volume features and price features, by adding a dummy variable for each workday). There is a breakdown of the coefficient sign distribution (i.e. positive and negative) across the stock universe for each workday, which is included under the presence proportion of each day-of-the-week (resulting from the stepwise regression). Monday is a notable day-of-the-week feature, which is consistently picked in the raw day-of-week model and in the historical dynamic and day-of-week model, where it is generally negatively correlated with the trading volume. The Monday coefficient is consistently negative despite the overnight return correction (i.e. dividing the overnight return by the number of intervening nights), which suggests that generally there is less trading activity on Mondays. This confirms the potential existence of a weekend effect on trading volumes. We also report predominantly negative coefficients for Fridays, although the Friday day-of-the-week feature is significantly picked only in the historical day-of-week model. Although the weekend

TABLE 9 Volatility feature presence across States A and B for the entire data set

Volatility feature	State A count	State B count
Intraday range	1,185	1,453
Intraday return abs	810	894
Intraday return absPos	649	646
Intraday return absNeg	840	734

TABLE 10 Volatility feature presence across States A and B for the sectional break data subsets

	Precrisis data (2000–2007)		Postcrisis data (2008–2015)	
Volatility feature	State A count	State B count	State A count	State B count
Intraday range	1,012	1,110	1,157	1,322
Intraday return abs	648	662	734	799
Intraday return absPos	488	485	549	558
Intraday return absNeg	543	493	802	726

TABLE 11 Day-of-the-week findings

	Observations percentage		
	Entire data set	Structural break subsets	
Hypothesis	2000–2015 (%)	2000–2007 (%)	2008–2015 (%)
Historical dynamic model is significantly better than the raw day-of-week model	99.87	99.44	99.79
The day-of-week features improve the historical dynamic model	90.57	88.29	88.61

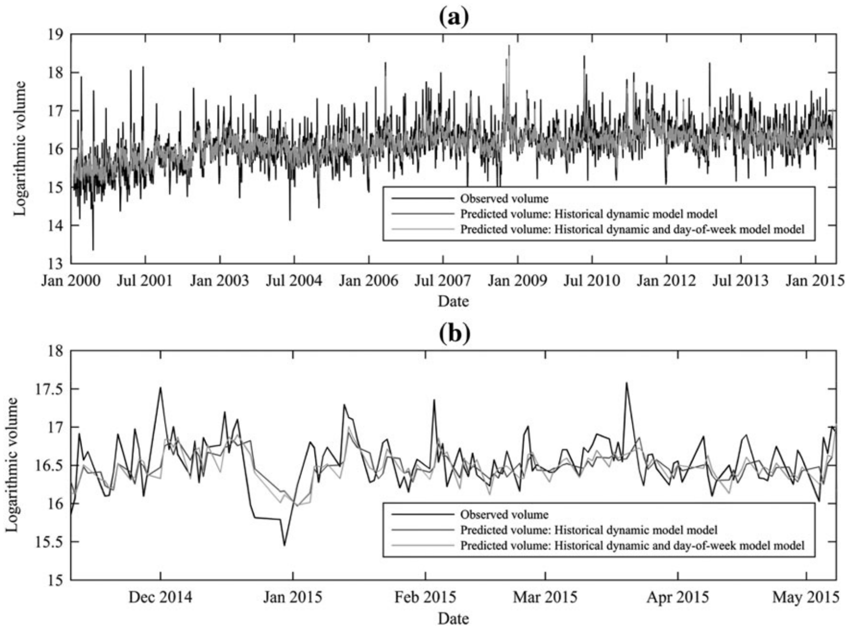

FIGURE 5 Day-of-week improvement over the historical dynamic model (7.03%) for E.ON SE (EONGn.DE). Panel (a) shows the complete time series between 24 January 2000 and 8 May 2015 (3,880 observations), whereas Panel (b) provides a zoomed view of the most recent 6 months

TABLE 12 Day-of-the-week feature selection—Presence percentage for each day of the week along with the distribution of coefficient signs

	Monday (%)	Tuesday (%)	Wednesday (%)	Thursday (%)	Friday (%)
Panel (a): Raw day-of-the-week model					
Occurrence	64.39	17.81	11.77	12.92	18.78
Positive coefficient	4.49	41.77	76.53	74.34	24.89
Negative coefficient	95.51	58.23	23.47	25.66	75.11
Panel (b): Historical dynamic and day-of-the-week model					
Occurrence	75.69	30.41	21.16	21.87	45.33
Positive coefficient	9.11	90.28	67.41	35.19	14.29
Negative coefficient	90.89	9.72	32.59	64.81	85.71

effect is documented as having higher than usual Friday returns and hence higher volumes (according to the literature on the volume–price relation), we observe a mostly negative Friday coefficient, which is associated with lower volumes.

The Monday and Friday feature presence and coefficient sign distribution for the structural break data subsets are outlined in Table 13. We observe a constant Monday effect for both day-of-the-week models throughout the precrisis and postcrisis periods. The

TABLE 13 Day-of-the-week feature selection for the structural break subsets

	Precrisis data (2000–2007)		Postcrisis data (2008–2015)	
	Monday (%)	Friday (%)	Monday (%)	Friday (%)
Panel (a): Raw day-of-the-week model				
Occurrence	58.56	21.81	63.54	14.02
Positive coefficient	4.22	18.29	4.62	53.03
Negative coefficient	95.78	81.71	95.38	46.97
Panel (b): Historical dynamic and day-of-the-week model				
Occurrence	70.73	44.48	72.09	35.11
Positive coefficient	8.49	11.83	9.58	21.04
Negative coefficient	91.51	88.17	90.42	78.96

Friday effect is more volatile though, and its coefficient becomes positive for more than 50% of occurrences in the raw day-of-the-week model trained on the postcrisis data.

The historical dynamic and day-of-week model was generally the most accurate model in this analysis. Figures 6 and 7 illustrate the predicted volume time series along with the observed volumes for two stocks (i.e. Royal Dutch Shell and Siemens); a zoomed plot for the most recent 6 months of modelling data accompanies these figures for better visualization.

Figure 8 depicts the error percentage change from the historical dynamic model to the raw day-of-the-week model, showing predominantly positive observations, meaning that the average MSE increases, making the model worse. We conclude that the traditional raw day-of-the-week model is inferior to the historical dynamic model. Similarly, Figure 9 illustrates the error percentage change between the historical dynamic model and the augmented model that adds day-of-the-week features on top of the historical dynamic model, with a dominantly negative distribution suggesting that the historical dynamic and day-of-week model lowers the average MSE and provides a better fit.

At this phase, the day-of-the-week analysis provides a discussion point, which leads to a further study on special events (e.g. cross-market holidays and stock index futures

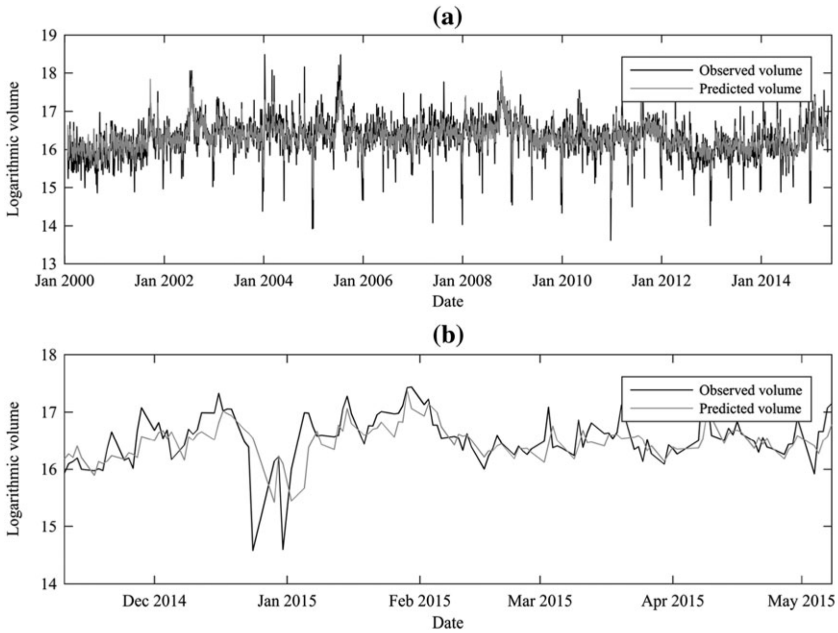

FIGURE 6 Observed volume and predicted volume using the historical dynamic and day-of-week model for Royal Dutch Shell PLC (RDSa.AS) for 3,909 daily observations (24 January 2000 to 8 May 2015). Panel (a) is a zoomed in plot of the most recent 6 months of data

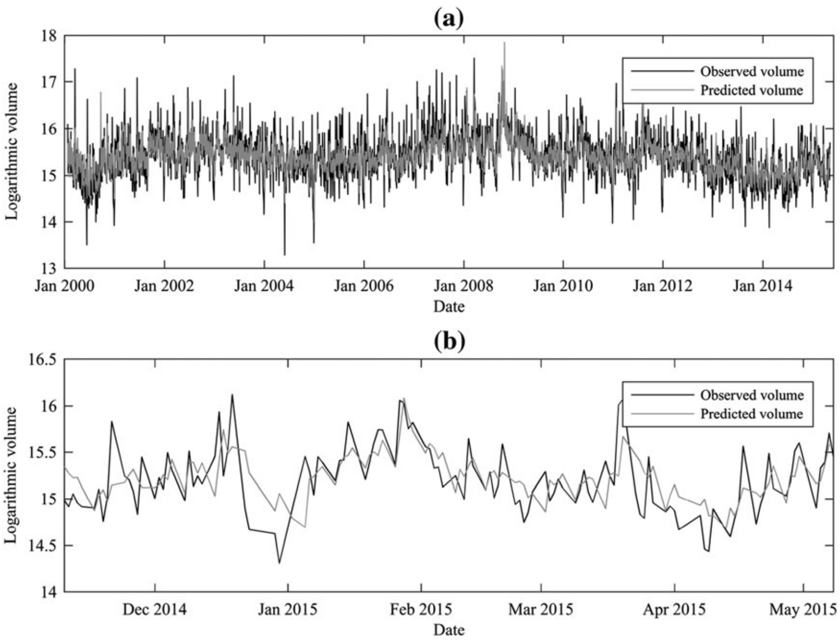

FIGURE 7 Observed volume and predicted volume using the historical dynamic and day-of-week model for Siemens AG (SIEGn.DE). Panel (a) illustrates the entire time period being studied, between 24 January 2000 and 8 May 2015 (3,880 trading days), whereas Panel (b) shows a magnified view of the most recent 6 months

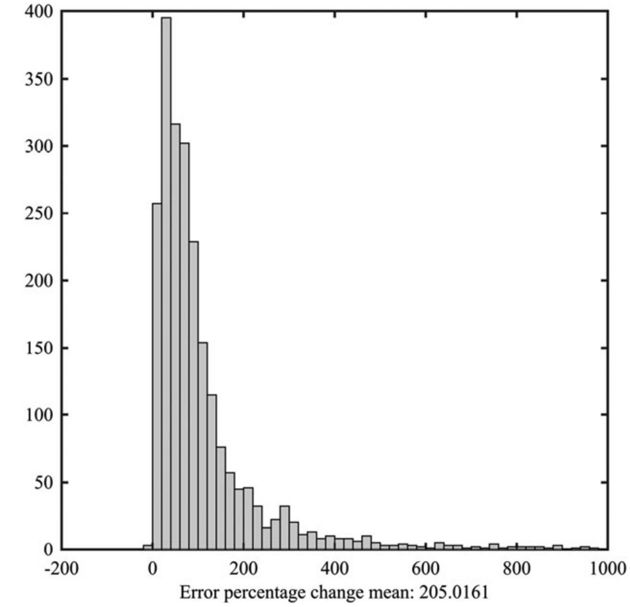

FIGURE 8 Histogram of error percentage change from historical dynamic model to raw day-of-week model for the entire stock universe (2,353 stocks)

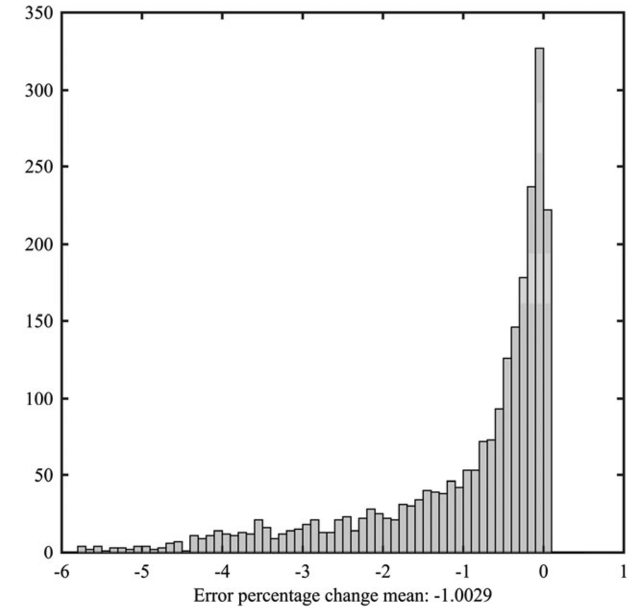

FIGURE 9 Histogram of error percentage change from historical dynamic model to historical dynamic and day-of-week model for the 2,353 stocks studied

expiries), which could potentially impact on the Friday and Monday volumes.

7 | DISCUSSION

This study provides a broad exploration of endogenous and exogenous factors driving trading volume, while investigating a number of relevant aspects, such as the volume–price relation asymmetry and the existence of structural breaks. The effect size is not part of the scope of this study mainly because we fit the models independently for each stock. The rationale is that there are strong stock-specific variability and magnitude levels that could not allow for a unified model across stocks. Instead, the aim is to identify the variables that help predict the trading volume of the following day. To the best of our knowledge, we provide the largest pan-European stock universe in any academic study. The extended data universe provides robust validation of our results.

We investigate potential structural breaks and nonstationarity around the financial crisis of 2007–2008 as a method of validating the results, which assume strong homogeneity. We split the data set into two folds: the precrisis data set (2000–2007) and the postcrisis data set (2008–2015).

The study considers single stock modelling and eventually aggregates the results across a data universe of 2,353 stocks. We provide empirical evidence of a significant improvement over the autoregressive volume model using volatility features (i.e. intraday range and intraday return for the previous trading day), more recent price information (i.e. overnight return as a proxy for the opening auction volume), and day-of-the-week features. The only constant day-of-the-week exerting a dominant influence over trading volumes is Monday, which improves the historical dynamic model in over 75% of the sample stocks. The coefficients are predominantly negative for Mondays, even though we divide the overnight return by the number of intervening nights; Monday's coefficient is not a corrective factor, and it suggests that there is less activity on Monday. Friday is the second most selected day-of-the-week feature, but it improves the model in only 45% of the times; its regression coefficient is mostly negative as well, although it is positive in more than 50% of the observed models for the raw day-of-the-week model using the postcrisis data subset.

The empirical evidence suggests a stronger day-of-the-week effect in conjunction with the endogenous variables. More notably, there is a Monday effect and a less salient Friday effect, both days exhibiting negative returns. This confirms the weekend effect literature with regard to lower Monday returns. However, the Friday returns tend to be negatively correlated with the volume. We have not specifically addressed special events in the context of this analysis. The reasoning behind this decision consists of the insufficient number of observations of special events in each fold (i.e. futures expiries and cross-market holidays, which have very few observations anyway), given the context of our stock-specific modelling. The day-of-the-week modelling provides a discussion point, which leads to a separate investigation of the special events, that is, whether cross-market holidays affect the Monday volumes and whether stock index futures expiries can potentially influence the Friday volumes.

We also examined the accuracy of the raw day-of-the-week model, which is traditionally employed in the weekend effect and day-of-the-week literature. It constantly underperformed compared to the historical dynamic model by a large factor, and this evidence suggests that fitting a dummy variable model for a particular effect in complete isolation from other variables, especially endogenous, would provide questionable results.

Another interesting aspect being analysed in this study is the asymmetry of the price–volume relation. We proposed a different framework for exploring the price–volume relation asymmetry. The empirical results suggest that we first need to discriminate between two types of price–volume relations; there is an intraday return (i.e. the previous day price returns)—volume relation, which manifests asymmetry in approximately 60% of the stocks, and there is an overnight return (i.e. newer data)—volume relation, which exhibits asymmetry in 54% of the stocks. Combining these two relations, we find an overall asymmetry in approximately 70% of the analysed stocks. We report a structural break with regard to the overnight return asymmetry, which is salient for the pre-crisis data, but then it becomes rather neutral and reaches similar performance as the symmetric overnight return. Apart from the Friday effect and the overnight asymmetry, where we found notable structural breaks, the study confirms data homogeneity for all of the other aspects being examined throughout the entire sample period, that is, 2000–2015.

The empirical results show that more recent information (i.e. overnight return) improves the volume prediction model and having the overnight return being used as a proxy for the opening auction volume confirms the price–volume relation. This could be further improved by performing an intraday volume prediction model to further analyse the price–volume relation based on tick data.

ACKNOWLEDGEMENTS

This research has been funded by the Engineering and Physical Sciences Research Council (EPSRC) UK.

ORCID

Bogdan Batrinca http://orcid.org/0000-0002-5281-0457
Christian W. Hesse http://orcid.org/0000-0003-4719-2153
Philip C. Treleaven http://orcid.org/0000-0001-8803-8297

REFERENCES

Abraham, A., & Ikenberry, D. L. (1994). The individual investor and the weekend effect. *The Journal of Financial and Quantitative Analysis*, *29*(2), 263–277.

Al-Deehani, T. M. (2007). Modelling asymmetry in the price–volume relation: Evidence from nine stock markets. *Investment Management and Financial Innovations*, *4*(4), 8–15.

Assogbavi, T., Khoury, N., & Yourougou, P. (1995). Short interest and the asymmetry of the price–volume relationship in the Canadian stock market. *Journal of Banking & Finance*, *19*(8), 1341–1358.

Assogbavi, T., & Osagie, J. E. (2006). Equity valuation process and price–volume relationship on emerging markets. *International Business & Economics Research Journal*, *5*(9), 7–18.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research*, *6*, 67–92.

Berument, H., & Kiymaz, H. (2001). The day of the week effect on stock market volatility. *Journal of Economics and Finance*, *25*(2), 181–193.

Berument, H., & Kiymaz, H. (2003). The day of the week effect on stock market volatility and volume: International evidence. *Review of Financial Economics*, *12*(4), 363–380.

Brusa, J., Liu, P., & Schulman, C. (2000). The weekend effect, 'reverse' weekend effect, and firm size. *Journal of Business Finance & Accounting*, *27*(1), 555–574.

Cross, F. (1973). The behavior of stock prices on Fridays and Mondays. *Financial Analysts Journal*, *29*(6), 67–69.

Dubois, M., & Louvet, P. (1996). The day-of-the-week effect: The international evidence. *Journal of Banking & Finance*, *20*(9), 1463–1484.

Edgington, E. S. (1964). Randomisation tests. *The Journal of Psychology*, *57*(2), 445–449.

Epps, T. W. (1975). Security price changes and transaction volumes: Theory and evidence. *The American Economic Review*, *65*(4), 586–597.

Epps, T. W. (1977). Security price changes and transaction volumes: Some additional evidence. *The Journal of Financial and Quantitative Analysis*, *12*(1), 141–146.

Fama, E. F. (1969). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, *25*(2), 383–417.

Flannery, M. J., & Protopadakis, A. A. (1988). From T-Bills to common stocks: Investigating the generality of intra-week return seasonality. *The Journal of Finance*, *43*(2), 431–450.

French, K. R. (1980). Stock returns and the weekend effect. *Journal of Financial Economics*, *8*(1), 55–69.

Gerety, M. S., & Mulherin, J. H. (1994). Price formation on stock exchanges: The evolution of trading within the day. *The Review of Financial Studies*, *7*(3), 609–629.

Gibbons, M. R., & Hess, P. (1981). Day of the week effects and asset returns. *The Journal of Business*, *54*(4), 579–596.

Godfrey, M. D., Granger, C. W. J., & Morgenstern, O. (1964). The random walk hypothesis of stock market Behavior. *Kyklos*, *17*, 1–30.

Harris, L. (1986). A transaction data study of weekly and intradaily patterns in stock returns. *Journal of Financial Economics*, *16*(1), 99–117.

Harris, M., & Raviv, A. (1993). Differences of opinion make a horse race. *The Review of Financial Studies*, *6*(3), 473–506.

Hong, H., & Stein, J. C. (2007). Disagreement and the stock market. *Journal of Economic Perspectives*, *21*(2), 109–128.

Jaffe, J., & Westerfield, R. (1985). The week-end effect in common stock returns: The international evidence. *The Journal of Finance*, *40*(2), 433–454.

Jain, P. C., & Joh, G.-H. (1988). The dependence between hourly prices and trading volume. *The Journal of Financial and Quantitative Analysis*, *23*(3), 269–283.

Karpoff, J. M. (1986). A theory of trading volume. *The Journal of Finance*, *41*(5), 1069–1087.

Karpoff, J. M. (1987). The relation between price changes and trading volume: A survey. *The Journal of Financial and Quantitative Analysis*, *22*(1), 109–126.

Keim, D. B., & Stambaugh, R. F. (1984). A further investigation of the weekend effect in stock returns. *The Journal of Finance*, *39*(3), 819–835.

Lakonishok, J., & Levi, M. (1982). Weekend effects on stock returns: A note. *The Journal of Finance*, *37*(3), 883–889.

Lakonishok, J., & Maberly, E. (1990). The weekend effect: Trading patterns of individual and institutional investors. *The Journal of Finance*, *45*(1), 231–243.

Moosa, I. A., Silvapulle, P., & Silvapulle, M. (2003). Testing for temporal asymmetry in the price–volume relationship. *Bulletin of Economic Research*, *55*(4), 373–389.

NYSE Euronext, 2016. *NYSE Group Turnover*. [Online] Available at: http%3A%2F%2Fwww.nyxdata.com%2Fnysedata%2Fasp%2Ffactbook%2Fviewer_edition.asp%3Fmode%3Dtable%26amp%3Bkey%3D3307%26amp%3Bcategory%3D3 [Accessed 23 03 2016].

Pettengill, G. N. (2003). A survey of the monday effect literature. *Quarterly Journal of Business and Economics*, *42*(3), 3–27.

Rogalski, R. J. (1984). New findings regarding day-of-the-week returns over trading and non-trading periods: A note. *The Journal of Finance*, *39*(5), 1603–1614.

Sias, R. W., & Starks, L. T. (1995). The day-of-the-week anomaly: The role of institutional investors. *Financial Analysts Journal*, *51*(3), 58–67.

Smirlock, M., & Starks, L. (1985). A further examination of stock price changes and transactions volume. *Journal of Financial Research*, *8*(3), 217–226.

Smirlock, M., & Starks, L. (1986). Day-of-the-week and intraday effects in stock returns. *Journal of Financial Economics*, *17*(1), 197–210.

Steeley, J. M. (2001). A note on information seasonality and the disappearance of the weekend effect in the UK stock market. *Journal of Banking & Finance*, *25*(10), 1941–1956.

Wang, K., Li, Y., & Erickson, J. (1997). A new look at the Monday effect. *The Journal of Finance*, *52*(5), 2171–2186.

Woord, R. A., McInish, T. H., & Ord, K. J. (1985). An investigation of transactions data for NYSE stocks. *The Journal of Finance*, *40*(3), 723–739.

World Federation of Exchanges, 2012. *2011 WFE Market Highlights*. [Online] Available at: http://www.world-exchanges.org/files/file/stats%20and%20charts/2011%20WFE%20Market%20Highlights.pdf [Accessed 20 05 2014].

World Federation of Exchanges, 2013. *2012 WFE Market Highlights*. [Online] Available at: http://www.world-exchanges.org/files/statistics/2012%20WFE%20Market%20Highlights.pdf [Accessed 20 05 2014].

Ying, C. C. (1966). Stock market prices and volumes of sales. *Econometrica*, *34*(3), 676–685.

How to cite this article: Batrinca B, Hesse CW, Treleaven PC. Examining drivers of trading volume in European markets. *Int J Fin Econ*. 2018;23:134–154. https://doi.org/10.1002/ijfe.1608