HEC MONTRÉAL

Forecasting volatility using liquidity measures in a high frequency returns model

par

Guillaume Trudel

Sciences de la gestion
(Option ingénierie financière)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

Juin 2019

Résumé

Dans cet article, nous explorons des possibles améliorations aux prévisions de la volatilité réalisée quotidienne, calculée à partir de rendements à haute fréquence, en faisant usage de l'information contenue dans certaines mesures de liquidité et d'activité du marché. Plus particulièrement, nous insérons les écarts d'offre et demande et les volumes transactionnels en dollars ou en actions comme termes additionnels dans un modèle autorégressif linéaire de volatilité réalisée quotidienne à mémoire approximativement longue, que nous nommons HAR-RV-LIQ. Nous construisons un échantillon de 10 actions américaines composantes du S&P500 durant la période 2005-2013 et analysons pour chacune des actions individuellement la performance de notre modèle. En utilisant une méthode d'estimation de type moindres carrés pour les diverses spécifications de notre modèle, nous comparons la qualité de l'ajustement en échantillon et la performance des prédictions hors échantillon en termes de RMSE, suivi des quelques diagnostiques des erreurs de prédictions. Malgré que les gains de l'ajout des termes d'écart d'offre et demande ou de volume dans notre spécification du modèle de volatilité sont selon nos résultats marginalement faibles, nous trouvons qu'il existe une certaine quantité d'information dans les mesures de liquidité et d'activité qui peut bel et bien améliorer les prédictions de volatilité. De plus, même si nous observons que les écarts d'offre et demande ou le volume peuvent améliorer dans la plupart des cas les prédictions de volatilité pour une action particulière, nous n'identifions pas de choix unique évident applicable à l'ensemble de l'échantillon.

Mots-clés

prévision de volatilité, liquidité, données à haute fréquence, volatilité réalisée, dépendance à long terme

Abstract

In this paper, we investigate possible improvements in the forecasting of daily realized volatility computed from high frequency returns by using information in market liquidity and activity measures. In particular, we look at the proportional and effective bid-ask spreads as well as the trading volumes in dollars and in shares as additional terms in a linear autoregressive approximate long memory model of realized volatility, dubbed HAR-RV-LIQ. To this end, we construct a sample of 10 U.S. stocks that were components of the S&P500 during the 2005-2013 periods and analyse the performance of our model for each individually. Using least square estimation for the various specifications of our model, we compare in-sample fits and out-of-sample prediction performance in terms RMSE and various diagnostic of prediction errors. Despite the gains of including spread and volume in our current daily realized volatility specification are small at the margin according to our results, we do find that there exist some information in measures of activity and liquidity that can be harnessed to improve volatility forecast performance. Furthermore, even though we find that a choice of volume or spread term in most cases can improve volatility forecasting performance, we do not identify an obvious unique choice applicable to the entire sample.

Keywords

volatility forecast, liquidity, high-frequency data, realized volatility, long memory

Contents

List of Tables

List of Figures

List of acronyms

ACF Autocorrelation function

AR Autoregressive

ARCH Autoregressive conditional heteroskedasticity

ARFIMA Autoregressive fractionally integrated moving average

BSX Boston Scientific Corporation

CMS CMS Energy Corporation

CVX Chevron Corporation

d Daily

ESPR Effective bid-ask spread

GARCH Generalized autoregressive conditional heteroskedasticity

HAC Heteroskedasticity and autocorrelation consistent

HAR-RV Heterogeneous autoregressive model for realized volatility

HAR-RV-LIQ Heterogeneous autoregressive model with liquidity predictors for realized volatility

HPQ HP, Inc. (formerly Hewlett-Packard Company)

Kurt. Kurtosis

LLTC Linear Technology Corporation

m Monthly

MON Monsanto Company

NKE NIKE, Inc.

NUE Nucor Corporation

NYSE New York Stock Exchange

OLS Ordinary least squares

PQSPR Proportional quoted bid-ask spread

PESPR Proportional effective bid-ask spread

QSPR Quoted bid-ask spread

RMSE Root mean square (prediction) error

RMSPE Root mean square percentage (prediction) error

RV Realized volatility

Skew. Skewness

SSE Sum of square errors

Std. Standard deviation

Sub Subsampling estimator

TAQ New York Stock Exchange Trade and Quote

US United States of America

VIF Variance inflation factor

VLMD Trading volume in dollars

VLMN Trading volume in number of shares

w Weekly

WMB The Williams Companies, Inc.

WRDS Wharton Research Data Services

X United States Steel Corporation

Acknowledgements

I wish to first and foremost thank Geneviève Gauthier, Ph.D., full professor at HEC Montréal, for her sustained support, supervision and patience throughout my entire research process. Credits for the research idea itself is due to Geneviève Gauthier and Diego Amaya, Ph.D., assistant professor at Wilfrid Laurier University, who also contributed with data collection and guidance.

I also wish to express my gratitude to my loved ones who have only showed support and understanding from beginning to end.

Chapter 1

Introduction

Volatility is at the heart of the ever ongoing quest of more accurately predicting the distribution of future stock prices. It is in a simplistic way the expected magnitude of an average surprise beyond the expected return. It is a rough frame around the elements of the market that we fail to measure or understand. The smaller the volatility, the more we anticipate the market to pursue its current trend undisturbed, while a high volatility makes it impossible to pinpoint future returns with small bounds of confidence. A better understanding of this key component of returns models has a wide array of applications, such as a more accurate pricing of derivative instruments and more informed risk management decisions, all of which can have a significant impact on the bottom line of a financial firm.

In more recent years, the use of intraday high frequency returns to obtain a non-parametric estimator of stock returns volatility has gained attention, as it was demonstrated to asymptotically converge to the implied volatility of the returns underlying stochastic process (Barndorff-Nielsen and Shephard, 2002; Andersen et al., 2001, 2003). It is well known that volatility tends to persist and is a strongly autocorrelated process. Corsi (2009) shows that a simple linear model using lagged values of realized volatility can take advantage of the autocorrelated nature of volatility to yield decent predictions of future volatility. This alternative to GARCH models has the advantage to exhibit apparent long memory, i.e. long term persistence of shocks in volatility and a very slow decay of their autocorrelation function. Absolute returns and volatility have been documented to show evidence of long memory (Ding et al., 1993; Granger and Ding, 1995). To incorporate a semblance of long memory, Corsi's model simply defines the daily realized volatility as the sum of the past day, the past week's average and past month's average realized volatility. These last forcibly introduce long lasting autocorrelation in daily returns.

Beyond trying to predict the trend of volatility using its own past values and in returns themselves, we seek to find whether there exist other time series that can shed additional light on the current trend of volatility. Other market time series informative on the behaviour of agents in the market could also be informative on volatility, which is in essence itself connected to what we cannot measure about this very mass of market agents.

In the context of the study of stock returns dynamics, the market liquidity and intensity of activity are often overlooked. The concept of liquidity refers to the depth of the market, in other words the ease at which one agent can readily find another to transact with. Activity refers to the frequency of transactions for a particular asset. Should we know more about the current behaviour of market agents and include this factors into a correctly specified model, the remaining unknown and noise should be lessened and volatility better estimated. Both concepts of liquidity and activity have readily observable and available market time series that measure their effect. We thus turn our attention to the bid-ask spread as a proxy for liquidity and to trading volume as a natural proxy for activity. We hope to show evidence that the consideration of such time series can improve volatility forecasting.

In the next section, we delve a bit deeper in the ongoing literature surrounding long memory realized volatility models, as well as the connection between liquidity and volatility. In section 3.1, we present the construction of our sample of 10 US companies used in this study of volatility as well as our selected measures of liquidity and activity. In section 3.2, we investigate stylized facts and summary statistics of realized volatility, bid-ask spread and volume time series, mainly in terms of moments, cross-correlations and autocorrelations. In section 3.3, we present our selected HAR-RV-LIQ linear forecasting model of realized volatility, which is itself an extension of Corsi's HAR-RV model that includes spread and volume terms and which is estimated by least square techniques. We opt to present the literature surrounding realized volatility itself alongside the formulation of our model, as it is better explained in the context of their equations. In section 4.1, we estimate our models for all 10 selected stocks over the 2005-2013 period and present the results and their analysis. In section 4.2, we analyse the forecasting performance of the HAR-RV-LIQ model out-of-sample. Finally, in section 4.3, we do further analysis of the distribution of residuals and of the fit for each firm's model with the best out-of-sample performance.

Chapter 2

Literature review

2.1 Long Memory Realized Volatility Models

"Long memory" in stock returns generally refers to the persistence of positive autocorrelations in absolute returns over very long lags and similarity in returns distribution when aggregating over various time intervals (Ding et al., 1993). More formally, this refers to *apparent* long memory, while true long memory refers to a specific class of time series model with hyperbolic autocorrelation decay and self-similarity obtained from using a fractional differencing operator on a white noise (Granger, 1980; Baillie, 1996). It has been documented that absolute returns, which are closely related to volatility, exhibit a slow decay in their autocorrelation function (Granger and Ding, 1995), which implies that a large shock in volatility at a given point in time will tend to leave slowly fainting echoes throughout returns for very long periods of times. In a model with long memory, an extremely volatile day would cause subsequent days to also be highly volatile, which in absence of new large exogenous shocks would only very slowly over the course of weeks or months stabilize back to normal levels. Such property can mimic well the behaviour of persistently turbulent markets following a financial crash.

It has also been shown that the implied volatility of a stochastic model can be non-parametrically approximated using discretely sampled high frequency squared returns

(Andersen et al., 2001; Barndorff-Nielsen and Shephard, 2002), which has opened the door to an ever growing field of high frequency realized volatility models. Such realized volatility time series exhibit long-run persistence in autocorrelation, which has led authors to advocate for true long-memory models using fractionally-integrated models (see for example Baillie et al. (1996)). The analytical complexity of such models has led researches to seek simpler alternatives that can still produce long-run dependencies within realized volatility time series.

Corsi (2009) proposed the linear HAR-RV model, a constrained AR model of realized volatility with daily, weekly and monthly lag components. Corsi shows that the HAR(3)-RV not only exhibits approximate long-memory, but also outperforms the corresponding unconstrained AR(22) model in goodness-of-fit tests and provides similar out-of-sample volatility forecast performance to the fractionally integrated ARFIMA(5,d,0) on the S&P500. Realized volatility, which literature and pitfalls we cover in more details in section 3.1.4, is in essence a non-parametric estimate of the volatility of returns over a short period of time (e.g. a day) obtained using only higher frequency returns observed within that period of time. As a result, it reacts extremely quickly to changing market conditions and is not constrained by a choice of parametric distribution. Realized volatility estimates have been shown to converge to the time-dependent implied volatility of a supposed stochastic dynamic of stock returns (Barndorff-Nielsen and Shephard, 2002; Andersen et al., 2001, 2003).

2.2 Liquidity and Volatility

The idea that a stock liquidity might contains useful information for predicting stock returns or stock volatility has long been studied. For example, Karpoff (1987) reviews the early research in the price-volume relationship for individual stocks, covering notably some evidences of an existing relationship between higher moments of returns and trading activity. Schwert (1989) finds that unexpected aggregate stock market volume and volatility shocks tend to occur together on the NYSE, but finds little evidence of correlation between volatility and lagged values of volume. Among possible factors linking volume and volatility, Schwert mentions new information flow, causing both trading and price changes through heterogeneous beliefs of investors, and trading noise, as market agents follow on a price shift and thus further increase trading activity and volatility both. Copeland and Galai (1983) proposed a theory, at the time verified by empirical literature, that conceptualizes the bid-ask spread as a straddle on the stock, thus a positive function of the volatility expected by market makers.

It is through the intricate web of transactions that the price is allowed to fluctuate. The volatility of price returns is a measure of the intensity of this fluctuation, while the liquidity is a measure of the frequency and ease at which transactions happen. To observe a change in price, there must be transactions. Thus, when we measure volatility and liquidity, we are in fact measuring two linked consequences of the very same set of complex market mechanisms. Campbell et al. (1993), in their study of market transactions and information, argue that high trading volume occurs jointly with large shifts in demand from noninformational traders. In the landmark article from Black (1986) about the nature of noise, he similarly states that there cannot be liquidity without the noise of noninformational traders, as those who trade on more complete information may generally not desire to trade with one another. We therefore adopt the view that the activity of noninformational traders may give rise to both high volume and high price volatility.

We do not suppose volatility to be the cause of liquidity or vice-versa, but that they are both related by the same underlying forces. In particular, we wish to test the existence of persistent information within recent past values of liquidity measures that is not already carried by recent past values of volatility, which can in turn be used to infer on the direction of future volatility. The main hypothesis of our preliminary research is for this relationship to be positive and suited to a linear representation. It is however entirely possible that the complex mechanisms underlying the relation between volatility and liquidity cannot be approximated by a linear form, that the relationship is entirely concurrent or that the information set is fully overlapped, in which case an autoregressive predictive model of volatility cannot be improved by the linear addition of liquidity factors.

We opt to investigate the potential of bid-ask spread and trading volume to improve predictions of realized volatility. Alternate measures of liquidity and activity that we did not select for our study are numerous in the literature. For example, Dufour and Engle (2000) suggests a measure of price impact based on a vector autoregression model of trade sign and price change following a trade. Amihud (2002) suggests a measure of illiquidity based on the ratio of absolute returns and trading volume in dollars averaged over time and shows that this measure explains part of the variance of excess stock returns. In theory investors demand higher expected returns when they anticipate higher risk. Should liquidity relate to the ex-ante risk premium, then it should also relate to the ex-ante risk, which itself is conceptually connected to ex-ante volatility. Chordia et al. (2000) also find evidence of common drivers underlying the dynamics of various measures of liquidity and activity, but leave open the questions as to what those drivers are. We must thus take care not to arbitrarily include too many measures of liquidity or activity in our linear models, for it increases the possibilities of inflated variance estimator due to multicollinearity. As such, we are motivated to open our study with the simplest and most common measures of liquidity, the bid-ask spread and the trading volume.

Chapter 3

Data and Methodology

3.1 High Frequency Stock Price and Liquidity Data

We study the relationship between volatility and measures of liquidity and trading activity for 10 companies that were components of the S&P500 from the beginning of 2005 to the end of 2013.

3.1.1 Sample Description

The time series discussed in this section are sampled at the one minute interval from the NYSE Daily TAQ database, for each trading day from the beginning of 2005 to the end of 2013, excluding half-days due to U.S. holidays. We sampled data for each company recognized within the database as a component of the S&P500. We plot in figure 3.1 the number of stocks for each day that the Daily TAQ database includes within the S&P500. As expected, our sample consists of roughly 500 companies per day, with the notable exception of August 1st, 2012, with only 459 stocks available. However, this date corresponds to a market day plagued by a trading glitch, caused by the brokerage firm Knight Capital Group, which resulted in abnormal trading volume and price volatility levels for close to 150 stocks (Scaggs, 2012). As such, it is not warranted to exclude this abnormal trading day from our sample, as its data possibly includes information on the price-volume relationship.

For each stock of the S&P500 with available data and each day, we sampled from the NYSE Daily TAQ 5 time series. First, the series of midquotes were extracted and serve as the observed market price of our stocks. Second, we selected the measures of liquidity quoted spread (QSPR) and effective spread (ESPR), as well as the measures of trading activity volume in number of shares (VLMN) and volume in US dollars (VLMD). We present in equations (3.1) to 3.3 the standard methods to compute these measures from bid, ask and transaction price quotes.

$$\mathrm{Mid}_{i,t} = (\mathrm{Bid}_{i,t} + \mathrm{Ask}_{i,t})/2 \tag{3.1}$$

$$\mathrm{QSPR}_{i,t} = \mathrm{Ask}_{i,t} - \mathrm{Bid}_{i,t} \tag{3.2}$$

$$\mathrm{ESPR}_{i,t} = |P_{i,t} - \mathrm{Mid}_{i,t}| \times 2 \tag{3.3}$$

where i and t are indices along the company and time dimensions respectively, and $P_{i,t}$ is the last transaction price. The quoted spread is the transaction cost of immediately buying and selling a share and represents the immediate liquidity of a stock, absent price impact. The effective spread measures the same cost using instead the difference between midquote $\mathrm{Mid}_{i,t}$ and transaction price $P_{i,t}$ as the true (symmetric) spread paid by an investor in an hypothetical instantaneous transaction.

We choose to express those costs relative to price, on a scale akin to returns. We calculate the proportional quoted spread (PQSPR) and proportional effective spread (PESPR) from the preceding two measures by dividing every observation by the prevailing midquote.

$$\mathrm{PQSPR}_{i,t} = \mathrm{QSPR}_{i,t}/\mathrm{Mid}_{i,t} \tag{3.4}$$

$$\mathrm{PESPR}_{i,t} = \mathrm{ESPR}_{i,t}/\mathrm{Mid}_{i,t} \tag{3.5}$$

3.1.2 Company Selection

To reduce the dimensionality of our study, we select what we consider to be 10 representative companies within the index. Since we desire uninterrupted time series throughout

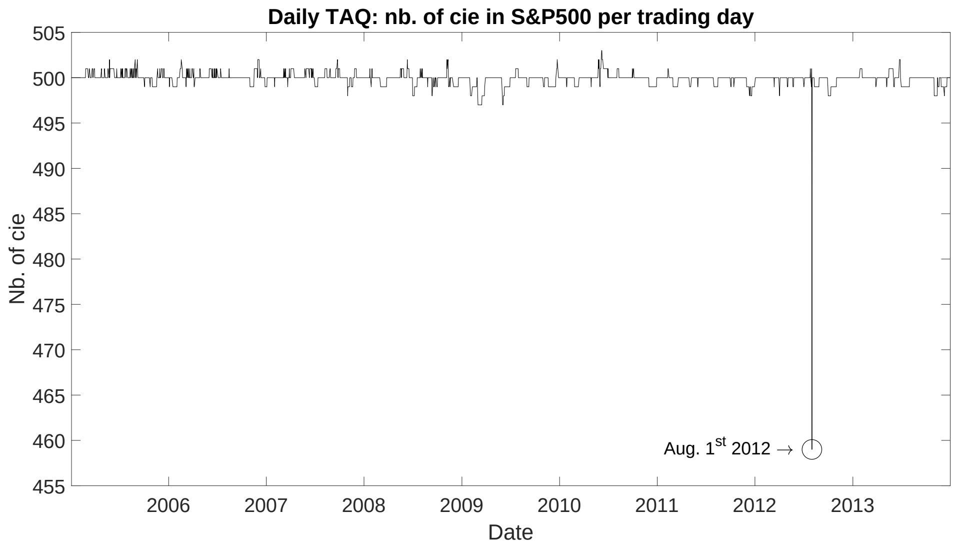

Figure 3.1: **S&P500 stocks and the Daily TAQ data availability**
Quantity of stocks per day in our dataset, following a query on the daily TAQ database for stocks flagged as components of the S&P500. Our dataset roughly includes 500 firms per day as being part of the index, with very little variation in the count from day to day, except notably for August 1^{st}, 2012, with a count of 459 stocks.

our sample, some survivorship bias cannot be avoided. To each day t corresponds a set C_t of approximately 500 company tickers. As a first cut, we have excluded from our analysis any ticker not member of the set $\cap_{t=1}^{T} C_t$, meaning that we exclude any company that for any given day within the full sample period were not a component of the S&P500 or for which there were one or more days without a single minute of data.

We identified 237 such companies with uninterrupted data presence within the index. We then separated these into 10 deciles, according to their market capitalization in January 2005 at the beginning of the sample period, obtained from the WRDS database. We randomly picked one company from each decile, allowing us to potentially observe how our models perform for firms of different size. We present the selected companies in table 3.1 and all stocks are hereafter referred to by their ticker symbol. In figure 3.2, we plot for each of these firms the log of the monthly average market capitalization, which allows us to see the evolution of the firm size ordering through time. The 2005 ranking in market capitalization was not preserved throughout the sample period, which will make inferences on the impact of firm size on the performance of our models difficult.

Decile	Mkt Cap. (2005-01)	Ticker	Name	Industry
1	$2,023,203,140	CMS	CMS Energy Corp.	Utilities
2	$5,603,878,680	X	United States Steel Corp.	Steel
3	$8,026,140,950	NUE	Nucor Corp.	Steel
4	$8,641,699,560	WMB	The Williams Cos., Inc.	Energy
5	$11,701,370,620	LLTC	Linear Technology Corp.	Technology
6	$14,525,120,250	MON	Monsanto Co.	Biochems
7	$16,912,052,580	NKE	NIKE, Inc.	Cons. goods
8	$28,957,116,330	BSX	Boston Scientific Corp.	Healthcare
9	$63,568,873,950	HPQ	HP, Inc.	Technology
10	$107,837,350,800	CVX	Chevron Corp.	Energy

Table 3.1: **Selected companies**
Set of selected stocks for our study. Selected companies have continuous daily presence within the S&P500 index from January 2005 to December 2013. The set of stocks with continuous daily data availability within the Daily TAQ database were ranked by their beginning of January 2005 market capitalization and 1 firm was randomly selected from each resulting decile.

3.1.3 Treatment of Missing Data, Outliers and Daily Aggregation

We have structured the data into daily blocks of 390 minutes. Missing data for any given minutes were filled by interpolation. Missing values for volume in shares and dollars were set to 0, under our assumption that the majority of missing data corresponded to an absence of transactions. Price or spread time series were interpolated piece-wise constant with limit to the left or exceptionally limit to the right for any block of missing values that included the first minute of the day. Negative values for spread measures were treated the same as if missing. In total 0.31% of all data was interpolated. We break down the proportion of data interpolated by company and by measure in table 3.2. Notice that CMS has both the highest amount of missing data interpolated and the lowest average trading volume in dollars throughout the sample period (roughly 3 times lower than the 2^{nd} lowest, see figure 3.9 and table A.1.5), which is consistent with our hypothesis that the observed gaps in the database correspond to a lack of transactions.

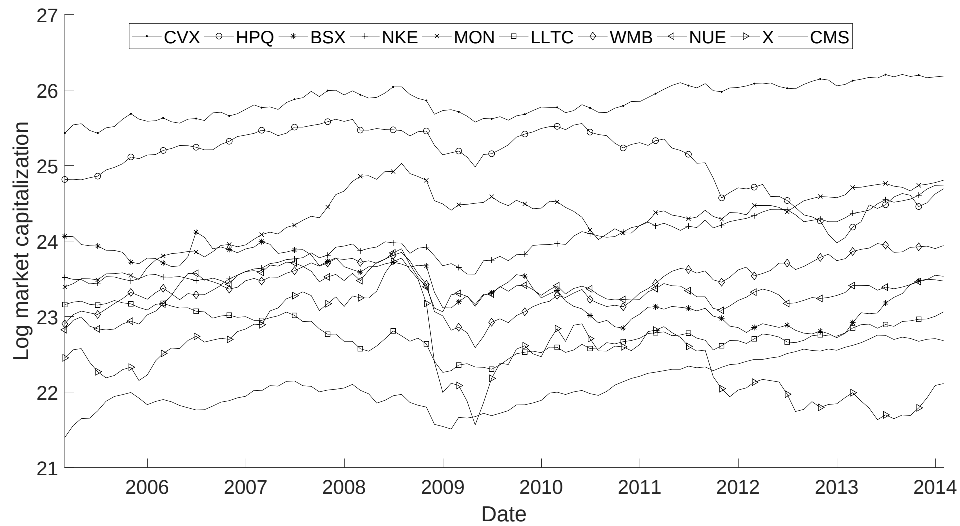

Figure 3.2: **Log market capitalisation, 2005-2012**
Log of the monthly average market capitalisation for each stock, for the purpose of viewing firm size ordering through time.

	Missing Data (%)					Neg. Spread (%)	
Ticker	**Midquote**	**QSPR**	**ESPR**	**VLMN**	**VLMD**	**QSPR**	**ESPR**
CMS	2.39	2.39	2.39	2.38	2.39	0.10	0.43
X	0.26	0.26	0.26	0.25	0.26	0.20	0.03
NUE	0.36	0.36	0.36	0.36	0.36	0.16	0.07
WMB	0.28	0.28	0.28	0.28	0.28	0.13	0.06
LLTC	0.64	0.64	0.64	0.64	0.64	0.67	0.16
MON	0.40	0.40	0.40	0.39	0.40	0.10	0.09
NKE	0.64	0.64	0.64	0.63	0.64	0.11	0.11
BSX	0.11	0.11	0.11	0.11	0.11	0.15	0.10
HPQ	0.02	0.02	0.02	0.02	0.02	0.16	0.01
CVX	0.08	0.08	0.08	0.07	0.08	0.18	0.00

Table 3.2: **Proportion of data interpolated**
Percentage of data interpolated per stock for their midquote, quoted (QSPR) and effective spread (ESPR) as well as volume in number of shares (VLMN) and in dollars (VLMD) time series. Interpolation was performed on a time series for any given minute with missing data or for non-positive elements of spread time series. Refer to section 3.1.3 for the methodology.

To prevent bias induced by outliers in our regression sample, we have applied a winsorization at the 0.01^{th} and 99.99^{th} percentiles to the interpolated time series, after ag-

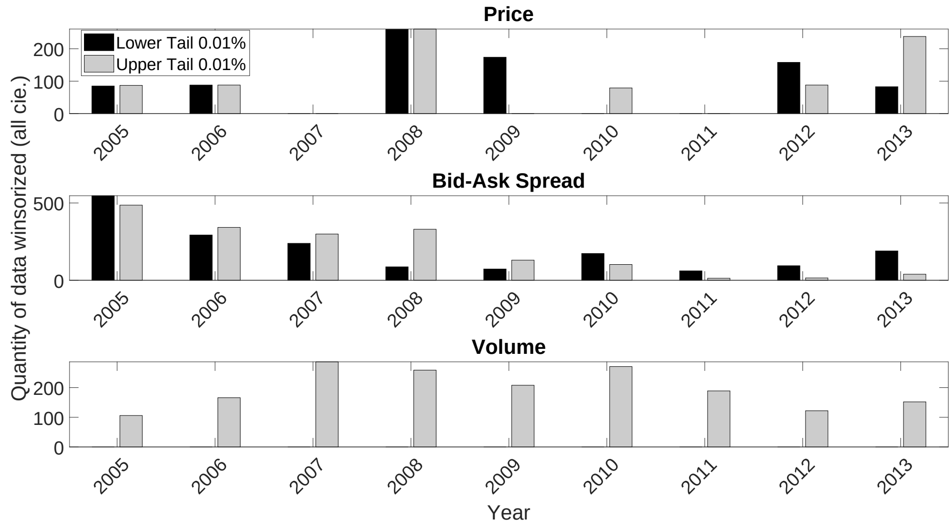

Figure 3.3: **Time distribution of tail values**
Distribution by year of tail values (0.01^{th} and 99.99^{th} percentiles in black and gray respectively) that were winsorized for the price, spread and volume time series. The counts are summed up for the proportional quoted (PQSPR) and effective spread (PESPR), as well as for the volume in number of shares (VLMN) and in dollars (VLMD).

gregation over all trading days. We show in figure 3.3 the time distribution of tail values affected by the winsorization, with the counts summed across all 10 companies. Different measures of spread and volume have been grouped. Intraday volume values of zero have not been adjusted by winsorization, as it consisted of more than 0.01% of the sample. The period of the financial Crisis (or pre-Crisis for spread measures) has been more impacted by the winsorization than other time periods, but not in proportions and quantities that we judge unacceptable for inference. Consider the price time series for illustration: in total, only less than 300 minutes in each tail over the entirety of 2008 were adjusted.

Intraday prices are aggregated into daily realized volatility time series. Spread and volume intraday time series are transformed into daily series by taking the daily average. All 10 companies had trading activity for at least 1 minute per day in sample, as per our company selection filter, which results in no single daily value of 0 for the volume series.

3.1.4 Realized Volatility

For a given stock, its continuous unobserved true log-price process is assumed to follow an Itô process,

$$dX_t = \mu_t \, dt + \sigma_t \, dW_t \tag{3.6}$$

where μ_t and σ_t are respectively the predictable drift and volatility stochastic processes, and $\{W_t\}_{t\geq 0}$ is a Wiener process, such that we can approximate the cumulated volatility $(\int_{t-1}^{t} \sigma^2(s)\, ds)^{\frac{1}{2}}$, over a time period of unit of interest such as a trading day, by an estimator known as the realized volatility (Andersen et al., 2001). For simplicity, we omit for this section only any indices relating to the stock itself, but it can be understood that every single stock has its own drift μ_t, volatility σ_t and noise W_t process.

For what follows, we use the notation $r_t = X_t - X_{t-j}$ for the discrete intraday log-returns, for some arbitrary time increment j. A basic estimator of daily realized volatility is the square root of the the sum of intraday squared log-returns,

$$\mathrm{RV}_t^{\mathrm{Basic}} = \sqrt{\sum_{j=1}^{M} r_j^2} \tag{3.7}$$

where $r_1, \ldots, r_M$ is a sequence of intraday log-returns corresponding to a fine time partition of the trading day. However, as M grows larger, this estimator is only convergent to the intraday volatility in the absence of microstructure noise. The larger M is, i.e. the higher the available data sampling frequency is, then more biased upward $\mathrm{RV}_t^{\mathrm{Basic}}$ will be. This is because a more significant share of total observed price variance is due to microstructure noise effects, for example bid-ask bounces, rather than due to the variance of the unobserved theoretical continuous price process (Hansen and Lunde, 2006).

We can correct for this bias by sampling at a lower frequency than available, which gives us the naive RV estimator,

$$\mathrm{RV}_t^{\mathrm{Naive}}(K, a) = \sqrt{\sum_{j=1}^{M_K} r_{jK-a}^2} \tag{3.8}$$

where K relates to the sampling frequency and $M_K = \lfloor M/K \rfloor$, such that every K^{th} return is sampled, starting from the $(K-a)^{\mathrm{th}}$ intraday return, $a \in \{0, ..., K-1\}$. This approach divides the available intraday sample into K non-overlapping grids $\{G_0, G_1, \ldots, G_{K-1}\}$ of daily returns, where only the grid G_a is used for the approximation, and the other $K-1$ grids are discarded. This naive estimator is well documented to be flawed, where on one hand at fast frequencies it is biased upwards by ignoring the effect of microstructure noise and on the other hand at slow frequencies it inefficiently discards significant portion of the collected data sample (Zhang et al., 2005; Aït-Sahalia et al., 2005).

Several methods are used in the literature to correct for the deficiencies of the estimator. Hansen and Lunde (2006), for example, investigate the use of various covariance kernel estimators of RV. Aït-Sahalia et al. (2005) show that if we explicitly model the observed log-price process $\{X_t^*\}_{t\geq 0}$ as a measurement with error of the efficient log-price process $\{X_t\}_{t\geq 0}$, i.e. $X_t^* = X_t + \varepsilon_t$, and derive an estimator robust to noise, then it becomes optimal to sample at the highest frequency possible and make use of all available data. Jacod et al. (2009) suggest the pre-averaging method, which attempts to approximate the latent efficient price process by averaging out successive noise terms in the observed price process. Barndorff-Nielsen et al. (2011) propose the realized kernel approach, which can also be used to estimate the covariance term between difference stock price processes. Zhang et al. (2005) show that it is also possible to reduce the variance of the estimator and avoid cumulating microstructure noise by computing the subsampling estimator, which is done very simply by taking the average of the sparsely sampled naive estimator over all K grids,

$$\begin{aligned} \mathrm{RV}_t^{\mathrm{Sub}}(K) &= \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\left(\sum_{j=1}^{M_K} r_{jK-k}^2\right)} \\ &= \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\mathrm{RV}_t^{\mathrm{Naive}}(K,k)^2} \end{aligned} \tag{3.9}$$

where K is the number of subsampling grids. Note that $\mathrm{RV}_t^{\mathrm{Sub}}(1) = \mathrm{RV}_t^{\mathrm{Naive}}(1,1) = \mathrm{RV}_t^{\mathrm{Basic}}$.

This simple to implement estimator of RV can approximate robust results without the added complexity of explicitly modelling the microstructure noise effects, by averaging out its impact over multiple grids. It is popular in the litterature and, for example, is the approach used in Corsi (2009). We thus select the subsampling estimator for its simplicity and for continuity in methodology with our benchmark model's author. We however wish our readers to be aware of the multiple competing approaches to measuring realized volatility and of possible refinements to our methodology. Nevertheless, Amaya et al. (2017), in the case of options on the S&P500 index, show that the $K = 5$ minutes subsampling estimator of realized variance offers superior robustness relative to the naive method and similar results to kernel-based or pre-average methods. Moreover, Liu et al. (2015) compare 400 different estimators of RV for 31 assets over 5 asset classes and conclude that there is little evidence that any robust approach significantly outperforms the naive 5-minutes estimator. Still, we consider it prudent to make use of all available data and use the more robust 5-minutes subsampling estimator rather than the 5 minutes naive one.

We use our sample to investigate the impact of the choice of the subsampling window K on the daily $\mathrm{RV}^{\mathrm{Sub}}$. Given K, for every day t in our sample and for every company, we compute $\mathrm{RV}_t^{\mathrm{Sub}}(K)$. For fixed K and company, we then calculate for the $\{\mathrm{RV}_t^{\mathrm{Sub}}(K)\}_{t=1}^{T}$ time series its sample mean, standard deviation, median and quantiles 25% and 75%. The results are plotted in figure 3.4. Note that as K goes to 1 minute, the subsampling estimator approaches the naive estimator, because our data is limited to 1 observation per minute. As expected, for all 10 companies, the sample mean and quartiles of $\{\mathrm{RV}_t^{\mathrm{Sub}}(K)\}_{t=1}^{T}$ steadily increases as K decreases, because longer windows fail to capture the volatility underlying any price change that reverts itself within a period of length K and because shorter windows increasingly capture market microstructure noise. The sample standard deviation also rises inversely with K, steadily in some cases (e.g. BSX), or very sharply at low levels of K. For example, X and LLTC from $K = 10$ to $K = 1$ see a large increase in standard deviation of 0.1, while none of the company with a more steady curve see a range of standard deviation larger than 0.035 for the entire curve. In the cases of X and LLTC, this seems indicative of large spikes in the second moment of the estimator at low very frequencies (i.e. lower than $K = 5$) and thus undesirable. This provides evidence that in the case of individual liquid stocks and per-minute data, a subsampling window lower than $K = 5$ may yields an estimator that is both biased upward and inefficient. In view of the literature and of our analysis, we find the 5 minutes subsampling window to be adequate for our data.

3.2 Stylized Facts

Before implementing our predictive models, we first do a summary analysis of stylized facts exhibited by our sampled realized volatility and liquidity series. The period studied for this section is from January 2005 to December 2013. The daily RV is calculated using the 5 minutes subsampling estimator.

3.2.1 Graphical Investigation and Summary Statistics

For each of the 10 companies, the daily log-realized volatility time series in figure 3.5 visibly exhibit strong persistence of shocks. For example, we observe for each company a slow reversion to long term levels between the 4$^{\text{th}}$ quarter of 2008 and the end of 2010. The increase in volatility throughout the later half of 2008 takes the major part of the year 2009 to stabilize back to the lower levels observed during the following years. Several clusters of higher volatility are synchronized across all firms, among them notably the financial Crisis period, the 2010 Flash Crash and the 2011 Black Monday. LLTC and HPQ, both in the electronics industry, exhibit a severe and very brief volatility spike during the 2010 Flash Crash. These phenomenons suggests the existence of systemic and industry related factors driving regimes of volatility across all asset classes, as well as volatility jumps across industries or asset clusters. From the summary statistics in table A.1.1, we can observe that the unconditional empirical distribution of log-RV only moderately varies from one firm to the next, and indeed leptokurtic (sample kurtosis ranging between 3.89 and 5.37) and slightly right-skewed across all of them.

Measures of spread, such as the daily close proportional quoted spread (figure 3.6, equations (3.2) and (3.4)) or proportional effective spread (figure 3.7, equations (3.3) and

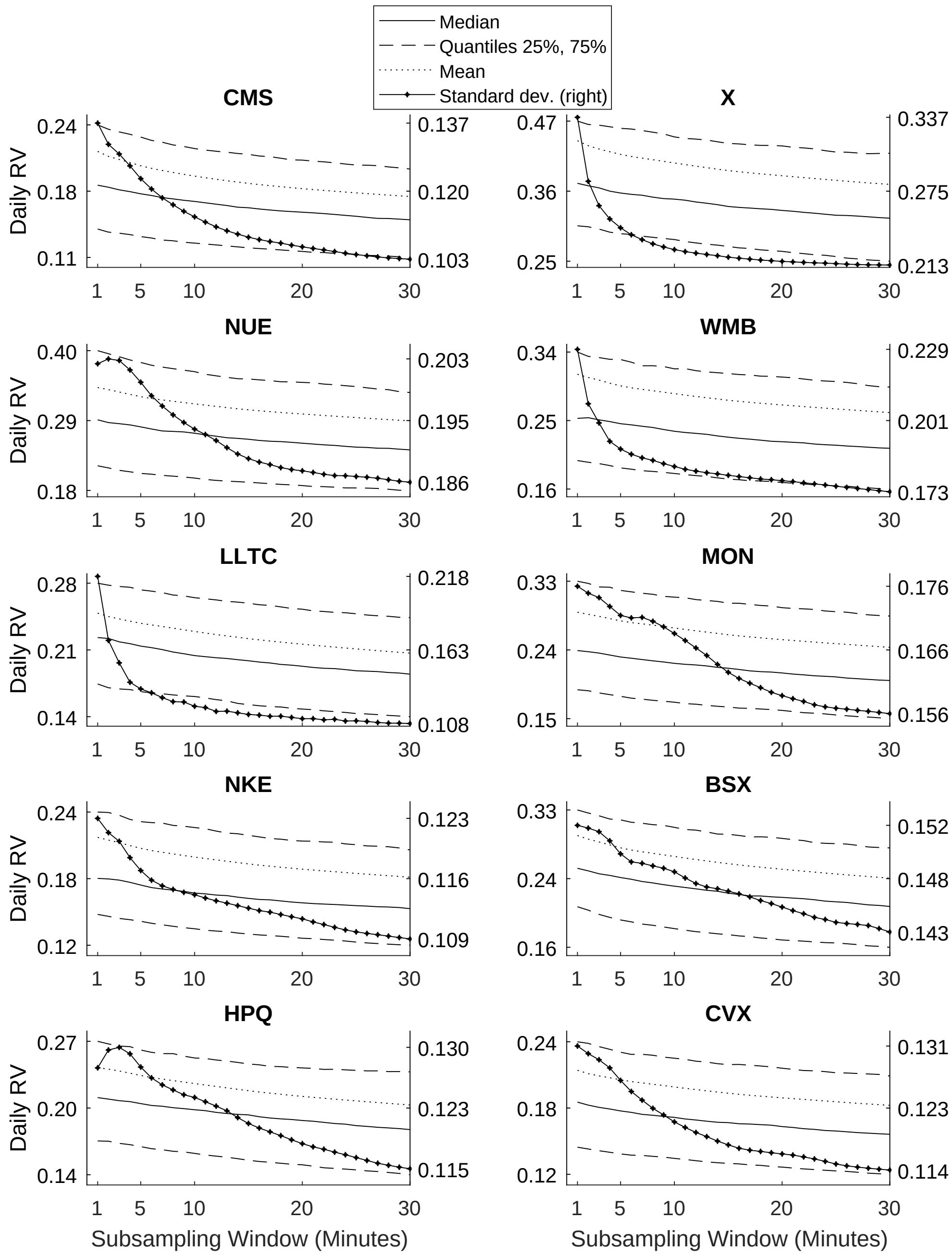

Figure 3.4: **Distribution of realized volatility by subsampling window, 2005-2013**
For each stock, we plot the mean, quartiles and standard deviation of annualized daily RV^{Sub} over the 2005-2013 sample period with respect to change in the subsampling window, as defined in equation (3.9).

(3.5)), also clearly display shock persistence, suggesting that they could be integrated processes. The market wide factors driving volume seem at times dwarfed by idiosyncratic factors, as the commonality in each plotted spread time series is less apparent than in realized volatility figures. Yet again, the financial Crisis of 2008-2009 stands out clearly as a period of lower liquidity across all firms. The main visually identifiable difference between quoted and effective spread series appears to be the occasional presence of strong outliers in the later, reflecting outliers in the transaction price processes themselves, which for stocks are well documented to show evidence of high kurtosis and jumps (Eraker et al., 2003; Andersen et al., 2007; Lee and Mykland, 2008). Looking at the descriptive statistics in tables A.1.2 and A.1.3, we indeed observe for all firms a higher sample kurtosis for PESPR time series than that of PQSPR. The most extreme case being LLTC, with a PESPR kurtosis of 1663.58, as opposed to 10.05 for PQSPR, mostly due to the disproportionate impact of the 2010 Flash Crash. With the notable exception of BSX, we can also perceive a loose downward trend in the mean spread as the market capitalization ranking increases (ranked in 2005). Firms with a relatively low average quoted spread tend to have leptokurtic and right-skewed unconditional distribution, while those of two firms in our sample with the highest average quoted spread (BSX and CMS) instead appear to be platykurtic, with a sample kurtosis of 1.75 and 2.71 respectively.

Daily volume series in number of shares and in dollars, in figures 3.8 and 3.9, not unlike the spread series, also appear autocorrelated with some systemic dependence structure. For most firms, the intraday counting process underlying the volume dynamic seems to have frequent isolated extreme values.

3.2.2 Autocorrelations and Cross-Correlations

In the graphical analysis, the persistence of shocks was already evident for all time series studied. We now look at the autocorrelation functions (ACF) to further quantify this earlier observation. Figures A.1.1 to A.1.5 show the sample ACFs up to a full year (252 days).

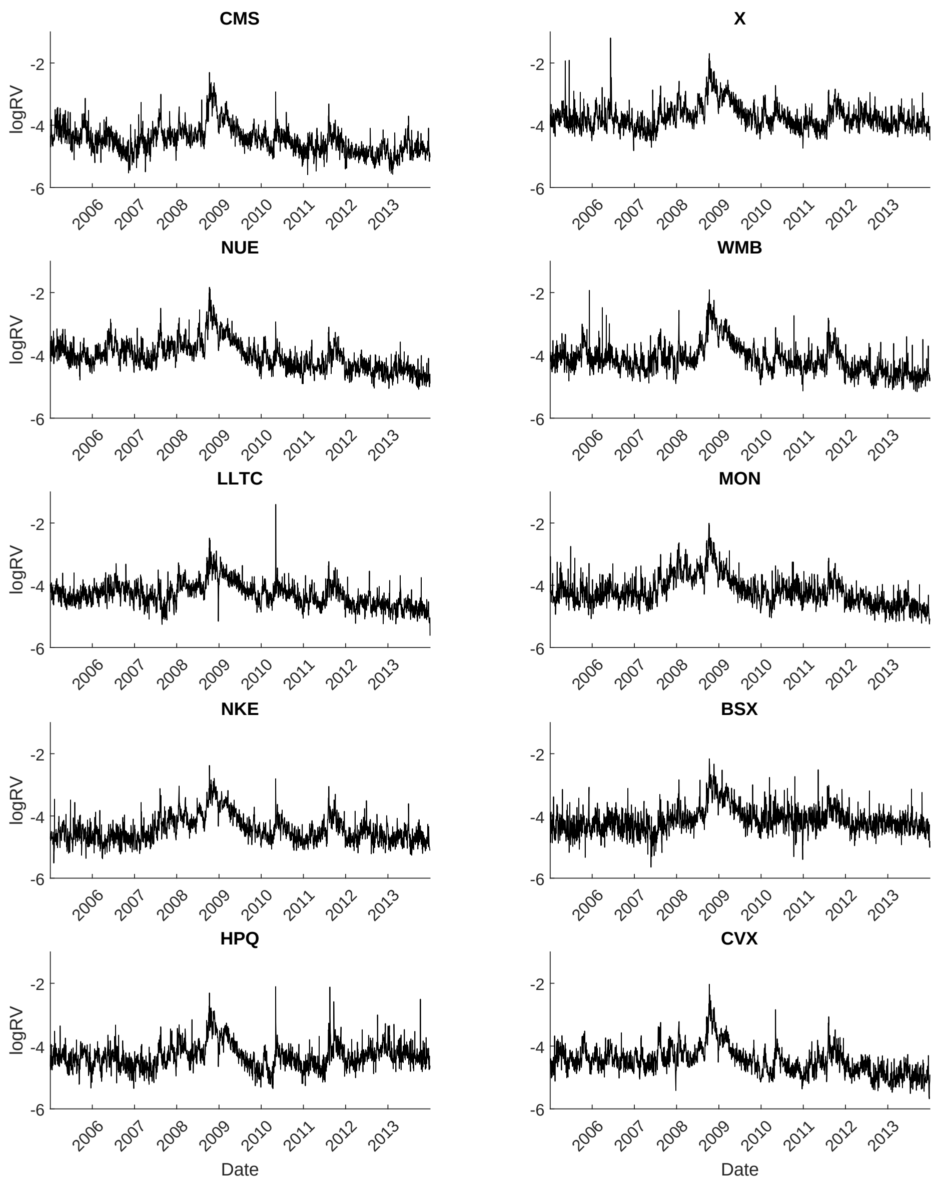

Figure 3.5: **Timeplot of log(RV), 2005-2013**
Timeplot of the daily log of the 5 minutes sumbsampling realized volatility estimator (equation (3.9) with $K = 5$) for the years 2005 to 2013 inclusively and for each of the selected stocks, after winsorization and interpolation.

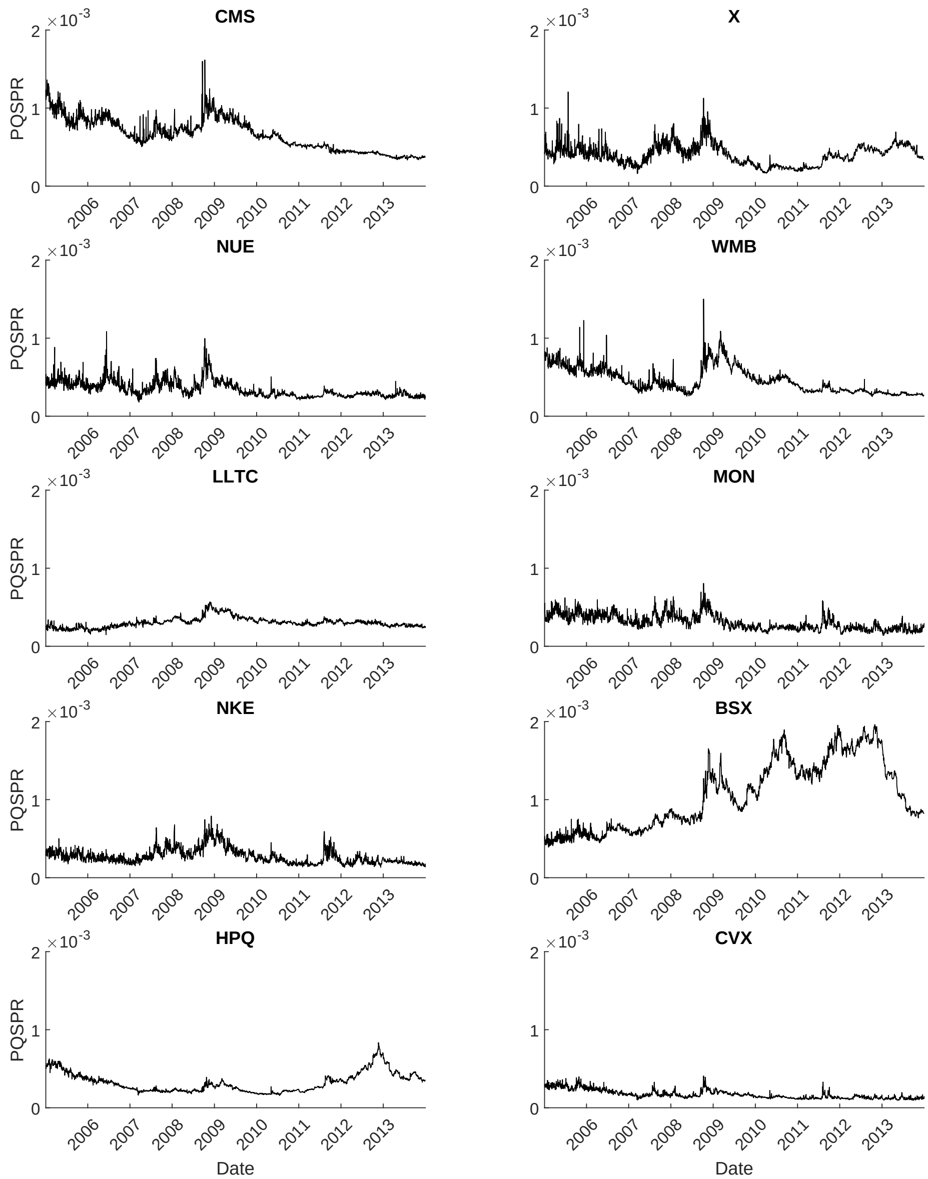

Figure 3.6: **Timeplot of PQSPR, 2005-2013**
Timeplot of the daily proportional quoted spread (equations (3.2) and (3.4)) for the years 2005 to 2013 inclusively and for each of the selected stocks, after winsorization and interpolation.

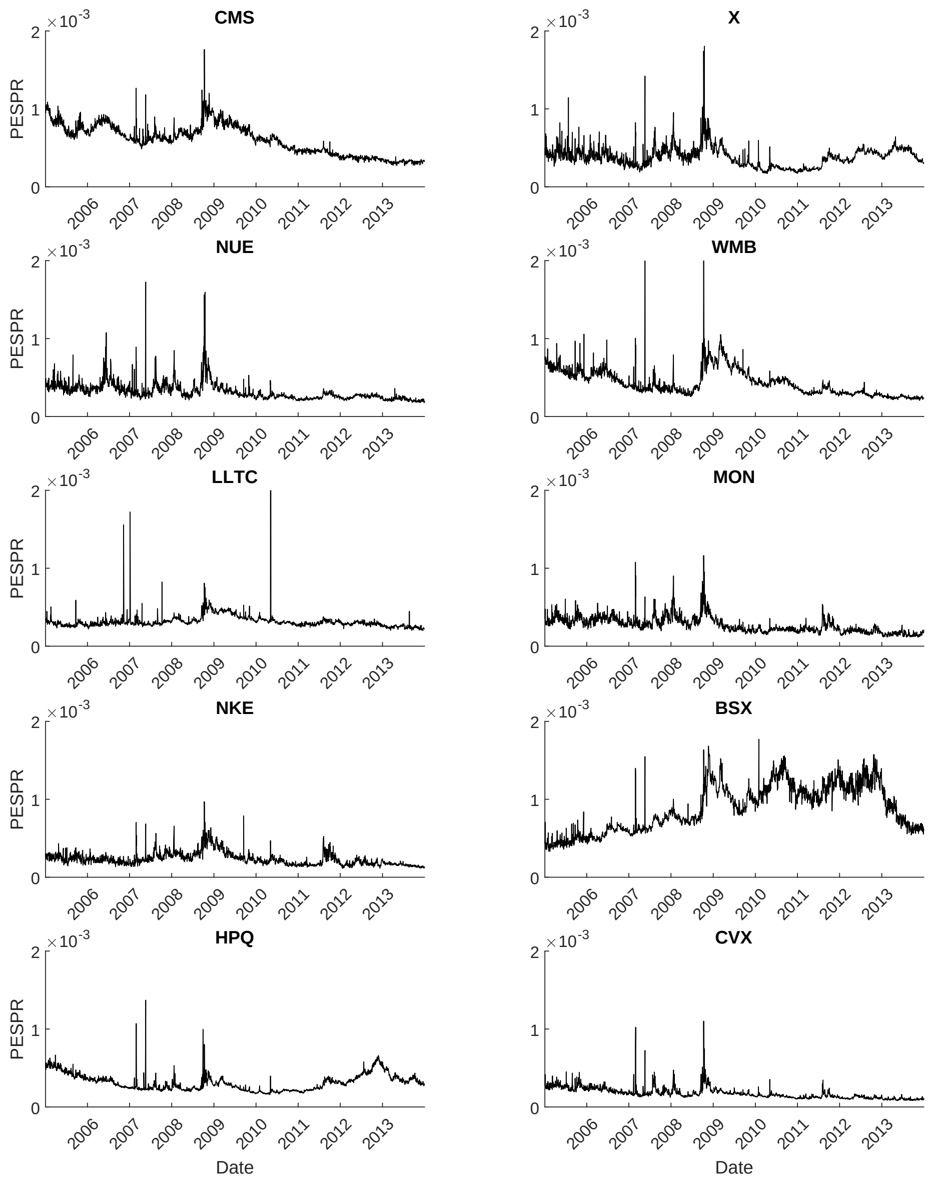

Figure 3.7: **Timeplot of PESPR, 2005-2013**
Timeplot of the daily proportional effective spread (equations (3.3) and (3.5)) for the years 2005 to 2013 inclusively and for each of the selected stocks, after winsorization and interpolation.

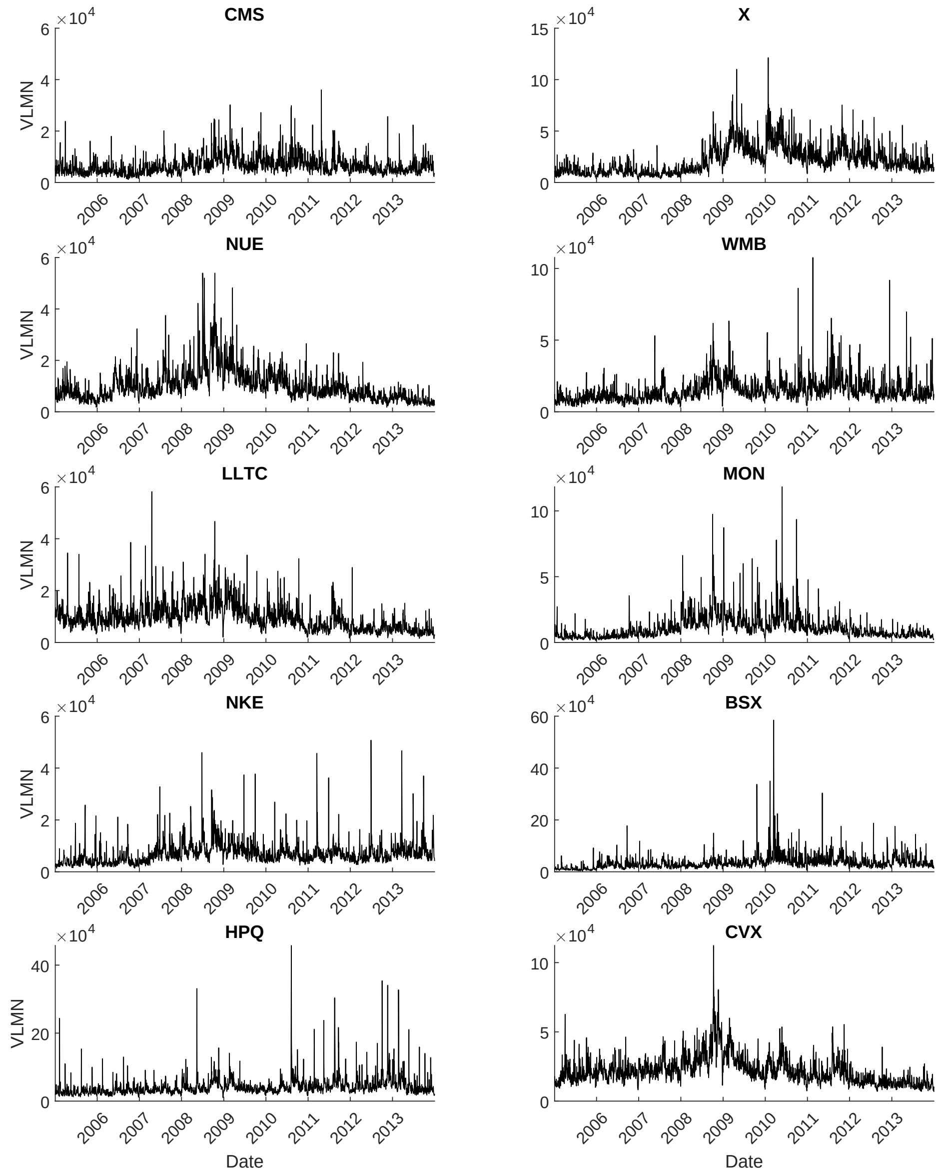

Figure 3.8: **Timeplot of VLMN, 2005-2013**
Timeplot of the daily trade volume in number of shares for the years 2005 to 2013 inclusively and for each of the selected stocks, after winsorization and interpolation.

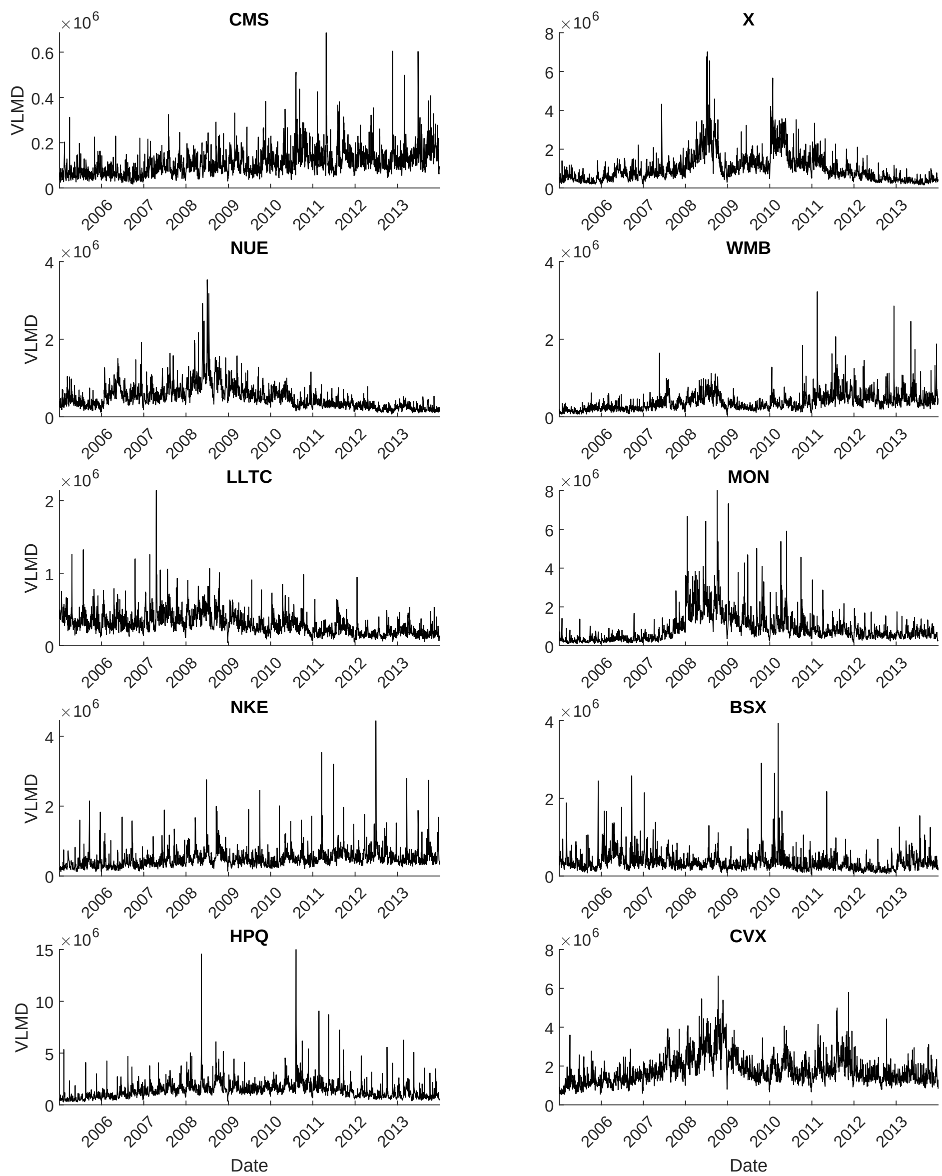

Figure 3.9: **Timeplot of VLMD, 2005-2013**
Timeplot of the daily trade volume in US dollars for the years 2005 to 2013 inclusively and for each of the selected stocks, after winsorization and interpolation.

For daily log-RV, PQSPR and PESPR, we observe evidence of long memory, as shocks persist through each time series generally up to 100 lags or more of statistically significant autocorrelations. For volume measures, firms generally have a slow ACF decay rate, yet not quite as consistently reach 100 lags of statistically significant autocorrelations at the 95% level[1]. Furthermore, some cases exhibit clear quarterly patterns of seasonality, in particular LLTC and HPQ from the electronics technology industry and NKE from consumer cyclical.

Looking at cross-correlations between these 5 variables for each of the 10 firms in figures A.1.6 to A.1.15, we first observe strong correlations across all first 21 lags between the two spread measures, PQSPR and PESPR. The cross-correlations are positive and decreasing as the lag increase, while still higher than 0.6 for all firms at the 21$^{\text{th}}$ lag, except for LLTC, due to the impact on PESPR of the May 2010 Flash Crash. Apart from LLTC at 0.43 and CVX at 0.84, the contemporary cross-correlation between PQSPR and PESPR is above 90% for each other firm. Similarly for measures of activity, with the exceptions of X at 0.60 and BSX at 0.75, the contemporary cross-correlations between VLMN and VLMD are above 0.80 for all other firms. We thus expect that including both measures of spread or activity in a linear model will induce strong multicollinearity and degrade the variance of the estimator.

We summarize contemporary and one-day lagged correlations over the full sample period between log-RV and the four measures of liquidity and activity in table 3.3. All measures of spread and volume have a positive contemporary correlation with log-RV, the lowest being 12% for CMS between log-RV and VLMD and highest 81% for CVX between log-RV and VLMN. For spread or activity measures with a one-day lag, the cross-correlations for each company between log-RV and these lagged measures are all positive and statistically significant at the 95% level assuming asymptotic normality[2],

[1]For a time series with T observations, we test the statistical significance of autocorrelations using the asymptotic standard error $\sqrt{\frac{1+2\sum_{t=1}^{\text{lag}-1}\hat{\rho}_t}{T}}$ (Tsay, 2005), from which is derived the upper confidence bound for each lag presented in figures A.1.1 to A.1.5.

[2]For a time series of T observations, we test the statistical significance of cross-correlations using the asymptotic standard error $\frac{1}{\sqrt{T}}$ (Tsay, 2005)

with only one correlation bordering its critical value (VLMD$_t$ and log-RV$_{t+1}$ for CMS at 2%).

	Correlation with log-RV$_{i,t}$				Correlation with log-RV$_{i,t+1}$			
Ticker	**PQSPR$_{i,t}$**	**PESPR$_{i,t}$**	**VLMN$_{i,t}$**	**VLMD$_{i,t}$**	**PQSPR$_{i,t}$**	**PESPR$_{i,t}$**	**VLMN$_{i,t}$**	**VLMD$_{i,t}$**
CMS	0.65	0.68	0.48	0.12	0.63	0.65	0.38	0.02
X	0.51	0.59	0.48	0.37	0.45	0.52	0.39	0.30
NUE	0.73	0.72	0.77	0.70	0.69	0.66	0.69	0.63
WMB	0.63	0.67	0.49	0.17	0.59	0.60	0.36	0.05
LLTC	0.51	0.36	0.70	0.53	0.50	0.21	0.58	0.41
MON	0.59	0.68	0.61	0.65	0.53	0.61	0.50	0.56
KNE	0.77	0.79	0.54	0.36	0.73	0.72	0.43	0.24
BSX	0.23	0.42	0.43	0.30	0.23	0.39	0.22	0.07
HPQ	0.19	0.34	0.50	0.36	0.19	0.29	0.32	0.21
CVX	0.56	0.67	0.81	0.65	0.51	0.60	0.72	0.57

Table 3.3: **Contemporary and one-day lagged correlations with log-RV**
Correlations over the 2005-2013 period between daily log-RV and selected daily measures of liquidity (proportional quoted, PQSPR, and effective spread, PESPR) and activity (volume in number of shares, VLMN, or dollars, VLMD). All correlations are statistically significant assuming asymptotic normality, with a sample size of 2246 observations for contemporary correlations and 2245 for one-day lagged ones.

3.3 Realized Volatility Model

Corsi (2009) proposed the simple Heterogeneous Autoregressive model of Realized Volatility (HAR-RV) for forecasting the implied stock volatility. He shows that this model is able to mimic long memory and fat tails properties apparent in stock returns time series as well as to provide good forecasting performance. Its advantageous linear form allows considerable flexibility in incorporating additional predictors. As such, our aim is to start with the HAR-RV as our benchmark model and investigate whether volume or bid-ask spread time series can improve volatility forecasting.

3.3.1 Benchmark Model

The linear HAR-RV model can be written as constrained autoregressive (AR) model of realized volatility, such that several lag terms are grouped into fewer moving average terms. Our aim is to use this model to forecast future implied volatility with past values of implied volatilities. However, since implied volatility is unobservable, we substitute their past values by their estimator, the realized volatility, which has some degree of measurement error.

Corsi (2009) suggests the use of daily, weekly and monthly lag components, although the model can flexibly be adapted to any time scales of interest. Using a year convention of 252 days with 21 days months, we construct the weekly and monthly lag components by averaging the daily subsampling RV estimator over 5 and 21 days respectively. Unlike in Corsi (2009) and like other authors (Bee et al. (2016), for example), we use the log-specification of the model to maintain volatility positive:

$$\log(\mathrm{RV}^{(\mathrm{d})}_{i,t+1}) = \alpha + \beta^{(\mathrm{d})}_{\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{i,t}) + \beta^{(\mathrm{w})}_{\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{i,t}) + \beta^{(\mathrm{m})}_{\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{i,t}) + \varepsilon_{i,t}, \quad (3.10)$$

where $\varepsilon_{i,t}$ is a white noise and:

$$\begin{aligned} \mathrm{RV}^{(\mathrm{d})}_{i,t} &= \mathrm{RV}^{\mathrm{Sub}}_{i,t}(K) \\ \mathrm{RV}^{(\mathrm{w})}_{i,t} &= \frac{1}{5}\left(\mathrm{RV}^{(\mathrm{d})}_{i,t} + \ldots + \mathrm{RV}^{(\mathrm{d})}_{i,t-4}\right) \\ \mathrm{RV}^{(\mathrm{m})}_{i,t} &= \frac{1}{21}\left(\mathrm{RV}^{(\mathrm{d})}_{i,t} + \ldots + \mathrm{RV}^{(\mathrm{d})}_{i,t-20}\right) \end{aligned} \quad (3.11)$$

for some subsampling window K and any given stock i and day t.

3.3.2 HAR-RV-LIQ Model

We expand the benchmark model in the HAR-RV-LIQ to include measures of liquidity and activity as additional predictors. Our measures of liquidity are the proportional quoted bid-ask spread (PQSPR) and the proportional effective bid-ask spread (PESPR); our measures of activity are the volume in number of shares (VLMN) and volume in dollars (VLMD). Our motivations for these additional predictors are presented in section 2.2. Our hypothesis is that measures of liquidity or activity provide useful information for 1-day ahead predictions of volatility, however we are unsure as to the appropriate specification of this relationship. Our study is limited to linear autoregressive and differenced specifications. We thus include these measures both in level and in difference, as well as their weekly and monthly lags.

$$
\begin{aligned}
\log(\mathrm{RV}^{(\mathrm{d})}_{i,t+1}) = \alpha_i &+ \beta^{(\mathrm{d})}_{i,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{i,t}) + \beta^{(\mathrm{w})}_{i,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{i,t}) + \beta^{(\mathrm{m})}_{i,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{i,t}) \\
&+ \beta^{(\mathrm{d})}_{i,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{m})}_{i,t} \\
&+ \beta^{(\mathrm{d})}_{i,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{m})}_{i,t} \\
&+ \beta^{(\mathrm{d})}_{i,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{m})}_{i,t} \\
&+ \beta^{(\mathrm{d})}_{i,\mathrm{VLMD}} \mathrm{VLMD}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\mathrm{VLMD}} \mathrm{VLMD}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\mathrm{VLMD}} \mathrm{VLMD}^{(\mathrm{m})}_{i,t} \\
&+ \beta^{(\mathrm{d})}_{i,\Delta\mathrm{PQSPR}} \Delta\mathrm{PQSPR}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\Delta\mathrm{PQSPR}} \Delta\mathrm{PQSPR}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\Delta\mathrm{PQSPR}} \Delta\mathrm{PQSPR}^{(\mathrm{m})}_{i,t} \\
&+ \beta^{(\mathrm{d})}_{i,\Delta\mathrm{PESPR}} \Delta\mathrm{PESPR}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\Delta\mathrm{PESPR}} \Delta\mathrm{PESPR}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\Delta\mathrm{PESPR}} \Delta\mathrm{PESPR}^{(\mathrm{m})}_{i,t} \\
&+ \beta^{(\mathrm{d})}_{i,\Delta\mathrm{VLMN}} \Delta\mathrm{VLMN}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\Delta\mathrm{VLMN}} \Delta\mathrm{VLMN}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\Delta\mathrm{VLMN}} \Delta\mathrm{VLMN}^{(\mathrm{m})}_{i,t} \\
&+ \beta^{(\mathrm{d})}_{i,\Delta\mathrm{VLMD}} \Delta\mathrm{VLMD}^{(\mathrm{d})}_{i,t} + \beta^{(\mathrm{w})}_{i,\Delta\mathrm{VLMD}} \Delta\mathrm{VLMD}^{(\mathrm{w})}_{i,t} + \beta^{(\mathrm{m})}_{i,\Delta\mathrm{VLMD}} \Delta\mathrm{VLMD}^{(\mathrm{m})}_{i,t} \\
&+ \varepsilon_{i,t},
\end{aligned}
\tag{3.12}
$$

where $\varepsilon_{i,t}$ is a white noise and:

$$
\begin{aligned}
\mathrm{Y}^{(\mathrm{w})}_{i,t} &= \frac{1}{5}\left(\mathrm{Y}^{(\mathrm{d})}_{i,t} + ... + \mathrm{Y}^{(\mathrm{d})}_{i,t-4}\right) \\
\mathrm{Y}^{(\mathrm{m})}_{i,t} &= \frac{1}{21}\left(\mathrm{Y}^{(\mathrm{d})}_{i,t} + ... + \mathrm{Y}^{(\mathrm{d})}_{i,t-20}\right) \\
\Delta\mathrm{Y}^{(h)}_{i,t} &= \mathrm{Y}^{(h)}_{i,t} - \mathrm{Y}^{(h)}_{i,t-1},
\end{aligned}
\tag{3.13}
$$

for h in $\{\mathrm{d},\mathrm{w},\mathrm{m}\}$, any given stock i and day t and for $\{\mathrm{Y}^{(h)}_{i,t}\}_{t\geq 0}$ in $\left\{\{\mathrm{PQSPR}^{(h)}_{i,t}\}_{t\geq 0}, \{\mathrm{PESPR}^{(h)}_{i,t}\}_{t\geq 0}, \{\mathrm{VLMN}^{(h)}_{i,t}\}_{t\geq 0}, \{\mathrm{VLMD}^{(h)}_{i,t}\}_{t\geq 0}\right\}$.

While the full model specification includes all components of interest, we will consider only proper subsets of these factors when estimating and evaluating the model. In particular, we include only either PQSPR or PESPR and only either VLMN or VLMD in any given nested model, due to their strong collinearity. We will thus estimate and evaluate the performance of various models nested within the general HAR-RV-LIQ. The benchmark model is also obviously nested within the HAR-RV-LIQ.

Let $\mathscr{M}_i$ be the set of models nested within the HAR-RV-LIQ model that we select for our study of a given company i. In order to investigate only a limited amount of models built from combinations of the spread or volume measures, we impose that for any given company i, models in $\mathscr{M}_i$ must follow the following rules:

1. Every model considered is an extension of the benchmark;
2. Quoted and effective spread are never both included in a model (multicollinearity);
3. Volume in shares or dollars are never both included in a model (multicollinearity);
4. A given spread or volume measure is included in level or in differences, not both;
5. Shorter explanatory lags have precedence over longer lags, e.g. the weekly lag of a liquidity measure is only included if its daily lag already is;
6. Variables a priori seemingly uncorrelated with the dependent variable are excluded. Any selected predictor must have an absolute correlation over the 2005-2013 period of at least 20% with the dependent variable $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$;
7. For a given company, we may retain only one of the two spread measures (or volume) if its correlation with the dependent variable is much larger than the other one's correlation.
8. Variables with high marginal multicollinearity are excluded, using the usual variance inflation factor (VIF)[3] cut-off of 10.

The rule 1 imposes that the benchmark model is nested in every model, in accordance with our initial objective of verifying whether this benchmark can be improved by additional terms. Rules 2 to 5 impose a hierarchical structure to our factors. We established our hypothesis that measures of liquidity (PQSPR, PESPR) or measures of activity (VLMN, VLMD) may help better predict volatility, but we do not know whether they do

[3]Let $\mathrm{X}_t = [\mathrm{X}_{1,t}, \ldots, \mathrm{X}_{p,t}]$ be a multivariate stochastic process and M be the linear model defined by the equation $\mathrm{Y}_t = \alpha + \beta_1 \mathrm{X}_{1,t} + \cdots + \beta_p \mathrm{X}_{p,t} + \varepsilon_t$. For $k \in \{1, \ldots, p\}$, the variance inflation factor (VIF) corresponding to factor $\{\mathrm{X}_{k,t}\}_{t\geq 0}$ in model M, written $\mathrm{VIF}\Big(M, \{\mathrm{X}_{k,t}\}_{t\geq 0}\Big)$, can be obtained from the k^{th} element of the diagonal of $\rho(\mathrm{X}_t)^{-1}$, where $\rho(\mathrm{X}_t)$ is the correlation matrix of the random vector X_t, i.e. $\mathrm{VIF}\Big(M, \{\mathrm{X}_{k,t}\}_{t\geq 0}\Big) = \Big(\rho(\mathrm{X}_t)^{-1}\Big)_{kk}$ (Belsley et al., 2004).

so when included in level or in difference, yet we refrain from included both levels and differences within the same model (rule 4). We recognized in our correlation analysis that different spread (or volume) measures were highly collinear and could degrade the variance of the estimator if both included (rules 2 and 3), which is not surprising when we consider that they essentially estimate more or less the same thing. The time lag structure imposed by rule 5 may weed out a few interesting models, but it prevents their unnecessary multiplication and prioritizes fresh information over lagged information.

While predictors that are uncorrelated with the dependent variables may be used as control variables or may become correlated after correcting for the variation of control variables, we have however not developed or employed an economic theory supporting the inclusion of such variables in our model. We thus consider variables $\mathrm{Y}_{i,t}^{(\mathrm{d})}$ to be unlikely to be good linear predictors if their correlation with $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$ over the entire sample period is too low. In rule 6, we use an absolute correlation $|\rho|$ of 0.2 as an arbitrary yet reasonable cut-off for low correlations. Rule 7 is our most subjective decision rule, but it allows us to eliminate a large amount of models that a priori are unlikely to be the best performing in predictions, which also reduces the chances of in-sample overfitting. For this decision criteria, we evaluate the correlations for the entire sample period as well as sub-periods. If one of the dual variables (PQSPR & PESPR, VLMN & VLMD) appear to clearly dominate its counterpart with respect to their linear correlation with $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$, then the variable with the weaker relationship will simply be eliminated from the analysis for this company. We do not precisely define "dominate" and we will instead make a judgement call based on the observed sample correlations. The correlations analysis for rules 6 and 7 will be presented prior to estimation results in the next section.

With rule 8, we take care to eliminate models with excess multicollinearity between factors. For every model that conform to the first 7 criteria and for every factor within that model, we calculate the variance inflation factor (VIF), which (asymptotically) measures how the collinearity affects the variance of the estimator. As is common in the literature, we exclude from our analysis any model for which one of its factors has a VIF larger than 10, with the exception of RV time series, for it would go against our rule 1. Coefficients of models characterized by a high degrees of multicollinearity between factors are difficult to estimate even with large samples. Such models are a priori less likely to provide good predictive performance, unless backed by sound economic theory. In consideration of the vast amount of tables already present in this paper, we opt not to report VIF values in the next section and present only the estimation results for models which have passed this final filter.

3.3.3 Estimation Procedure

The HAR-RV-LIQ model can be estimated via ordinary least squares (OLS). However, due to measurement errors (unobservable true volatility) and lagged values of the dependent variable as predictors, we cannot assume the error terms $\{\varepsilon_{i,t}\}_{t\geq 0}$ to be independent and Gaussian. It is thus necessary to correct the covariance matrix of the estimator to account for heteroscedasticity and serial autocorrelation (HAC effects) in the error terms. As done in Corsi (2009), we choose to compute the Newey-West robust standard errors. We use the automatic bandwidth h selection $h = 4\,(T/100)^{2/9}$ (Newey and West, 1994; Tsay, 2005), where T is the time series regression sample size. Furthermore, to avoid scaling issues causing numerically singular matrices, we individually center and normalize to unit length all non-RV predictors with respect to their own values over the whole sample period. Following our discussion in section 3.1.4, we choose $K = 5$ minutes as our subsampling window for the computation of the subsampling estimator of RV, presented in equation (3.9).

Chapter 4

Results

4.1 In-Sample Analysis

In this section, we estimate for our data a subset of models nested within the HAR-RV-LIQ previously presented. We first analyse in 4.1.1 the correlations between the dependent variable and the chosen set of potential predictors. Following this correlation study and our decisions rules presented in section 3.3.2, we define in 4.1.2 for each of the 10 company a set of models to be estimated and evaluated, for which the regression results analysis can be found in 4.1.3.

4.1.1 Correlation Analysis

For each stock i in our sample, we compute the correlation between the dependent variable $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$ and each of the HAR-RV-LIQ factors in equation (3.12). The correlations for 2005-2013 period are presented in table 4.1. We partitioned our sample into three sub-periods, the pre-Crisis (2005-2006), the Financial Crisis (2007-2009) and the post-Crisis (2010-2013), for which we also compute the correlations, which are found in appendix tables A.1.6, A.1.7 and A.1.8.

For the 2005-2006 period, which is the one that in most part preserves the 2005 market capitalisation ordering between companies presented in table 3.1 and as seen in 3.2, we observe no clear trends in correlations as market capitalisation increase for any of the factors. The strength of the linear relationships between the next day's realized volatility and our various daily measures of liquidity and activity does not seem to depend on the size of the firm in a trivially identifiable manner. The correlations between $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$ and the various factors of the HAR-RV-LIQ model, including lagged measures of realized volatility, are generally weakest during this 2005-2006 period and strongest during the 2007-2010 Financial Crisis period. Refer to equation (3.13) for how lagged measures of realized volatility, spread or volume are constructed.

For the full sample period as well as every subperiod, we observe no daily and almost no weekly differenced series with an absolute correlation of 0.2 or larger. For all 10 companies, the correlation with $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$ of differenced series, ΔPQSPR, ΔPESPR, ΔVLMN and ΔVLMD, increases as their time interval increases (from daily (d) to monthly (m)). Since a daily time series of monthly differences is expected to show some collinearity with the daily time series taken in level, it is not surprising that we observe this increasing pattern in correlations as the differencing interval increases. According to our model selection rules established in section 3.3.2, we therefore exclude for all companies all differenced time series from further analysis, due to their weaker linear relationship with $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$ as opposed to the series taken in level.

For CMS, NUE, WMB and NKE, the quoted and effective spreads have similar levels of correlation with the dependent variable, varying from 0.5 to 0.75 depending on the lag and firm (table 4.1). For HPQ and CVX, while their effective spread shows a correlation almost 10% superior compared to their quoted spread's for the 2005-2013 full sample period, they however have a correlation 15% and 20% inferior respectively to that of their quoted spread during the Financial Crisis (table A.1.7). Consequently, for these 6 firms, we choose to include both measures of spread in our sets of models to be estimated and analysed. For X, MON and BSX, the effective spread shows a larger (by 7%, 8% and 16% respectively) correlation than the quoted spread in the full sample period, which is not reversed in any subperiod. We choose to only investigate the predictive performance of PESPR for these three firms. For LLTC, we observe instead a stronger correlation for the quoted spread. The daily effective spread's correlation during the 2010-2013 period is at a low 0.11 (only occurrence among all firms below 20%), likely due to the impact of the large 2010 Flash Crash outlier (see figure 3.7). However, it is not quite obvious that such outliers should be excluded from the PESPR time series. While the main difference between PESPR and PQSPR is in PESPR's more frequent extreme values, we still observe strong correlations between $\text{PESPR}_{i,t}^{(\text{d})}$ and $\log(\text{RV}_{i,t+1}^{(\text{d})})$ for most firms.

The volume in dollars appear uncorrelated for with the next day's realized volatility for CMS, WMB and BSX and more weakly correlated than the volume in number of shares for X, NUE, LLTC, NKE, HPQ and CVX. We therefore include the volume in dollars only within MON's model set.

4.1.2 Model Sets

We now are able to construct for each firm a set of models to be further investigated with in-sample and out-of-sample analysis. In the previous section, we have used our decision rule 6 and 7 (see section 3.3.2) to determine which measures were going to be included in our model sets. Considering the large amount of possible factor combinations and the necessity to compute variance inflation factors (VIF) for every factor of every model having satisfying rules 1 to 7 for every firm, we opt not to report their specific value. There was not a single occurrence in which combining a daily measure of spread and a daily measure of volume alongside daily, weekly and monthly realized volatilities resulted in a VIF larger than 10 for the daily spread or volume. However note that in some cases weekly or monthly lags are excluded from a firm's model set due to a VIF larger than 10, indicating excess multicollinearity and a potentially high variance estimator. Specifically, for all firms, monthly lags of spread measures failed to pass the VIF cut-off whenever included in a model alongside daily and weekly measures of spread.

For each firm i, we next present the general model equation for which every model in $\mathscr{M}_i$ is also nested within. The specific models within each set are easily identifiable within regression results tables 4.2 to 4.11.

For CMS, the 1th firm, there are 12 models in $\mathscr{M}_1$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{1,t+1}) = \alpha_1 &+ \beta^{(\mathrm{d})}_{1,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{d})}_{1,t}) + \beta^{(\mathrm{w})}_{1,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{w})}_{1,t}) + \beta^{(\mathrm{m})}_{1,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{m})}_{1,t})\\ &+ \beta^{(\mathrm{d})}_{1,\mathrm{PQSPR}}\mathrm{PQSPR}^{(\mathrm{d})}_{1,t} + \beta^{(\mathrm{d})}_{1,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{d})}_{1,t}\\ &+ \beta^{(\mathrm{d})}_{1,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{d})}_{1,t} + \beta^{(\mathrm{w})}_{1,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{w})}_{1,t} + \beta^{(\mathrm{m})}_{1,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{m})}_{1,t}\\ &+ \varepsilon_{1,t}\end{aligned} \tag{4.1}$$

For X, the 2nd firm, there are 9 models in $\mathscr{M}_2$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{2,t+1}) = \alpha_2 &+ \beta^{(\mathrm{d})}_{2,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{d})}_{2,t}) + \beta^{(\mathrm{w})}_{2,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{w})}_{2,t}) + \beta^{(\mathrm{m})}_{2,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{m})}_{2,t})\\ &+ \beta^{(\mathrm{d})}_{2,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{d})}_{2,t} + \beta^{(\mathrm{w})}_{2,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{w})}_{2,t}\\ &+ \beta^{(\mathrm{d})}_{2,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{d})}_{2,t} + \beta^{(\mathrm{w})}_{2,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{w})}_{2,t} + \beta^{(\mathrm{m})}_{2,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{m})}_{2,t}\\ &+ \varepsilon_{2,t}\end{aligned} \tag{4.2}$$

For NUE, the 3rd firm, there are 12 models in $\mathscr{M}_3$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{3,t+1}) = \alpha_3 &+ \beta^{(\mathrm{d})}_{3,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{d})}_{3,t}) + \beta^{(\mathrm{w})}_{3,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{w})}_{3,t}) + \beta^{(\mathrm{m})}_{3,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{m})}_{3,t})\\ &+ \beta^{(\mathrm{d})}_{3,\mathrm{PQSPR}}\mathrm{PQSPR}^{(\mathrm{d})}_{3,t} + \beta^{(\mathrm{d})}_{3,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{d})}_{3,t} + \beta^{(\mathrm{w})}_{3,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{w})}_{3,t}\\ &+ \beta^{(\mathrm{d})}_{3,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{d})}_{3,t} + \beta^{(\mathrm{w})}_{3,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{w})}_{3,t} + \varepsilon_{3,t}\end{aligned} \tag{4.3}$$

For WMB, the 4th firm, there are 16 models in $\mathscr{M}_4$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{4,t+1}) = \alpha_4 &+ \beta^{(\mathrm{d})}_{4,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{d})}_{4,t}) + \beta^{(\mathrm{w})}_{4,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{w})}_{4,t}) + \beta^{(\mathrm{m})}_{4,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{m})}_{4,t})\\ &+ \beta^{(\mathrm{d})}_{4,\mathrm{PQSPR}}\mathrm{PQSPR}^{(\mathrm{d})}_{4,t} + \beta^{(\mathrm{d})}_{4,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{d})}_{4,t} + \beta^{(\mathrm{w})}_{4,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{w})}_{4,t}\\ &+ \beta^{(\mathrm{d})}_{4,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{d})}_{4,t} + \beta^{(\mathrm{w})}_{4,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{w})}_{4,t} + \beta^{(\mathrm{m})}_{4,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{m})}_{4,t}\\ &+ \varepsilon_{4,t}\end{aligned} \tag{4.4}$$

For LLTC, the 5th firm, there are 8 models in $\mathscr{M}_5$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{5,t+1}) = \alpha_5 &+ \beta^{(\mathrm{d})}_{5,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{d})}_{5,t}) + \beta^{(\mathrm{w})}_{5,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{w})}_{5,t}) + \beta^{(\mathrm{m})}_{5,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{m})}_{5,t})\\ &+ \beta^{(\mathrm{d})}_{5,\mathrm{PQSPR}}\mathrm{PQSPR}^{(\mathrm{d})}_{5,t}\\ &+ \beta^{(\mathrm{d})}_{5,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{d})}_{5,t} + \beta^{(\mathrm{w})}_{5,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{w})}_{5,t} + \beta^{(\mathrm{m})}_{5,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{m})}_{5,t}\\ &+ \varepsilon_{5,t}\end{aligned} \tag{4.5}$$

For MON, the 6$^{\text{th}}$ firm, there are 18 models in $\mathscr{M}_6$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{6,t+1}) = \alpha_6 &+ \beta^{(\mathrm{d})}_{6,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{6,t}) + \beta^{(\mathrm{w})}_{6,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{6,t}) + \beta^{(\mathrm{m})}_{6,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{6,t}) \\ &+ \beta^{(\mathrm{d})}_{6,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{6,t} + \beta^{(\mathrm{w})}_{6,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{w})}_{6,t} \\ &+ \beta^{(\mathrm{d})}_{6,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{6,t} + \beta^{(\mathrm{w})}_{6,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{6,t} + \beta^{(\mathrm{m})}_{6,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{m})}_{6,t} \\ &+ \beta^{(\mathrm{d})}_{6,\mathrm{VLMD}} \mathrm{VLMD}^{(\mathrm{d})}_{6,t} + \beta^{(\mathrm{w})}_{6,\mathrm{VLMD}} \mathrm{VLMD}^{(\mathrm{w})}_{6,t} + \varepsilon_{6,t}\end{aligned} \tag{4.6}$$

For NKE, the 7$^{\text{th}}$ firm, there are 12 models in $\mathscr{M}_7$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{7,t+1}) = \alpha_7 &+ \beta^{(\mathrm{d})}_{7,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{7,t}) + \beta^{(\mathrm{w})}_{7,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{7,t}) + \beta^{(\mathrm{m})}_{7,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{7,t}) \\ &+ \beta^{(\mathrm{d})}_{7,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{d})}_{7,t} + \beta^{(\mathrm{d})}_{7,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{7,t} \\ &+ \beta^{(\mathrm{d})}_{7,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{7,t} + \beta^{(\mathrm{w})}_{7,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{7,t} + \beta^{(\mathrm{m})}_{7,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{m})}_{7,t} \\ &+ \varepsilon_{7,t}\end{aligned} \tag{4.7}$$

For BSX, the 8$^{\text{th}}$ firm, there are 6 models in $\mathscr{M}_8$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{8,t+1}) = \alpha_8 &+ \beta^{(\mathrm{d})}_{8,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{8,t}) + \beta^{(\mathrm{w})}_{8,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{8,t}) + \beta^{(\mathrm{m})}_{8,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{8,t}) \\ &+ \beta^{(\mathrm{d})}_{8,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{8,t} \\ &+ \beta^{(\mathrm{d})}_{8,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{8,t} + \beta^{(\mathrm{w})}_{8,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{8,t} + \varepsilon_{8,t}\end{aligned} \tag{4.8}$$

For HPQ, the 9$^{\text{th}}$ firm, there are 16 models in $\mathscr{M}_9$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{9,t+1}) = \alpha_9 &+ \beta^{(\mathrm{d})}_{9,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{9,t}) + \beta^{(\mathrm{w})}_{9,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{9,t}) + \beta^{(\mathrm{m})}_{9,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{9,t}) \\ &+ \beta^{(\mathrm{d})}_{9,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{d})}_{9,t} + \beta^{(\mathrm{d})}_{9,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{9,t} + \beta^{(\mathrm{w})}_{9,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{w})}_{9,t} \\ &+ \beta^{(\mathrm{d})}_{9,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{9,t} + \beta^{(\mathrm{w})}_{9,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{9,t} + \beta^{(\mathrm{m})}_{9,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{m})}_{9,t} \\ &+ \varepsilon_{9,t}\end{aligned} \tag{4.9}$$

For CVX, the 10$^{\text{th}}$ firm, there are 12 models in $\mathscr{M}_{10}$ and all are nested within:

$$\begin{aligned}\log(\mathrm{RV}^{(\mathrm{d})}_{10,t+1}) = \alpha_{10} &+ \beta^{(\mathrm{d})}_{10,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{10,t}) + \beta^{(\mathrm{w})}_{10,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{10,t}) + \beta^{(\mathrm{m})}_{10,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{10,t}) \\ &+ \beta^{(\mathrm{d})}_{10,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{d})}_{10,t} + \beta^{(\mathrm{d})}_{10,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{10,t} + \beta^{(\mathrm{w})}_{10,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{w})}_{10,t} \\ &+ \beta^{(\mathrm{d})}_{10,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{10,t} + \beta^{(\mathrm{w})}_{10,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{10,t} + \varepsilon_{10,t}\end{aligned} \tag{4.10}$$

NUE and CVX have the same model set, so do CMS and NKE and so do WMB and HPQ. None of these pairs are from the same industry sector. However, WMB and CVX are both in the energy industry and only differ in model set by the inclusion of the monthly volume in number of shares for WMB. LLTC's model set is a subset of HPQ's, both in the electronics technology industry. Similarly for the two firms in the steel industry sector, X's model set is also a subset of NUE's. While our firm sample isn't large enough to draw any meaningful generalization, we do observe some degree of consistency in correlations between next day's RV and daily spread or volume measures within the same industry, considering correlations was an important criteria in building model sets.

Factor	CMS	X	NUE	WMB	LLTC	MON	NKE	BSX	HPQ	CVX
$\log(\mathrm{RV}_{i,t}^{(d)})$	0.82	0.77	0.86	0.81	0.79	0.81	0.82	0.68	0.74	0.84
$\log(\mathrm{RV}_{i,t}^{(w)})$	0.83	0.78	0.87	0.82	0.80	0.82	0.82	0.68	0.75	0.85
$\log(\mathrm{RV}_{i,t}^{(m)})$	0.79	0.72	0.83	0.79	0.75	0.81	0.79	0.67	0.70	0.78
$\mathrm{PQSPR}_{i,t}^{(d)}$	0.63	0.45	0.69	0.59	0.50	0.53	0.73	0.23	0.19	0.51
$\mathrm{PQSPR}_{i,t}^{(w)}$	0.63	0.46	0.70	0.58	0.50	0.54	0.73	0.22	0.18	0.50
$\mathrm{PQSPR}_{i,t}^{(m)}$	0.61	0.44	0.68	0.53	0.47	0.53	0.70	0.20	0.16	0.45
$\mathrm{PESPR}_{i,t}^{(d)}$	0.65	0.52	0.66	0.60	0.21	0.61	0.72	0.39	0.29	0.60
$\mathrm{PESPR}_{i,t}^{(w)}$	0.65	0.53	0.71	0.62	0.40	0.62	0.74	0.38	0.28	0.63
$\mathrm{PESPR}_{i,t}^{(m)}$	0.63	0.49	0.69	0.58	0.56	0.60	0.71	0.35	0.25	0.58
$\mathrm{VLMN}_{i,t}^{(d)}$	0.38	0.39	0.69	0.36	0.58	0.50	0.43	0.22	0.32	0.72
$\mathrm{VLMN}_{i,t}^{(w)}$	0.42	0.37	0.73	0.41	0.62	0.53	0.49	0.20	0.33	0.76
$\mathrm{VLMN}_{i,t}^{(m)}$	0.41	0.30	0.71	0.42	0.62	0.55	0.55	0.18	0.34	0.72
$\mathrm{VLMD}_{i,t}^{(d)}$	0.02	0.30	0.63	0.05	0.41	0.56	0.24	0.07	0.21	0.57
$\mathrm{VLMD}_{i,t}^{(w)}$	-0.04	0.28	0.66	0.02	0.44	0.59	0.25	0.00	0.19	0.59
$\mathrm{VLMD}_{i,t}^{(m)}$	-0.14	0.27	0.67	0.00	0.44	0.60	0.28	-0.08	0.16	0.57
$\Delta\mathrm{PQSPR}_{i,t}^{(d)}$	0.03	0.02	0.04	0.04	0.03	0.04	0.04	0.04	0.04	0.07
$\Delta\mathrm{PQSPR}_{i,t}^{(w)}$	0.07	0.06	0.09	0.11	0.09	0.08	0.12	0.11	0.09	0.13
$\Delta\mathrm{PQSPR}_{i,t}^{(m)}$	0.15	0.14	0.17	0.24	0.18	0.13	0.18	0.23	0.16	0.24
$\Delta\mathrm{PESPR}_{i,t}^{(d)}$	0.02	0.04	0.04	0.04	0.00	0.04	0.05	0.06	0.02	0.04
$\Delta\mathrm{PESPR}_{i,t}^{(w)}$	0.07	0.10	0.09	0.08	0.00	0.09	0.13	0.10	0.08	0.10
$\Delta\mathrm{PESPR}_{i,t}^{(m)}$	0.16	0.18	0.16	0.19	0.02	0.15	0.18	0.21	0.16	0.21
$\Delta\mathrm{VLMN}_{i,t}^{(d)}$	0.04	0.08	0.05	0.04	0.07	0.06	0.05	0.06	0.07	0.07
$\Delta\mathrm{VLMN}_{i,t}^{(w)}$	0.09	0.14	0.11	0.07	0.09	0.07	0.07	0.09	0.11	0.13
$\Delta\mathrm{VLMN}_{i,t}^{(m)}$	0.14	0.24	0.16	0.11	0.18	0.12	0.10	0.10	0.13	0.25
$\Delta\mathrm{VLMD}_{i,t}^{(d)}$	0.04	0.06	0.05	0.04	0.06	0.06	0.05	0.06	0.06	0.06
$\Delta\mathrm{VLMD}_{i,t}^{(w)}$	0.07	0.09	0.08	0.05	0.07	0.07	0.07	0.09	0.09	0.11
$\Delta\mathrm{VLMD}_{i,t}^{(m)}$	0.11	0.09	0.09	0.05	0.15	0.11	0.08	0.08	0.11	0.19

Table 4.1: **Factor correlations with dependent variable, 2005-2013**

Correlation between the dependent variable $\log(\mathrm{RV}_{i,t+1}^{(d)})$ and the various factors of the HAR-RV-LIQ model for each company (see table 3.1), evaluated for the 2005-2013 full sample period (2224 observations). The upper blocks showcase correlations for realized volatility (RV), proportional quoted or effective spread (PQSPR and PESPR) and for volume in number of shares (VLMN) or in dollars (VLMD) in level, while the bottom blocks are for the same factors taken in difference (Δ). Each daily (d) factor is also included in lagged weekly (w) or monthly(m) moving averages. Realized volatility is calculated using the 5 minutes subsampling estimator.

4.1.3 Estimation Results

For the 2005-2013 sample period, we run a set of regressions with 2224 observations for each firm in accordance to the previously defined model sets with dependent variable $\log(\mathrm{RV}^{(\mathrm{d})}_{i,t+1})$. Again, all realized volatility are estimated with the 5 minutes subsampling estimator using per-minute intraday data. We present the resulting coefficients, Newey-West t-statistics and adjusted R^2 in tables 4.2 to 4.11. Coefficients with absolute t-statistics larger than 1.96 (95% confidence level critical value) are in bold and starred. Any model's adjusted R^2 that is superior to the benchmark's is also in bold characters.

One immediately striking global result is that all components of the benchmark model, i.e. the daily, weekly and monthly realized volatility terms, all have significant t-statistics for every company and every model. Since the standard deviation of daily RV ranges from 0.40 to 0.47 between companies (see table A.1.1), if we were to normalize to unit length RV time series to bring their coefficients on a scale comparable to the normalized spread and volume series, the RV coefficients would still generally dwarf the spread or volume coefficients. Indeed, normalizing the RV time series is equivalent to multiplying the coefficients by the standard deviation. The coefficients for normalized daily RV thus range between 0.11 and 0.19 across all models and firms; those for weekly RV between 0.07 and 0.18 and those for monthly RV between 0.05 and 0.17. In contrast, the highest absolute volume or spread coefficient across all firms and models is a daily volume coefficient of 0.06 in one of United States Steel Corp's (X) models (see model 8 in table 4.3). Furthermore, across all companies and models combined, the highest increase in adjusted R^2 with respect to the benchmark's is only 0.0054 (also X's model 8, table 4.3). As previously observed in the autocorrelation functions in figure A.1.1, realized volatility is a strongly autocorrelated process. As such, it is not too surprising to see spread and volume time series only increase in very small margins the fit beyond the contribution of daily, weekly and monthly lags of RV. Since the various models for a given firm are mostly indistinguishable in terms of adjusted R^2, we will focus most of our attention on the significance of t-statistics. Consequently, we exclusively and exhaustively judge models with a full set of significant coefficients worthy of out-of-sample performance analysis.

For CMS (table 4.2), the daily volume in shares (VLMN) is only significant whenever at least 1 measure of spread or the weekly VLMN is included within the model. This is the case for most other firms, including the daily volume in dollars (VLMD) for MON, with the exception of WMB, BSX and HPQ. The addition of monthly VLMN seems unnecessary for a good fit, as its coefficient is never significant and in some cases degrades the significance of the weekly term. In the case of CMS, the daily and weekly VLMN coefficients, when both significant, have similar magnitudes but opposing signs. We observe a similar pattern for most other firms, although only CMS has an estimated model with both daily and weekly volume coefficients significant. This indicates that excess daily volume with respect to the past 5 days's average could potentially be used as a predictor rather than separate daily and weekly terms to predict the next day's volatility.

Across all firms and models, there is not a single occurrence of a statistically significant weekly effective spread (PESPR) coefficient. For CMS, WMB, LLTC, NKE and BSX, whenever the daily quoted (PQSPR) or effective spread is included, its coefficient is significant. For X and MON, the daily spread coefficient is always significant provided the weekly spread is not included within the model.

WMB and HPQ (tables 4.5 and 4.10), on the other hand, have no single model with all significant t-statistics other than the benchmark model. It does not seem that a linear specification incorporating spread or volume terms in level provides a better fit than the simple HAR-RV model. Microscopic adjusted R^2 gains do no seem to justify the inclusion of non-significant terms. We will therefore not analyse the out-of-sample performance for these firms, as no investigated model appears to be a better fit than the benchmark. We cannot conclude that liquidity and activity bear no information of predictive usefulness for these firms' volatility, but more so that the measures of liquidity and activity that we have sampled, in their current form, are not well suited for a static linear specification. It is worth investigating in the future whether a transformation of these measures or alternate proxies for liquidity and activity can produce better results.

	Regression results: CMS – Model set $\mathscr{M}_1$											
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Intercept	-0.24	-0.44	-0.45	-0.28	-0.56	-0.56	-0.22	-0.47	-0.48	-0.18	-0.44	-0.44
	(-3.35*)	**(-4.17*)**	**(-4.17*)**	**(-3.79*)**	**(-4.71*)**	**(-4.61*)**	**(-2.89*)**	**(-3.77*)**	**(-3.76*)**	**(-2.35*)**	**(-3.17*)**	**(-3.17*)**
$\log(\mathrm{RV}^{(d)})$	0.38	0.37	0.37	0.36	0.34	0.34	0.33	0.33	0.32	0.33	0.33	0.32
	(12.51*)	**(12.41*)**	**(12.35*)**	**(12.10*)**	**(11.56*)**	**(11.48*)**	**(10.77*)**	**(10.65*)**	**(10.54*)**	**(10.81*)**	**(10.66*)**	**(10.54*)**
$\log(\mathrm{RV}^{(w)})$	0.34	0.33	0.33	0.34	0.33	0.33	0.39	0.37	0.38	0.37	0.36	0.36
	(7.71*)	**(7.76*)**	**(7.75*)**	**(7.72*)**	**(7.79*)**	**(7.78*)**	**(8.54*)**	**(8.07*)**	**(8.18*)**	**(7.42*)**	**(7.42*)**	**(7.48*)**
$\log(\mathrm{RV}^{(m)})$	0.23	0.20	0.20	0.24	0.20	0.20	0.23	0.20	0.20	0.26	0.22	0.22
	(6.14*)	**(5.06*)**	**(5.02*)**	**(6.31*)**	**(5.10*)**	**(5.09*)**	**(6.01*)**	**(5.04*)**	**(4.99*)**	**(6.15*)**	**(4.71*)**	**(4.71*)**
$\mathrm{PQSPR}^{(d)}$		0.02			0.03			0.03			0.02	
		(3.15*)			**(3.66*)**			**(2.99*)**			**(2.60*)**	
$\mathrm{PESPR}^{(d)}$			0.02			0.03			0.03			0.02
			(3.11*)			**(3.50*)**			**(2.98*)**			**(2.60*)**
$\mathrm{VLMN}^{(d)}$				0.01	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.03
				(1.85)	**(2.74*)**	**(2.59*)**	**(3.13*)**	**(3.18*)**	**(3.19*)**	**(3.07*)**	**(3.14*)**	**(3.15*)**
$\mathrm{VLMN}^{(w)}$							-0.03	-0.02	-0.02	-0.01	-0.01	-0.01
							(-2.87*)	**(-2.00*)**	**(-2.20*)**	(-1.05)	(-1.02)	(-1.08)
$\mathrm{VLMN}^{(m)}$										-0.02	-0.01	-0.01
										(-1.67)	(-0.81)	(-0.89)
R^2 (adj.)	0.7276	**0.7287**	**0.7287**	**0.7279**	**0.7296**	**0.7295**	**0.7289**	**0.7300**	**0.7301**	**0.7292**	**0.7300**	**0.7301**

Table 4.2: **In-sample estimation results for CMS, 2005-2013**

Regression coefficient results for the CMS regression model set $\mathscr{M}_1$ with general equation $\log(\mathrm{RV}^{(d)}_{1,t+1}) = \alpha_1 + \beta^{(d)}_{1,\mathrm{RV}} \log(\mathrm{RV}^{(d)}_{1,t}) + \beta^{(w)}_{1,\mathrm{RV}} \log(\mathrm{RV}^{(w)}_{1,t}) + \beta^{(m)}_{1,\mathrm{RV}} \log(\mathrm{RV}^{(m)}_{1,t}) + \beta^{(d)}_{1,\mathrm{PQSPR}} \mathrm{PQSPR}^{(d)}_{1,t} + \beta^{(d)}_{1,\mathrm{PESPR}} \mathrm{PESPR}^{(d)}_{1,t} + \beta^{(d)}_{1,\mathrm{VLMN}} \mathrm{VLMN}^{(d)}_{1,t} + \beta^{(w)}_{1,\mathrm{VLMN}} \mathrm{VLMN}^{(w)}_{1,t} + \beta^{(m)}_{1,\mathrm{VLMN}} \mathrm{VLMN}^{(m)}_{1,t} + \varepsilon_{1,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: X – Model set $\mathscr{M}_2$								
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Intercept	-0.30	-0.40	-0.37	-0.39	-0.64	-0.32	-0.67	-0.56	-0.57
	(-3.87*)	**(-4.69*)**	**(-4.00*)**	**(-4.53*)**	**(-4.99*)**	**(-3.55*)**	**(-5.11*)**	**(-4.37*)**	**(-4.15*)**
$\log(\mathrm{RV}^{(d)})$	0.38	0.37	0.36	0.36	0.31	0.31	0.32	0.28	0.28
	(13.27*)	**(12.39*)**	**(11.73*)**	**(12.13*)**	**(9.25*)**	**(9.42*)**	**(9.41*)**	**(8.08*)**	**(8.02*)**
$\log(\mathrm{RV}^{(w)})$	0.33	0.32	0.33	0.32	0.31	0.39	0.30	0.36	0.36
	(6.43*)	**(6.20*)**	**(6.47*)**	**(6.08*)**	**(6.03*)**	**(7.03*)**	**(5.55*)**	**(6.35*)**	**(5.79*)**
$\log(\mathrm{RV}^{(m)})$	0.22	0.22	0.22	0.22	0.21	0.22	0.21	0.22	0.22
	(4.62*)	**(4.58*)**	**(4.70*)**	**(4.65*)**	**(4.71*)**	**(4.61*)**	**(4.70*)**	**(4.64*)**	**(4.71*)**
$\mathrm{PESPR}^{(d)}$		0.02		0.02	0.03		0.02	0.03	0.03
		(2.55*)		(1.43)	**(3.67*)**		(1.61)	**(3.20*)**	(1.72)
$\mathrm{PESPR}^{(w)}$				-0.00			0.01		0.00
				(-0.33)			(0.85)		(0.19)
$\mathrm{VLMN}^{(d)}$			0.02		0.03	0.06	0.03	0.06	0.06
			(2.39*)		**(3.75*)**	**(4.97*)**	**(3.87*)**	**(5.19*)**	**(5.18*)**
$\mathrm{VLMN}^{(w)}$						-0.05		-0.04	-0.04
						(-4.41*)		**(-3.50*)**	**(-3.35*)**
R^2 (adj.)	0.6471	**0.6480**	**0.6482**	**0.6479**	**0.6512**	**0.6505**	**0.6512**	**0.6525**	**0.6524**

Table 4.3: **In-sample estimation results for X, 2005-2013**

Regression coefficient results for the X regression model set $\mathscr{M}_2$ with general equation $\log(\mathrm{RV}^{(d)}_{2,t+1}) = \alpha_2 + \beta^{(d)}_{2,\mathrm{RV}}\log(\mathrm{RV}^{(d)}_{2,t}) + \beta^{(w)}_{2,\mathrm{RV}}\log(\mathrm{RV}^{(w)}_{2,t}) + \beta^{(m)}_{2,\mathrm{RV}}\log(\mathrm{RV}^{(m)}_{2,t}) + \beta^{(d)}_{2,\mathrm{PESPR}}\mathrm{PESPR}^{(d)}_{2,t} + \beta^{(w)}_{2,\mathrm{PESPR}}\mathrm{PESPR}^{(w)}_{2,t} + \beta^{(d)}_{2,\mathrm{VLMN}}\mathrm{VLMN}^{(d)}_{2,t} + \beta^{(w)}_{2,\mathrm{VLMN}}\mathrm{VLMN}^{(w)}_{2,t} + \beta^{(m)}_{2,\mathrm{VLMN}}\mathrm{VLMN}^{(m)}_{2,t} + \varepsilon_{2,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: NUE – Model set $\mathscr{M}_3$											
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Intercept	-0.17	-0.28	-0.30	-0.35	-0.63	-0.24	-0.55	-0.29	-0.59	-0.52	-0.51	-0.46
	(-3.15*)	**(-3.95*)**	**(-4.09*)**	**(-4.65*)**	**(-5.97*)**	**(-3.21*)**	**(-5.03*)**	**(-3.50*)**	**(-4.89*)**	**(-4.92*)**	**(-4.16*)**	**(-3.95*)**
$\log(\mathrm{RV}^{(\mathrm{d})})$	0.41	0.40	0.39	0.37	0.33	0.37	0.34	0.35	0.32	0.33	0.33	0.32
	(13.33*)	**(12.49*)**	**(12.14*)**	**(11.70*)**	**(9.76*)**	**(11.44*)**	**(9.92*)**	**(10.44*)**	**(9.29*)**	**(9.53*)**	**(9.42*)**	**(8.90*)**
$\log(\mathrm{RV}^{(\mathrm{w})})$	0.35	0.35	0.34	0.34	0.32	0.36	0.32	0.38	0.34	0.33	0.34	0.36
	(8.17*)	**(8.04*)**	**(7.87*)**	**(8.03*)**	**(7.70*)**	**(7.65*)**	**(7.60*)**	**(8.24*)**	**(7.37*)**	**(7.16*)**	**(7.22*)**	**(6.86*)**
$\log(\mathrm{RV}^{(\mathrm{m})})$	0.20	0.19	0.20	0.21	0.19	0.20	0.21	0.21	0.19	0.21	0.21	0.21
	(5.94*)	**(5.73*)**	**(5.88*)**	**(6.23*)**	**(5.79*)**	**(6.13*)**	**(6.19*)**	**(6.19*)**	**(5.79*)**	**(6.33*)**	**(6.17*)**	**(6.33*)**
$\mathrm{PQSPR}^{(\mathrm{d})}$		0.02			0.03				0.03			
		(2.04*)			**(3.59*)**				**(3.32*)**			
$\mathrm{PESPR}^{(\mathrm{d})}$			0.02			0.03	0.03			0.03	0.02	0.03
			(2.72*)			**(2.45*)**	**(2.84*)**			**(2.22*)**	**(2.66*)**	**(2.17*)**
$\mathrm{PESPR}^{(\mathrm{w})}$						-0.02				-0.01		-0.01
						(-1.29)				(-0.59)		(-0.70)
$\mathrm{VLMN}^{(\mathrm{d})}$				0.03	0.04		0.03	0.04	0.05	0.03	0.04	0.04
				(3.08*)	**(3.93*)**		**(3.36*)**	**(3.26*)**	**(3.55*)**	**(3.33*)**	**(3.11*)**	**(3.09*)**
$\mathrm{VLMN}^{(\mathrm{w})}$								-0.02	-0.01		-0.01	-0.01
								(-1.52)	(-0.78)		(-0.81)	(-0.91)
R^2 (adj.)	0.7836	**0.7840**	**0.7844**	**0.7849**	**0.7863**	**0.7845**	**0.7862**	**0.7851**	**0.7862**	**0.7861**	**0.7862**	**0.7861**

Table 4.4: **In-sample estimation results for NUE, 2005-2013**

Regression coefficient results for the NUE regression model set $\mathscr{M}_3$ with general equation $\log(\mathrm{RV}^{(\mathrm{d})}_{3,t+1}) = \alpha_3 + \beta^{(\mathrm{d})}_{3,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{d})}_{3,t}) + \beta^{(\mathrm{w})}_{3,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{w})}_{3,t}) + \beta^{(\mathrm{m})}_{3,\mathrm{RV}}\log(\mathrm{RV}^{(\mathrm{m})}_{3,t}) + \beta^{(\mathrm{d})}_{3,\mathrm{PQSPR}}\mathrm{PQSPR}^{(\mathrm{d})}_{3,t} + \beta^{(\mathrm{d})}_{3,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{d})}_{3,t} + \beta^{(\mathrm{w})}_{3,\mathrm{PESPR}}\mathrm{PESPR}^{(\mathrm{w})}_{3,t} + \beta^{(\mathrm{d})}_{3,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{d})}_{3,t} + \beta^{(\mathrm{w})}_{3,\mathrm{VLMN}}\mathrm{VLMN}^{(\mathrm{w})}_{3,t} + \varepsilon_{3,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: WMB – Model set $\mathscr{M}_4$															
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Intercept	-0.24	-0.29	-0.28	-0.26	-0.35	-0.31	-0.33	-0.25	-0.34	-0.37	-0.32	-0.24	-0.34	-0.36	-0.31	-0.36
	(-3.80*)	**(-3.03*)**	**(-2.86*)**	**(-4.05*)**	**(-3.36*)**	**(-3.04*)**	**(-3.03*)**	**(-3.76*)**	**(-3.21*)**	**(-3.27*)**	**(-2.80*)**	**(-3.54*)**	**(-2.89*)**	**(-2.98*)**	**(-2.56*)**	**(-2.82*)**
$\log(\mathrm{RV}^{(d)})$	0.37	0.37	0.37	0.36	0.35	0.38	0.35	0.36	0.35	0.36	0.35	0.36	0.35	0.36	0.35	0.36
	(11.81*)	**(11.59*)**	**(11.39*)**	**(10.94*)**	**(10.28*)**	**(11.46*)**	**(10.25*)**	**(10.46*)**	**(10.05*)**	**(10.57*)**	**(10.02*)**	**(10.42*)**	**(10.08*)**	**(10.25*)**	**(10.05*)**	**(10.25*)**
$\log(\mathrm{RV}^{(w)})$	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.32	0.32	0.30	0.32	0.31	0.31	0.30	0.31	0.30
	(6.69*)	**(6.70*)**	**(6.69*)**	**(6.65*)**	**(6.64*)**	**(6.13*)**	**(6.62*)**	**(5.77*)**	**(5.65*)**	**(5.98*)**	**(5.59*)**	**(5.16*)**	**(5.17*)**	**(4.93*)**	**(5.15*)**	**(4.66*)**
$\log(\mathrm{RV}^{(m)})$	0.26	0.25	0.25	0.27	0.26	0.25	0.26	0.26	0.25	0.25	0.26	0.28	0.26	0.25	0.26	0.26
	(5.64*)	**(5.22*)**	**(5.43*)**	**(5.76*)**	**(5.32*)**	**(5.50*)**	**(5.54*)**	**(5.45*)**	**(5.18*)**	**(5.60*)**	**(5.34*)**	**(4.74*)**	**(4.17*)**	**(5.43*)**	**(4.45*)**	**(4.53*)**
$\mathrm{PQSPR}^{(d)}$		0.01			0.01				0.01				0.01			
		(0.86)			(1.33)				(1.31)				(1.18)			
$\mathrm{PESPR}^{(d)}$			0.01			-0.01	0.01			-0.01	0.01			-0.01	0.01	-0.01
			(0.76)			(-0.37)	(0.98)			(-0.38)	(0.91)			(-0.38)	(0.83)	(-0.39)
$\mathrm{PESPR}^{(w)}$						0.01				0.02				0.02		0.02
						(0.99)				(1.17)				(1.12)		(1.12)
$\mathrm{VLMN}^{(d)}$				0.01	0.01		0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
				(0.85)	(1.21)		(1.05)	(0.85)	(0.92)	(1.16)	(0.86)	(0.85)	(0.92)	(0.88)	(0.86)	(0.88)
$\mathrm{VLMN}^{(w)}$								-0.00	-0.00		-0.00	-0.00	-0.00	-0.00	-0.00	-0.00
								(-0.43)	(-0.12)		(-0.22)	(-0.00)	(-0.02)	(-0.09)	(-0.02)	(-0.01)
$\mathrm{VLMN}^{(m)}$												-0.01	-0.00		-0.00	-0.00
												(-0.50)	(-0.11)		(-0.24)	(-0.10)
R^2 (adj.)	0.7102	0.7102	0.7102	**0.7102**	**0.7104**	0.7102	**0.7103**	0.7101	**0.7102**	**0.7103**	0.7102	0.7100	0.7101	0.7102	0.7100	0.7101

Table 4.5: **In-sample estimation results for WMB, 2005-2013**

Regression coefficient results for the WMB regression model set $\mathscr{M}_4$ with general equation $\log(\mathrm{RV}^{(d)}_{4,t+1}) = \alpha_4 + \beta^{(d)}_{4,\mathrm{RV}} \log(\mathrm{RV}^{(d)}_{4,t}) + \beta^{(w)}_{4,\mathrm{RV}} \log(\mathrm{RV}^{(w)}_{4,t}) + \beta^{(m)}_{4,\mathrm{RV}} \log(\mathrm{RV}^{(m)}_{4,t}) + \beta^{(d)}_{4,\mathrm{PQSPR}} \mathrm{PQSPR}^{(d)}_{4,t} + \beta^{(d)}_{4,\mathrm{PESPR}} \mathrm{PESPR}^{(d)}_{4,t} + \beta^{(w)}_{4,\mathrm{PESPR}} \mathrm{PESPR}^{(w)}_{4,t} + \beta^{(d)}_{4,\mathrm{VLMN}} \mathrm{VLMN}^{(d)}_{4,t} + \beta^{(w)}_{4,\mathrm{VLMN}} \mathrm{VLMN}^{(w)}_{4,t} + \beta^{(m)}_{4,\mathrm{VLMN}} \mathrm{VLMN}^{(m)}_{4,t} + \varepsilon_{4,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: LLTC – Model set $\mathscr{M}_5$							
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Intercept	-0.29	-0.41	-0.45	-0.58	-0.45	-0.57	-0.40	-0.52
	(-3.78*)	**(-4.37*)**	**(-4.32*)**	**(-4.78*)**	**(-3.32*)**	**(-3.81*)**	**(-2.66*)**	**(-3.20*)**
$\log(\mathrm{RV}^{(\mathrm{d})})$	0.38	0.38	0.34	0.34	0.34	0.34	0.34	0.34
	(9.41*)	**(9.58*)**	**(8.35*)**	**(8.45*)**	**(7.17*)**	**(7.26*)**	**(7.20*)**	**(7.30*)**
$\log(\mathrm{RV}^{(\mathrm{w})})$	0.32	0.32	0.31	0.31	0.31	0.32	0.30	0.30
	(3.94*)	**(4.00*)**	**(3.87*)**	**(3.92*)**	**(3.16*)**	**(3.24*)**	**(2.97*)**	**(3.01*)**
$\log(\mathrm{RV}^{(\mathrm{m})})$	0.23	0.21	0.24	0.21	0.24	0.21	0.27	0.25
	(4.67*)	**(4.41*)**	**(4.80*)**	**(4.55*)**	**(4.73*)**	**(4.46*)**	**(4.30*)**	**(4.14*)**
$\mathrm{PQSPR}^{(\mathrm{d})}$		0.02		0.02		0.02		0.02
		(2.52*)		**(2.59*)**		**(2.63*)**		**(2.73*)**
$\mathrm{VLMN}^{(\mathrm{d})}$			0.02	0.02	0.02	0.02	0.02	0.02
			(2.54*)	**(2.57*)**	**(2.13*)**	**(2.24*)**	**(2.12*)**	**(2.23*)**
$\mathrm{VLMN}^{(\mathrm{w})}$					-0.00	-0.00	0.01	0.01
					(-0.04)	(-0.13)	(0.42)	(0.45)
$\mathrm{VLMN}^{(\mathrm{m})}$							-0.01	-0.01
							(-0.81)	(-0.98)
R^2 (adj.)	0.6777	**0.6785**	**0.6792**	**0.6800**	**0.6791**	**0.6799**	**0.6790**	**0.6799**

Table 4.6: **In-sample estimation results for LLTC, 2005-2013**

Regression coefficient results for the LLTC regression model set $\mathscr{M}_5$ with general equation $\log(\mathrm{RV}^{(\mathrm{d})}_{5,t+1}) = \alpha_5 + \beta^{(\mathrm{d})}_{5,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{5,t}) + \beta^{(\mathrm{w})}_{5,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{5,t}) + \beta^{(\mathrm{m})}_{5,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{5,t}) + \beta^{(\mathrm{d})}_{5,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{d})}_{5,t} + \beta^{(\mathrm{d})}_{5,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{5,t} + \beta^{(\mathrm{w})}_{5,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{5,t} + \beta^{(\mathrm{m})}_{5,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{m})}_{5,t} + \varepsilon_{5,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: MON – Model set $\mathscr{M}_6$																	
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Intercept	-0.18	-0.30	-0.26	-0.32	-0.29	-0.52	-0.63	-0.23	-0.29	-0.54	-0.66	-0.50	-0.64	-0.22	-0.52	-0.70	-0.54	-0.57
	(-2.74*)	**(-3.67*)**	**(-3.81*)**	**(-4.62*)**	**(-3.58*)**	**(-5.18*)**	**(-6.13*)**	**(-3.17*)**	**(-4.13*)**	**(-5.23*)**	**(-6.42*)**	**(-4.64*)**	**(-5.79*)**	**(-2.76*)**	**(-4.32*)**	**(-5.65*)**	**(-4.26*)**	**(-4.14*)**
$\log(\mathrm{RV}^{(d)})$	0.38	0.37	0.36	0.35	0.36	0.32	0.30	0.34	0.34	0.32	0.31	0.31	0.31	0.34	0.32	0.32	0.31	0.32
	(13.79*)	**(13.19*)**	**(12.07*)**	**(11.84*)**	**(13.21*)**	**(10.31*)**	**(9.87*)**	**(10.82*)**	**(10.85*)**	**(10.62*)**	**(10.23*)**	**(9.80*)**	**(9.57*)**	**(10.85*)**	**(9.75*)**	**(9.65*)**	**(9.80*)**	**(9.72*)**
$\log(\mathrm{RV}^{(w)})$	0.22	0.21	0.22	0.22	0.21	0.20	0.19	0.25	0.23	0.19	0.18	0.20	0.18	0.24	0.20	0.16	0.22	0.21
	(5.04*)	**(4.70*)**	**(4.97*)**	**(4.97*)**	**(4.73*)**	**(4.35*)**	**(4.27*)**	**(5.07*)**	**(4.86*)**	**(4.11*)**	**(3.95*)**	**(3.93*)**	**(3.62*)**	**(4.73*)**	**(3.42*)**	**(2.93*)**	**(4.09*)**	**(3.53*)**
$\log(\mathrm{RV}^{(m)})$	0.36	0.36	0.36	0.36	0.36	0.37	0.36	0.36	0.36	0.36	0.36	0.37	0.36	0.36	0.36	0.36	0.35	0.34
	(9.04*)	**(9.19*)**	**(8.98*)**	**(8.92*)**	**(9.26*)**	**(9.23*)**	**(9.15*)**	**(8.89*)**	**(8.91*)**	**(9.25*)**	**(9.12*)**	**(9.12*)**	**(9.11*)**	**(7.42*)**	**(9.19*)**	**(9.11*)**	**(7.25*)**	**(7.35*)**
$\mathrm{PESPR}^{(d)}$		0.02			0.02	0.03	0.03			0.02	0.03	0.03	0.04		0.02	0.02	0.03	0.03
		(2.63*)			(1.33)	**(3.88*)**	**(4.37*)**			(1.54)	(1.58)	**(3.63*)**	**(4.25*)**		(1.53)	(1.50)	**(3.56*)**	(1.55)
$\mathrm{PESPR}^{(w)}$					-0.00					0.01	0.01				0.01	0.01		0.01
					(-0.21)					(0.43)	(0.76)				(0.32)	(0.82)		(0.45)
$\mathrm{VLMN}^{(d)}$			0.02			0.03		0.03		0.03		0.03		0.03	0.03		0.03	0.03
			(2.79*)			**(4.19*)**		**(3.19*)**		**(4.22*)**		**(3.38*)**		**(3.19*)**	**(3.28*)**		**(3.42*)**	**(3.30*)**
$\mathrm{VLMN}^{(w)}$								-0.02				-0.00		-0.01	-0.00		-0.01	-0.01
								(-1.54)				(-0.42)		(-0.86)	(-0.27)		(-0.77)	(-0.68)
$\mathrm{VLMN}^{(m)}$														-0.00			0.01	0.01
														(-0.21)			(0.71)	(0.80)
$\mathrm{VLMD}^{(d)}$				0.02			0.04		0.03		0.04		0.04			0.03		
				(3.58*)			**(5.36*)**		**(3.29*)**		**(5.49*)**		**(3.56*)**			**(3.36*)**		
$\mathrm{VLMD}^{(w)}$									-0.01				0.00			0.01		
									(-0.86)				(0.26)			(0.49)		
R^2 (adj.)	0.7194	**0.7200**	**0.7200**	**0.7207**	**0.7199**	**0.7216**	**0.7228**	**0.7201**	**0.7206**	**0.7215**	**0.7228**	**0.7215**	**0.7227**	**0.7200**	**0.7214**	**0.7227**	**0.7215**	**0.7214**

Table 4.7: **In-sample estimation results for MON, 2005-2013**

Regression coefficient results for the MON regression model set $\mathscr{M}_6$ with general equation $\log(\mathrm{RV}^{(d)}_{6,t+1}) = \alpha_6 + \beta^{(d)}_{6,\mathrm{RV}} \log(\mathrm{RV}^{(d)}_{6,t}) + \beta^{(w)}_{6,\mathrm{RV}} \log(\mathrm{RV}^{(w)}_{6,t}) + \beta^{(m)}_{6,\mathrm{RV}} \log(\mathrm{RV}^{(m)}_{6,t}) + \beta^{(d)}_{6,\mathrm{PESPR}} \mathrm{PESPR}^{(d)}_{6,t} + \beta^{(w)}_{6,\mathrm{PESPR}} \mathrm{PESPR}^{(w)}_{6,t} + \beta^{(d)}_{6,\mathrm{VLMN}} \mathrm{VLMN}^{(d)}_{6,t} + \beta^{(w)}_{6,\mathrm{VLMN}} \mathrm{VLMN}^{(w)}_{6,t} + \beta^{(m)}_{6,\mathrm{VLMN}} \mathrm{VLMN}^{(m)}_{6,t} + \beta^{(d)}_{6,\mathrm{VLMD}} \mathrm{VLMD}^{(d)}_{6,t} + \beta^{(w)}_{6,\mathrm{VLMD}} \mathrm{VLMD}^{(w)}_{6,t} + \varepsilon_{6,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: NKE – Model set $\mathscr{M}_7$											
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Intercept	-0.22	-0.57	-0.54	-0.26	-0.79	-0.69	-0.21	-0.77	-0.66	-0.22	-0.82	-0.71
	(-3.43*)	**(-5.59*)**	**(-5.65*)**	**(-3.78*)**	**(-6.67*)**	**(-6.29*)**	**(-2.81*)**	**(-5.70*)**	**(-5.37*)**	**(-2.84*)**	**(-5.83*)**	**(-5.49*)**
$\log(\mathrm{RV}^{(\mathrm{d})})$	0.41	0.39	0.38	0.40	0.35	0.34	0.38	0.35	0.34	0.38	0.34	0.34
	(13.43*)	**(12.75*)**	**(12.11*)**	**(12.19*)**	**(10.53*)**	(10.25*)	**(11.43*)**	**(10.32*)**	**(10.00*)**	**(11.41*)**	**(10.22*)**	**(9.90*)**
$\log(\mathrm{RV}^{(\mathrm{w})})$	0.25	0.24	0.23	0.25	0.22	0.22	0.28	0.23	0.23	0.28	0.24	0.25
	(5.86*)	**(5.60*)**	**(5.52*)**	**(5.82*)**	**(5.40*)**	**(5.33*)**	**(5.65*)**	**(4.73*)**	**(4.73*)**	**(5.33*)**	**(4.66*)**	**(4.66*)**
$\log(\mathrm{RV}^{(\mathrm{m})})$	0.29	0.25	0.27	0.30	0.25	0.28	0.29	0.25	0.28	0.29	0.23	0.26
	(7.48*)	**(6.26*)**	**(7.00*)**	**(7.58*)**	**(6.38*)**	**(7.24*)**	**(7.28*)**	**(6.32*)**	**(7.07*)**	**(6.09*)**	**(4.86*)**	**(5.52*)**
$\mathrm{PQSPR}^{(\mathrm{d})}$		0.04			0.05			0.05			0.05	
		(4.01*)			**(4.99*)**			**(4.68*)**			**(4.75*)**	
$\mathrm{PESPR}^{(\mathrm{d})}$			0.03			0.04			0.04			0.04
			(4.08*)			**(4.60*)**			**(4.37*)**			**(4.42*)**
$\mathrm{VLMN}^{(\mathrm{d})}$				0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
				(1.54)	**(3.32*)**	**(2.95*)**	**(2.45*)**	**(3.22*)**	**(2.94*)**	**(2.46*)**	**(3.27*)**	**(2.99*)**
$\mathrm{VLMN}^{(\mathrm{w})}$							-0.01	-0.00	-0.00	-0.02	-0.01	-0.01
							(-1.58)	(-0.37)	(-0.57)	(-1.33)	(-0.82)	(-0.96)
$\mathrm{VLMN}^{(\mathrm{m})}$										0.00	0.01	0.01
										(0.27)	(0.92)	(0.91)
R^2 (adj.)	0.7184	**0.7204**	**0.7204**	**0.7185**	**0.7216**	**0.7213**	**0.7187**	**0.7215**	**0.7212**	**0.7186**	**0.7215**	**0.7212**

Table 4.8: **In-sample estimation results for NKE, 2005-2013**

Regression coefficient results for the NKE regression model set $\mathscr{M}_7$ with general equation $\log(\mathrm{RV}^{(\mathrm{d})}_{7,t+1}) = \alpha_7 + \beta^{(\mathrm{d})}_{7,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{7,t}) + \beta^{(\mathrm{w})}_{7,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{7,t}) + \beta^{(\mathrm{m})}_{7,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{7,t}) + \beta^{(\mathrm{d})}_{7,\mathrm{PQSPR}} \mathrm{PQSPR}^{(\mathrm{d})}_{7,t} + \beta^{(\mathrm{d})}_{7,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{7,t} + \beta^{(\mathrm{d})}_{7,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{7,t} + \beta^{(\mathrm{w})}_{7,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{7,t} + \beta^{(\mathrm{m})}_{7,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{m})}_{7,t} + \varepsilon_{7,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: BSX – Model set $\mathscr{M}_8$					
Factor	**1**	**2**	**3**	**4**	**5**	**6**
Intercept	-0.32	-0.42	-0.30	-0.41	-0.28	-0.41
	(-3.31*)	**(-3.75*)**	**(-3.07*)**	**(-3.74*)**	**(-2.83*)**	**(-3.71*)**
$\log(\mathrm{RV}^{(\mathrm{d})})$	0.35	0.35	0.36	0.37	0.35	0.34
	(12.50*)	**(12.28*)**	**(12.25*)**	**(12.32*)**	**(11.19*)**	**(11.00*)**
$\log(\mathrm{RV}^{(\mathrm{w})})$	0.17	0.17	0.17	0.17	0.19	0.21
	(3.57*)	**(3.56*)**	**(3.61*)**	**(3.61*)**	**(3.64*)**	**(3.97*)**
$\log(\mathrm{RV}^{(\mathrm{m})})$	0.41	0.39	0.40	0.37	0.39	0.35
	(9.68*)	**(9.19*)**	**(9.29*)**	**(8.44*)**	**(9.02*)**	**(7.90*)**
$\mathrm{PESPR}^{(\mathrm{d})}$		0.01		0.02		0.02
		(2.17*)		**(2.70*)**		**(3.24*)**
$\mathrm{VLMN}^{(\mathrm{d})}$			-0.01	-0.01	-0.00	-0.00
			(-1.29)	(-1.82)	(-0.12)	(-0.17)
$\mathrm{VLMN}^{(\mathrm{w})}$					-0.01	-0.02
					(-1.31)	**(-2.12*)**
R^2 (adj.)	0.5361	**0.5368**	**0.5363**	**0.5374**	**0.5364**	**0.5380**

Table 4.9: **In-sample estimation results for BSX, 2005-2013**

Regression coefficient results for the BSX regression model set $\mathscr{M}_8$ with general equation $\log(\mathrm{RV}^{(\mathrm{d})}_{8,t+1}) = \alpha_8 + \beta^{(\mathrm{d})}_{8,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{d})}_{8,t}) + \beta^{(\mathrm{w})}_{8,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{w})}_{8,t}) + \beta^{(\mathrm{m})}_{8,\mathrm{RV}} \log(\mathrm{RV}^{(\mathrm{m})}_{8,t}) + \beta^{(\mathrm{d})}_{8,\mathrm{PESPR}} \mathrm{PESPR}^{(\mathrm{d})}_{8,t} + \beta^{(\mathrm{d})}_{8,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{d})}_{8,t} + \beta^{(\mathrm{w})}_{8,\mathrm{VLMN}} \mathrm{VLMN}^{(\mathrm{w})}_{8,t} + \varepsilon_{8,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: HPQ – Model set $\mathscr{M}_9$															
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Intercept	-0.39	-0.41	-0.42	-0.35	-0.37	-0.41	-0.38	-0.28	-0.30	-0.38	-0.31	-0.29	-0.30	-0.30	-0.31	-0.31
	(-4.61*)	**(-4.52*)**	**(-4.41*)**	**(-4.15*)**	**(-4.09*)**	**(-4.54*)**	**(-3.96*)**	**(-3.06*)**	**(-3.11*)**	**(-4.07*)**	**(-3.02*)**	**(-3.10*)**	**(-3.12*)**	**(-3.09*)**	**(-3.08*)**	**(-3.13*)**
$\log(\mathrm{RV}^{(d)})$	0.38	0.38	0.38	0.39	0.40	0.37	0.39	0.36	0.36	0.39	0.36	0.36	0.36	0.35	0.36	0.35
	(13.42*)	**(13.42*)**	**(13.26*)**	**(12.99*)**	**(13.02*)**	**(13.03*)**	**(12.80*)**	**(10.68*)**	**(10.66*)**	**(12.57*)**	**(10.57*)**	**(10.61*)**	**(10.61*)**	**(10.46*)**	**(10.52*)**	**(10.43*)**
$\log(\mathrm{RV}^{(w)})$	0.29	0.29	0.29	0.30	0.29	0.29	0.29	0.36	0.36	0.30	0.36	0.37	0.37	0.36	0.36	0.37
	(6.34*)	**(6.33*)**	**(6.31*)**	**(6.44*)**	**(6.43*)**	**(6.49*)**	**(6.40*)**	**(6.53*)**	**(6.59*)**	**(6.59*)**	**(6.52*)**	**(6.12*)**	**(6.11*)**	**(6.74*)**	**(6.07*)**	**(6.31*)**
$\log(\mathrm{RV}^{(m)})$	0.24	0.24	0.24	0.23	0.23	0.24	0.23	0.22	0.21	0.23	0.22	0.21	0.21	0.22	0.21	0.21
	(6.69*)	**(6.62*)**	**(6.66*)**	**(6.45*)**	**(6.37*)**	**(6.69*)**	**(6.42*)**	**(5.90*)**	**(5.78*)**	**(6.45*)**	**(5.88*)**	**(4.85*)**	**(4.83*)**	**(5.90*)**	**(4.85*)**	**(4.85*)**
$\mathrm{PQSPR}^{(d)}$		0.01			0.01				0.01				0.01			
		(0.93)			(1.03)				(1.28)				(1.24)			
$\mathrm{PESPR}^{(d)}$			0.01			0.01	0.01			0.01	0.01			0.01	0.01	0.01
			(1.06)			(0.56)	(1.05)			(0.55)	(1.06)			(0.61)	(1.04)	(0.61)
$\mathrm{PESPR}^{(w)}$						-0.01				-0.01				-0.01		-0.01
						(-0.29)				(-0.28)				(-0.34)		(-0.34)
$\mathrm{VLMN}^{(d)}$				-0.01	-0.01		-0.01	0.01	0.01	-0.01	0.01	0.01	0.01	0.01	0.01	0.01
				(-1.89)	**(-2.04*)**		(-1.91)	(1.14)	(1.16)	(-1.90)	(1.16)	(1.15)	(1.16)	(1.18)	(1.16)	(1.18)
$\mathrm{VLMN}^{(w)}$								-0.03	-0.03		-0.03	-0.03	-0.03	-0.03	-0.03	-0.03
								(-3.37*)	**(-3.47*)**		**(-3.39*)**	**(-2.85*)**	**(-2.82*)**	**(-3.42*)**	**(-2.81*)**	**(-2.86*)**
$\mathrm{VLMN}^{(m)}$												0.00	0.00		0.00	0.00
												(0.33)	(0.12)		(0.25)	(0.26)
R^2 (adj.)	0.6092	0.6092	0.6092	0.6094	0.6094	0.6091	0.6094	0.6109	0.6111	0.6093	0.6110	0.6108	0.6109	0.6108	0.6108	0.6107

Table 4.10: **In-sample estimation results for HPQ, 2005-2013**

Regression coefficient results for the HPQ regression model set $\mathscr{M}_9$ with general equation $\log(\mathrm{RV}^{(d)}_{9,t+1}) = \alpha_9 + \beta^{(d)}_{9,\mathrm{RV}} \log(\mathrm{RV}^{(d)}_{9,t}) + \beta^{(w)}_{9,\mathrm{RV}} \log(\mathrm{RV}^{(w)}_{9,t}) + \beta^{(m)}_{9,\mathrm{RV}} \log(\mathrm{RV}^{(m)}_{9,t}) + \beta^{(d)}_{9,\mathrm{PQSPR}} \mathrm{PQSPR}^{(d)}_{9,t} + \beta^{(d)}_{9,\mathrm{PESPR}} \mathrm{PESPR}^{(d)}_{9,t} + \beta^{(w)}_{9,\mathrm{PESPR}} \mathrm{PESPR}^{(w)}_{9,t} + \beta^{(d)}_{9,\mathrm{VLMN}} \mathrm{VLMN}^{(d)}_{9,t} + \beta^{(w)}_{9,\mathrm{VLMN}} \mathrm{VLMN}^{(w)}_{9,t} + \beta^{(m)}_{9,\mathrm{VLMN}} \mathrm{VLMN}^{(m)}_{9,t} + \varepsilon_{9,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

	Regression results: CVX – Model set $\mathscr{M}_{10}$											
Factor	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Intercept	-0.25	-0.29	-0.34	-0.59	-0.74	-0.32	-0.70	-0.52	-0.70	-0.71	-0.66	-0.66
	(-3.91*)	**(-3.88*)**	**(-3.89*)**	**(-6.42*)**	**(-6.62*)**	**(-3.74*)**	**(-5.74*)**	**(-5.10*)**	**(-5.41*)**	**(-6.19*)**	**(-4.79*)**	**(-5.00*)**
$\log(\mathrm{RV}^{(d)})$	0.41	0.41	0.40	0.36	0.34	0.40	0.34	0.35	0.33	0.34	0.33	0.33
	(13.38*)	**(13.06*)**	**(12.48*)**	**(11.40*)**	**(10.48*)**	**(12.17*)**	**(10.36*)**	**(10.94*)**	**(10.37*)**	**(10.00*)**	**(10.18*)**	**(9.74*)**
$\log(\mathrm{RV}^{(w)})$	0.40	0.40	0.39	0.39	0.38	0.40	0.38	0.41	0.39	0.38	0.40	0.39
	(8.90*)	**(8.82*)**	**(8.68*)**	**(8.71*)**	**(8.55*)**	**(8.54*)**	**(8.43*)**	**(8.54*)**	**(8.06*)**	**(8.12*)**	**(7.96*)**	**(7.59*)**
$\log(\mathrm{RV}^{(m)})$	0.13	0.13	0.13	0.12	0.12	0.13	0.12	0.13	0.12	0.12	0.13	0.12
	(4.04*)	**(4.01*)**	**(4.07*)**	**(3.82*)**	**(3.66*)**	**(4.15*)**	**(3.86*)**	**(3.92*)**	**(3.73*)**	**(3.87*)**	**(3.93*)**	**(3.94*)**
$\mathrm{PQSPR}^{(d)}$		0.01			0.02				0.02			
		(1.27)			**(2.84*)**				**(2.64*)**			
$\mathrm{PESPR}^{(d)}$			0.01			0.02	0.01			0.01	0.01	0.01
			(1.98*)			(1.50)	**(2.07*)**			(0.97)	(1.95)	(0.92)
$\mathrm{PESPR}^{(w)}$						-0.01				0.00		0.00
						(-0.58)				(0.15)		(0.14)
$\mathrm{VLMN}^{(d)}$				0.04	0.05		0.04	0.05	0.05	0.04	0.05	0.05
				(5.13*)	**(5.62*)**		**(5.11*)**	**(3.86*)**	**(3.97*)**	**(5.26*)**	**(3.65*)**	**(3.72*)**
$\mathrm{VLMN}^{(w)}$								-0.01	-0.01		-0.01	-0.01
								(-0.97)	(-0.48)		(-0.61)	(-0.60)
R^2 (adj.)	0.7488	**0.7488**	**0.7491**	**0.7514**	**0.7521**	**0.7490**	**0.7519**	**0.7514**	**0.7520**	**0.7518**	**0.7518**	**0.7517**

Table 4.11: **In-sample estimation results for CVX, 2005-2013**

Regression coefficient results for the CVX regression model set $\mathscr{M}_{10}$ with general equation $\log(\mathrm{RV}^{(d)}_{10,t+1}) = \alpha_{10} + \beta^{(d)}_{10,\mathrm{RV}} \log(\mathrm{RV}^{(d)}_{10,t}) + \beta^{(w)}_{10,\mathrm{RV}} \log(\mathrm{RV}^{(w)}_{10,t}) + \beta^{(m)}_{10,\mathrm{RV}} \log(\mathrm{RV}^{(m)}_{10,t}) + \beta^{(d)}_{10,\mathrm{PQSPR}} \mathrm{PQSPR}^{(d)}_{10,t} + \beta^{(d)}_{10,\mathrm{PESPR}} \mathrm{PESPR}^{(d)}_{10,t} + \beta^{(w)}_{10,\mathrm{PESPR}} \mathrm{PESPR}^{(w)}_{10,t} + \beta^{(d)}_{10,\mathrm{VLMN}} \mathrm{VLMN}^{(d)}_{10,t} + \beta^{(w)}_{10,\mathrm{VLMN}} \mathrm{VLMN}^{(w)}_{10,t} + \varepsilon_{10,t}$. Newey-west t-statistics are displayed in parenthesis below the coefficient estimate, bold and starred when significant at the 95% confidence level. Adjusted R^2 is set to bold when greater than the benchmark's adjusted R^2 (column 1). Volume and spread time series have been centered and normalized to unit length.

4.2 Out-of-Sample Analysis

For every model estimated in the previous section for which all coefficients have their t-statistics individually significant at the 95% confidence level, we pursue an out-of-sample performance analysis. We do not report results of WMB and HPQ, for which only the benchmark remains in the model set after the t-statistics filter, because our focus is on the relative performance of HAR-RV-LIQ models that appear to be good a fit, with respect to the benchmark's. Our measure of performance is the root mean square prediction error (RMSE)[1] of the annualized daily realized volatility, $\sqrt{252}\mathrm{RV}_{i,t+1}^{(\mathrm{d})}$ using the HAR-RV-LIQ model for predicting the logarithm of the daily realized volatility, $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$. All remaining models, all firms combined, are nested within the following one:

$$\begin{aligned}\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})}) = \alpha_i &+ \beta_{i,\mathrm{RV}}^{(\mathrm{d})}\log(\mathrm{RV}_{i,t}^{(\mathrm{d})}) + \beta_{i,\mathrm{RV}}^{(\mathrm{w})}\log(\mathrm{RV}_{i,t}^{(\mathrm{w})}) + \beta_{i,\mathrm{RV}}^{(\mathrm{m})}\log(\mathrm{RV}_{i,t}^{(\mathrm{m})}) \\ &+ \beta_{i,\mathrm{PQSPR}}^{(\mathrm{d})}\mathrm{PQSPR}_{i,t}^{(\mathrm{d})} + \beta_{i,\mathrm{PESPR}}^{(\mathrm{d})}\mathrm{PESPR}_{i,t}^{(\mathrm{d})} \\ &+ \beta_{i,\mathrm{VLMN}}^{(\mathrm{d})}\mathrm{VLMN}_{i,t}^{(\mathrm{d})} + \beta_{i,\mathrm{VLMN}}^{(\mathrm{w})}\mathrm{VLMN}_{i,t}^{(\mathrm{w})} + \beta_{i,\mathrm{VLMD}}^{(\mathrm{d})}\mathrm{VLMD}_{i,t}^{(\mathrm{d})} \\ &+ \varepsilon_{i,t}.\end{aligned} \tag{4.11}$$

We are interested to see how the model performs in different recent historical periods. From our original sample, we construct three overlapping subsamples of 1250 consecutive daily observations of $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$, for which the first 3 years (750 days) are used for in-sample estimation and for which the remaining 2 years (500 days) are used to compute the prediction errors. The three 2-years prediction periods are mostly non-overlapping and corresponds roughly to the 2008-2009, 2010-2011 and 2012-2013 pairs of years, as presented in table 4.12. The first period corresponds to the depth of the Financial Crisis and the other periods are part of the aftermath.

We use two different method to gather a time series of coefficient estimates for our daily predictions. For the first method, which we refer to as static, we only estimate the model once, using the 3 years of historical data prior to the first prediction, and keep the coefficients constant for all future predictions. For the second method, dynamic, we re-estimate the model every 25 days and keep the coefficients constant for the next 24 days.

[1]Given a time series of T prediction errors $\{e_t\}_{t=1}^{T}$ from a linear model with p estimated parameters, $\mathrm{RMSE} = \sum_{t=1}^{T} e_t^2/(T-p)$.

The most recent 750 observations are used every time the model is estimated. As a consequence, the daily implied volatility during the Financial Crisis will mostly be predicted using coefficients estimated during a period low volatility and high growth. Assuming a linear autoregressive model with constant coefficients is not a correct specification of the real underlying volatility process, we should observe a higher RMSE for the benchmark during this period for the majority of firms. Following the results, we will be able to observe whether incorporating volume or spread terms can help reduce the RMSE in periods when it is the highest. The out-of-sample prediction results are presented in tables 4.13 to 4.20 (RMSE) and A.2.1 to A.2.8 (RMSPE). The RMSE for models with additional spread or volume terms is presented as a percentage change from the RMSE of the benchmark model, i.e. for the model j,

$$\text{Performance}_j = 100\% \times (\text{RMSE}_j - \text{RMSE}_{\text{bm}})/\text{RMSE}_{\text{bm}}, \tag{4.12}$$

where RMSE_{bm} is the RMSE of the benchmark. We follow the same presentation format for the root mean square percentage prediction error (RMSPE)[2] in appendix, i.e. for the model j,

$$\text{Relative Performance}_j = 100\% \times (\text{RMSPE}_j - \text{RMSPE}_{\text{bm}})/\text{RMSPE}_{\text{bm}}, \tag{4.13}$$

where RMSPE_{bm} is the RMSE of the benchmark.

Period	**2008 - 2009**	**2010 - 2011**	**2012 - 2013**
First prediction date	February 6, 2008	January 7, 2010	December 27, 2011
Last prediction date	February 5, 2010	January 3, 2012	December 31, 2013
Number of observations	500	500	500

Table 4.12: **Out-of-sample analysis periods**
The out-of-sample performance analysis is performed for three subsample periods each covering roughly 2 years of observations.

We first observe as expected a higher RMSE for the benchmark in the 2008-2009 period with static estimation. Across all firms, LLTC excluded, the 2008-2009 RMSE with

[2] Given a time series of T prediction errors $\{e_t\}_{t=1}^{T}$ from a linear model with p estimated parameters and T observed true values of the predicted time series $\{y_t\}_{t=1}^{T}$, $\text{RMSPE} = \sum_{t=1}^{T}(\frac{e_t}{y_t})^2/(T-p)$.

static estimation ranges from 20% (BSX, table 4.19) to 150% (NUE, table 4.15) higher than the 2010-2011 RMSE and ranges from 130% (NKE, table 4.18) to 280% (CVX, table 4.20) higher than the 2012-2013 RMSE, LLTC included. The benchmark's root mean square percentage errors (RMSPE), with either static or dynamic estimation, are however surprisingly stable across all firms and periods, ranging only between 20% to 25%, with the exception of LLTC and BSX both spiking to 30% and 34% respectively during the 2010-2011 period. The simple autoregressive model thus yields higher absolute errors in periods of spiking volatility, which begs for improvements. The 2010-2011 period includes the 2010 Flash Clash, which most affected LLTC and was a one-off volatility spike (see figure 3.5) The autocorrelated and jump-less structure of the model likely caused severe estimation and prediction biases throughout the entire period.

Re-estimating the coefficients every month (dynamic) rather than once prior to 2 years of daily predictions (static) shows some reduction in RMSE during the 2008-2009 period for all firms, from -1.3% (MON, table 4.17) to -12.1% (X, table 4.14). However, during the same period, we generally observe small increases in RMSPE. For the other two periods, the variations in RMSE and RMSPE are both negligible between static and dynamic estimation. The RMSE puts an equal weight on square errors, while the RMSPE puts a weight inversely proportional to the volatility level. Therefore, a decrease in RMSE at the cost of a comparable increase in RMSPE is overall beneficial, as it suggests lower kurtosis of prediction errors, due to some reduction in large absolute errors when the volatility is high and some increases in small absolute errors when the volatility is low.

We now wish to identify the best predictive model for each firm for which we will further analyse estimation residuals and prediction errors. To analyse which model offers the best performance for a given firm, we will mostly focus on the RMSE reduction in periods of high RMSE for the benchmark. In our opinion, a model better than the benchmark would be one that reduces the RMSE in periods when it is the highest, while maintaining or improving it in other periods. A better model should result in consistently better performance in across different periods. Since small absolute errors in period of low volatility can inflate the RMSPE, we will only use the RMSPE performance to distinguish between two models with comparable RMSE performance. For each firm, we collect a pair of model to further analyse in the next section: the benchmark and one HAR-RV-LIQ model that displays the best predictive performance in terms of RMSE.

For CMS, tables 4.13 and A.2.1, all models with two or more terms in addition to the benchmark show an increase of at least 1% for RMSE and 2% for RMSPE for some period and some estimation method. They also offer next to no improvements under static estimation for any period. The models with either only quoted spread or effective spread have similar performance both in terms of RMSE and RMSPE. Between PQSPR and PESPR, the latter (model 2), appears to yield a slightly more consistent performance, since it does better with dynamic estimation during the 2008-2009 period. In fact, the daily PQSPR and PESPR time series for CMS are near perfectly cross-correlated as can be observed in figure A.1.6, so which one is selected should have little impact as their overlap in information appears to be near complete. We select model 2, with the daily PQSPR term, as our model for CMS to further investigate in the next section.

For X, tables 4.14 and A.2.2, the models that offer the best good improvements during the 2008-2009 period with static estimation all offer modest or severe degradation with dynamic improvements. The lack of consistency when the coefficients are more frequently estimated lets us believe that the 2008-2009 performance with static estimation was a one-off product of chance. All models with the volume term show this distortion. We therefore opt for model 2, with only the effective spread and the most consistent performance, despite its generally poorer performance than the benchmark.

For NUE, tables 4.15 and A.2.3, the model 2 with the quoted spread, mostly maintains the benchmark's performance, with minor 2008-2009 improvements of 0.05% with static estimation and 0.32% with dynamic estimation, while all other models generally degrade the benchmark's performance.

For LLTC, tables 4.16 and A.2.4, the model 4 with the quoted spread and volume in number of shares terms, shows the best improvements during the 2010-2011 period, LLTC's worst performance for the benchmark, with 1.0% and 1.2% reductions in RMSE with static and dynamic estimation respectively, with improvements elsewhere except static 2008-2009, with an increase in RMSE of 0.63%. However, no model with static estimation improves upon the benchmark's performance during the 2008-2009 period.

For MON, tables 4.17 and A.2.5, no model offers a better performance than the benchmark. Model 2 with the effective spread seems to more consistently and closely replicate the benchmark's performance, with no decrease in RMSE larger than 1% for either static or dynamic estimation.

For NKE, tables 4.18 and A.2.6, the inclusion of the volume term in addition to either the quoted or effective spread significantly degrades the predictive performance during the 2008-2009 period, with increases in static RMSE 3.2% and 2.4% respectively. This performance is however slightly improved when the coefficients are re-estimated every month, yet still inferior to models without the volume term. Furthermore, no model improves upon the benchmark during the period when it performs the worst. Similarly to the case of CMS, the performance of the models with either the PQSPR or PESPR term are hardly distinguishable in terms of either RMSE or RMSPE. The concurrent cross-correlation between these two time series is near unity, as seen in figure A.1.12. We select model 3, with PESPR, for its (barely) lowest RMSE during 2008-2009 period.

There is only one model to compare with the benchmark for BSX, tables 4.19 and A.2.7, which includes a proportional effective spread term. This model improves upon the benchmark during the 2008-2009 period, its worst in terms of RMSE, by 2.9% and 1.1% for the static and dynamic estimation respectively, while only slightly underperforming in other periods of lower benchmark RMSE, with increases in RMSE varying from 0.02% to 0.25%.

Finally, for CVX, tables 4.20 and A.2.8, the models 4 and 5 with volume or quoted spread and volume respectively improve upon the benchmark in all periods for both methods of estimation. Model 4, with only the volume, also improves upon the benchmark in terms of RMSPE for all periods and for both methods, while also showing the best dynamic RMSE performance for 2008-2009. We thus select it for further analysis in the next section.

We have eliminated models with volume terms for all firms but 2, LLTC and CVX, and a model with a spread component has passed the various stages of our analysis for 7 out of 10 firms.

	RMSE of annualized daily RV predictions – CMS					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.0962	0.0527	0.0293	0.0938	0.0531	0.0293
[2] PQSPR$^{(d)}$	+0.22%	**-0.17%**	**-0.42%**	+0.01%	**-0.19%**	+0.69%
[3] PESPR$^{(d)}$	**-0.26%**	**-0.12%**	**-0.32%**	+0.19%	**-0.30%**	+0.69%
[5] PQSPR$^{(d)}$, VLMN$^{(d)}$	**-0.17%**	+0.09%	+0.27%	+0.91%	**-0.04%**	+1.22%
[6] PESPR$^{(d)}$, VLMN$^{(d)}$	**-0.87%**	+0.22%	+0.40%	+0.96%	**-0.17%**	+1.27%
[7] VLMN$^{(d)}$, VLMN$^{(w)}$	+0.33%	+0.24%	+0.59%	+1.24%	+0.09%	+0.51%
[8] PQSPR$^{(d)}$, VLMN$^{(d)}$, VLMN$^{(w)}$	+0.44%	+0.08%	+0.22%	+1.29%	**-0.10%**	+1.06%
[9] PESPR$^{(d)}$, VLMN$^{(d)}$, VLMN$^{(w)}$	+0.10%	+0.18%	+0.32%	+1.40%	**-0.24%**	+1.11%

Table 4.13: **Out-of-sample performance results for CMS**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of CMS, $\sqrt{252}\mathrm{RV}^{(\mathrm{d})}_{\mathrm{CMS},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.2, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – X					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.1986	0.0903	0.0820	0.1745	0.0907	0.0815
[2] PESPR$^{(d)}$	+0.51%	**-0.32%**	+3.49%	+0.67%	**-0.38%**	**-0.03%**
[3] VLMN$^{(d)}$	**-4.77%**	**-0.08%**	**-0.13%**	+0.89%	+0.04%	**-0.28%**
[5] PESPR$^{(d)}$, VLMN$^{(d)}$	**-2.49%**	**-0.86%**	+6.26%	+8.00%	**-0.90%**	**-0.56%**
[6] VLMN$^{(d)}$, VLMN$^{(w)}$	**-1.00%**	**-0.65%**	**-0.39%**	+2.73%	**-0.43%**	**-0.43%**
[8] PESPR$^{(d)}$, VLMN$^{(d)}$, VLMN$^{(w)}$	**-1.01%**	**-1.17%**	+5.69%	+6.64%	**-1.09%**	**-0.51%**

Table 4.14: **Out-of-sample performance results for X**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of X, $\sqrt{252}\mathrm{RV}^{(\mathrm{d})}_{\mathrm{X},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.3, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – NUE					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.1629	0.0657	0.0389	0.1582	0.0660	0.0388
[2] $\text{PQSPR}^{(\text{d})}$	**-0.05%**	+0.05%	+0.57%	**-0.32%**	+0.09%	**-0.04%**
[3] $\text{PESPR}^{(\text{d})}$	+1.05%	**-0.21%**	+2.82%	+0.24%	**-0.21%**	+0.33%
[4] $\text{VLMN}^{(\text{d})}$	+2.65%	**-0.24%**	**-0.40%**	+1.41%	**-0.09%**	**-0.65%**
[5] $\text{PQSPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	+3.52%	**-0.15%**	+0.10%	+1.32%	+0.13%	**-0.69%**
[7] $\text{PESPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	+4.41%	**-0.46%**	+1.99%	+2.75%	**-0.28%**	**-0.41%**

Table 4.15: **Out-of-sample performance results for NUE**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of NUE, $\sqrt{252}\text{RV}^{(\text{d})}_{\text{NUE},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.4, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – LLTC					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.0917	0.1734	0.0380	0.0900	0.1735	0.0378
[2] $\text{PQSPR}^{(\text{d})}$	+0.28%	**-0.56%**	**-0.62%**	+0.14%	**-0.59%**	**-0.33%**
[3] $\text{VLMN}^{(\text{d})}$	+0.53%	**-0.26%**	+0.30%	**-0.59%**	**-0.42%**	+0.07%
[4] $\text{PQSPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	+0.63%	**-0.97%**	**-0.77%**	**-0.35%**	**-1.19%**	**-0.42%**

Table 4.16: **Out-of-sample performance results for LLTC**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of LLTC, $\sqrt{252}\text{RV}^{(\text{d})}_{\text{LLTC},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.6, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – MON					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.1420	0.0710	0.0374	0.1402	0.0718	0.0373
[2] $\text{PESPR}^{(\text{d})}$	+0.41%	**-0.22%**	+0.47%	+0.27%	**-0.84%**	+0.43%
[3] $\text{VLMN}^{(\text{d})}$	**-0.22%**	+0.44%	**-0.06%**	+1.07%	+0.43%	**-0.10%**
[4] $\text{VLMD}^{(\text{d})}$	+5.10%	+0.45%	**-0.03%**	+1.64%	+0.66%	**-0.03%**
[6] $\text{PESPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	+1.07%	+0.15%	+0.48%	+3.05%	**-0.48%**	+0.29%
[7] $\text{PESPR}^{(\text{d})}$, $\text{VLMD}^{(\text{d})}$	+5.37%	**-0.04%**	+0.56%	+2.53%	**-0.40%**	+0.30%

Table 4.17: **Out-of-sample performance results for MON**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of MON, $\sqrt{252}\text{RV}^{(\text{d})}_{\text{MON},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.7, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – NKE					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.0833	0.0574	0.0387	0.0806	0.0579	0.0386
[2] $\text{PQSPR}^{(\text{d})}$	+0.57%	**-0.61%**	**-1.01%**	+0.50%	**-0.70%**	**-1.00%**
[3] $\text{PESPR}^{(\text{d})}$	+0.30%	**-0.21%**	**-1.42%**	+0.49%	**-0.54%**	**-1.23%**
[5] $\text{PQSPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	+3.21%	**-0.65%**	**-0.66%**	+0.69%	**-0.71%**	**-0.62%**
[6] $\text{PESPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	+2.37%	**-0.15%**	**-1.25%**	+1.10%	**-0.49%**	**-0.93%**

Table 4.18: **Out-of-sample performance results for NKE**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of NKE, $\sqrt{252}\text{RV}^{(\text{d})}_{\text{NKE},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.8, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – BSX					
	Static			**Dynamic**		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.1425	0.1159	0.0626	0.1314	0.1158	0.0623
[2] $\text{PESPR}^{(d)}$	**-2.86%**	+0.25%	+0.02%	**-1.10%**	+0.18%	+0.04%

Table 4.19: **Out-of-sample performance results for BSX**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of BSX, $\sqrt{252}\text{RV}^{(d)}_{\text{BSX},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.9, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – CVX					
	Static			**Dynamic**		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.1106	0.0588	0.0290	0.1030	0.0592	0.0288
[3] $\text{PESPR}^{(d)}$	+0.28%	**-0.00%**	**-0.14%**	+0.07%	**-0.01%**	**-0.27%**
[4] $\text{VLMN}^{(d)}$	**-1.86%**	**-0.20%**	**-0.31%**	**-0.89%**	**-0.02%**	**-0.51%**
[5] $\text{PQSPR}^{(d)}$, $\text{VLMN}^{(d)}$	**-2.02%**	**-0.33%**	**-0.26%**	**-0.28%**	**-0.02%**	**-0.53%**
[7] $\text{PESPR}^{(d)}$, $\text{VLMN}^{(d)}$	**-1.13%**	**-0.20%**	**-0.40%**	+0.62%	**-0.02%**	**-0.71%**

Table 4.20: **Out-of-sample performance results for CVX**
Root mean square prediction errors (RMSE) for the annualized daily realized volatility of CVX, $\sqrt{252}\text{RV}^{(d)}_{\text{CVX},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.11, followed by the list of their additional terms. Their RMSE is expressed in percentage (%) change with respect to the benchmark's RMSE, with improvements in bold. Refer to section 4.2 and equation (4.12) for details on the period samples and prediction methodology (static vs. dynamic).

4.3 Further Analysis of Selected Models

Throughout our out-of-sample analysis, we have identified for each firm a seemingly better performing HAR-RV-LIQ model for further analysis of estimated coefficients and residuals distribution. We summarize in table 4.21 which predictors of daily log-realized volatility were selected for each firm, in addition to the three HAR-RV predictors, i.e. the log of the daily, weekly and monthly lagged 5-minutes subsampling estimators of realized volatility.

Ticker	**Additional Predictors**
CMS	$\mathrm{PQSPR}^{(\mathrm{d})}$
X	$\mathrm{PESPR}^{(\mathrm{d})}$
NUE	$\mathrm{PQSPR}^{(\mathrm{d})}$
LLTC	$\mathrm{PQSPR}^{(\mathrm{d})}$, $\mathrm{VLMN}^{(\mathrm{d})}$
MON	$\mathrm{PESPR}^{(\mathrm{d})}$
NKE	$\mathrm{PESPR}^{(\mathrm{d})}$
BSX	$\mathrm{PESPR}^{(\mathrm{d})}$
CVX	$\mathrm{VLMN}^{(\mathrm{d})}$

Table 4.21: **Selected predictors for each stock**
Predictors included within the linear prediction models of log-realized volatility, in addition to the 1 day, 5 days and 21 days lagged log-realized volatility estimators. The only predictors with sufficient in and out-of-sample performance to be retained for further analysis for at least one firm are the daily proportional (PQSPR) or effective (PESPR) spreads and the daily volume in number of shares (VLMN). All models are thus nested within:
$\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})}) = \alpha_i + \beta_{i,\mathrm{RV}}^{(\mathrm{d})} \log(\mathrm{RV}_{i,t}^{(\mathrm{d})}) + \beta_{i,\mathrm{RV}}^{(\mathrm{w})} \log(\mathrm{RV}_{i,t}^{(\mathrm{w})}) + \beta_{i,\mathrm{RV}}^{(\mathrm{m})} \log(\mathrm{RV}_{i,t}^{(\mathrm{m})}) + \beta_{i,\mathrm{PQSPR}}^{(\mathrm{d})} \mathrm{PQSPR}_{i,t}^{(\mathrm{d})} + \beta_{i,\mathrm{PESPR}}^{(\mathrm{d})} \mathrm{PESPR}_{i,t}^{(\mathrm{d})} + \beta_{i,\mathrm{VLMN}}^{(\mathrm{d})} \mathrm{VLMN}_{i,t}^{(\mathrm{d})} + \varepsilon_{i,t}$.

For each of these models, we look at the out-of-sample errors and coefficients from our dynamic estimation scheme over the 2008-2013 period. We recall that for the dynamic scheme, the model is re-estimated every 25 days using the most recent 750 daily observations.

We plot in figure 4.1 the evolution of the t-statistics of autoregressive RV terms for the benchmark and for the selected model. The coefficient of daily log-RV lag is always statistically significant at the 95% level for the entire sample period and for all 8 stocks.

The weekly lag is generally significant for all firms and the monthly lag is generally significant or close to significant for most firms. The selected models do not differ much in terms of the statistical significance of the coefficients of autoregressive log-RV terms. For most stocks, there are periods where the inclusion of an additional liquidity term is sufficient to bring the weekly or monthly log-RV lag below its critical value (e.g. CMS from the middle of 2010 to the of 2013). The severe bucket-shaped drop in the level of weekly log-RV t-statistic for LLTC corresponds to the period where the May 2010 outlier is included within the 750 days rolling window sample.

In figure 4.2, we plot the evolution of the t-statistic of the coefficients of additional terms of selected models with respect to the benchmark. For firms with at least one period of time where the t-statistic of the spread or volume term is above its 95% critical value, we see that this period is not often shorter than the width of the estimation rolling window. Thus, the significance of the t-statistic may be in large part due to a better fit against a few very large outliers close-by in time.

To assess how the inclusion of liquidity terms affects prediction errors through time, we plot in figure 4.3 the partial sum of squared prediction errors (partial SSE)[3] of daily log-RV for the selected models and the benchmark, as well as their difference. When the latter is negatively sloped, it implies that the selected model is performing worse than the benchmark in terms of prediction errors. The partial SSE at the end of the period is also equal to the usual prediction SSE. Thus, a partial SSE difference curve ending below the zero red dotted line at the end of 2013 indicate that the selected model's cumulative performance was superior to the benchmark's. We also plot the squared prediction errors of log-RV themselves in figure 4.5. It can be seen that jumps in the selected model's partial SSE corresponds to an extremely large squared error and increases in slope corresponds to a cluster of above average squared errors. For CMS, LLTC, NKE, BSX and CVX, the selected model is overall performing better out-of-sample, with a negative SSE difference at the end of the period. We notably observe that in periods where the t-statistics of the spread or volume is large often coincides with periods of favorable jumps in rel-

[3]Given a time series of prediction errors $\{e_t\}_{t=1}^{T}$, partial $\mathrm{SSE}_t = \sum_{j=1}^{t} e_j^2$ for $t = 1, \ldots, T$.

ative performance for selected models, e.g. end of 2008 for BSX and CVX or the May 2010 flash Crash for LLTC. In fact, only for X is there a steady degradation of predictive performance through time during a period where the t-statistic of the spread coefficient is significant. This is evidence that our model in its current form does help to improve performance, however only at small marginal levels. However, the performance for X is also evidence that spread measures can also decrease performance for specific firms. The addition of industry or firm-specific control variates could potentially resolve degradation in performance for specific firms. Nonetheless, the cumulative difference in performance is overall small, with both partial SSE curves mostly indistinguishable. This is reflective of the very minor differences for in-sample R-squared and out-of-sample RMSE between most studied linear models.

We also include for each stock in figure 4.4 the prediction error plots of log-RV for selected models and in figure 4.6 the plot of the difference between the absolute errors of selected models and the absolute errors of the benchmark. The latter shows as expected clusters of high absolute error difference volatility synchronously to when the t-statistics of additional spread or volume terms are significant (figure 4.2). Those clusters are however fairly symmetric around the zero axis, which is consistent with the previous observation that for a given firm the partial SSE curves of the selected models and benchmarks are near indistinguishable from each others.

To further comment on the general fit of the model, we test for conditional heteroscedasticity of residuals using the Engel ARCH test, for normality of residuals using the Kolmogorov-Smirnov test and for autocorrelation of residuals using the Durbin-Watson test[4]. The statistics of these tests are summarized in table 4.22. The Angel ARCH test only rejects the null hypothesis of conditional homoscedasticity in residuals for NUE, LLTC, NKE and BSX stocks, although other stocks may also display what appears to be occasional clusters of slightly higher volatility in their squared error plots (e.g CVX in 2008). The Durbin-Watson test for autocorrelation in residuals is generally inconclu-

[4]Refer to Durbin and Watson (1950) for the Durbin-Watson test of residual autocorrelation, to Engle (1982) for the Engle ARCH test and to Massey (1951) for the Kolmogorov-Smirnov test of normality.

sive, because the value of the test statistics is very close to 2.0 for all stocks. Finally, the Kolmogorov-Smirnov test rejects the null hypothesis of normally distribution residuals for all stocks. Also in table 4.22, we present the skewness and kurtosis of residuals as well as the proportion of positive and negative errors. In agreement with the Kolmogorov-Smirnov test, the residuals exhibit positive skewness and excess kurtosis for all stocks. For all stocks, the predictive model has a tendency to more often overestimate the next day's log-RV, as show with the proportion of positive errors between 44 and 48% for all stocks. This observation combined with the positive skewness, which is indicative of a higher likelihood of large underestimations than large overestimations, and the autoregressive nature of the model specification, we can easily deduct that the model tends to occasionally undershoot extremely high but short bursts of volatility in the market, which is then turned into consecutive overshooting due to the autoregressive specification. We showcase an example of such behaviour in figure 4.7. During the Flash Crash of May 2010, the model obviously failed to predict the large spike in volatility and then proceeded to overshoot the next few days volatility.

	Engel ARCH Test		Kolmogorov-Smirnov Test		Durbin-Watson Test	Central Moments		Proportion	
Ticker	**p-Value**	**Statistic**	**p-Value**	**Statistic**	**Statistic**	**Skewness**	**Kurtosis**	**Positive**	**Negative**
CMS	0.126	2.347	0.000	0.327	2.047	0.63	5.26	0.47	0.53
X	0.202	1.628	0.000	0.323	1.922	0.41	3.83	0.46	0.54
NUE	0.009	6.918	0.000	0.321	1.987	0.37	3.79	0.46	0.54
LLTC	0.000	292.2	0.000	0.327	1.994	0.48	11.37	0.48	0.52
MON	0.325	0.970	0.000	0.313	1.968	0.66	4.48	0.44	0.56
NKE	0.002	9.646	0.000	0.322	1.991	0.35	4.87	0.48	0.52
BSX	0.045	4.027	0.000	0.297	2.006	0.66	6.09	0.46	0.54
CVX	0.780	0.078	0.000	0.318	2.030	0.48	4.99	0.46	0.54

Table 4.22: **Residuals diagnostic**
The table first presents various tests on the residuals of the log-RV regression for each stock's selected model. Each test uses a sample size of 1474 daily prediction errors from the dynamic estimation method. The 95% confidence level critical values of the Engle ARCH Test for conditional heteroscedasticity and of the Kolmogorov-Smirnov test for the normality of residuals are 3.841 and 0.035 respectively. For the Durbin-Watson test for autocorrelation, which test statistic is bounded between 0 and 4, a value not close to 2 generally indicates evidence of autocorrelation in residuals. Following those test results, we present the sample skewness and kurtosis of residuals as well as the proportion of positive and negative errors.

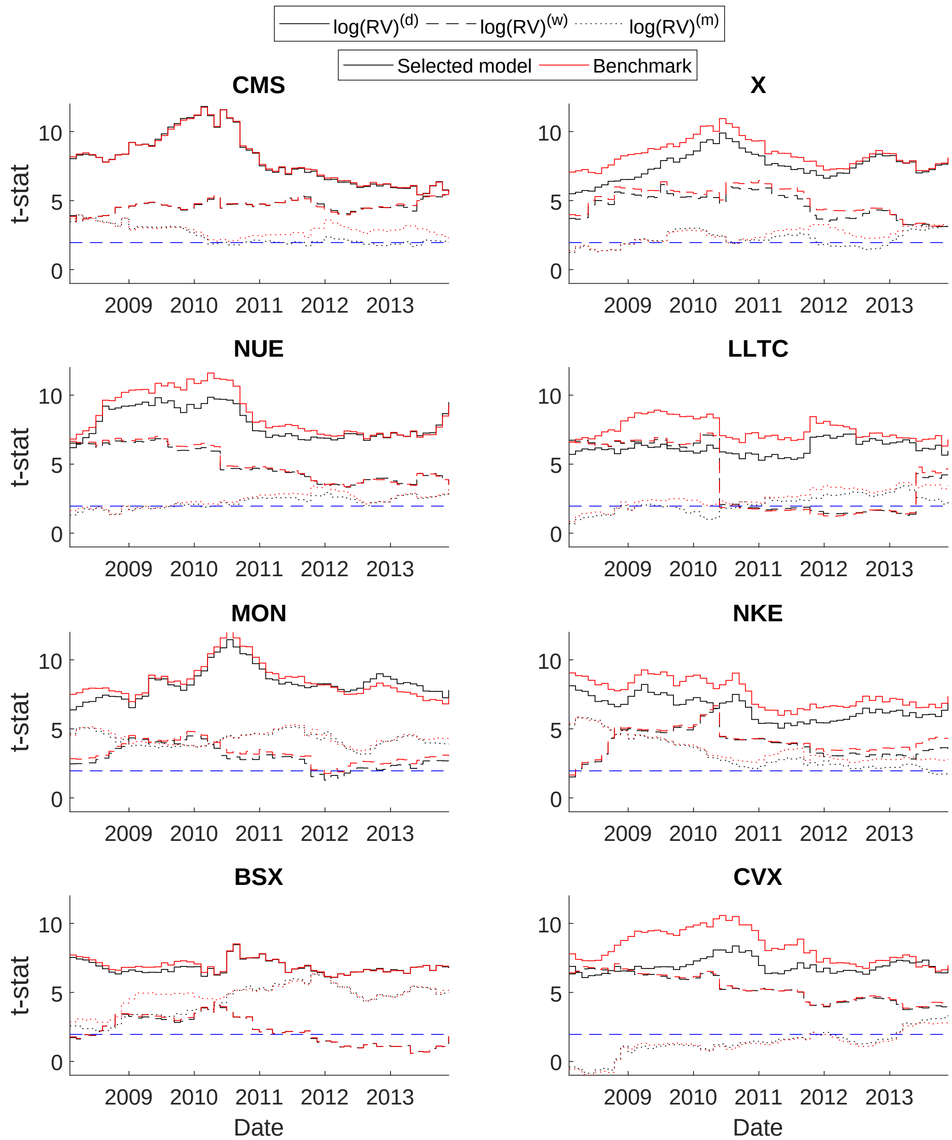

Figure 4.1: **Plot of t-statistics of RV terms**
Plot of the t-statistics from dynamic estimation of lagged log-RV coefficients for selected models and benchmark model and for each stock. Dynamic estimation is done by re-estimating the model every 25 days with the 750 observations. The blue dotted line is at the level of the 95% asymptotic critical value (1.96).

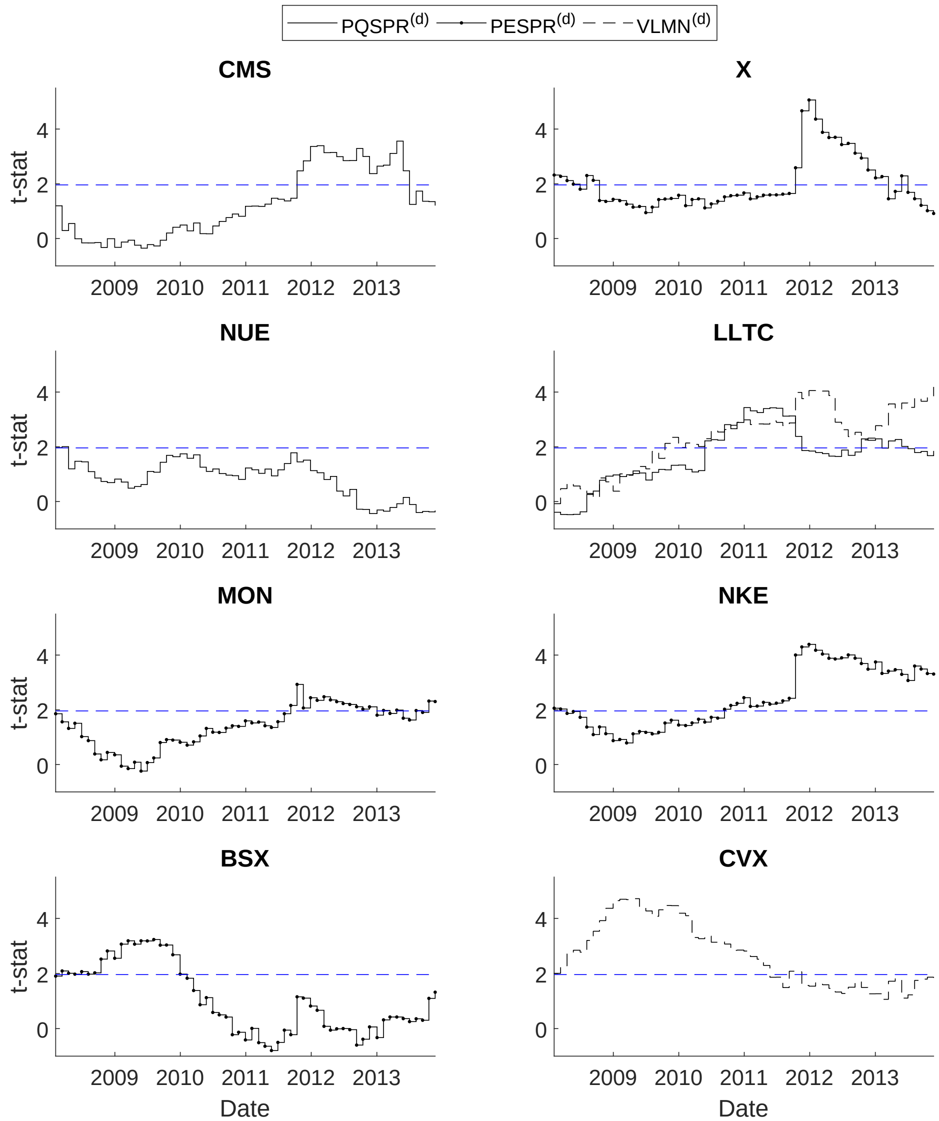

Figure 4.2: **Plot of t-statistics of liquidity terms**
Plot of the t-statistics from dynamic estimation of spread or volume coefficients for selected models and for each stock. Dynamic estimation is done by re-estimating the model every 25 days with the 750 observations. The blue dotted line is at the level of the 95% asymptotic critical value (1.96).

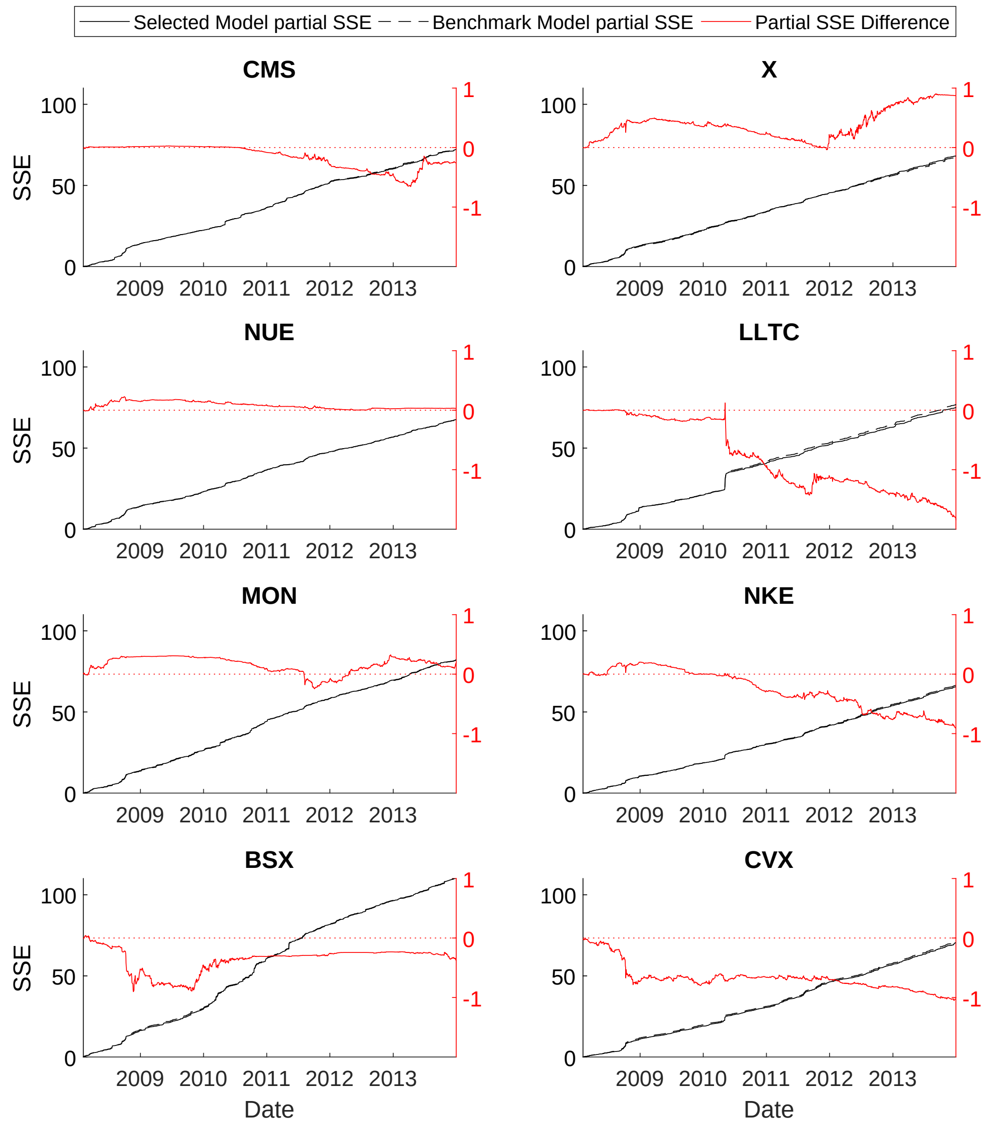

Figure 4.3: **Partial sum of squared errors**
Plot of partial sum of squared prediction errors (partiel SSE) of the daily log-RV from dynamic estimation. Given a time series of prediction errors $\{e_t\}_{t=1}^{T}$, partial $\text{SSE}_t = \sum_{j=1}^{t} e_j^2$ for $t = 1, \ldots, T$. Dynamic estimation is done by re-estimating the model every 25 days with the 750 observations. The red curve (right axis) is the difference between the selected model's partial SSE (full line) and the benchmark's partial SSE (dotted line). A red curve at the end of 2013 below the red dotted line indicates a better cumulative performance from the selected model over the benchmark.

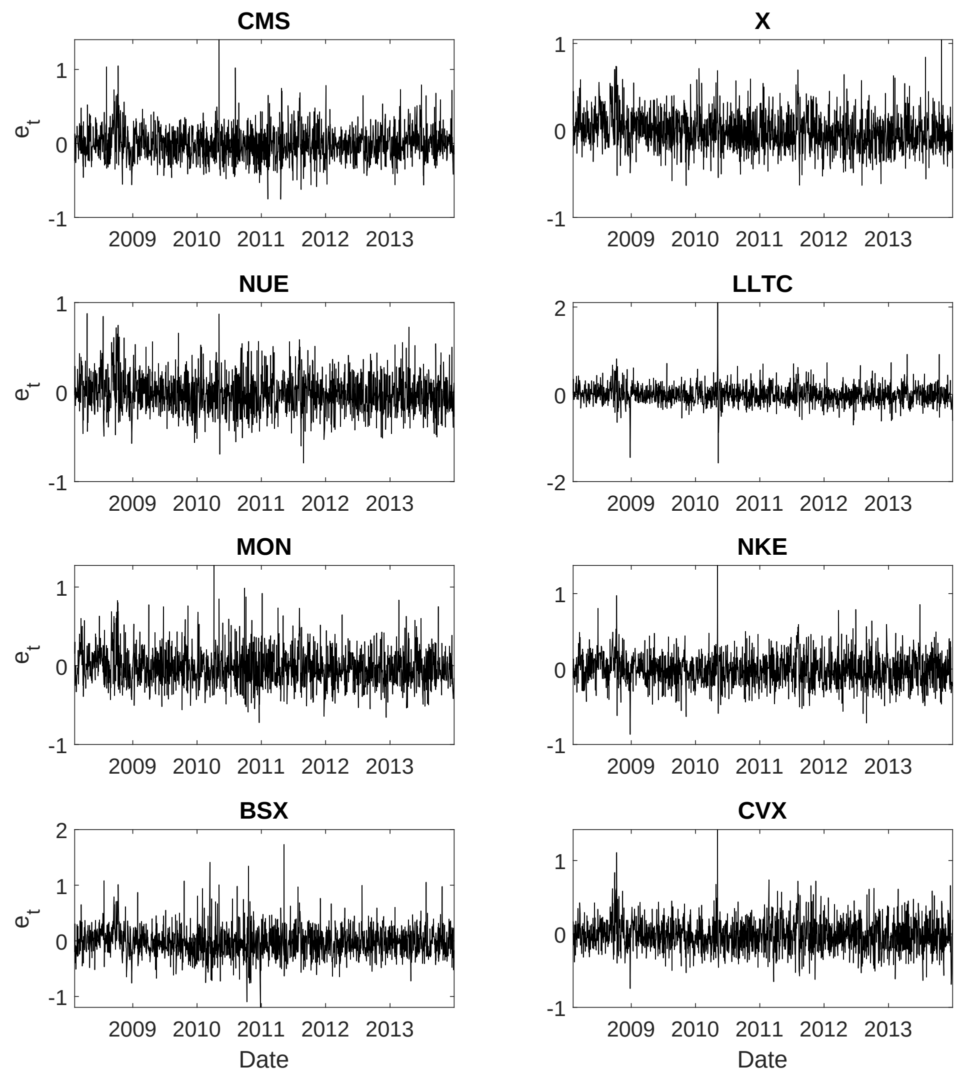

Figure 4.4: **Prediction errors plot**
Plot of log-RV prediction errors from dynamic estimation of selected models for each stock. Dynamic estimation is done by re-estimating the model every 25 days with the 750 observations.

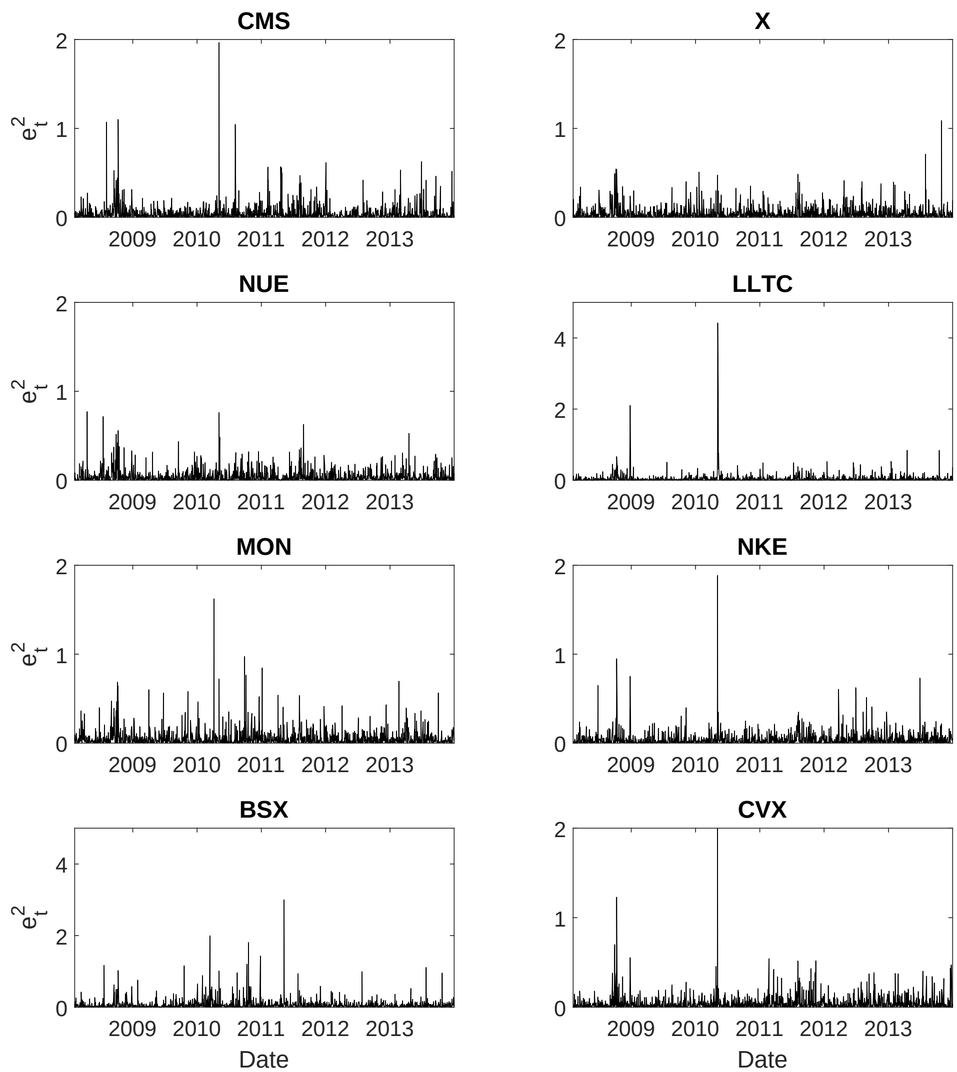

Figure 4.5: **Prediction squared errors plot**
Plot of log-RV prediction squared errors from dynamic estimation of selected models for each stock. Dynamic estimation is done by re-estimating the model every 25 days with the 750 observations.

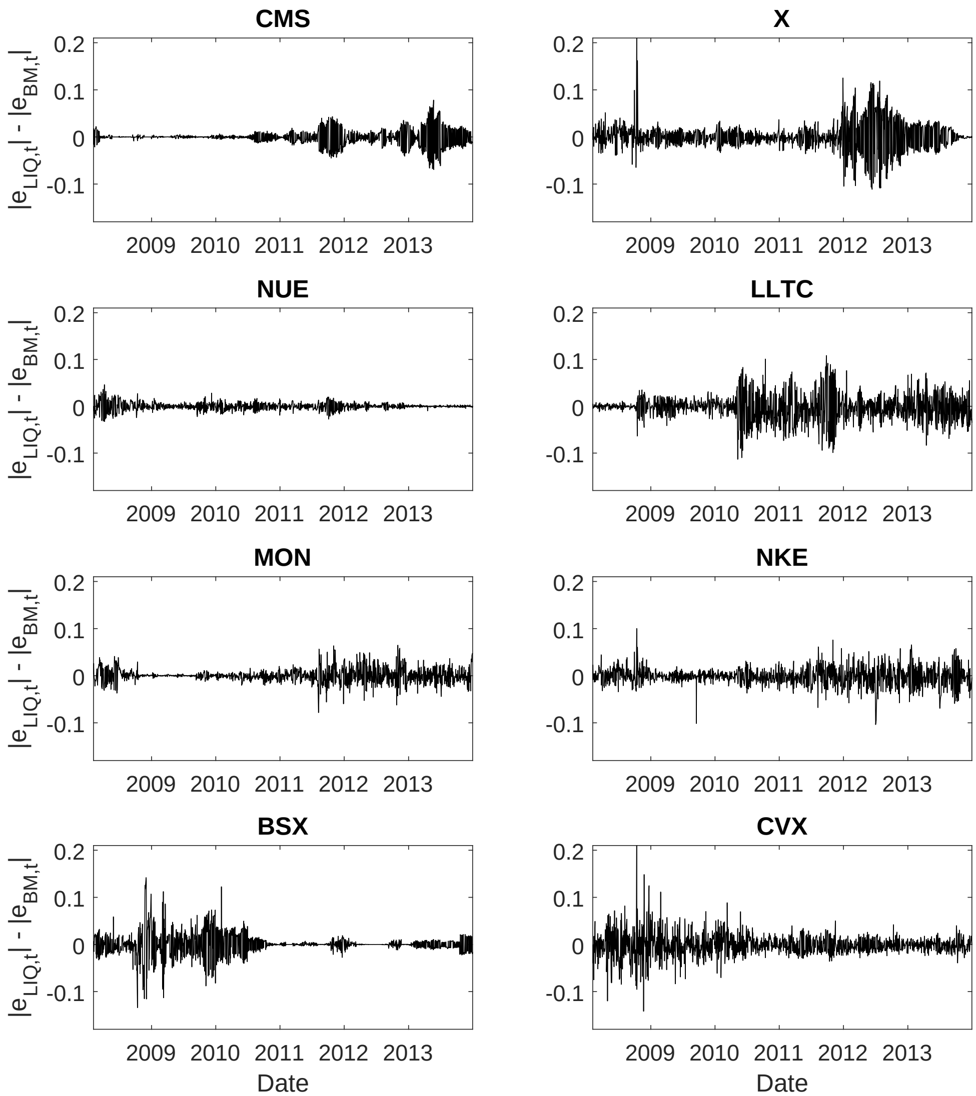

Figure 4.6: **Prediction absolute errors in excess benchmark's**
Plot of log-RV prediction absolute errors from dynamic estimation of selected models minus that of benchmark, for each stock. Dynamic estimation is done by re-estimating the model every 25 days with the 750 observations.

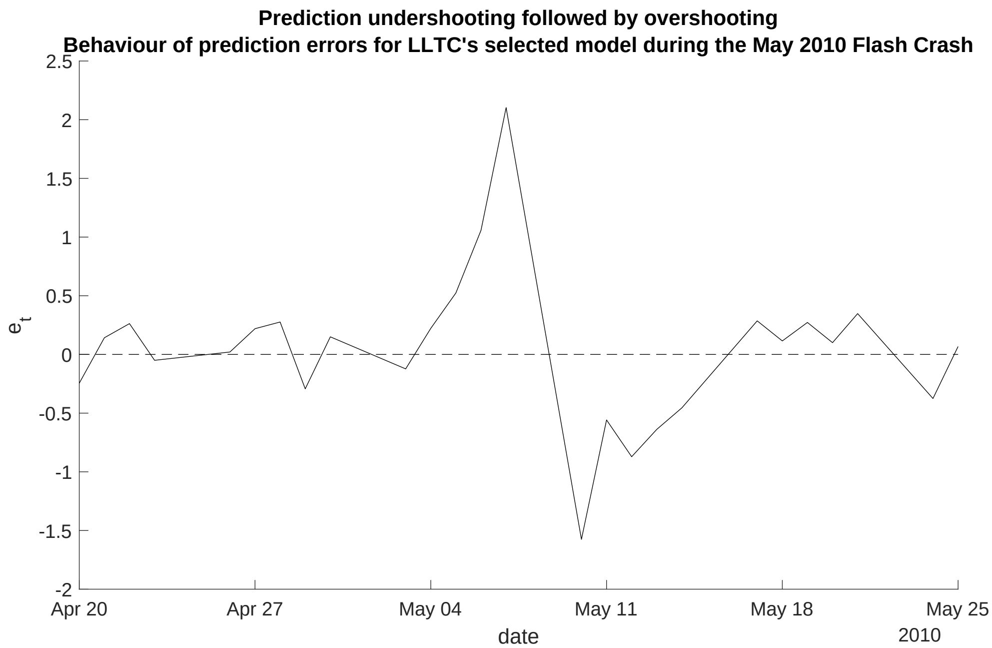

Figure 4.7: **Showcase of undershooting-overshooting**
The LLTC stocks during the May 2010 Flash Crash is an example of model behaviour described in section 4.3. The model residuals exhibit positive skewness yet a slightly higher proportion of negative errors. It can be interpreted as the model occasionally undershoots tail events of extremely high log-RV and then proceeding to overshoot predictions of log-RV for several following days due to the autoregressive specification.

Conclusion

We set out with the objective to verify whether information present in the daily volume of transactions and the daily spread between ask and bid prices for a given stock could improve predictions of the next day's realized volatility, in margin to whatever predictive capacity is already carried by previous values of realized volatility. In other words, does the timing of the relationship between measures liquidity and realized volatility goes beyond concurrent? If it does indeed, properly understanding it and then estimating a model that approximates well their relationship could help us better forecast stock volatility. As a start to model and measure this relationship, we have used the popular autoregressive HAR-RV model (Corsi, 2009) as our benchmark, to which we have added various measures of volume or spread to form our HAR-RV-LIQ model. We have also found the spread or volume of a given stock to be in general significantly positively correlated with its daily log-realized volatility at various lags.

For the 10 stocks, we have measured the predictive performance of several variants of HAR-RV-LIQ models and have found for half the select stocks at least one such model yielding improvements, although fairly small. These minute reductions in prediction RMSE are evidence that exist some quantity of information within past days' volume and spread not already carried by past days' realized volatility that can be used to infer on the next day's realized volatility. In particular, we have found the measures of proportional spread to more consistently yield improved or at the very least non-degraded performance, compared to volume in number of shares only occasionally improving predictive performance and volume in dollars performing the worse as an additional predictor. Most recent values of spread or volume in level, rather than weekly or monthly moving average of several recent values or rather than their differenced values, have showed the best performance. Our research could also be expanded to include other measures of liquidity and activity, such as the Amihud measure (Amihud, 2002), which is a transformation of daily absolute returns and dollar volume, or price impact (see Dufour and Engle (2000) for an example of measure of the impact on price of a large trade).

We have started with Corsi's simple yet well performing linear autoregressive specification, but it begs the question as to whether the proper relationship between realized volatility and the selected measures of activity and liquidity is best approximated by a linear form. Furthermore, using OLS with Newey-West error covariance prevents us from having any sort of proper theoretical prediction interval bounds. Our predictive model as presented in section 3.3.2 did not include any industry or market factors (e.g. the VIX), nor leverage effects that can capture the asymetric response of volatility to negative and positive returns. Some other authors have extended the HAR-RV model to include such market, jumps or leverage components (see for example Corsi and Reno (2009), Wang et al. (2012), Huang et al. (2013) and Patton and Sheppard (2015)). It could be worthwhile to consider in future research the inclusion of liquidity and activity factors alongside leverage, market and industry factors. Additional significant factors could provide a more complete specification and help correct biased estimation of spread or volume coefficients.

Bibliography

Amaya, Diego, Jean-François Bégin, and Geneviève Gauthier (2017). «Extracting latent states from high frequency option prices», *Available at SSRN 2975355*.

Amihud, Yakov (2002). «Illiquidity and stock returns: cross-section and time-series effects», *Journal of Financial Markets*, vol. 5, no. 1, pp. 31–56.

Andersen, Torben G., Tim Bollerslev, and Francis X. Diebold (2007). «Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility», *The Review of Economics and Statistics*, vol. 89, no. 4, pp. 701–720.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Heiko Ebens (2001). «The distribution of realized stock return volatility», *Journal of Financial Economics*, vol. 61, no. 1, pp. 43–76.

Andersen, Torben G., Tim Bollerslev, Francis X. Diebold, and Paul Labys (2003). «Modeling and forecasting realized volatility», *Econometrica*, vol. 71, no. 2, pp. 579–625.

Aït-Sahalia, Yacine, Per A. Mykland, and Lan Zhang (2005). «How often to sample a continuous-time process in the presence of market microstructure noise», *The Review of Financial Studies*, vol. 18, no. 2, pp. 351–416.

Baillie, Richard T. (1996). «Long memory processes and fractional integration in econometrics», *Journal of Econometrics*, vol. 73, no. 1, pp. 5–59.

Baillie, Richard T., Tim Bollerslev, and Hans O. Mikkelsen (1996). «Fractionally integrated generalized autoregressive conditional heteroskedasticity», *Journal of Econometrics*, vol. 74, no. 1, pp. 3–30.

Barndorff-Nielsen, Ole E., Peter R. Hansen, Asger Lunde, and Neil Shephard (2011). «Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading», *Journal of Econometrics*, vol. 162, no. 2, pp. 149–169.

Barndorff-Nielsen, Ole E. and Neil Shephard (2002). «Econometric analysis of realized volatility and its use in estimating stochastic volatility models», *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 64, no. 2, pp. 253–280.

Bee, Marco, Debbie J. Dupuis, and Luca Trapin (2016). «Realizing the extremes: Estimation of tail-risk measures from a high-frequency perspective», *Journal of Empirical Finance*, vol. 36, pp. 86–99.

Belsley, David A., Edwin Kuh, and Roy E. Welsch (2004). *Regression diagnostics: identifying influential data and sources of collinearity*, Wiley, Hoboken, N.J.

Black, Fischer (1986). «Noise», *The Journal of Finance*, vol. 41, no. 3, pp. 529–543.

Campbell, John Y., Sanford J. Grossman, and Jiang Wang (1993). «Trading volume and serial correlation in stock returns», *The Quarterly Journal of Economics*, vol. 108, no. 4, pp. 905–939.

Chordia, Tarun, Richard Roll, and Avanidhar Subrahmanyam (2000). «Commonality in liquidity», *Journal of Financial Economics*, vol. 56, no. 1, pp. 3–28.

Copeland, Thomas E. and Dan Galai (1983). «Information effects on the bid-ask spread», *The Journal of Finance*, vol. 38, no. 5, pp. 1457–1469.

Corsi, Fulvio (2009). «A simple approximate long-memory model of realized volatility», *Journal of Financial Econometrics*, vol. 7, no. 2, pp. 174–196.

Corsi, Fulvio and Roberto Reno (2009). «HAR volatility modelling with heterogeneous leverage and jumps», *Available at SSRN 1316953*.

Ding, Zhuanxin, Clive W. J. Granger, and Robert F. Engle (1993). «A long memory property of stock market returns and a new model», *Journal of empirical finance*, vol. 1, no. 1, pp. 83–106.

Dufour, Alfonso and Robert F. Engle (2000). «Time and the price impact of a trade», *The Journal of Finance*, vol. 55, no. 6, pp. 2467–2498.

Durbin, J. and G. S. Watson (1950). «Testing for serial correlation in least squares regression: I», *Biometrika*, vol. 37, no. 3/4, pp. 409–428.

Engle, Robert F. (1982). «Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation», *Econometric*, vol. 50, no. 4, pp. 987–1007.

Eraker, Bjørn, Michael Johannes, and Nicholas Polson (2003). «The impact of jumps in volatility and returns», *The Journal of Finance*, vol. 58, no. 3, pp. 1269–1300.

Granger, C. W. J. (1980). «Long memory relationships and the aggregation of dynamic models», *Journal of Econometrics*, vol. 14, no. 2, pp. 227–238.

Granger, C. W. J. and Zhuanxin Ding (1995). «Some properties of absolute return: An alternative measure of risk», *Annales d'Économie et de Statistique*, , no. 40, pp. 67–91.

Hansen, Peter R. and Asger Lunde (2006). «Realized variance and market microstructure noise», *Journal of Business & Economic Statistics*, vol. 24, no. 2, pp. 127–161.

Huang, Chuangxia, Xu Gong, Xiaohong Chen, and Fenghua Wen (2013). «Measuring and forecasting volatility in chinese stock market using HAR-CJ-M model», *Abstract and Applied Analysis*, vol. 2013, pp. 1–13.

Jacod, Jean, Yingying Li, Per A. Mykland, Mark Podolskij, and Mathias Vetter (2009). «Microstructure noise in the continuous case: The pre-averaging approach», *Stochastic Processes and their Applications*, vol. 119, no. 7, pp. 2249–2276.

Karpoff, Jonathan M. (1987). «The relation between price changes and trading volume: A survey», *Journal of Financial and Quantitative Analysis*, vol. 22, no. 1, pp. 109–126.

Lee, Suzanne S. and Per A. Mykland (2008). «Jumps in financial markets: A new non-parametric test and jump dynamics», *The Review of Financial Studies*, vol. 21, no. 6, pp. 2535–2563.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard (2015). «Does anything beat 5-minute RV? a comparison of realized measures across multiple asset classes», *Journal of Econometrics*, vol. 187, no. 1, pp. 293–311.

Massey, F. J. (1951). «The kolmogorov-smirnov test for goodness of fit», *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78.

Newey, Whitney K. and Kenneth D. West (1994). «Automatic lag selection in covariance matrix estimation», *The Review of Economic Studies*, vol. 61, no. 4, pp. 631–653.

Patton, Andrew J. and Kevin Sheppard (2015). «Good volatility, bad volatility: Signed jumps and the persistence of volatility», *Review of Economics and Statistics*, vol. 97, no. 3, pp. 683–697.

Scaggs, Alexandra (2012). «Floor traders praise NYSE officials' handling of trading glitch», *The Wall Street Journal* [ONLINE].
URL: *https://blogs.wsj.com/marketbeat/2012/08/01/floor-traders-praise-nyse-officials-handling-of-trading-glitch/*

Schwert, G. W. (1989). «Why does stock market volatility change over time?», *The Journal of Finance*, vol. 44, no. 5, pp. 1115–1153.

Tsay, Ruey S. (2005). *Analysis of financial time series*, J. Wiley & Sons, Hoboken, N.J, 2nd ed.

Wang, Tianyi, Zhuo Huang, et al. (2012). «The relationship between volatility and trading volume in the chinese stock market: a volatility decomposition perspective», *Annals of economics and finance*, vol. 13, no. 1, pp. 211–236.

Zhang, Lan, Per A. Mykland, and Yacine Aït-Sahalia (2005). «A tale of two time scales: Determining integrated volatility with noisy high-frequency data», *Journal of the American Statistical Association*, vol. 100, no. 472, pp. 1394–1411.

Appendix A – Additional tables and figures

A.1 Descriptive statistics and correlations

	Moments				Quantiles								
Ticker	**Mean**	**Std.**	**Skew.**	**Kurt.**	**1%**	**5%**	**10%**	**25%**	**Med.**	**75%**	**90%**	**95%**	**99%**
CMS	-4.47	0.45	0.82	4.45	-5.32	-5.10	-4.99	-4.78	-4.50	-4.22	-3.92	-3.66	-3.01
X	-3.73	0.40	1.12	5.37	-4.44	-4.26	-4.16	-4.00	-3.79	-3.54	-3.19	-2.95	-2.48
NUE	-4.01	0.47	0.70	3.92	-4.89	-4.68	-4.57	-4.36	-4.05	-3.73	-3.42	-3.17	-2.63
WMB	-4.11	0.47	0.93	4.43	-4.96	-4.76	-4.64	-4.44	-4.17	-3.88	-3.52	-3.24	-2.66
LLTC	-4.29	0.40	0.64	4.80	-5.07	-4.89	-4.77	-4.56	-4.31	-4.06	-3.79	-3.60	-3.20
MON	-4.18	0.47	0.78	3.92	-5.05	-4.84	-4.72	-4.50	-4.24	-3.91	-3.54	-3.31	-2.75
NKE	-4.43	0.43	0.93	3.89	-5.17	-4.99	-4.90	-4.74	-4.51	-4.21	-3.82	-3.57	-3.16
BSX	-4.12	0.41	0.75	4.43	-4.94	-4.71	-4.59	-4.40	-4.16	-3.91	-3.59	-3.33	-2.85
HPQ	-4.31	0.40	0.83	4.93	-5.14	-4.89	-4.78	-4.58	-4.35	-4.12	-3.79	-3.57	-3.06
CVX	-4.46	0.44	0.83	4.67	-5.31	-5.10	-4.99	-4.76	-4.50	-4.23	-3.91	-3.68	-3.06

Table A.1.1: **Summary statistics of log(RV), 2005-2013**
Summary statistics of the daily 5 minutes subsampling realized volatility estimator (equation (3.9) with $K = 5$) for the beginning of 2005 to the end of 2013 period and for each of the selected stocks, after winsorization and interpolation.

	Moments				Quantiles								
Ticker	**Mean**	**Std.**	**Skew.**	**Kurt.**	**1%**	**5%**	**10%**	**25%**	**Med.**	**75%**	**90%**	**95%**	**99%**
CMS	6.62	2.09	0.44	2.71	3.53	3.70	3.94	4.78	6.47	8.21	9.33	10.16	11.76
X	4.03	1.31	0.67	4.29	1.81	2.15	2.32	3.02	3.99	4.84	5.59	6.18	7.90
NUE	3.48	1.06	1.56	6.79	2.24	2.36	2.44	2.69	3.15	4.08	4.84	5.44	7.13
WMB	4.61	1.64	0.98	3.82	2.66	2.77	2.88	3.22	4.19	5.72	7.00	7.69	9.21
LLTC	3.04	0.64	1.04	4.79	1.90	2.10	2.27	2.65	2.97	3.28	3.76	4.53	5.01
MON	3.11	1.00	0.80	3.24	1.63	1.85	2.01	2.32	2.88	3.80	4.53	4.97	5.84
NKE	2.74	0.98	1.34	4.95	1.48	1.64	1.74	2.01	2.49	3.21	4.05	4.88	5.84
BSX	10.55	4.39	0.32	1.75	4.45	4.91	5.30	6.37	9.74	14.00	17.12	17.81	18.98
HPQ	3.24	1.27	1.05	3.59	1.72	1.83	1.99	2.18	2.89	3.93	5.12	5.73	6.93
CVX	1.71	0.57	1.11	3.62	1.02	1.10	1.15	1.27	1.54	2.03	2.60	2.88	3.31
Scale	$\times 10^4$		$\times 1$		$\times 10^4$								

Table A.1.2: **Summary statistics of PQSPR, 2005-2013**

Summary statistics of the daily proportional quoted spread (equations (3.2) and (3.4)) for the beginning of 2005 to the end of 2013 period and for each of the selected stocks, after winsorization and interpolation.

	Moments				Quantiles								
Ticker	**Mean**	**Std.**	**Skew.**	**Kurt.**	**1%**	**5%**	**10%**	**25%**	**Med.**	**75%**	**90%**	**95%**	**99%**
CMS	6.01	1.94	0.26	2.75	2.94	3.18	3.37	4.27	6.13	7.41	8.48	9.12	10.27
X	3.83	1.29	2.07	17.90	1.86	2.11	2.27	2.92	3.75	4.52	5.15	5.84	7.71
NUE	3.17	1.14	3.51	30.13	1.92	2.09	2.19	2.45	2.86	3.63	4.39	5.07	7.35
WMB	4.35	1.71	2.08	18.01	2.25	2.42	2.61	3.04	3.91	5.40	6.58	7.44	9.38
LLTC	3.18	2.01	38.12	1663.58	2.20	2.40	2.53	2.75	2.99	3.32	4.00	4.57	5.31
MON	2.71	1.04	1.71	10.08	1.24	1.45	1.62	1.97	2.48	3.31	3.98	4.46	5.93
NKE	2.34	0.85	1.89	9.32	1.26	1.43	1.50	1.73	2.17	2.68	3.38	4.02	5.24
BSX	8.84	3.01	0.15	1.88	3.85	4.43	4.90	6.19	8.53	11.37	12.87	13.55	15.04
HPQ	3.14	1.08	1.46	8.78	1.72	1.87	2.03	2.25	2.97	3.69	4.65	5.22	6.01
CVX	1.75	0.76	3.21	28.33	0.87	0.95	1.04	1.23	1.58	2.11	2.66	2.98	4.18
Scale	$\times 10^4$		$\times 1$		$\times 10^4$								

Table A.1.3: **Summary statistics of PESPR, 2005-2013**

Summary statistics of the daily proportional effective spread (equations (3.3) and (3.5)) for the beginning of 2005 to the end of 2013 period and for each of the selected stocks, after winsorization and interpolation.

	Moments				Quantiles								
Ticker	**Mean**	**Std.**	**Skew.**	**Kurt.**	**1%**	**5%**	**10%**	**25%**	**Med.**	**75%**	**90%**	**95%**	**99%**
CMS	0.67	0.36	2.03	10.22	0.21	0.27	0.32	0.43	0.58	0.80	1.10	1.34	1.98
X	2.16	1.37	1.46	6.52	0.50	0.64	0.76	1.10	1.83	2.84	3.98	4.78	6.56
NUE	1.02	0.61	1.97	9.58	0.27	0.36	0.44	0.60	0.88	1.26	1.76	2.15	3.30
WMB	1.50	0.84	2.81	19.44	0.46	0.61	0.74	0.96	1.31	1.80	2.43	2.95	4.71
LLTC	0.97	0.53	1.79	10.05	0.26	0.34	0.42	0.60	0.88	1.22	1.62	1.91	2.75
MON	1.12	0.90	3.42	25.54	0.23	0.32	0.38	0.54	0.88	1.41	2.08	2.72	4.56
NKE	0.70	0.42	3.34	24.69	0.20	0.26	0.32	0.45	0.62	0.85	1.13	1.40	2.18
BSX	3.54	2.90	6.24	83.03	0.58	0.89	1.28	2.02	2.95	4.28	6.05	7.93	13.52
HPQ	4.28	2.84	5.44	53.33	1.53	1.96	2.24	2.81	3.67	4.81	6.68	8.20	15.35
CVX	2.22	0.99	1.68	8.77	0.85	1.04	1.17	1.54	2.04	2.66	3.48	4.16	5.54
Scale	$\times 10^{-4}$		$\times 1$		$\times 10^{-4}$								

Table A.1.4: **Summary statistics of VLMN, 2005-2013**
Summary statistics of the daily trade volume in number of shares for the beginning of 2005 to the end of 2013 period and for each of the selected stocks, after winsorization and interpolation.

	Moments				Quantiles								
Ticker	**Mean**	**Std.**	**Skew.**	**Kurt.**	**1%**	**5%**	**10%**	**25%**	**Med.**	**75%**	**90%**	**95%**	**99%**
CMS	1.14	0.62	2.09	12.70	0.30	0.41	0.50	0.71	1.01	1.43	1.87	2.29	3.21
X	10.17	7.50	2.14	10.56	2.18	2.83	3.37	4.97	8.07	12.82	19.53	25.47	35.21
NUE	5.08	3.15	2.18	13.35	1.29	1.64	2.01	2.83	4.35	6.57	8.91	10.68	15.31
WMB	3.90	2.37	2.98	23.85	0.89	1.37	1.70	2.33	3.40	4.93	6.51	7.88	11.66
LLTC	3.07	1.60	2.23	15.97	0.91	1.20	1.41	1.98	2.78	3.76	5.02	5.83	8.22
MON	8.90	7.57	2.77	15.94	1.52	2.13	2.60	4.11	6.71	11.07	17.76	23.05	37.03
NKE	5.05	2.84	4.16	37.66	1.66	2.23	2.63	3.43	4.52	5.87	7.74	9.27	16.04
BSX	3.62	2.56	4.27	37.53	0.89	1.32	1.57	2.18	3.01	4.31	6.16	7.79	13.01
HPQ	14.41	9.56	5.86	82.44	3.71	5.18	6.23	8.75	12.92	17.66	23.15	27.61	47.44
CVX	18.07	7.49	1.44	6.11	6.59	9.01	10.52	13.02	16.39	21.27	28.02	33.00	43.01
Scale	$\times 10^{-5}$		$\times 1$		$\times 10^{-5}$								

Table A.1.5: **Summary statistics of VLMD, 2005-2013**
Summary statistics of the daily trade volume in US dollars for the beginning of 2005 to the end of 2013 period and for each of the selected stocks, after winsorization and interpolation.

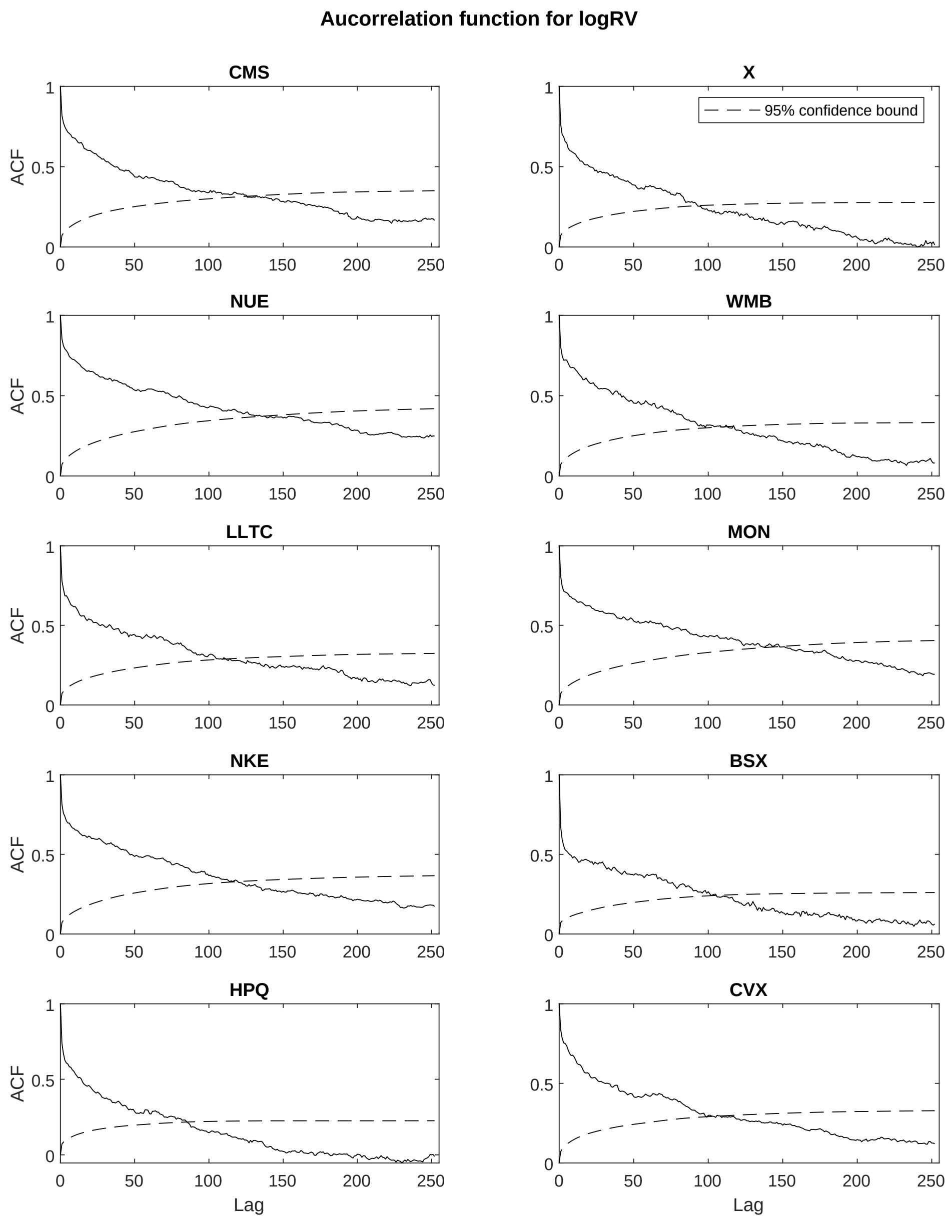

Figure A.1.1: **ACF of log-realized volatility**
Sample autocorrelation function of daily log-realized volatility for each firm with 1 year worth of lags. The 95% critical value is computed assuming asymptotic normality. The standard error estimator is presented in footnote 1.

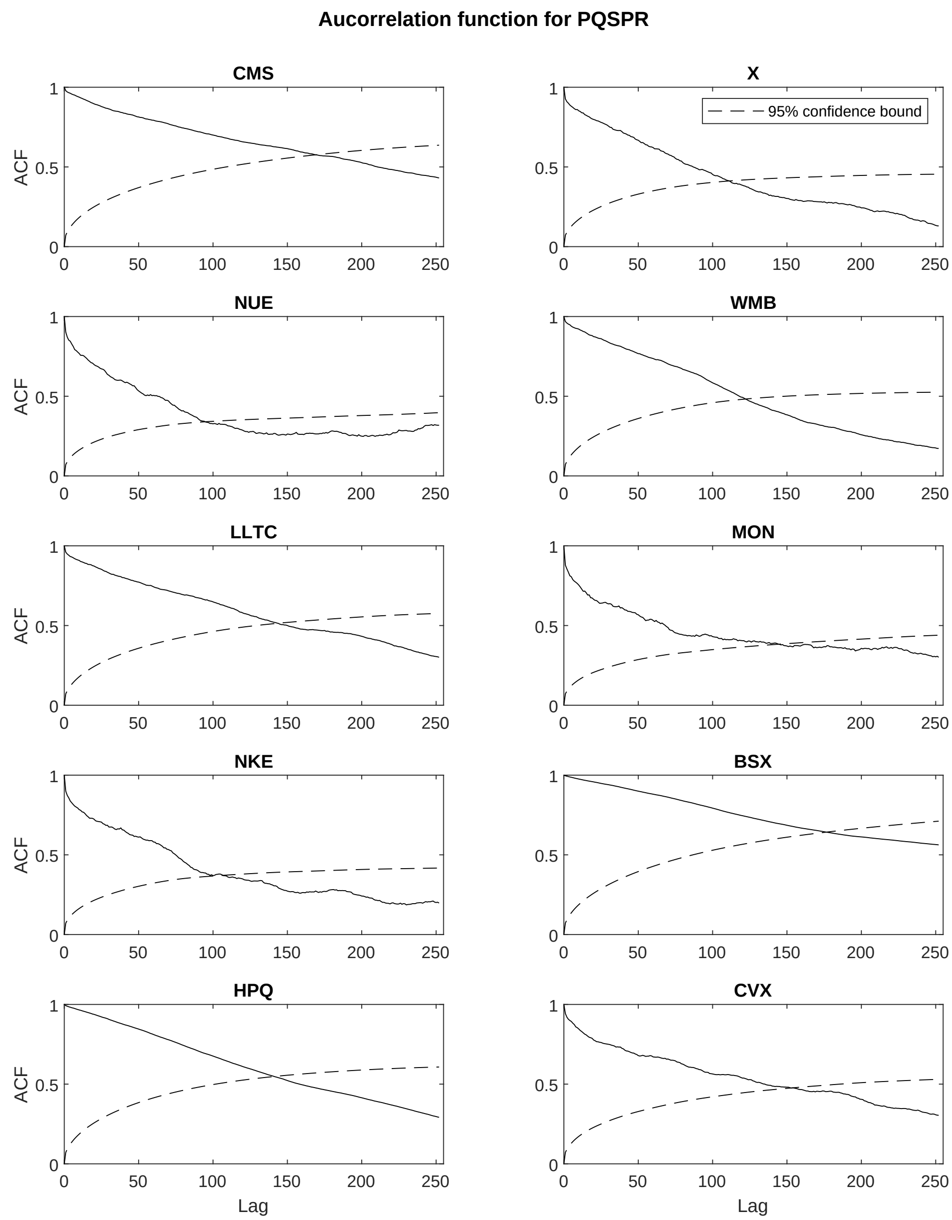

Figure A.1.2: **ACF of PQSPR**
Sample autocorrelation function of proportional quoted spread for each firm with 1 year worth of lags. The 95% critical value is computed assuming asymptotic normality. The standard error estimator is presented in footnote 1.

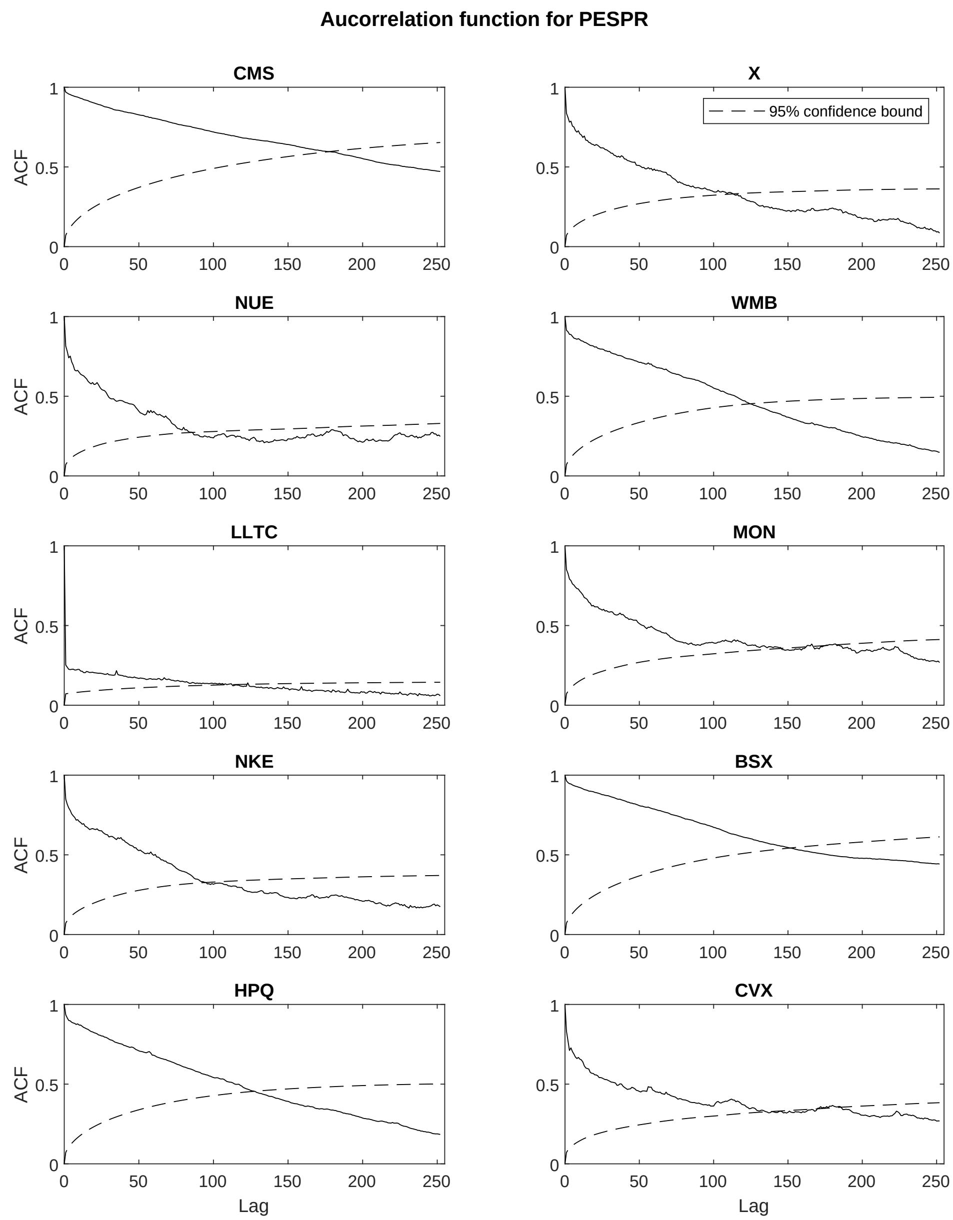

Figure A.1.3: **ACF of PESPR**
Sample autocorrelation function of proportional effective spread for each firm with 1 year worth of lags. The 95% critical value is computed assuming asymptotic normality. The standard error estimator is presented in footnote 1.

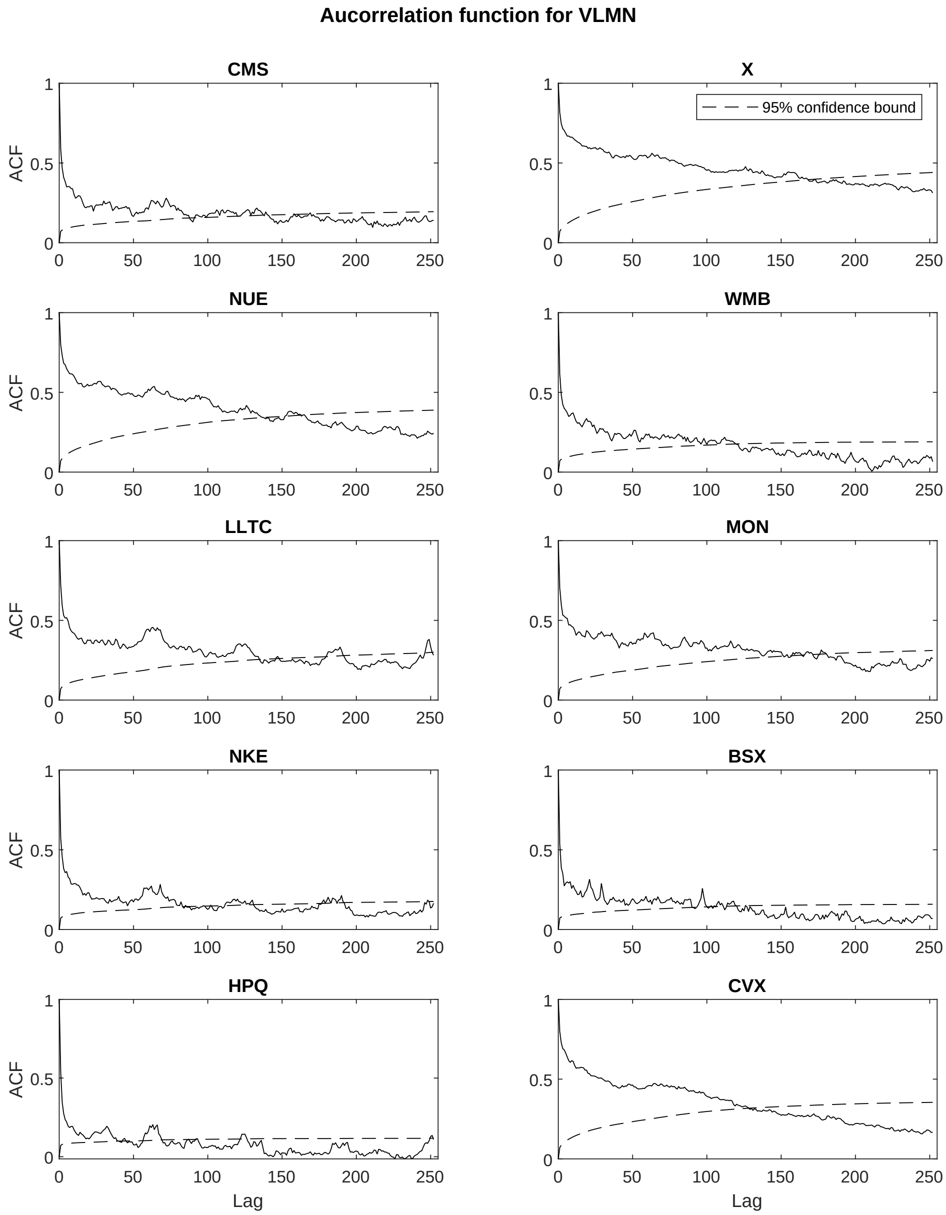

Figure A.1.4: **ACF of VLMN**
Sample autocorrelation function of volume in number of shares for each firm with 1 year worth of lags. The 95% critical value is computed assuming asymptotic normality. The standard error estimator is presented in footnote 1.

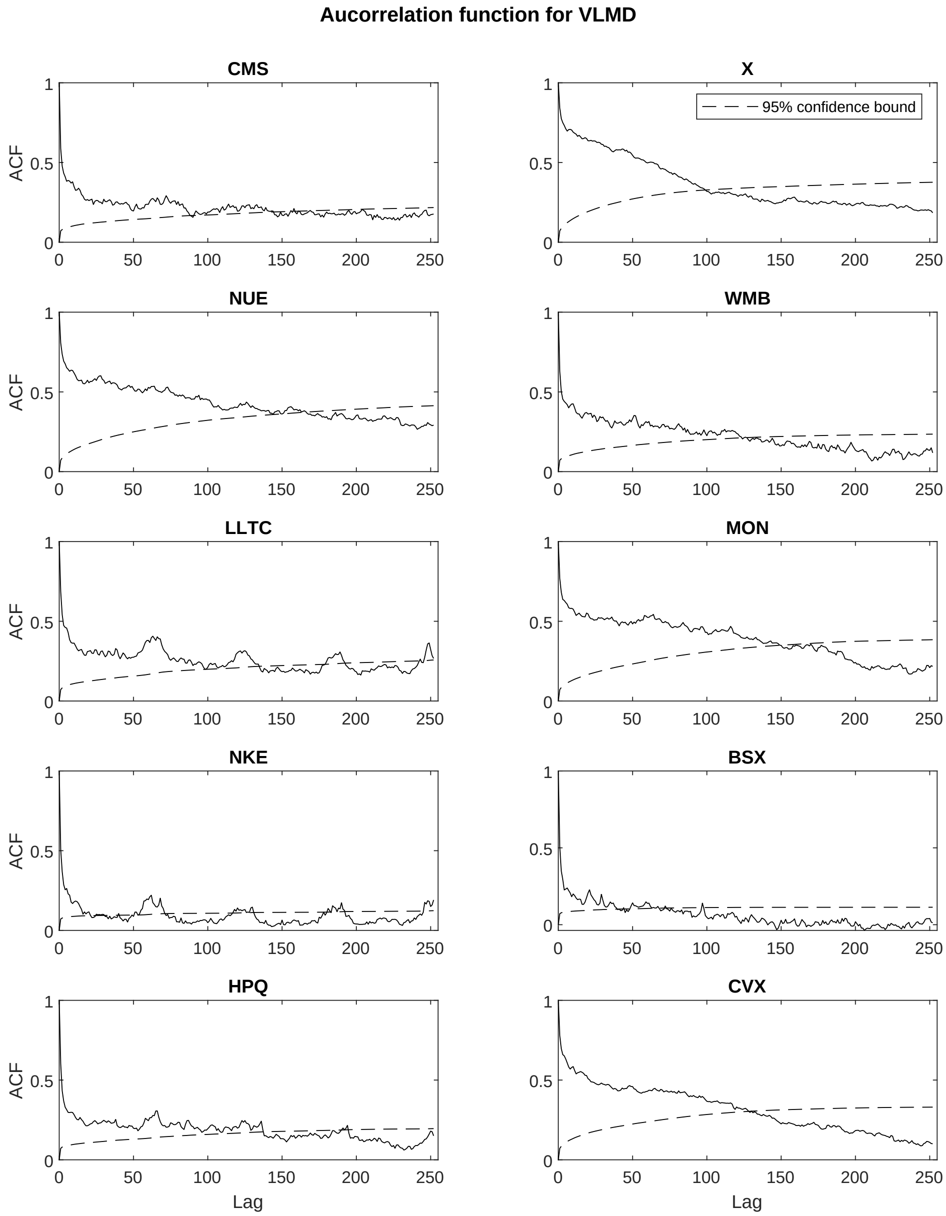

Figure A.1.5: **ACF of VLMD**
Sample autocorrelation function of volume in dollars for each firm with 1 year worth of lags. The 95% critical value is computed assuming asymptotic normality. The standard error estimator is presented in footnote 1.

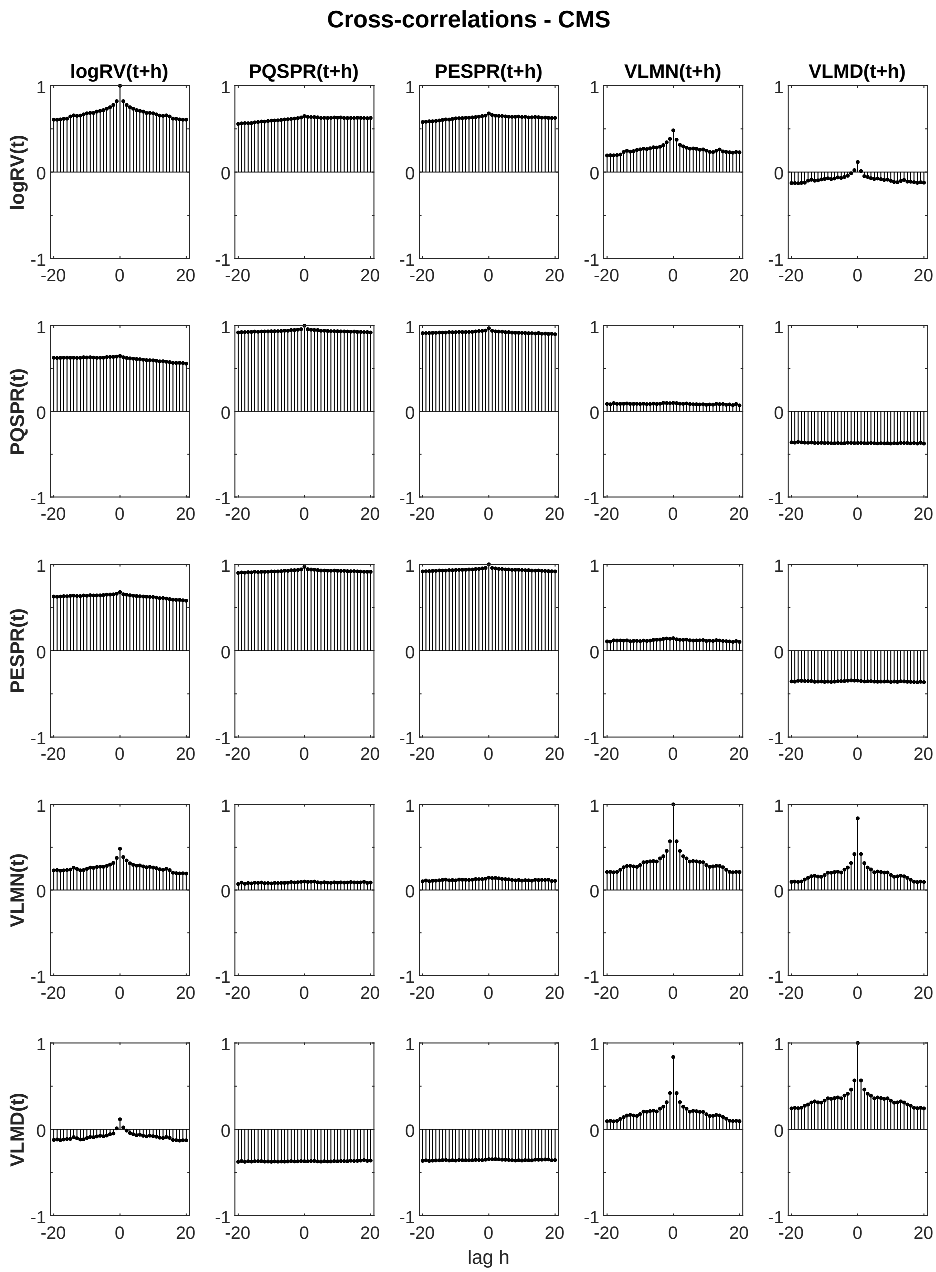

Figure A.1.6: **Cross-correlations for CMS, 2005-2013**
Sample cross correlations for the firm CMS between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

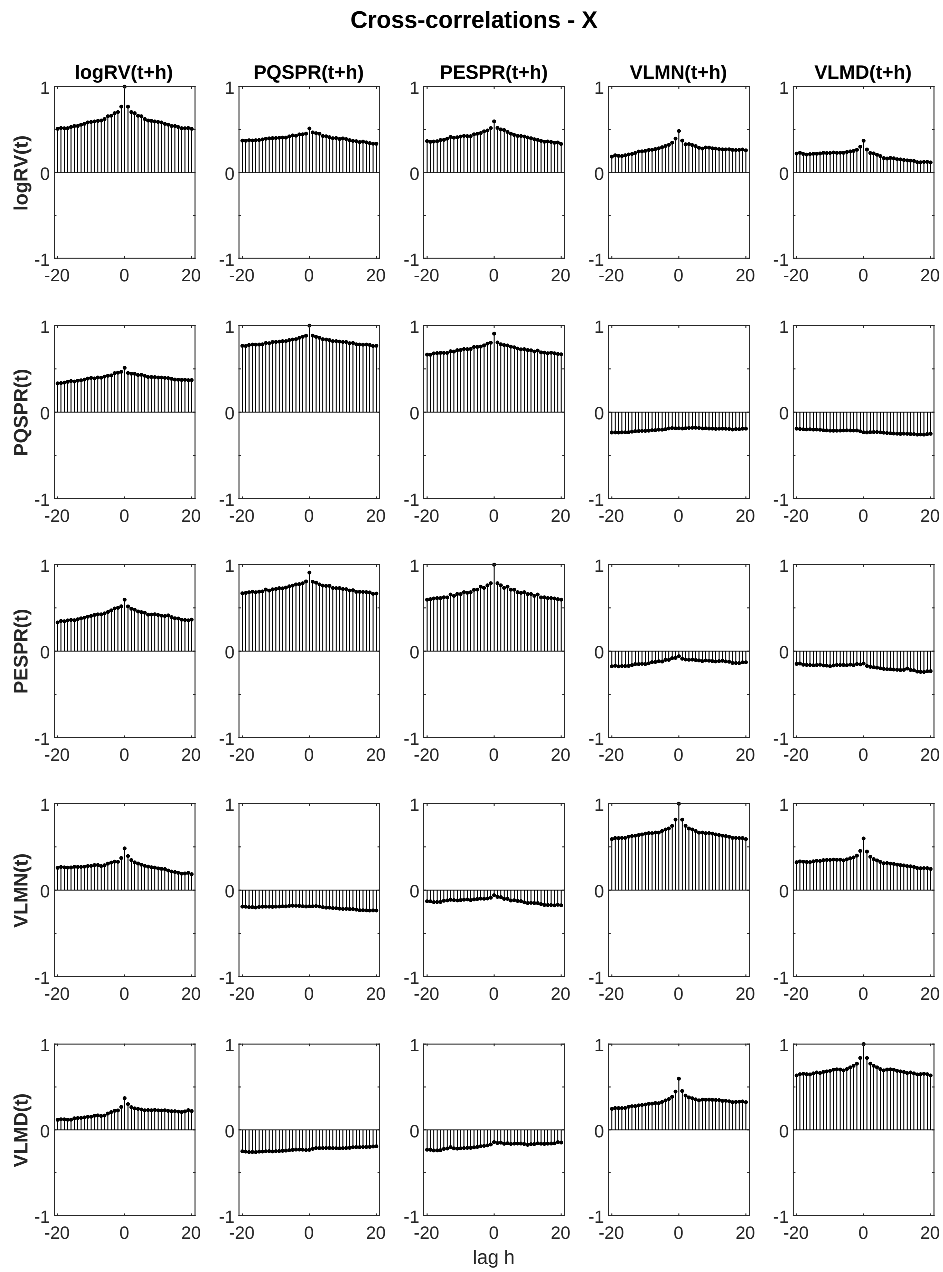

Figure A.1.7: **Cross-correlations for X, 2005-2013**
Sample cross correlations for the firm X between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

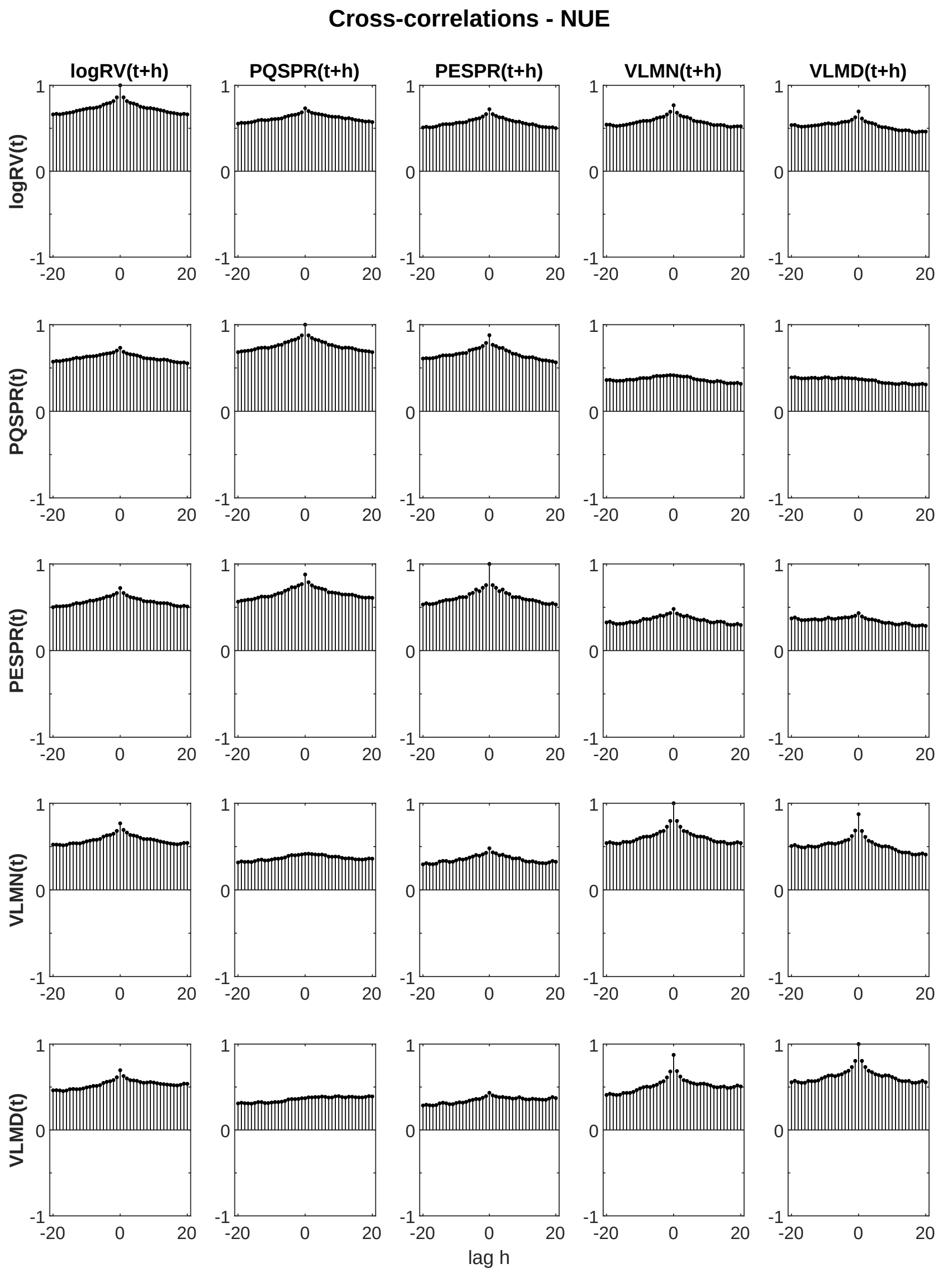

Figure A.1.8: **Cross-correlations for NUE, 2005-2013**
Sample cross correlations for the firm NUE between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

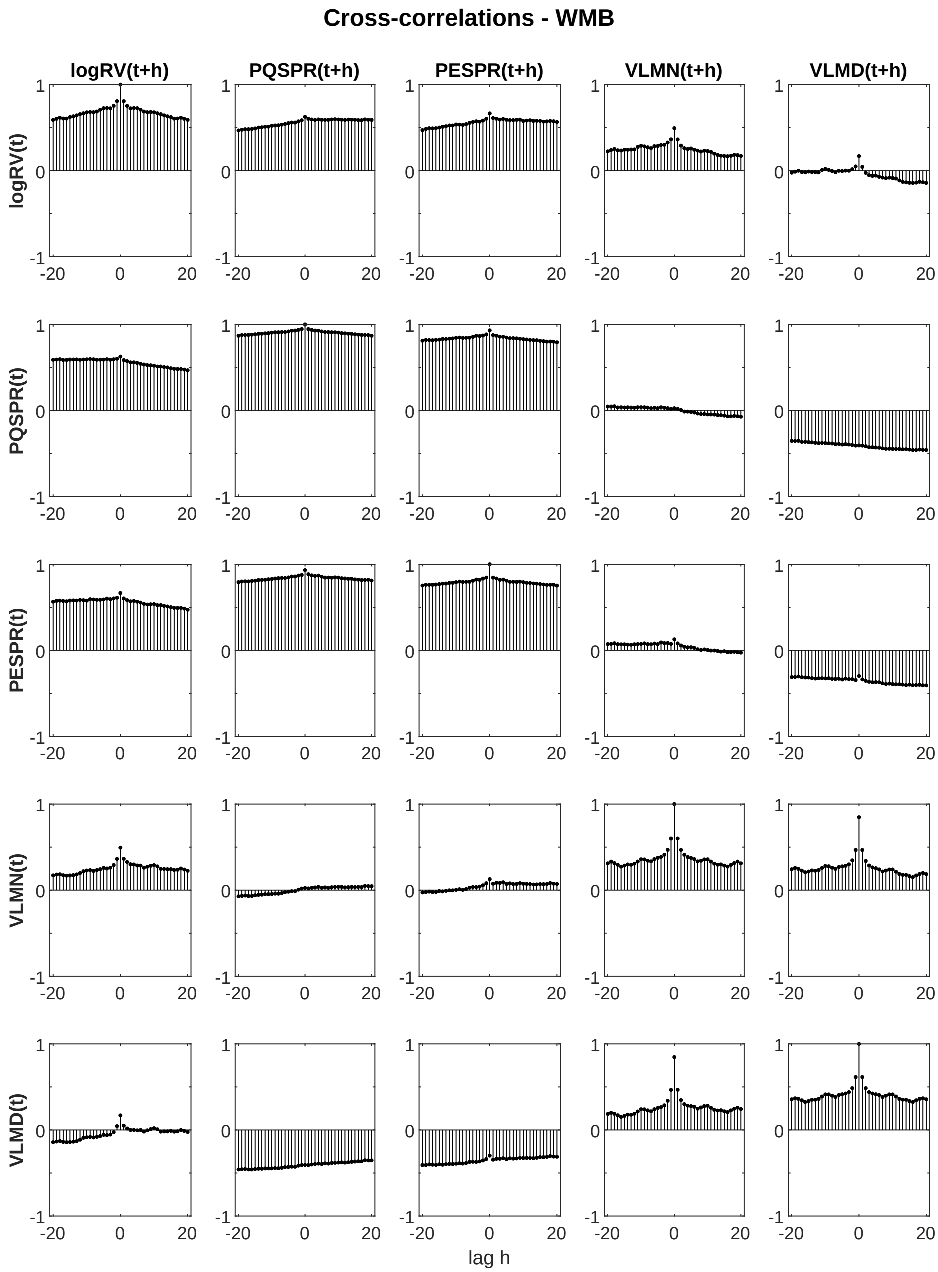

Figure A.1.9: **Cross-correlations for WMB, 2005-2013**
Sample cross correlations for the firm WMB between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

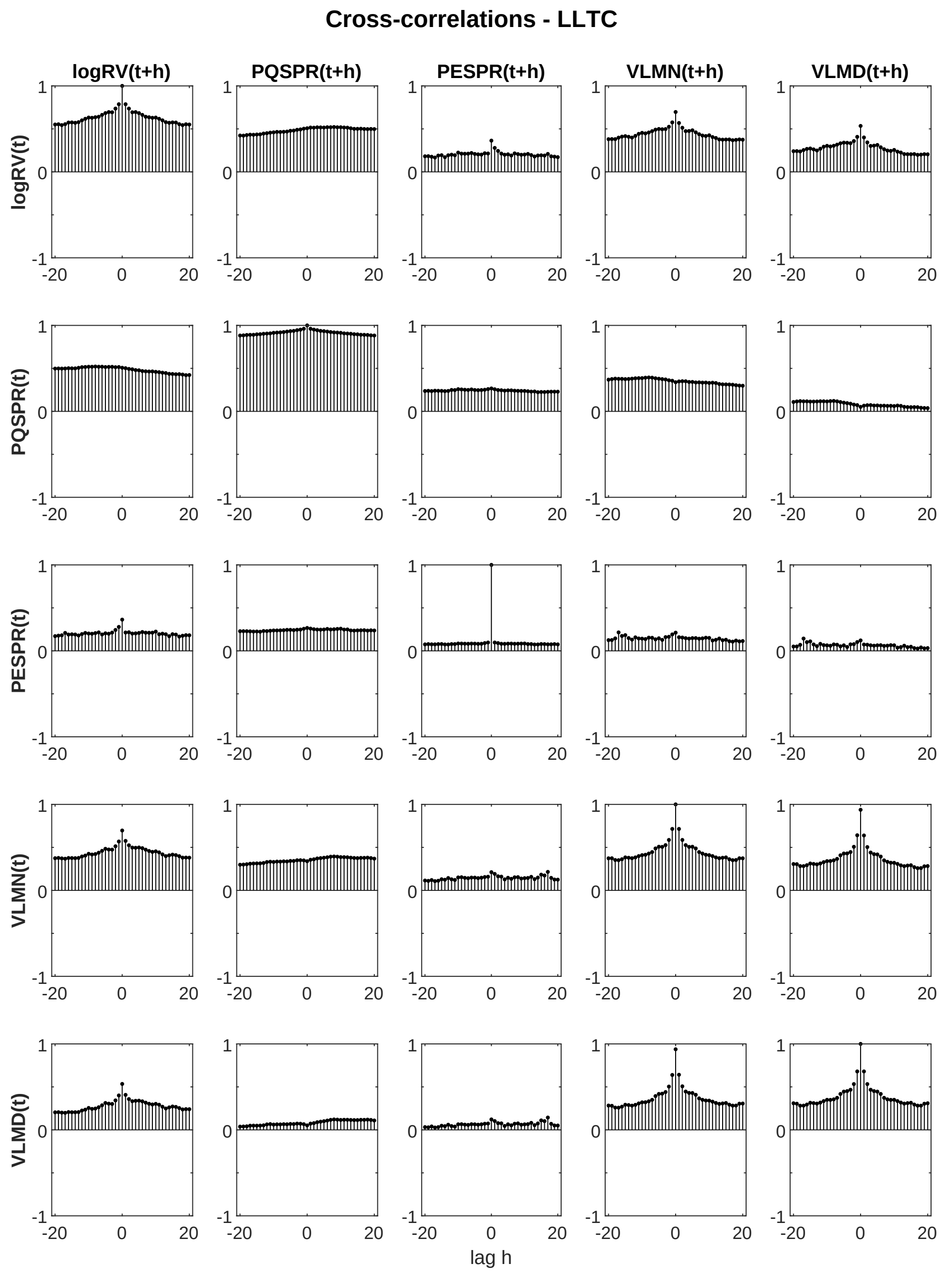

Figure A.1.10: **Cross-correlations for LLTC, 2005-2013**
Sample cross correlations for the firm LLTC between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

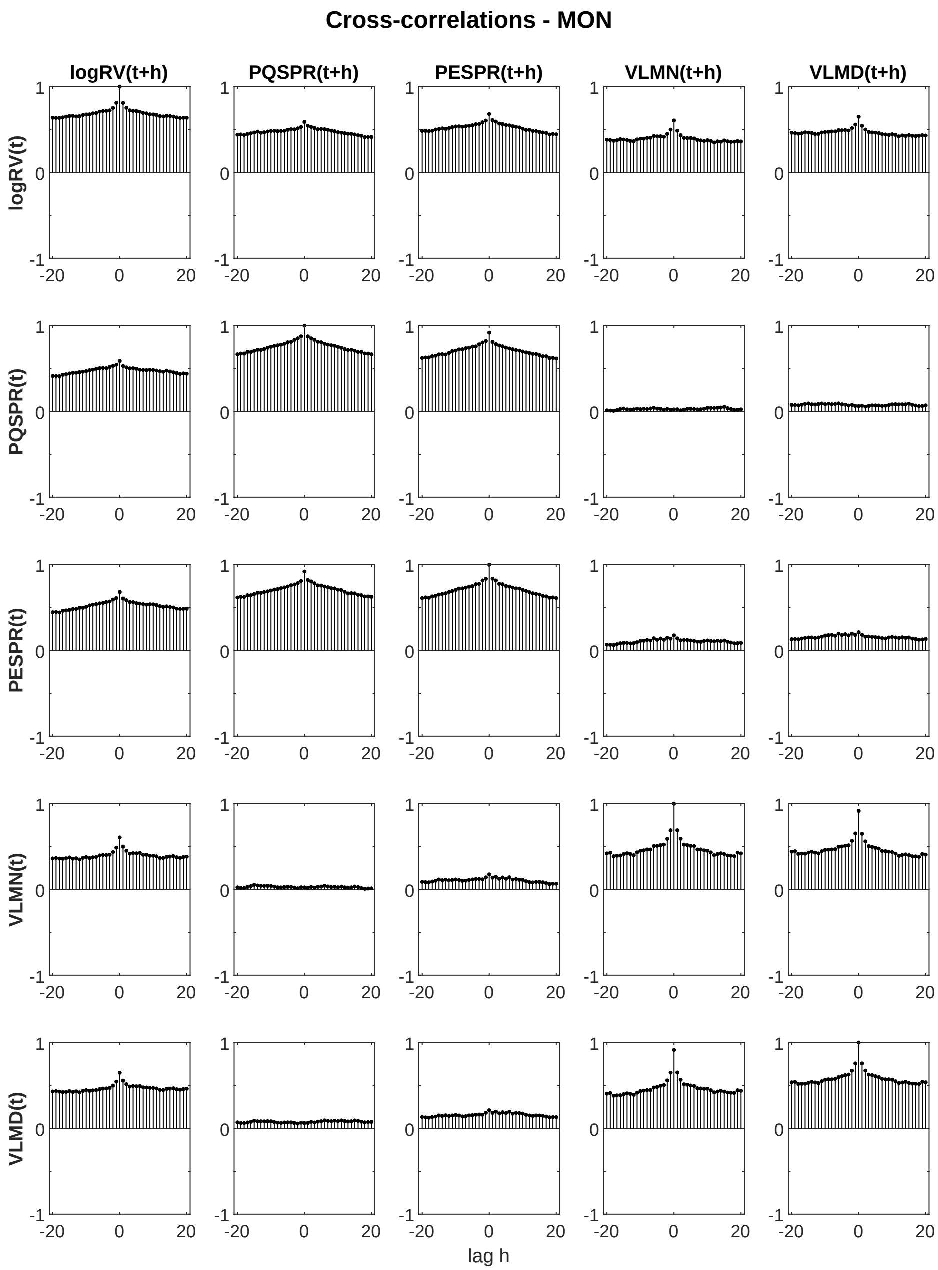

Figure A.1.11: **Cross-correlations for MON, 2005-2013**
Sample cross correlations for the firm MON between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

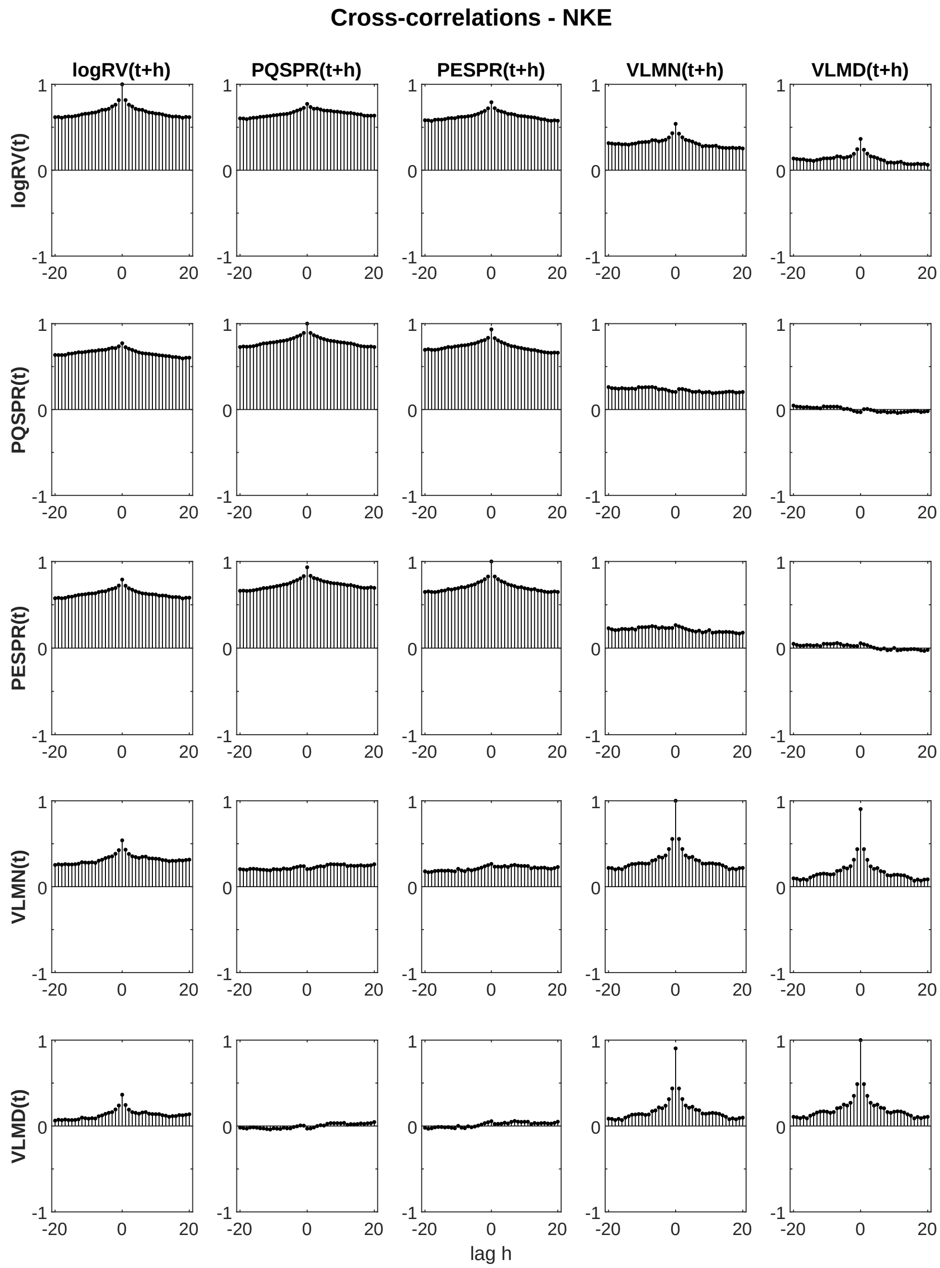

Figure A.1.12: **Cross-correlations for NKE, 2005-2013**
Sample cross correlations for the firm NKE between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

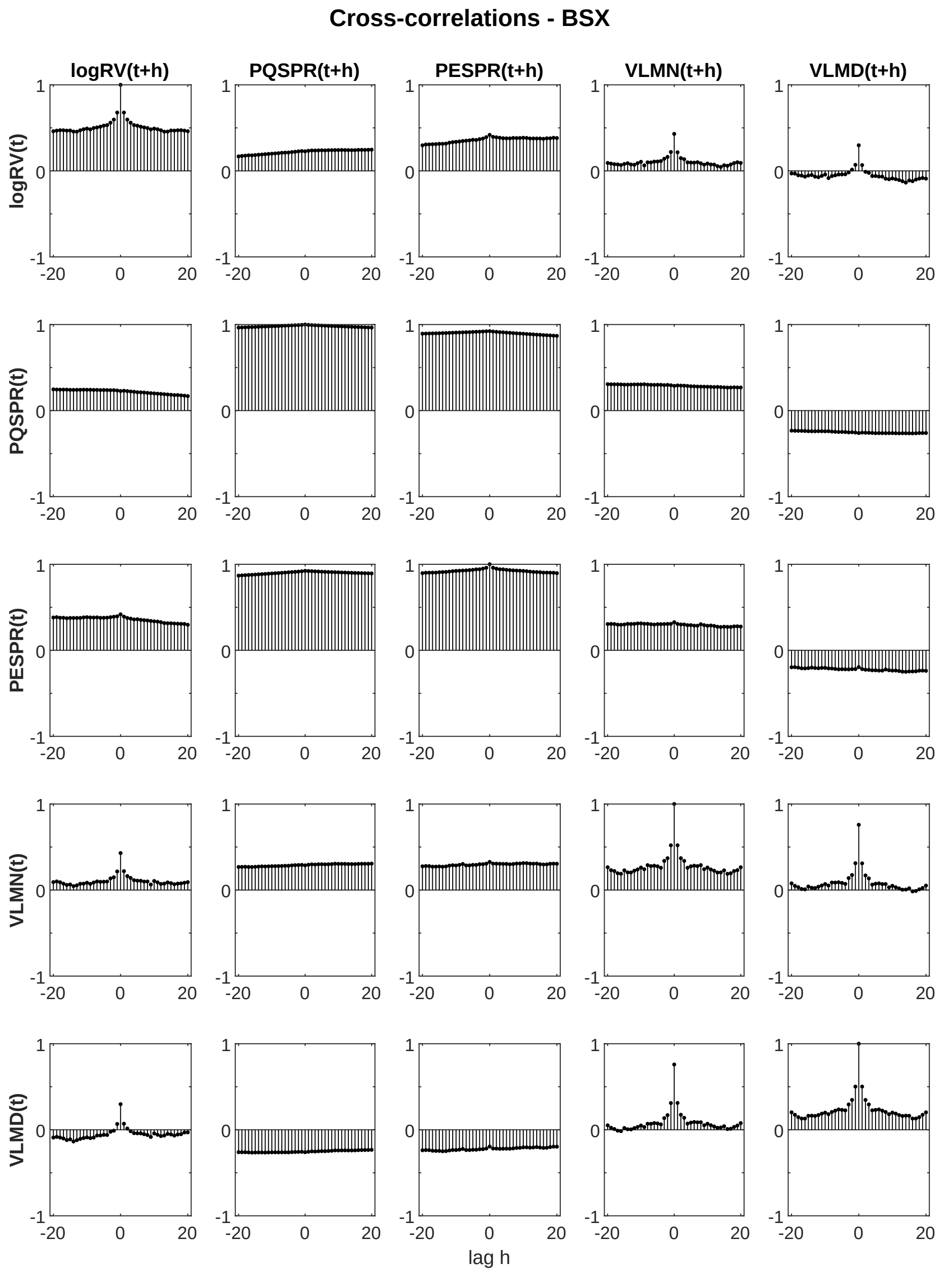

Figure A.1.13: **Cross-correlations for BSX, 2005-2013**
Sample cross correlations for the firm BSX between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

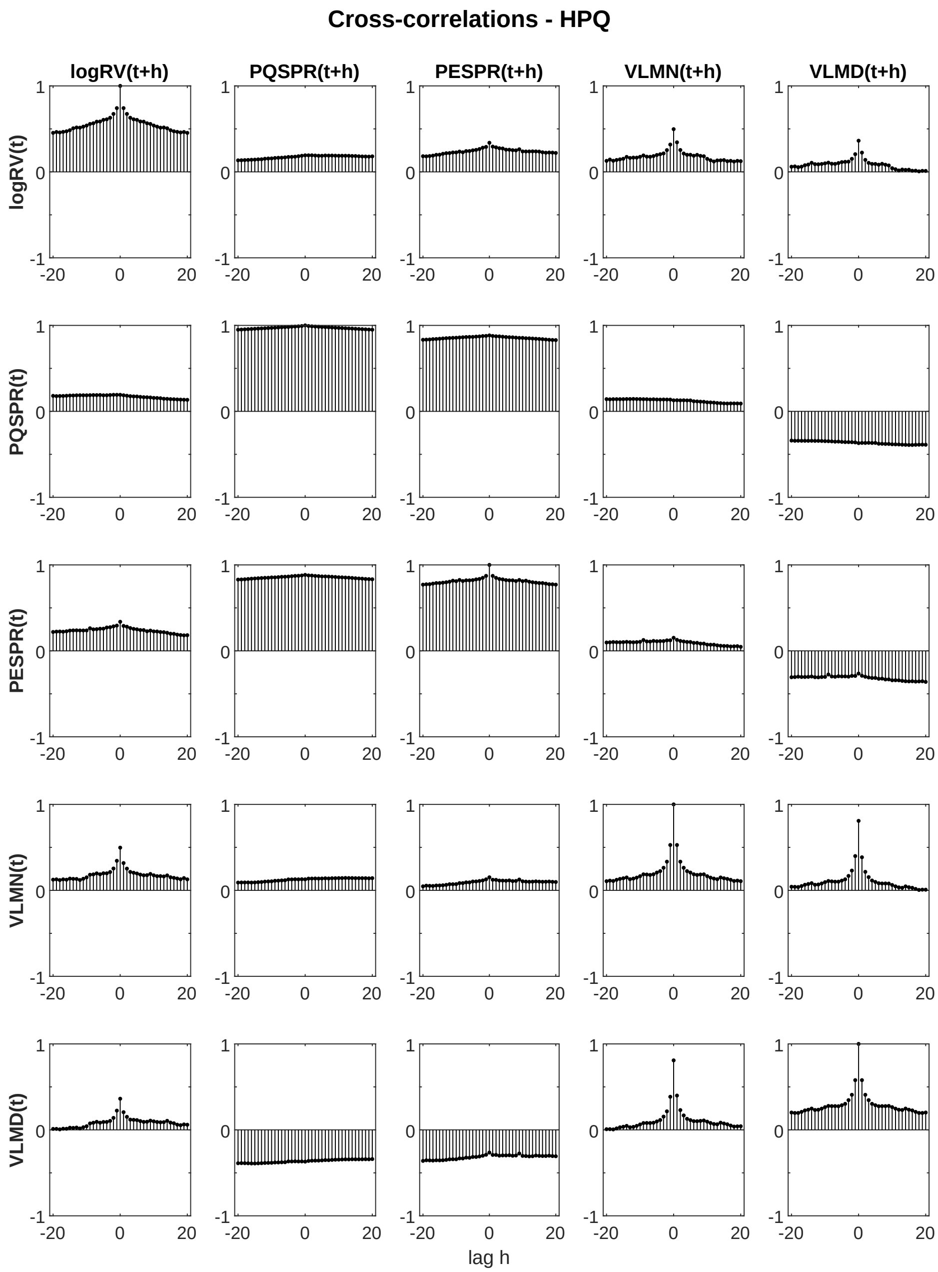

Figure A.1.14: **Cross-correlations for HPQ, 2005-2013**
Sample cross correlations for the firm HPQ between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

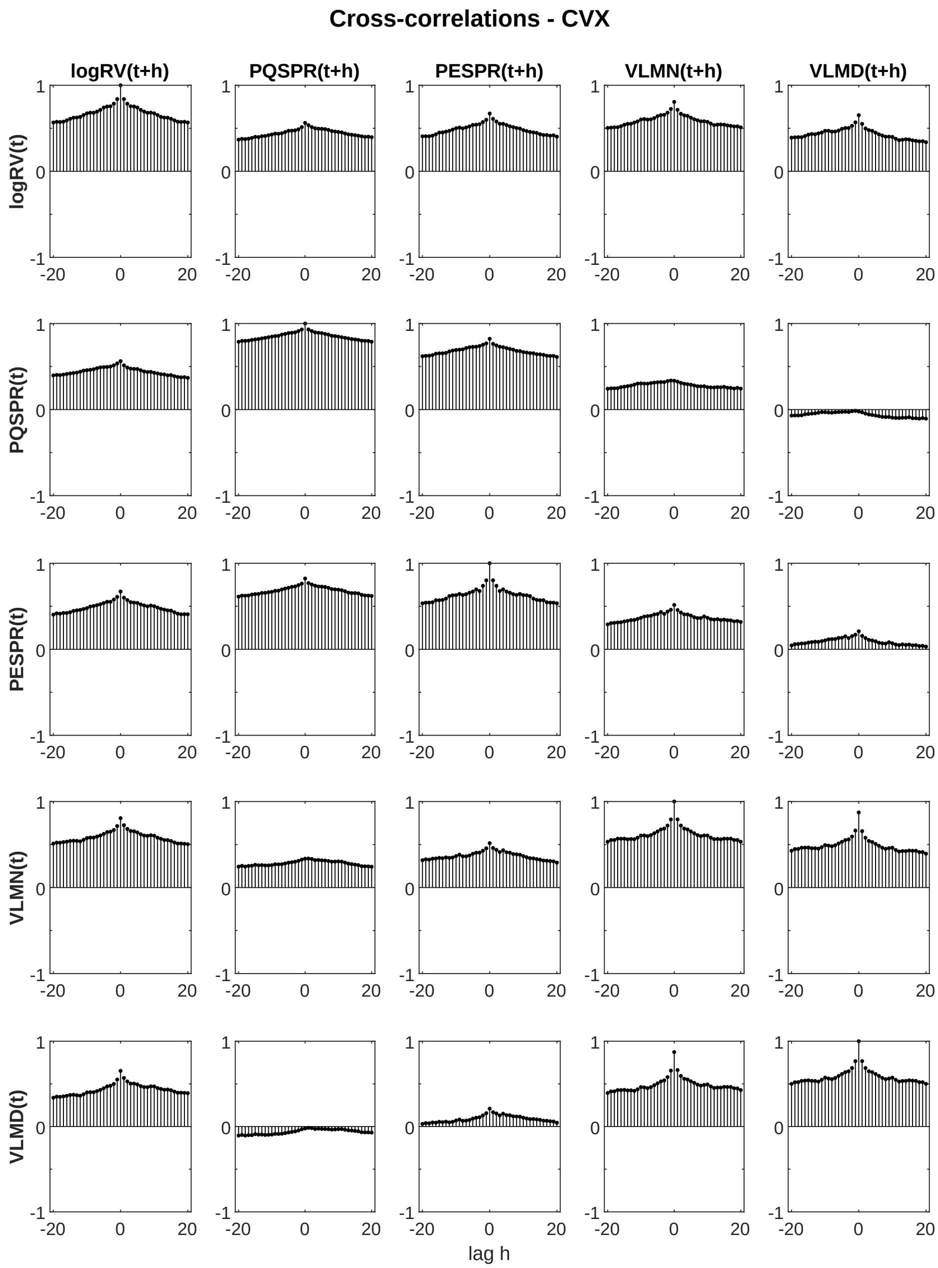

Figure A.1.15: **Cross-correlations for CVX, 2005-2013**
Sample cross correlations for the firm CVX between its daily log-RV, daily proportional quoted (PQSPR) or effective spread (PESPR) and daily volume in number of shares (VLMN) or dollars (VLMD), for up to one month worth of lags forward and backward.

Factor	CMS	X	NUE	WMB	LLTC	MON	NKE	BSX	HPQ	CVX
$\log(\mathrm{RV}_{i,t}^{(d)})$	0.66	0.41	0.58	0.43	0.51	0.35	0.38	0.37	0.50	0.56
$\log(\mathrm{RV}_{i,t}^{(w)})$	0.69	0.40	0.64	0.42	0.55	0.29	0.26	0.25	0.49	0.56
$\log(\mathrm{RV}_{i,t}^{(m)})$	0.67	0.20	0.52	0.39	0.44	0.27	0.24	0.16	0.35	0.40
$\mathrm{PQSPR}_{i,t}^{(d)}$	0.61	0.17	0.33	0.30	0.12	0.18	0.17	0.15	0.25	0.30
$\mathrm{PQSPR}_{i,t}^{(w)}$	0.62	0.19	0.34	0.32	0.16	0.15	0.20	0.08	0.24	0.28
$\mathrm{PQSPR}_{i,t}^{(m)}$	0.59	0.11	0.26	0.23	0.18	0.16	0.17	0.05	0.21	0.16
$\mathrm{PESPR}_{i,t}^{(d)}$	0.48	0.29	0.43	0.30	0.15	0.27	0.21	0.22	0.30	0.36
$\mathrm{PESPR}_{i,t}^{(w)}$	0.49	0.29	0.44	0.33	0.27	0.24	0.20	0.15	0.28	0.35
$\mathrm{PESPR}_{i,t}^{(m)}$	0.45	0.17	0.37	0.25	0.23	0.22	0.16	0.11	0.24	0.21
$\mathrm{VLMN}_{i,t}^{(d)}$	0.27	0.41	0.50	0.23	0.28	0.20	0.24	0.20	0.11	0.33
$\mathrm{VLMN}_{i,t}^{(w)}$	0.23	0.33	0.51	0.26	0.30	0.12	0.14	0.14	0.03	0.35
$\mathrm{VLMN}_{i,t}^{(m)}$	0.28	0.19	0.39	0.27	0.19	0.04	0.11	0.18	-0.07	0.20
$\mathrm{VLMD}_{i,t}^{(d)}$	0.22	0.35	0.47	0.15	0.21	0.22	0.22	0.21	0.02	0.27
$\mathrm{VLMD}_{i,t}^{(w)}$	0.15	0.28	0.48	0.17	0.20	0.15	0.12	0.14	-0.10	0.28
$\mathrm{VLMD}_{i,t}^{(m)}$	0.16	0.17	0.46	0.17	0.03	0.10	0.09	0.16	-0.21	0.15
$\Delta\mathrm{PQSPR}_{i,t}^{(d)}$	0.04	0.01	0.09	0.03	-0.04	0.06	-0.03	0.09	0.04	0.09
$\Delta\mathrm{PQSPR}_{i,t}^{(w)}$	0.08	0.08	0.09	0.06	0.02	0.09	0.05	0.12	0.08	0.17
$\Delta\mathrm{PQSPR}_{i,t}^{(m)}$	0.11	0.12	0.19	0.21	0.03	0.10	0.09	0.14	0.24	0.31
$\Delta\mathrm{PESPR}_{i,t}^{(d)}$	0.01	0.06	0.09	0.03	0.01	0.07	0.02	0.11	0.06	0.10
$\Delta\mathrm{PESPR}_{i,t}^{(w)}$	0.09	0.17	0.13	0.08	0.03	0.14	0.13	0.14	0.14	0.21
$\Delta\mathrm{PESPR}_{i,t}^{(m)}$	0.13	0.20	0.21	0.19	0.06	0.17	0.15	0.20	0.28	0.36
$\Delta\mathrm{VLMN}_{i,t}^{(d)}$	0.10	0.16	0.10	0.02	0.09	0.12	0.12	0.05	0.07	0.13
$\Delta\mathrm{VLMN}_{i,t}^{(w)}$	0.13	0.24	0.19	0.05	0.08	0.11	0.18	0.10	0.07	0.13
$\Delta\mathrm{VLMN}_{i,t}^{(m)}$	0.10	0.30	0.28	0.11	0.25	0.18	0.15	0.04	0.08	0.32
$\Delta\mathrm{VLMD}_{i,t}^{(d)}$	0.10	0.13	0.09	0.01	0.08	0.12	0.12	0.06	0.06	0.12
$\Delta\mathrm{VLMD}_{i,t}^{(w)}$	0.13	0.21	0.12	0.03	0.07	0.11	0.17	0.11	0.07	0.11
$\Delta\mathrm{VLMD}_{i,t}^{(m)}$	0.10	0.27	0.18	0.07	0.24	0.19	0.14	0.06	0.06	0.28

Table A.1.6: **Factor correlations with dependent variable, 2005-2006**

Correlation between the dependent variable $\log(\mathrm{RV}_{i,t+1}^{(d)})$ and the various factors of the HAR-RV-LIQ model for each company (see table 3.1), evaluated for the 2005-2006 subsample period (478 observations). The upper blocks showcase correlations for realized volatility (RV), proportional quoted or effective spread (PQSPR and PESPR) and for volume in number of shares (VLMN) or in dollars (VLMD) in level, while the bottom blocks are for the same factors taken in difference (Δ). Each daily (d) factor is also included in lagged weekly (w) or monthly(m) moving averages. Realized volatility is calculated using the 5 minutes subsampling estimator.

Factor	CMS	X	NUE	WMB	LLTC	MON	NKE	BSX	HPQ	CVX
$\log(\mathrm{RV}_{i,t}^{(\mathrm{d})})$	0.84	0.88	0.86	0.87	0.81	0.84	0.86	0.81	0.85	0.86
$\log(\mathrm{RV}_{i,t}^{(\mathrm{w})})$	0.83	0.88	0.86	0.88	0.83	0.85	0.87	0.82	0.86	0.86
$\log(\mathrm{RV}_{i,t}^{(\mathrm{m})})$	0.77	0.83	0.78	0.82	0.77	0.80	0.82	0.79	0.81	0.77
$\mathrm{PQSPR}_{i,t}^{(\mathrm{d})}$	0.66	0.68	0.73	0.66	0.68	0.65	0.77	0.66	0.67	0.72
$\mathrm{PQSPR}_{i,t}^{(\mathrm{w})}$	0.68	0.71	0.74	0.65	0.68	0.67	0.78	0.66	0.66	0.72
$\mathrm{PQSPR}_{i,t}^{(\mathrm{m})}$	0.63	0.70	0.68	0.58	0.63	0.66	0.74	0.62	0.57	0.63
$\mathrm{PESPR}_{i,t}^{(\mathrm{d})}$	0.66	0.69	0.63	0.61	0.53	0.62	0.74	0.69	0.47	0.56
$\mathrm{PESPR}_{i,t}^{(\mathrm{w})}$	0.70	0.76	0.69	0.65	0.65	0.66	0.79	0.69	0.61	0.65
$\mathrm{PESPR}_{i,t}^{(\mathrm{m})}$	0.65	0.74	0.64	0.59	0.64	0.64	0.76	0.65	0.61	0.60
$\mathrm{VLMN}_{i,t}^{(\mathrm{d})}$	0.52	0.54	0.68	0.66	0.44	0.56	0.46	0.25	0.47	0.75
$\mathrm{VLMN}_{i,t}^{(\mathrm{w})}$	0.58	0.54	0.73	0.75	0.49	0.61	0.58	0.27	0.57	0.78
$\mathrm{VLMN}_{i,t}^{(\mathrm{m})}$	0.58	0.48	0.72	0.76	0.49	0.62	0.67	0.23	0.63	0.72
$\mathrm{VLMD}_{i,t}^{(\mathrm{d})}$	0.31	0.26	0.40	0.24	0.15	0.61	0.30	-0.01	0.33	0.67
$\mathrm{VLMD}_{i,t}^{(\mathrm{w})}$	0.35	0.26	0.43	0.27	0.12	0.66	0.38	-0.08	0.40	0.70
$\mathrm{VLMD}_{i,t}^{(\mathrm{m})}$	0.36	0.28	0.43	0.30	0.02	0.66	0.49	-0.24	0.48	0.68
$\Delta\mathrm{PQSPR}_{i,t}^{(\mathrm{d})}$	0.03	0.02	0.04	0.05	0.03	0.03	0.04	0.05	0.06	0.07
$\Delta\mathrm{PQSPR}_{i,t}^{(\mathrm{w})}$	0.10	0.06	0.14	0.16	0.12	0.11	0.14	0.10	0.18	0.15
$\Delta\mathrm{PQSPR}_{i,t}^{(\mathrm{m})}$	0.24	0.14	0.27	0.34	0.27	0.17	0.19	0.27	0.37	0.32
$\Delta\mathrm{PESPR}_{i,t}^{(\mathrm{d})}$	0.02	0.03	0.05	0.04	0.00	0.03	0.05	0.06	0.00	0.03
$\Delta\mathrm{PESPR}_{i,t}^{(\mathrm{w})}$	0.09	0.08	0.11	0.09	0.01	0.11	0.12	0.11	0.07	0.10
$\Delta\mathrm{PESPR}_{i,t}^{(\mathrm{m})}$	0.23	0.17	0.23	0.22	0.11	0.18	0.17	0.27	0.15	0.25
$\Delta\mathrm{VLMN}_{i,t}^{(\mathrm{d})}$	0.03	0.06	0.05	0.03	0.08	0.06	0.03	0.03	0.05	0.07
$\Delta\mathrm{VLMN}_{i,t}^{(\mathrm{w})}$	0.09	0.12	0.11	0.07	0.10	0.11	0.03	0.09	0.07	0.15
$\Delta\mathrm{VLMN}_{i,t}^{(\mathrm{m})}$	0.18	0.24	0.18	0.12	0.21	0.20	0.07	0.16	0.15	0.30
$\Delta\mathrm{VLMD}_{i,t}^{(\mathrm{d})}$	0.03	0.05	0.05	0.02	0.07	0.06	0.03	0.02	0.05	0.05
$\Delta\mathrm{VLMD}_{i,t}^{(\mathrm{w})}$	0.08	0.06	0.09	0.02	0.07	0.09	0.02	0.08	0.05	0.11
$\Delta\mathrm{VLMD}_{i,t}^{(\mathrm{m})}$	0.13	-0.01	0.10	0.00	0.15	0.16	0.04	0.13	0.09	0.21

Table A.1.7: **Factor correlations with dependent variable, 2007-2009**

Correlation between the dependent variable $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$ and the various factors of the HAR-RV-LIQ model for each company (see table 3.1), evaluated for the 2007-2009 subsample period (747 observations). The upper blocks showcase correlations for realized volatility (RV), proportional quoted or effective spread (PQSPR and PESPR) and for volume in number of shares (VLMN) or in dollars (VLMD) in level, while the bottom blocks are for the same factors taken in difference (Δ). Each daily (d) factor is also included in lagged weekly (w) or monthly(m) moving averages. Realized volatility is calculated using the 5 minutes subsampling estimator.

Factor	CMS	X	NUE	WMB	LLTC	MON	NKE	BSX	HPQ	CVX
$\log(\mathrm{RV}_{i,t}^{(\mathrm{d})})$	0.68	0.65	0.72	0.72	0.71	0.71	0.69	0.47	0.64	0.75
$\log(\mathrm{RV}_{i,t}^{(\mathrm{w})})$	0.70	0.64	0.71	0.73	0.70	0.71	0.68	0.44	0.63	0.75
$\log(\mathrm{RV}_{i,t}^{(\mathrm{m})})$	0.63	0.56	0.65	0.68	0.65	0.69	0.60	0.42	0.57	0.67
$\mathrm{PQSPR}_{i,t}^{(\mathrm{d})}$	0.45	0.13	0.36	0.45	0.52	0.47	0.61	0.16	0.39	0.49
$\mathrm{PQSPR}_{i,t}^{(\mathrm{w})}$	0.45	0.11	0.36	0.43	0.52	0.49	0.58	0.16	0.38	0.46
$\mathrm{PQSPR}_{i,t}^{(\mathrm{m})}$	0.44	0.08	0.31	0.39	0.50	0.47	0.52	0.13	0.38	0.40
$\mathrm{PESPR}_{i,t}^{(\mathrm{d})}$	0.49	0.20	0.56	0.52	0.11	0.61	0.64	0.25	0.44	0.65
$\mathrm{PESPR}_{i,t}^{(\mathrm{w})}$	0.48	0.17	0.53	0.49	0.22	0.62	0.61	0.23	0.43	0.62
$\mathrm{PESPR}_{i,t}^{(\mathrm{m})}$	0.47	0.12	0.48	0.45	0.41	0.60	0.55	0.20	0.41	0.56
$\mathrm{VLMN}_{i,t}^{(\mathrm{d})}$	0.38	0.44	0.57	0.38	0.55	0.40	0.24	0.23	0.32	0.60
$\mathrm{VLMN}_{i,t}^{(\mathrm{w})}$	0.44	0.41	0.57	0.44	0.59	0.40	0.18	0.19	0.29	0.62
$\mathrm{VLMN}_{i,t}^{(\mathrm{m})}$	0.41	0.32	0.54	0.51	0.59	0.45	0.09	0.15	0.28	0.57
$\mathrm{VLMD}_{i,t}^{(\mathrm{d})}$	0.22	0.21	0.49	0.21	0.46	0.30	0.27	0.16	0.06	0.57
$\mathrm{VLMD}_{i,t}^{(\mathrm{w})}$	0.21	0.16	0.49	0.19	0.50	0.28	0.24	0.08	-0.04	0.63
$\mathrm{VLMD}_{i,t}^{(\mathrm{m})}$	0.11	0.12	0.47	0.18	0.51	0.34	0.20	0.02	-0.16	0.64
$\Delta\mathrm{PQSPR}_{i,t}^{(\mathrm{d})}$	0.05	0.08	0.03	0.08	0.08	0.06	0.11	0.02	0.03	0.11
$\Delta\mathrm{PQSPR}_{i,t}^{(\mathrm{w})}$	0.03	0.09	0.12	0.16	0.13	0.09	0.19	0.11	0.03	0.15
$\Delta\mathrm{PQSPR}_{i,t}^{(\mathrm{m})}$	0.05	0.27	0.20	0.17	0.14	0.18	0.24	0.22	0.02	0.23
$\Delta\mathrm{PESPR}_{i,t}^{(\mathrm{d})}$	0.05	0.10	0.08	0.12	0.00	0.08	0.13	0.05	0.10	0.13
$\Delta\mathrm{PESPR}_{i,t}^{(\mathrm{w})}$	0.05	0.12	0.18	0.18	-0.01	0.11	0.20	0.09	0.10	0.17
$\Delta\mathrm{PESPR}_{i,t}^{(\mathrm{m})}$	0.10	0.30	0.26	0.23	0.01	0.20	0.26	0.16	0.14	0.29
$\Delta\mathrm{VLMN}_{i,t}^{(\mathrm{d})}$	0.05	0.10	0.06	0.07	0.07	0.09	0.08	0.10	0.09	0.07
$\Delta\mathrm{VLMN}_{i,t}^{(\mathrm{w})}$	0.10	0.18	0.16	0.10	0.11	0.06	0.13	0.12	0.16	0.15
$\Delta\mathrm{VLMN}_{i,t}^{(\mathrm{m})}$	0.16	0.29	0.19	0.15	0.19	0.06	0.16	0.10	0.16	0.26
$\Delta\mathrm{VLMD}_{i,t}^{(\mathrm{d})}$	0.05	0.07	0.06	0.07	0.07	0.09	0.08	0.10	0.08	0.07
$\Delta\mathrm{VLMD}_{i,t}^{(\mathrm{w})}$	0.09	0.14	0.15	0.08	0.10	0.08	0.12	0.12	0.13	0.14
$\Delta\mathrm{VLMD}_{i,t}^{(\mathrm{m})}$	0.15	0.18	0.14	0.11	0.17	0.05	0.15	0.08	0.15	0.22

Table A.1.8: **Factor correlations with dependent variable, 2010-2013**

Correlation between the dependent variable $\log(\mathrm{RV}_{i,t+1}^{(\mathrm{d})})$ and the various factors of the HAR-RV-LIQ model for each company (see table 3.1), evaluated for the 2010-2013 subsample period (997 observations). The upper blocks showcase correlations for realized volatility (RV), proportional quoted or effective spread (PQSPR and PESPR) and for volume in number of shares (VLMN) or in dollars (VLMD) in level, while the bottom blocks are for the same factors taken in difference (Δ). Each daily (d) factor is also included in lagged weekly (w) or monthly(m) moving averages. Realized volatility is calculated using the 5 minutes subsampling estimator.

A.2 Relative prediction performance

	RMSPE of annualized daily RV predictions – CMS					
	Static			**Dynamic**		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.2038	0.2491	0.2046	0.2073	0.2476	0.2028
[2] $\text{PQSPR}^{(d)}$	+0.18%	**-0.48%**	**-1.72%**	**-0.18%**	**-1.25%**	**-1.51%**
[3] $\text{PESPR}^{(d)}$	+0.12%	**-0.36%**	**-1.44%**	**-0.11%**	**-1.53%**	**-1.46%**
[5] $\text{PQSPR}^{(d)}$, $\text{VLMN}^{(d)}$	+1.14%	+0.35%	**-1.40%**	+2.39%	**-1.38%**	**-1.42%**
[6] $\text{PESPR}^{(d)}$, $\text{VLMN}^{(d)}$	+0.94%	+0.66%	**-1.14%**	+2.29%	**-1.70%**	**-1.40%**
[7] $\text{VLMN}^{(d)}$, $\text{VLMN}^{(w)}$	**-0.40%**	+0.45%	+0.27%	+1.99%	**-0.04%**	+0.04%
[8] $\text{PQSPR}^{(d)}$, $\text{VLMN}^{(d)}$, $\text{VLMN}^{(w)}$	**-0.05%**	+0.05%	**-1.46%**	+2.04%	**-1.31%**	**-1.52%**
[9] $\text{PESPR}^{(d)}$, $\text{VLMN}^{(d)}$, $\text{VLMN}^{(w)}$	**-0.22%**	+0.30%	**-1.23%**	+1.96%	**-1.64%**	**-1.51%**

Table A.2.1: **Out-of-sample relative performance results for CMS**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of CMS, $\sqrt{252}\text{RV}^{(d)}_{\text{CMS},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.2, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSPE of annualized daily RV predictions – X					
	Static			**Dynamic**		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.2040	0.2166	0.2098	0.2080	0.2178	0.2082
[2] $\text{PESPR}^{(d)}$	-2.42%	-1.06%	+16.71%	-0.48%	-0.58%	+3.53%
[3] $\text{VLMN}^{(d)}$	+0.30%	+0.16%	-0.48%	+5.02%	+0.23%	-1.21%
[5] $\text{PESPR}^{(d)}$, $\text{VLMN}^{(d)}$	-1.81%	-1.04%	+22.58%	+7.21%	-0.69%	+3.46%
[6] $\text{VLMN}^{(d)}$, $\text{VLMN}^{(w)}$	-0.25%	-0.54%	-0.09%	+3.84%	-0.17%	-0.61%
[8] $\text{PESPR}^{(d)}$, $\text{VLMN}^{(d)}$, $\text{VLMN}^{(w)}$	-2.92%	-1.37%	+21.66%	+5.26%	-0.69%	+3.39%

Table A.2.2: **Out-of-sample relative performance results for X**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of X, $\sqrt{252}\text{RV}^{(d)}_{\text{X},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.3, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSE of annualized daily RV predictions – NUE					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.2106	0.2302	0.2053	0.2133	0.2297	0.2050
[2] $\text{PQSPR}^{(d)}$	**-1.61%**	+0.08%	+2.05%	**-0.66%**	**-0.34%**	+0.02%
[3] $\text{PESPR}^{(d)}$	**-1.54%**	**-0.27%**	+5.09%	**-0.72%**	**-0.40%**	+0.81%
[4] $\text{VLMN}^{(d)}$	+3.42%	**-0.28%**	**-0.40%**	+2.69%	**-0.55%**	**-1.29%**
[5] $\text{PQSPR}^{(d)}$, $\text{VLMN}^{(d)}$	+2.11%	**-0.10%**	+1.76%	+2.55%	**-1.05%**	**-1.22%**
[7] $\text{PESPR}^{(d)}$, $\text{VLMN}^{(d)}$	+2.11%	**-0.56%**	+4.26%	+2.24%	**-0.90%**	**-0.82%**

Table A.2.3: **Out-of-sample relative performance results for NUE**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of NUE, $\sqrt{252}\text{RV}^{(d)}_{\text{NUE},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.4, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSPE of annualized daily RV predictions – LLTC					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.2271	0.3041	0.2292	0.2405	0.3090	0.2244
[2] $\text{PQSPR}^{(d)}$	**-0.24%**	**-3.83%**	+0.78%	+1.48%	**-3.92%**	**-1.04%**
[3] $\text{VLMN}^{(d)}$	+0.20%	**-2.03%**	+1.69%	**-0.55%**	**-3.22%**	+0.09%
[4] $\text{PQSPR}^{(d)}$, $\text{VLMN}^{(d)}$	+0.08%	**-7.28%**	+1.54%	+0.64%	**-8.90%**	**-1.26%**

Table A.2.4: **Out-of-sample relative performance results for LLTC**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of LLTC, $\sqrt{252}\text{RV}^{(d)}_{\text{LLTC},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.6, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSPE of annualized daily RV predictions – MON					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.2195	0.2486	0.2248	0.2257	0.2483	0.2220
[2] PESPR$^{(d)}$	**-1.93%**	**-0.21%**	+2.25%	+0.13%	**-0.35%**	+0.69%
[3] VLMN$^{(d)}$	+1.08%	+0.53%	**-0.21%**	+2.48%	+0.15%	**-0.26%**
[4] VLMD$^{(d)}$	+2.73%	+0.02%	**-0.01%**	+2.67%	**-0.09%**	**-0.05%**
[6] PESPR$^{(d)}$, VLMN$^{(d)}$	**-1.05%**	+0.27%	+2.09%	+3.03%	**-0.30%**	+0.17%
[7] PESPR$^{(d)}$, VLMD$^{(d)}$	+0.02%	**-0.51%**	+3.05%	+2.11%	**-0.74%**	+0.45%

Table A.2.5: **Out-of-sample relative performance results for MON**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of MON, $\sqrt{252}\text{RV}^{(d)}_{\text{MON},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.7, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSPE of annualized daily RV predictions – NKE					
	Static			Dynamic		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.1975	0.2159	0.2286	0.2024	0.2166	0.2273
[2] PQSPR$^{(d)}$	**-2.25%**	**-0.79%**	+1.47%	+0.10%	**-1.22%**	**-1.13%**
[3] PESPR$^{(d)}$	**-1.27%**	**-0.55%**	+0.16%	+0.32%	**-1.01%**	**-1.80%**
[5] PQSPR$^{(d)}$, VLMN$^{(d)}$	**-4.00%**	**-0.83%**	+2.08%	+0.69%	**-1.62%**	**-0.87%**
[6] PESPR$^{(d)}$, VLMN$^{(d)}$	**-1.92%**	**-0.56%**	+0.48%	+1.04%	**-1.12%**	**-1.56%**

Table A.2.6: **Out-of-sample relative performance results for NKE**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of NKE, $\sqrt{252}\text{RV}^{(d)}_{\text{NKE},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.8, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSPE of annualized daily RV predictions – BSX					
	Static			**Dynamic**		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.2297	0.3357	0.2444	0.2435	0.3347	0.2442
[2] $\text{PESPR}^{(\text{d})}$	**-0.18%**	+2.85%	**-0.35%**	+3.07%	+0.87%	**-0.36%**

Table A.2.7: **Out-of-sample relative performance results for BSX**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of BSX, $\sqrt{252}\text{RV}^{(\text{d})}_{\text{BSX},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.9, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).

	RMSPE of annualized daily RV predictions – CVX					
	Static			**Dynamic**		
Model	**2008-2009**	**2010-2011**	**2012-2013**	**2008-2009**	**2010-2011**	**2012-2013**
[1] Benchmark	0.1982	0.2339	0.2357	0.1971	0.2332	0.2326
[3] $\text{PESPR}^{(\text{d})}$	**-0.17%**	**-0.02%**	+0.27%	+0.06%	**-0.34%**	**-0.72%**
[4] $\text{VLMN}^{(\text{d})}$	**-1.10%**	**-0.10%**	**-0.14%**	**-0.18%**	**-0.45%**	**-0.78%**
[5] $\text{PQSPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	**-1.67%**	**-0.61%**	+0.25%	+0.33%	**-0.71%**	**-0.67%**
[7] $\text{PESPR}^{(\text{d})}$, $\text{VLMN}^{(\text{d})}$	**-1.22%**	**-0.09%**	+0.11%	+0.28%	**-0.57%**	**-1.33%**

Table A.2.8: **Out-of-sample relative performance results for CVX**
Root mean square prediction percentage errors (RMSPE) for the annualized daily realized volatility of CVX, $\sqrt{252}\text{RV}^{(\text{d})}_{\text{CVX},t+1}$. Non-benchmark models are referred to by their estimation results index from table 4.11, followed by the list of their additional terms. Their RMSPE is expressed in percentage (%) change with respect to the benchmark's RMSPE, with improvements in bold. Refer to section 4.2 and equation (4.13) for details on the period samples and prediction methodology (static vs. dynamic).