Imperial College London

Imperial College Business School

Dynamic Modelling of Irregular Times, Prices and Volumes at High Frequencies

Nikhil Shenai

A thesis submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy of the Imperial College London and the Diploma of Imperial College London.

This thesis is my own work except where referenced.

ABSTRACT

This thesis undertakes an investigation into time series at high frequency. The three main channels of information in high frequency data - irregular time intervals (durations), prices and volumes - are all explored and modelled to improve current understanding, while accounting for the long memory property, a crucial stylised fact found in the literature. In doing so, we make use of the theory of point processes, econometric techniques such as Whittle estimation and Kalman Filter forecasting, and also sophisticated computing architecture including database systems and programming languages across multiple software environments.

Acknowledgements

First and foremost I would like to thank the Infinite, for giving me the opportunity to do the PhD, my mother, for supporting me through it, my supervisor, Professor Paolo Zaffaroni, for his insights, and my good friend Dr Filip Žikeš, for his invaluable help along the way.

I greatly appreciated useful insights from Professor Clifford Hurvich. Helpful comments from Dr Emma McCoy, Dr Adam Goliński and Professor Richard Davis also benefited me.

I acknowledge useful programming support from Dr Kevin Sheppard, Dr Mengchen Hsieh and Dr Marcelo Perlin, whose programs aided creation of FIACD, LMSD and ACD programs respectively, and the general support of Imperial College Business School.

Finally, I would like to thank my family and friends for being there for me throughout.

Table of Contents

CHAPTER 1: INTRODUCTION AND LITERATURE SURVEY

1.1 Introduction

Time series data has traditionally been assumed to arrive at relatively low frequencies (such as a day) and at regular time intervals. However, in Finance most events (e.g. asset transactions) occur at higher frequencies than previously analysed and at irregular time intervals. As Engle and Russell (1998) argue, new models for high frequency / transaction data and increases in computing power may make it possible to use the extra information in the timing of asset transactions to analyse and predict relevant quantities (such as volatility) better.

Further, as Manganelli (2002) discusses, time is just one of three major channels through which one can gain information, the others being volume and price. For example, in the market microstructure model of Easley and O'Hara (1992), long intervals between trades imply that there is no private information which informed traders can exploit so they do not trade. This also means that volumes will be low (only driven by "noise" traders). Conversely, when volumes are high, dealers may infer that new private information has arisen, leading to them raising prices. More generally, Lo and Wang (2000) argue that "price and quantity" [equivalent to volume] "are the fundamental building blocks of any theory of market interactions", giving further importance to the channels of volume and price.

Research has been conducted on modelling the irregular time process. The set of models proffered are sometimes termed "Time Deformation" models, within which two approaches might be distinguished, 1) the method of random time changes proposed by

Clark (1973) and 2) duration modelling as advocated by Engle and Russell. For now, the latter approach will be followed as it is more relevant to Econometrics. Engle and Russell specified the first model, an ARCH-like process for the conditional mean of the intervals. Bauwens and Veredas (1999) proposed an alternative which mimics a Stochastic Volatility (SV) model, while Jasiak (1999) and Deo, Hsieh and Hurvich (2010) extended both sets of models to allow for long memory. As Engle and Russell acknowledge though, their model only deals with the irregular spacing of transactions, and could be extended to incorporate variables which are associated with the random intervals, such as volumes and prices.

Research has also been done on the price-volume relationship. Karpoff (1987) conducted a review of the early literature. A key idea is the Mixture of Distributions Hypothesis (MDH) proposed in articles by Clark, Epps and Epps (1976), and Tauchen and Pitts (1983), according to which asset prices and trading volumes are driven by the same underlying news process, so that positive correlation between absolute price changes and volume should exist. Apart from Clark's article though, this strand of the literature does not explicitly model the randomness in time.

Overall, the intended contribution of this PhD is to develop further understanding of the three channels of time, price and volume. A key theme will be to understand if long memory exists in the data and can be incorporated into models, in order to make models as general as possible while remaining stationary. The roadmap is as follows. This first chapter contains a survey of the current literature. The second chapter examines the time process via a comparison of two existing long-memory models with an adaptation of Calvet and Fisher's Markov Switching Multifractal model. The third chapter provides an exploration of high frequency volumes and some of the features therein, as little work has been done on modelling volume on its own. The fourth chapter examines the implications of a theoretical high-frequency framework of Hurvich and other authors for fractional cointegration between assets for returns and realized volatility and extends it to volumes. Finally, the fifth chapter concludes and suggests further avenues for research.

1.2 Terminology for Transaction Data

This section introduces terminology associated with transaction data, based on Engle and Russell (1998) and Deo, Hurvich, Soulier and Wang (2009).

1.2.1 Point Processes

The most basic event for an asset is the quote revision. The time at which each revision occurs is called an *event time* or *arrival time* and is a random variable. Let the set of event times be denoted as $\{t_i\}_{i\in\mathbb{Z}} : \ldots < t_{-1} < t_0 \leq 0 < t_1 < t_2 < \ldots$.

The *trading process* is the process formed by the set of trades over time; since the time between each trade event is stochastic, the trading process is a *point process* formed by the trade event times.

Associated with each trade event time are other random variables, called *marks*, such as price, bid-ask spread and volume. These are *dependent point processes* (with respect to the trading process) or *marked point processes*.

1.2.2 Measures of Time

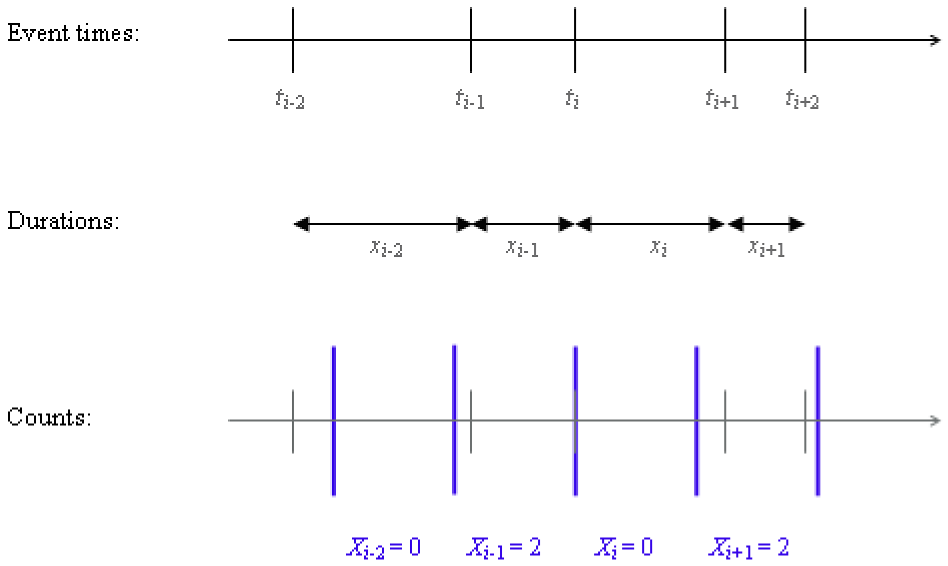
Additional processes can be defined in terms of event times in order to model different aspects of the trading process. A *duration* is the length of the time interval between two events. Let the set of durations be denoted as $\{x_i\}_{i\in\mathbb{Z}} : x_i = t_i - t_{i-1}$. A set of durations after time zero can also be defined as:

$$\{u_i\}_{i\in\mathbb{Z}^+} : u_i = \begin{cases} t_1 & \text{if } i = 1 \\ \\ x_i & \text{if } i \geq 2 \end{cases}$$

The *counting process* is the number of events to have occurred in the time interval $(0, t]$, and is denoted $N(t) = \max\{s : \sum_{i=1}^{s} u_i \leq t\}$. For any fixed time interval $\Delta t > 0$, the *count* or *short-term count* is the number of events in the ith time interval (where $i \in \mathbb{Z}^+$), denoted by $X_i = \Delta N_i = N(i\Delta t) - N((i-1)\Delta t)$. The *aggregated count* or *long-term count* is the number of events in some number of consecutive time intervals, i.e. $\sum_{i=1}^{n} X_i$. An illustration of event times, durations and counts for the same set of events is shown in Figure 1.1 below.

Figure 1.1: Equivalent Event Times, Durations and Counts

This diagram shows how the sets of event times $\{t_i\}$, durations $\{x_i\}$ and counts $\{X_i\}$ relate to each other.



{Note: X_{i-1} also includes an event at t_i on its right-hand boundary}

In fact, event times, durations and counts are equivalent in the sense that knowing one process enables construction of the others (as long as the fixed time interval defining a count can be shortened as much as is necessary to recover individual trades); durations and counts are both derived from the event times. So a common theoretical feature which affects all three quantities is that time is an increasing process. This means that the set of event times is increasing, while durations and counts are nonnegative.

1.3 Stylised Features

In Table 1.1 below we assemble certain data features found in the literature on all three channels - since these are not all universally agreed, we term these “Stylised Features” as opposed to “Stylised Facts”. More detail is provided in the following subsection. Note also that Appendix A contains verification of the stylised features for times (SF1, 2, 3.1-2, 4.1 and 5.1) based on IBM data from the Trades, Orders, Reports, and Quotes (TORQ) dataset used by Engle in 2000 (obtained as ”uhf.zip” from Engle’s website), and also foreign exchange data from EBS Dealing Resources, provided by Imperial College Business School, from 1997 to 2007, for the EUR/USD exchange rate.

Table 1.1: Stylised Features for Times, Prices and Volumes

Stylised Features on the major channels of information, categorised by specific measure. E.g. Event Times, Durations and Counts are separate measures related to the Time channel, and have specific data features associated with them.

Time	**Price**	**Volume**
Event Times	**Levels**	**Levels**
Irregular Spacing	Trend (I(1))	Trend (I($\geq$ 1))
Discreteness	Discreteness	Discreteness
Durations	**Returns**	
Clustering	Clustering	Clustering
Overdispersion	Leptokurtosis, Skewness	Skewness
Seasonality	Seasonality	Seasonality
Counts	**Volatility**	
Clustering - high-order in short term, lag-1 in long term	Clustering	
	Leverage	

Ideally we would wish to fit all stylised features, but this may prove to be too hard simultaneously. Since irregular spacing in time also underlies prices and volumes, it is a key feature to be modelled. Discreteness of variables, clustering and seasonality also exist in all three channels. However, as pointed out below, discreteness of variables is becoming less important as markets become more sophisticated, while we can adjust the raw data for seasonality after creating a core model. So the focus will be on irregular spacing in time and clustering. If possible without significant additional complexity, other stylised features will also be accounted for.

1.3.1 Detail on Stylised Features

Models for transaction data should be able to incorporate the following stylised features on times, prices and volumes.

SF1: ***Irregularly spaced events*****:** As mentioned before, this is the inherent property of transaction data; in general, marks can occur at any time, so durations and counts are not constant and may contain information.

SF2: ***Discreteness of variables*****:** At the level of individual trades, variables can no longer change by any amount. In particular:

SF2.1: ***Discreteness in times*****:** Event times are discrete as the timestamp of each is limited by data recording ability. Usually each trade is accurate to one second leading to the observation that multiple transactions may occur in a second, even at different prices. Equivalently, many zero-duration trades may exist at different prices. Pacurar (2006) cites this as a data issue with no standard method of resolution agreed. However, event times are being recorded increasingly more accurately; as Ng (2008) shows, some are already accurate to hundredths of seconds. Over time the issue should reduce.

SF2.2: ***Discreteness in marks*****:** Prices and volumes can only change by discrete amounts. Prices can only change by multiples of tick size, such as $0.01 on the NYSE. This means that returns are also discretely-valued, though the error from assuming continuity is less for higher-valued shares (e.g. IBM is valued at over $100, so the error is less than 0.01%). Hurvich and Wang (2010) argue that this indicates that the price process has a pure jump form (as opposed to diffusion or jump-diffusion forms). Volumes are also discrete; the smallest volume is currently 1 contract.

SF3: ***Clustering*:** Significant high-order correlation occurs in the observations of times and marks. This may be evidence of long memory, as will be detailed in Section 1.4.2. Specifically:

SF3.1: ***Duration Clustering*:** Engle and Russell (1998) found significant positive autocorrelations and partial autocorrelations for lags of up to order 15 in the durations from the Trades, Orders, Reports and Quotes (TORQ) dataset, suggesting some variant of a mixed time series model (e.g. ARMA) could model the durations. Deo, Hsieh and Hurvich (2010) also noted positive autocorrelation for the durations for all lags upto high orders ($\sim$500).

SF3.2: ***Autocorrelation in counts and aggregated counts*:** Deo, Hsieh and Hurvich show that counts for time intervals of 5 minutes and 30 minutes display significant positive correlation for lags of upto high orders ($\sim$500 for the former and $\sim$100 for the latter). They also show that counts for long time intervals such as 5 days have significant positive lag-1 correlation, as does realized volatility for corresponding intervals.

SF3.3: ***Clustering in Squared Returns and Volatility*:** Large price changes seem to come in bulks - leading to time periods when share return volatility is relatively low and other periods when it is high. Deo, Hsieh and Hurvich found high-order autocorrelation for squared returns and volatility similar to that for durations.

SF3.4: ***Clustering in Trading Volumes*:** Lo and Wang (2000) found significant sample correlations of upto order 10 for the portfolio turnover ratios they used as measures of trading volume. Bollerslev and Jubinsky also argue for clustering in volumes, citing evidence from Tauchen, Zhang and Liu (1996) in the form of semi-nonparametric density estimates.

SF4: ***Seasonality*****:** Seasonality also affects both times and marks:

SF4.1: ***Diurnal pattern of durations*****:** Tsay (2005) argues that the intensity of trades exhibits a U-shaped pattern every day, with trading heaviest at the beginning and the end of the day, and lightest during the lunch hour. So as Engle and Russell (1998) show using IBM data from the TORQ dataset, durations have a daily inverted U-shaped pattern.

SF4.2: ***Seasonality in volumes*****:** Using hourly volume data from the NYSE, Jain and Joh (1988) found that average trading volume also exhibits a U-shaped pattern during each day, while over the week, average trading volume displays an inverted U-shaped pattern. F-tests of the null hypothesis that the trading volumes did not show significant variation over days and weeks were rejected at the 1% significance level.

SF4.3: ***Seasonality in returns*****:** Jain and Joh did not find as clear-cut a pattern in hourly returns from the same dataset; they found significant evidence of variation of returns within days, and in particular, that the lowest average return occurred in the fifth trading hour, while the first and last trading hours featured relatively high returns. They also found evidence for a weekend effect - a tendency for the average return of the first hour of Mondays to be significantly negative and much larger in absolute terms than returns over the next five hours.

SF5: ***Distributional Features*****:** The distributions of times, prices and volumes have the following features:

SF5.1: ***Overdispersion of durations*****:** Engle and Russell (1998) argue that overdispersion often characterises duration data: the unconditional standard deviation of durations is greater than the unconditional mean. This has implications for the correct specification of the distribution of durations; distributions with standard deviation equal to mean ("equi-dispersion"), such as the exponential distribution will not fit unconditional durations, though may still be valid for conditional durations.

SF5.2: ***Leptokurtosis of returns*****:** The empirical distribution of logged share returns is observed to exhibit fat tails; there is a higher probability of a very high or very low return than would be expected using a normal distribution. At the same time, the curvature of the distribution is greater than under normality, resulting in a higher peak.

SF5.3: ***Skewness of returns*****:** According to Rydberg (2000), there is evidence that the distribution of returns is slightly negatively skewed, with the possible explanation that traders react more strongly to negative information; for instance, on Oct 19th, 1987, there was a single large crash in the share market, which required many smaller increases to get back to the previous level. Carr and Wu (2003) also support negative skewness, though Press (1967) and Kon (1984) found positive skewness. In either case, it appears that there is some asymmetry in the return distribution.

SF5.4: ***Leverage in returns*****:** Share price movements have been observed to be negatively correlated with volatility. According to Christie (1982), a rationale consistent with the evidence is that as a firm's income falls, its shareholder value falls, and at the same time, due to fixed costs (e.g. interest payments), its financial leverage rises, increasing its riskiness.

SF6: ***Trendedness of marks*** Prices and volumes exhibit the following trends:

SF6.1: ***Prices are I(1)*** If prices are martingales, their drift and variance should increase linearly, so that they are I(1).

SF6.2: ***Volumes are I($\geq$1)*** Bollerslev and Jubinsky (1999) assert that volumes are I(1) and extract a linear trend from their observations of volume. However, Lobato and Velasco (2000) found that volumes can have polynomial trends of higher order so proposed a data taper which removed trends of up to order 2 from their data.

1.4 Modelling Approaches

In this section detail is given on the current approaches to modelling the key stylised features - irregular spacing in time and clustering. In addition, a mechanism by which clustering can be transmitted from time to returns is introduced, as well as a possible statistical relationship between prices and volumes.

1.4.1 Irregular Time

Models of the time process which can feature irregularly spaced events typically specify a process for the durations as opposed to counts and event times. Pacurar (2006) provides an extensive survey of alternative duration models, although she is missing the LMSD process which will be detailed later. She essentially classifies models into 5 categories (using slightly different terminology) - 1) Baseline Duration models, 2) Latent Factor models, 3) Long Memory models, 4) Logarithmic models and 5) Regime-Switching models. In general, all the models can fit SF3.1 (duration clustering) and SF5.1 (duration overdispersion). Currently, Long Memory models explicitly model and fit SF3.2 (autocorrelation in counts and aggregated counts) as well, although Regime-Switching models may be able to fit SF3.2. The focus will be on Long Memory models. These are based on the first two types of models, so examples of these models will also be given.

1.4.1.1 Baseline Duration Models

Engle and Russell (1998) proposed the Autoregressive Conditional Duration (ACD) model, in which they suggest that the durations, x_t, follow a process such that:

$$x_t = \psi_t \varepsilon_t \qquad \varepsilon_t \sim IID(1, \sigma_\varepsilon^2) \tag{1.1}$$

$$\psi_t = \omega + \sum_{i=1}^{m} \alpha_i x_{t-i} + \sum_{i=1}^{q} \beta_i \psi_{t-i} \qquad \varepsilon_t \text{ and } \psi_t \text{ are independent} \tag{1.2}$$

The ACD model has order (m, q), abbreviated to ACD(m, q). ψ_t is the conditional duration, the conditional mean of x_t i.e. $E_{t-1}(x_t) = \psi_t$. ε_t is the standardised duration, $\frac{x_t}{\psi_t}$, and must have a distribution which ensures it is always positive, e.g. the Exponential distribution. ω, α_i and β_i are constants. Sufficient conditions for positive durations are that $\omega > 0$, $\alpha_i \geq 0$ and $\beta_i \geq 0$. Weak stationarity is guaranteed by $\sum_{i=1}^{m} \alpha + \sum_{i=1}^{q} \beta < 1$.

Overall, the model specification is similar to a GARCH model, except that the conditional mean is being modelled as opposed to the conditional volatility.

The first equation shows that the standardised duration is expected to be 1. The second equation shows that there is persistence in the conditional mean; the x_t term allows for limited m-term memory in the conditional mean, while the ψ_t term allows for longer-term memory. The combination enables duration clustering to be observed. To derive the autocorrelation function (ACF) of the process, a further distributional assumption is required so that the unconditional variance can be calculated.

Engle and Russell first proposed that ε_t is Exponentially distributed, but later versions of the ACD also use Generalised Gamma, Weibull and Burr distributions. Taking the Exponential case, the model is called the EACD(1,1) model, and its unconditional variance can be shown to be:

$$\begin{aligned} Var(x_t) = \sigma^2 &= \frac{1-\beta^2-2\alpha\beta}{1-2\alpha^2-\beta^2-2\alpha\beta}\mu^2 \qquad \text{where } \mu \text{ is the unconditional mean} \\ \Leftrightarrow \frac{\sigma^2}{\mu^2} &= \frac{1-\beta^2-2\alpha\beta}{1-2\alpha^2-\beta^2-2\alpha\beta} \end{aligned}$$

Since $\frac{\sigma}{\mu} > 1$ when $\alpha > 0$, the model can allow for SF5.1 (overdispersion of durations). Given the unconditional variance, it can then be shown that positive autocorrelation exists if $\alpha + \beta \geq 0$, hence the model can fit SF3.1 (duration clustering).

1.4.1.2 Latent Factor Models

Bauwens and Veredas (1999) proposed the the Stochastic Conditional Duration (SCD) model:

$$x_t = \varepsilon_t e^{\psi_t} \qquad \varepsilon_t \sim IID(\mu_\varepsilon, \sigma_\varepsilon^2) \qquad E(\varepsilon_t^r) := g_r \tag{1.3}$$

$$\psi_t = \omega + \beta\psi_{t-1} + u_t \qquad u_t \sim NID(0, \sigma_u^2) \qquad \varepsilon_t \text{ and } u_t \text{ are independent} \tag{1.4}$$

Here ω and β are constants. Unlike the ACD model, no conditions on parameters are required to ensure positive durations, since the exponential term is always positive for a real exponent. Also, weak stationarity is guaranteed as long as β is less than 1, which is a simpler condition than for the ACD model.

While the ACD has only one, observable random variable driving the system dynamics, the SCD model has an observable random variable driving the observed duration and a latent random variable, u_t, driving the conditional duration (now e^{ψ_t}) via an AR(1) process. The extra random variable enables richer dynamics as will be shown. Bauwens and

Veredas argue that it can model the random flow of information / the news process.

In the same way that ACD models are similar to GARCH models, SCD models are similar to another class of volatility models - SV (Stochastic Volatility) models. Now clustering is generated through the AR(1) process in the (logged) conditional duration equation.

The unconditional variance is given by:

$$Var(x_t) = \sigma_u^2 = \mu^2(\nu e^{\frac{\sigma_u^2}{1-\beta^2}} - 1) \text{ with } \nu := \frac{g_2}{g_1^2}$$
$$\Leftrightarrow \frac{\sigma_u^2}{\mu^2} = \nu e^{\frac{\sigma_u^2}{1-\beta^2}} - 1$$

As can be seen by the expression for the variance, the SCD model enables SF5.1 (duration overdispersion). For instance, if ε_t follows the Weibull distribution with shape parameter κ, a sufficient (but not necessary) condition is that $\kappa \leq 1$. Further detail on the conditions is given in Bauwens and Veredas' article.

Interestingly, Bauwens and Veredas point out that the parameters governing dispersion (σ) and persistence (β) are separated under the SCD model, whereas they are the same in the ACD model ($\alpha + \beta$), so enabling the SCD model to fit a greater variety of persistence-dispersion profiles.

1.4.2 Long Memory Models

1.4.2.1 The Long Memory Property

As discussed before, the ACD and SCD models are able to generate duration clustering by enabling positive serially correlated ACFs. Both their ACFs fade exponentially to zero / are bounded geometrically, i.e. $|\rho(\tau)| \leq Cr^{\tau}$ as τ increases (C and r are some constants). Let such processes be called *short memory* processes. There is evidence that perhaps a slower rate of decay, such as hyperbolic, is required. Such decay is shown by *long memory* processes.

Brockwell and Davis (1991) define long memory by the condition:

$$\lim_{\tau \to \infty} \rho(\tau) \sim C\tau^{2d-1}$$

where $\sim$ denotes asymptotic equivalence, C is a constant, and $d \in (0, \frac{1}{2})$ is known as the memory parameter. So long memory describes the set of stationary processes between a short memory time series, whose ACF always decays to zero (at least asymptotically), and an integrated (I(1)) time series, whose ACF does not die down.

Note that short memory processes have $d = 0$. Just as an I(1) series can be made into a short memory series by first differencing, a long memory process $\{Y_t\}$ can be converted into a short memory process by differencing to the fractional power d; $(1 - L)^d Y_t$ is now a short memory process, where $L(.)$ is the lag operator.

Brockwell and Davis also state that while a short memory process such as the ACD model can approximate long memory through suitably high orders of m and q, the orders required will be so large that parameter estimation will become difficult. Beran includes an alternate definition of long memory processes:

$$\lim_{n\to\infty} Var(\sum_{i=1}^{n} Y_i) \sim Cn^{2d+1}$$

Intuitively, both characterisations are similar in that they both specify the asymptotic decay of the stochastic process and both look at the combined influence of all observations. Extending to more than one stochastic process, two (or more) long memory series are fractionally cointegrated if there exists a linear combination of the series which reduces the memory parameter of the combined series to a level below the memory parameter of any of the original series.

1.4.2.2 Long Memory Duration Models

As for SF3.1 (duration clustering), Deo, Hsieh and Hurvich (2010) found significant correlation between distant durations. Furthermore, they tested for long memory in durations and counts using the Geweke-Porter-Hudak (GPH) estimator proposed by Geweke and Porter-Hudak (1983), which makes use of log-periodogram regression, and found significant evidence to support the null hypothesis of long memory.

Jasiak (1999) and Deo, Hsieh and Hurvich have adapted long memory volatility models to the duration setting to create the Fractionally Integrated ACD (FIACD) model and the Long Memory Stochastic Duration (LMSD) model respectively.

The FIACD Model

The ACD(m,q) model (equation (1.4)) can be expressed as:

$$\begin{aligned} \psi_t &= \omega + \alpha(L)x_t + \beta(L)\psi_t \\ [1 - \underbrace{(\alpha(L) + \beta(L))}_{:=\phi(L)}]x_t &= \omega + [1 - \beta(L)]\eta_t \qquad\qquad ; \eta_t = x_t - \psi_t \\ [1 - \phi(L)]x_t &= \omega + \beta(L)\eta_t \end{aligned}$$

where $\alpha(L)$ and $\beta(L)$ are the AR and MA polynomials respectively and $\phi(L)$ is their sum; $\alpha(L) = \sum\limits_{i=1}^{m} \alpha_i L^i$, $\beta(L) = \sum\limits_{i=1}^{q} \beta_i L^i$ and $\phi(L) = \sum\limits_{i=1}^{\max(m,q)} \phi_i L^i$.

To incorporate long memory, the durations are fractionally differenced using the long memory parameter d. Further rearranging:

$$
\begin{aligned}
&[1-\phi(L)](1-L)^d x_t = \omega + \beta(L)\eta_t \\
&\qquad [1-\beta(L)]\psi_t = \omega + \underbrace{[1-\beta(L)-[1-\phi(L)](1-L)^d]}_{:=\lambda(L)} x_t \\
&\qquad\qquad \psi_t = \underbrace{[1-\beta(1)]^{-1}\omega}_{:=\varpi} + \underbrace{[1-\beta(L)]^{-1}[1-\beta(L)+[1-\phi(L)](1-L)^d]}_{:=c(L)} x_t \\
&\qquad\qquad \psi_t = \varpi + c(L)x_t
\end{aligned}
$$

where $(1-L)^d = \sum\limits_{i=0}^{\infty} \pi_i L^i$ with $\pi_i = F(-d,1;1;1) = \begin{cases} 1 & i=0 \\ \prod\limits_{k=1}^{i} \frac{k-1-d}{k} & i>0 \end{cases}$

$F(a,b;c;z)$ is the Hypergeometric function as in e.g. Karanasos et al. (2004)

$\lambda(L)$ and $c(L)$ are infinite order lag polynomials; $\lambda(L) = \sum\limits_{i=1}^{\infty} \lambda_i L^i$ and $c(L) = \sum\limits_{i=1}^{\infty} c_i L^i$.

More persistence now exists because the conditional duration equation - equation (1.4) - has changed from an ARMA-type structure to an ARFIMA-type structure. In addition, Jasiak indicates that the unconditional mean is infinite, leading to heavy-tailed durations. Note that it is not proved that a stationary FIACD model exists for all values of d; so far only Randall, Roueff and Soulier (2008) have proved the existence of a strictly stationary and causal FIGARCH and hence FIACD. They showed this specifically for values of d between 0 and 1, but not close to 0; as d becomes small, the AR weights used in their proof become too large, causing failure of the existence condition they derive. The intuition is that the proof relies on casting a FIGARCH process as an IARCH(∞) process, so as d decreases from 1, the FIGARCH decays faster and is further away from being integrated. However, Randall, Roueff and Soulier also remark that it is not proven that FIGARCH models cannot exist for d close to 0, just that it has not yet been proven.

The LMSD Model

In the LMSD process, the (logged) conditional duration equation is replaced with:

$$\begin{aligned} \psi_t &= \omega + (1-L)^{-d} u_t \\ \Leftrightarrow \psi_t &= \omega + \sum_{i=0}^{\infty} b_i u_{t-i} \qquad\qquad ; b_i \sim C i^{d-1} \end{aligned}$$

Here there is more persistence because the (logged) conditional duration equation has changed from an AR(1) process to an infinite-order ARIMA process with slowly decaying coefficients.

To conclude, both the FIACD and LMSD models encompass the ACD and SCD models so they can incorporate SF3.1 and SF5.1 (duration clustering and overdispersion respectively). However, both can also fit SF3.2 (autocorrelation in counts and aggregated counts). Hurvich, Deo, Soulier and Wang (2009) have shown that long memory in durations suffices to generate 1) long memory in (short-term) counts (in their Theorem 2), and 2) lag-1 autocorrelation in long-term counts (in their Theorem 4). Overall then, long memory models can fit most of the stylised features of times reasonably.

1.4.2.3 Applications and Alternatives

Moving beyond times, Baillie, Bollerslev and Mikkelsen (1996) and Breidt, Crato and de Lima (1998) found evidence for long memory in the squared returns of exchange rates and market indices, supporting the idea of long memory in squared returns and return volatility. With respect to volumes, Bollerslev and Jubinsky (1999) found long memory around a linear trend in daily turnover ratios and in absolute returns (without a trend) for stocks in the S&P100 index, while Lobato and Velasco (2000) found evidence that log volumes of the Dow Jones Index have long memory after correcting for nonstationarity by tapering the data (an alternative to detrending which allows for nonlinear trends). Furthermore, Bollerslev and Jubinsky suggest that imposing a common long-memory component for volumes and volatility may improve long-term volatility forecasts and the accuracy of long-term financial contracts. So we require a framework which allows for long memory in times, prices and volumes.

Note that one possible objection exists to modelling variables as having long memory. Other modelling features can generate high order autocorrelation, such as structural breaks or regime-switching. These features are exploited in the formulation of Calvet and Fisher's formulation of their Markov Switching Multifractal model (2004), which was specified as an alternative to GARCH and SV models, and can exhibit hyperbolic decay upto a high order, mimicking long memory.

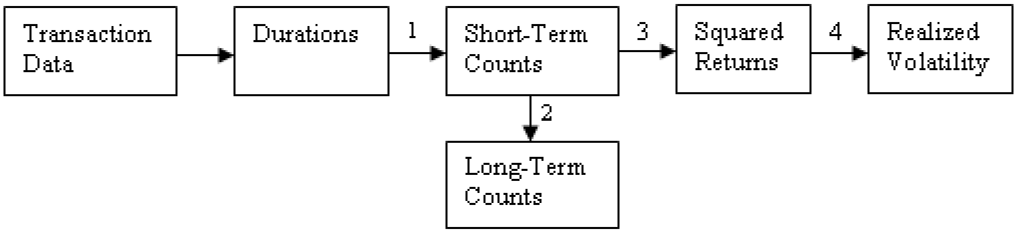
As Fleming and Kirby (2001) point out though with respect to return volatility, two counterarguments exist. First, even if structural breaks exist, long memory still models correlation dynamics because it enables the unconditional variance to change over time slowly. Second, since forecasting structural breaks is difficult, long memory enables the best volatility forecasts.

In addition, as Granger (2001) and Zaffaroni (2004) have shown, it is possible for many short memory relationships to aggregate to a univariate long memory series. So as Andersen and Bollerslev (1996) suggest, long memory in times, prices and volumes may arise from the interaction of a large number of underlying information processes. For now we avoid modelling these underlying processes, but bear them in mind for possible future research.

1.4.3 Extending from Time to Prices

Having determined a set of transaction models which can model times, it is desirable to extend them to explain prices and volumes. A series of papers has proposed one mechanism, by which durations from a LMSD process can influence prices, as in Figure 1.2 below.

Figure 1.2: A Dependence Transmission Mechanism
The same value of memory parameter can propagate from durations to realized volatility.



Following arrow 1, Deo, Hurvich, Soulier and Wang (2009) specify general conditions under which memory transmits from a durations process to short-term counts in their first theorem. They then compare two examples of durations processes - the LMSD process and the ACD(1,1) process. They show that the LMSD and ACD(1,1) processes are able to satisfy the general conditions specified by the first theorem, but while the LMSD process is able to transmit long memory from durations to counts, the ACD(1,1) is a short memory process so induced counts only have short memory. Finally, long memory in the short term counts is shown to yield lag-1 autocorrelation in long-term counts, no matter how large the time interval (Arrow 2).

In fact, Hurvich et al. do not consider the FIACD process for arrows 1 and 2, but their proof does not depend on a specific process, but instead on long memory, so at least arrow 2 can be extended to the FIACD process. It may also be extended to infinite-variance ACD models (as in Daley, Rolski and Vesilo (2000)); Hurvich et al. note that fat tails in the durations process is another means of generating long memory in counts.

Next, Hurvich et al. show how dependence transfers from counts over any time interval (short and long-term) to squared returns over the same interval (Arrow 3 - note that since short-term counts are more general than long-term, only one arrow has been drawn, from the former). These sum to realized volatility (Arrow 4). Interestingly, the same memory parameter propagates through the whole transmission mechanism.

Finally, Hurvich and Wang (2010) have extended the mechanism to the multivariate case by considering two assets at the same time. Hurvich and Wang incorporate SF2.2 (discreteness in marks) in a sense by defining their price processes to have pure jumps. They also retain generality by allowing for cointegration between their transaction-level data even if the sequences are asynchronous; if both assets have long memory, there is the possibility that they share a (fractional) cointegrating relationship if there is reason to believe that there is an underlying economic relationship. If a stable cointegrating relationship exists, then analysis of assets in pairs (or more generally, groups) may be a way to predict volatility better.

1.4.4 Extending to Volumes

Volumes are generally modelled in relation to prices, with the prevailing view being the Mixture of Distributions Hypothesis (MDH). However, some recent research has modelled volume as a process in its own right. We first introduce the MDH before proceeding to the univariate volume analysis.

1.4.4.1 The MDH

As mentioned above in SF5.2-3, the distribution of returns is relatively more peaked than the Normal distribution, with greater mass in the tails, and can exhibit more skewness. Two prominent alternative hypotheses can explain this stylised feature - the stable Paretian hypothesis of Mandelbrot (1963) or the Mixture of Distributions Hypothesis of Press (1967) (also attributed to Clark (1973), Epps and Epps (1976) and Tauchen and Pitts (1983)).

Mandelbrot observed leptokurtosis in commodity returns in 1963, and advanced the general class of stable Paretian (Lévy) distributions as an alternative to the Normal distribution (which is a special case where the tail index parameter $a = 2$). Fama (1965) rejected normality in his tests and found more evidence in support of Mandelbrot's hypothesis. However, some Paretian distributions suffer from having no finite moments of higher order than one. Also, their parameters are not easily estimable.

The alternative hypothesis is that the leptokurtosis arises because data are generated by a mixture of distributions that have different conditional variances. Press suggested a Poisson mixture of Normal distributions which could yield leptokurtosis and skewness, and by plotting theoretical against empirical cumulative distribution functions, found that such a distribution fitted some of the data well.

The MDH can also be more easily extended to jointly model prices and volumes. As Bollerslev and Jubinsky (1999) argue, prices and volumes move together because they are driven by the same underlying information or news-arrival variable. So rather than volume being a proxy for the rate of information flow, as in Clark (1973), it is also driven by a distinct underlying information variable. Then the arrival of unexpected good news results in a price increase and a volume increase, whereas bad news results in a price decrease and a volume increase. Bollerslev and Jubinsky conclude that absolute returns or volatility and trading volume should be positively correlated. A simple specifcation exists in Tauchen and Pitts which shows that squared price changes and volume should be positively correlated. Harris (1987) shows that other theoretical relationships between price and volume should possibly exist; price change kurtosis, volume skewness and correlation of squared price change with volume should all be positively correlated across securities.

The evidence in favour of the MDH has been mixed. According to Bollerslev and Jubinsky, the first set of tests of the MDH, such as those by Harris supported it, while later studies by Lamoureux and Lastrapes (1994) and Richardson and Smith (1994) found contradictory evidence; these latter studies indicated that the underlying information-arrival process is unable to describe short-run dependencies in volatility and volumes. However, Andersen (1996) found evidence in favour of a modified MDH, in which he 1) modelled the volume process as conditionally Poisson as opposed to Normal as previously assumed, and 2) included a constant term in the volume process to account for a constant level of noise or liquidity trading (in addition to random, information-driven trading). In fact, Bollerslev and Jubinsky's work builds on Andersen's paper by using the same specification and then studying the long-run correlation relationships between prices and volumes.

Furthermore, Bollerslev and Jubinsky conclude that the contradictory evidence (e.g. of

Lamoureux and Lastrapes) may reflect the fact that the MDH is a long-run relationship driven by an unobserved news arrival process with long memory. They therefore suggest a modified version of the MDH, which can be extended to show that squared returns and volumes should feature a cross-correlation structure similar to long memory. In fact, Bollerslev and Jubinsky found significant evidence that the value of long memory was the same between squared returns and volumes, and suggested that this may indicate the two series are fractionally cointegrated, although Lobato and Velasco (2000) rejected this latter hypothesis.

1.4.4.2 Univariate Volume Analysis

In contrast, Lo and Wang (2000) studied trading volumes on their own, providing definitions, data analysis and implications of the CAPM for volume turnover (volume divided by the number of shares outstanding); principally, page 268 argues that under the one-fund theorem (terminology as more widely used in e.g. Luenberger (1998), Lo and Wang themselves call the one-fund theorem "two-fund separation"), where investors only hold the market portfolio [1] (comprised of many stocks) and the risk-free asset, turnover should be identical for all stocks. However, if the one-fund theorem fails, resulting in an equilibrium with say K efficient funds (comprised of many more stocks), investors will have positions in the market portfolio, the risk-free asset and $K-1$ hedging portfolios. Then Lo and Wang argue on page 270 that turnover should follow an approximate K-factor structure. Lo and Wang (2001) extended the analysis to a dynamic setting, while they also explored the effect of transaction costs in Lo, Mamaysky & Wang (2001).

Darolles and le Fol (2003) decomposed volumes into market and specific components,

[1] Under the one-fund theorem, the efficient frontier, the set of all combinations of stocks with minimal variance for a given expected return, is a straight line passing from the risk-free asset and tangent to the minimum-variance frontier of stock combinations at a single point. This point is the market portfolio.

while Bialkowski, Darolles and le Fol (2008) showed that conventional time series models such as ARMA and SETAR can be applied to volume data. However, apart from these contributions, we know of no further econometric work on volumes as a process detached from prices, meaning the scope for further research is vast.

1.5 Research Gaps

Gaps in the literature and research required to investigate them further are suggested in this section. Ideally we would like to explore all of these, but have prioritised on three areas. On the theoretical side, many of the research gaps rest on having a theoretical joint model of times, prices and volumes, so we attempt to derive one in Chapter 4. We build up to this model by investigating the empirical side; there we find that there is a lack of comparison of alternative duration models, and also a lack of research on the volume process. We address these needs in Chapters 2 and 3 respectively. Further avenues for research can build on these chapters and are explored in Chapter 5 along with our conclusions.

1.5.1 Theory

Engle and Russell (1998) focussed on specifying the irregular time process, but at the end of their paper, advocated extending the analysis to marks through joint modelling of marks and events. Hurvich and Wang (2010) provided a model for irregular times and prices over more than one asset, but did not account for possible information in volumes. Finally, various authors have examined relationships between price and volume, but have not systematically included an explicit analysis of the irregular time process, or extended to multiple assets. A synthesis of all three strands - a joint model of times, prices and volumes across multiple assets does not exist yet. This is the aim of Chapter 4.

Refining to the price dimension, it also remains to be investigated whether an explicit model for volatility can be specified making use of Hurvich et al.'s long memory analysis. Without considering long memory, Engle (2000) suggested an ACD-GARCH framework as one approach; for instance, one version simply adds the reciprocal of durations into the conditional variance equation of a GARCH model. This principle might be extended to create a FIACD-FIGARCH or a LMSD-LMSV framework. However, such frameworks assume Granger causality from durations to volatility. It is also possible that durations are endogenous in the sense that the same information process may drive both durations and volatility - according to Renault and Werker (2011), this possibility is called "instantaneous causality" (following Pierce and Haugh (1977)). In fact, Renault and Werker found significant evidence that both effects exist, albeit in a continuous-time setting. They also state though, that it is not possible to disentangle the two causality effects in discrete-time. Therefore, further research into a long memory duration-volatility framework (e.g. LMSD-LMSV) would need to develop ways of accounting for the possible endogeneity of durations, such as lagging or instrumental variables - otherwise estimates may be biased and inconsistent.

Beyond a baseline model of times, prices and volumes, research can also be conducted to determine what derives from the model and what underlies it; what are the applications of irregular spacing and long memory in returns and volumes, and which processes drive the spacing of transaction data and its long memory?

In terms of the first question, Hurvich and Wang suggest their framework in particular can be extended to analyse the relationship between cointegration from a transaction level and option pricing, hedging, pairs trading and index tracking; this has not been done before based on pure-jump processes.

Addressing the second question, Deo, Hsieh and Hurvich (2010) suggest that an unobservable news arrival process underlies transaction data, and that this might be modelled as a duration process. Bauwens and Veredas (1999) make a similar suggestion in their creation of the SCD model. In fact, as mentioned before with reference to Granger (2001) and Zaffaroni (2004), the aggregation of many such news arrival processes may be driving the long memory in transaction data. Alternatively, Liesenfeld (2001) proposes that agents' (traders') sensitivity to information may also drive transaction dynamics, and allows for time inhomogeneity in the news arrival and news sensitivity processes.

Alternatives to a baseline model of times, prices and volumes may also be developed. For instance, there are alternatives to Hurvich et al.'s transmission mechanism. First, infinite variance duration processes as per Daley, Rolski and Vesilo (2000) can generate long memory. Second, Engle and Russell suggest that the duration process can be incorporated into the price process of an asset by making the latter a binomial tree with random spacings between nodes, with specific applications possible in derivative pricing and interest rate term structure modelling; Prigent, Renault and Scaillet (2000) have built on this idea for option pricing. It would be necessary to determine whether the latter approach can incorporate long memory and SF3.2 (autocorrelation in counts and aggregated counts) though.

Linking to another strand of literature, Engle and Russell state that transaction models fit within the wider setting of Time Deformation models (Bauwens and Veredas term the ACD and SCD models "accelerated time" models). Time as it is usually viewed flows regularly or evenly (1s, 2s, 3s, etc.) - this view is called "calendar / clock time". There is another concept of time though - "tick time", the sequence of event times, which increases randomly. In Time Deformation models a third concept of time exists - "business time" - which is related to the amount of business/trading activity in a given interval of calendar time. So business time speeds up and slows down as activity increases or decreases. In this manner, constant volatility in business time is able to generate stochastic volatility in calendar time. Further analysis could be conducted to assess the similarities and differences between tick time and business time, and how transaction models fit with continuous time models such as Heston's SV model, as presented by Carr and Wu (2002).

Finally, estimation techniques could be investigated further with respect to the impact of the point processes used to model irregular spacing in time. Hurvich and Wang estimate their cointegrating parameter using OLS, but admit that OLS only provides consistent estimates. They also suggest that semiparametric estimators could be considered. At the same time Engle and Russell indicate that running OLS on a point process will produce heteroskedastic residuals, so it would be useful to study the effects of a wrong assumption of fixed time intervals between events on estimation.

1.5.2 Empirics

The stylised features presented in this chapter have not been assembled together before, so a verification of all of them for the same dataset would be useful. This would require collection of a transactional dataset and creation of (computing) methods to calculate durations and counts over any interval. An attempt was made in Appendix A of this chapter, focussing on the stylised features relevant to the time process (SF1, 2, 3.1-2, 4.1 and 5.1).

Pacurar (2006) notes that there has been no systematic empirical comparison of all duration models on the same dataset; usually articles simply compare a new model specification with the baseline ACD model. So a comparison between the specification properties of the ACD, FIACD and LMSD processes would be original and would span the Long Memory model space. In addition, noting the alternative modelling approach to long memory in the form of regime shifts, an adaptation of the Markov Switching Multifractal model of Calvet and Fisher (2008) would enable further useful comparison. Finally, forecasting performance is not compared as standard in the existing literature (the initial articles which propose the ACD, FIACD and LMSD models do not include or outline forecasting), so analysis via measures such as Mean Square Error would also be important. This is the aim of Chapter 2.

Measurement of the long memory properties of durations, counts, squared returns, realized volatilities and trading volumes is required to test Hurvich et al.'s transmission mechanism and to see if a similar one exists for volumes. Cointegration vectors could also be tested.

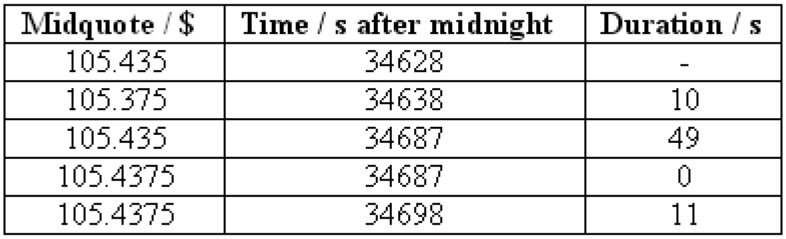
However, for our final chapter (Chapter 3), we wish to focus on investigating the volume process. Volume is almost invariably investigated as part of a joint relationship with prices or volatility e.g. as per the MDH - however it is possible that such a relationship is more complex than as indicated in Karpoff (1987) if it exists at all. So a stronger univariate understanding of volume properties may enable better modelling of volumes in the long run. We propose to conduct research using simple time series models as in Bialkowski, Darolles and le Fol (2008), but allowing for long memory, so implementing ARFIMA as opposed to ARMA models.

APPENDIX A: RESULTS ON STYLISED FEATURES

SF1:

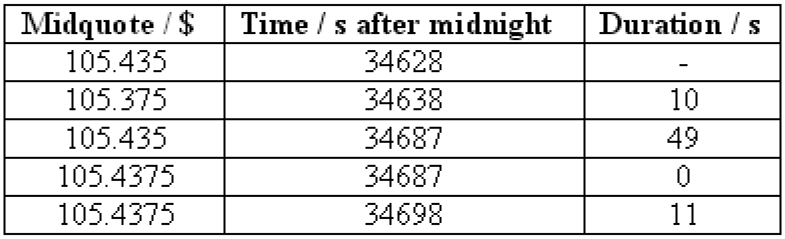
Tables A.1 and A.2 below show typical sequential transaction data. Since durations are not constant in either table, the data is irregularly spaced.

Table A.1: Sequential IBM Data



Midquote / $	Time / s after midnight	Duration / s
105.435	34628	-
105.375	34638	10
105.435	34687	49
105.4375	34687	0
105.4375	34698	11

Table A.2: Sequential EUR/USD Data



Midquote / $	Time / s after midnight	Duration / s
105.435	34628	-
105.375	34638	10
105.435	34687	49
105.4375	34687	0
105.4375	34698	11

SF2:

Tables A.1 and A.2 also show that the time measured for each transaction is discrete (to the nearest second). The IBM midquotes and EUR/USD bid and ask prices also move discretely.

SF4.1:

Figures A.1 and A.2 are histograms of the trades over the whole dataset against the time of day. As expected, the IBM histogram (Figure A.1) exhibits a clear U-shaped pattern.

Figure A.1: Diurnal Pattern for IBM Data

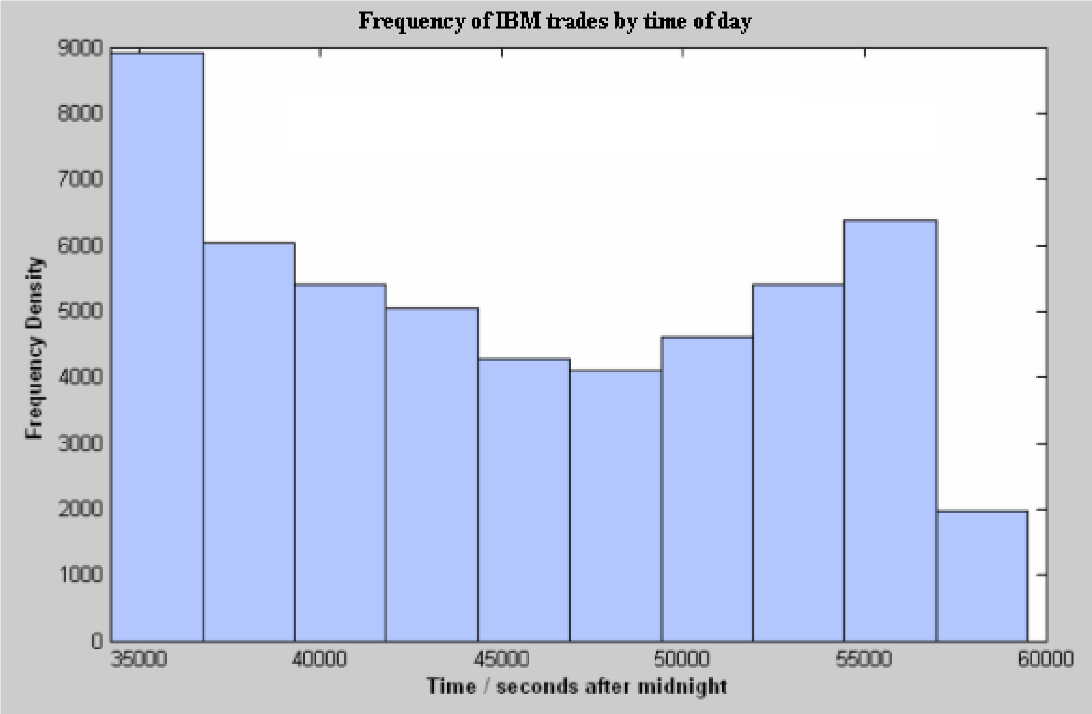
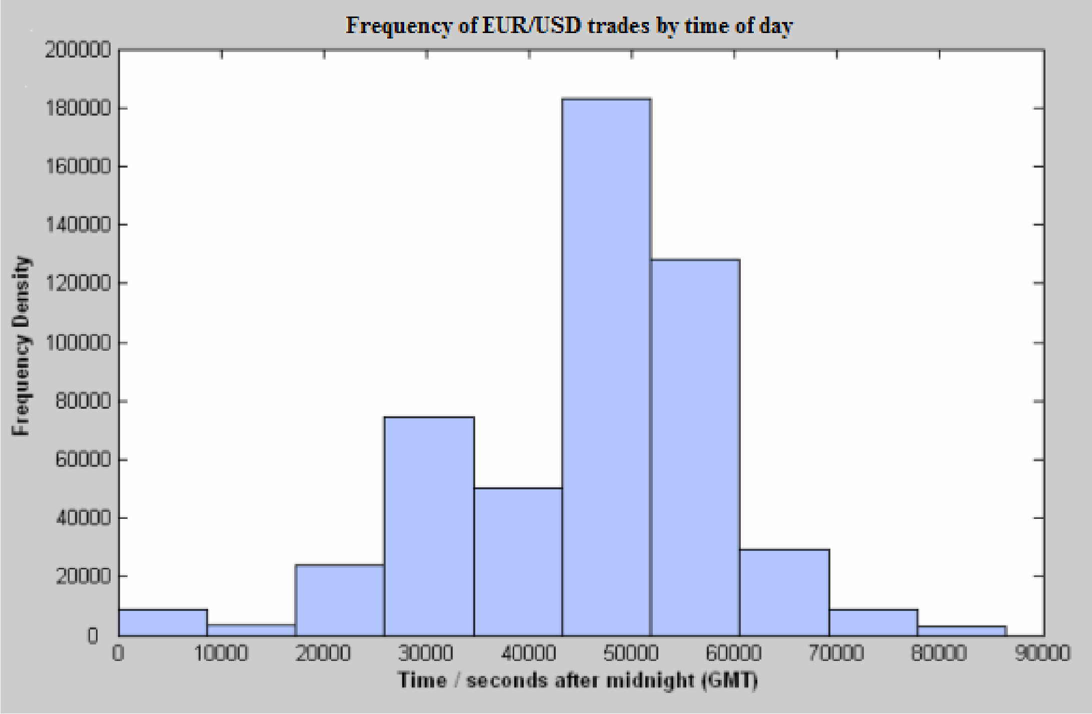


Figure A.2: Diurnal Pattern for EUR/USD Data

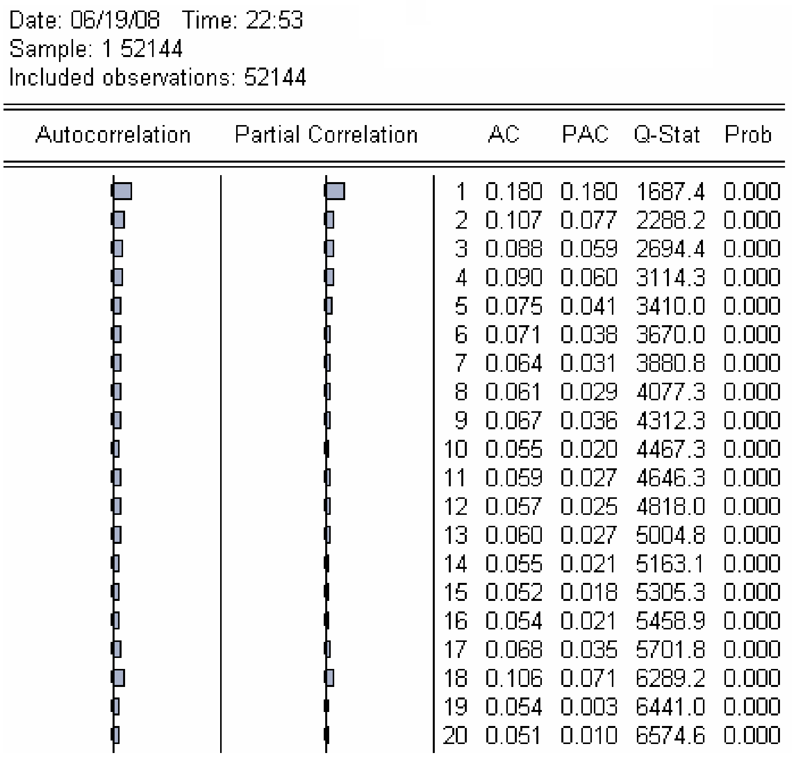


The shape in Figure A.2 is slightly more complicated, but essentially corresponds to separate U-shaped patterns over the course of the day. The three major forex markets are Tokyo, London and New York. Tokyo trading hours are from 1am to 9am (GMT) which corresponds to 3600 - 34200s after midnight, approximately the first 3 and a half bars (first U-shape). London runs from 9am to 5pm; 34200 - 61200s after midnight, which is roughly the next three bars (second U-shape). The second highest bar corresponds to the overlap in trading hours between London and New York between 2pm to 5pm; New York runs from 2pm to 10pm or 50400 - 79200s. There seems to be no U-shape for New York, this may be because traders there trade more in the morning than in the afternoon, as suggested by http://www.forexeconomiccalendar.com/forex-trading-hours.htm.

SF3.1-2:

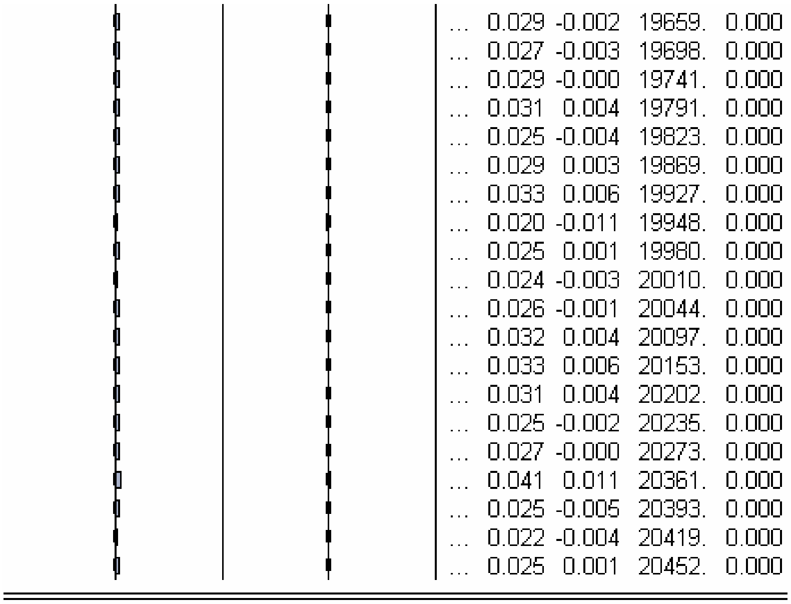
Following Deo, Hsieh and Hurvich (2010), we investigated correlations upto high lag order (200) for durations, short-term counts and long-term counts. The first and last 20 correlations and associated Q-statistics for the IBM data are shown in Figures A.3 and A.4 below. These demonstrate that there is significant correlation of up to order 200, supporting duration clustering and long memory. A similar result is found for EUR/USD data in Figures A.5 and A.6.

Figure A.3: First 20 (of 200) Autocorrelations and Partial Autocorrelations for IBM Durations



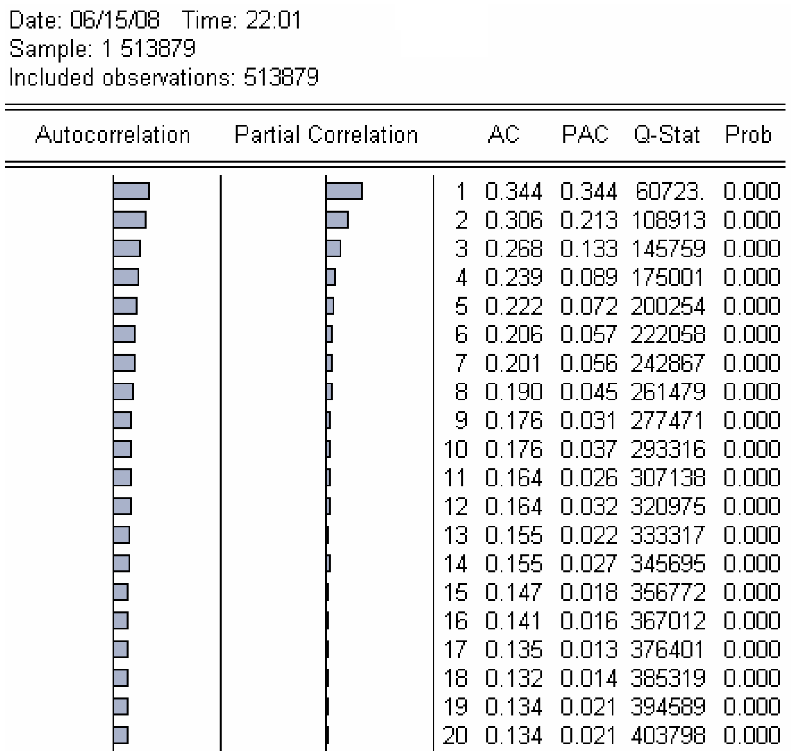
```
Date: 06/19/08   Time: 22:53
Sample: 1 52144
Included observations: 52144

Autocorrelation   Partial Correlation       AC     PAC    Q-Stat  Prob

                                         1  0.180  0.180  1687.4  0.000
                                         2  0.107  0.077  2288.2  0.000
                                         3  0.088  0.059  2694.4  0.000
                                         4  0.090  0.060  3114.3  0.000
                                         5  0.075  0.041  3410.0  0.000
                                         6  0.071  0.038  3670.0  0.000
                                         7  0.064  0.031  3880.8  0.000
                                         8  0.061  0.029  4077.3  0.000
                                         9  0.067  0.036  4312.3  0.000
                                        10  0.055  0.020  4467.3  0.000
                                        11  0.059  0.027  4646.3  0.000
                                        12  0.057  0.025  4818.0  0.000
                                        13  0.060  0.027  5004.8  0.000
                                        14  0.055  0.021  5163.1  0.000
                                        15  0.052  0.018  5305.3  0.000
                                        16  0.054  0.021  5458.9  0.000
                                        17  0.068  0.035  5701.8  0.000
                                        18  0.106  0.071  6289.2  0.000
                                        19  0.054  0.003  6441.0  0.000
                                        20  0.051  0.010  6574.6  0.000
```

Figure A.4: Last 20 (of 200) Autocorrelations and Partial Autocorrelations for IBM Durations



```
                                        ...  0.029 -0.002  19659.  0.000
                                        ...  0.027 -0.003  19698.  0.000
                                        ...  0.029 -0.000  19741.  0.000
                                        ...  0.031  0.004  19791.  0.000
                                        ...  0.025 -0.004  19823.  0.000
                                        ...  0.029  0.003  19869.  0.000
                                        ...  0.033  0.006  19927.  0.000
                                        ...  0.020 -0.011  19948.  0.000
                                        ...  0.025  0.001  19980.  0.000
                                        ...  0.024 -0.003  20010.  0.000
                                        ...  0.026 -0.001  20044.  0.000
                                        ...  0.032  0.004  20097.  0.000
                                        ...  0.033  0.006  20153.  0.000
                                        ...  0.031  0.004  20202.  0.000
                                        ...  0.025 -0.002  20235.  0.000
                                        ...  0.027 -0.000  20273.  0.000
                                        ...  0.041  0.011  20361.  0.000
                                        ...  0.025 -0.005  20393.  0.000
                                        ...  0.022 -0.004  20419.  0.000
                                        ...  0.025  0.001  20452.  0.000
```

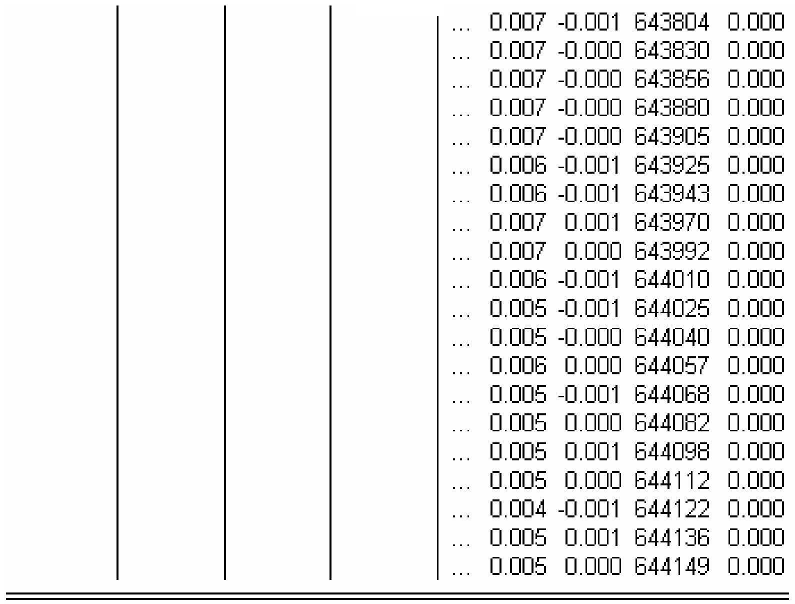
Figure A.5: First 20 (of 200) Autocorrelations and Partial Autocorrelations for EUR/USD Durations



Date: 06/15/08 Time: 22:01
Sample: 1 513879
Included observations: 513879

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.344	0.344	60723.	0.000
		2	0.306	0.213	108913	0.000
		3	0.268	0.133	145759	0.000
		4	0.239	0.089	175001	0.000
		5	0.222	0.072	200254	0.000
		6	0.206	0.057	222058	0.000
		7	0.201	0.056	242867	0.000
		8	0.190	0.045	261479	0.000
		9	0.176	0.031	277471	0.000
		10	0.176	0.037	293316	0.000
		11	0.164	0.026	307138	0.000
		12	0.164	0.032	320975	0.000
		13	0.155	0.022	333317	0.000
		14	0.155	0.027	345695	0.000
		15	0.147	0.018	356772	0.000
		16	0.141	0.016	367012	0.000
		17	0.135	0.013	376401	0.000
		18	0.132	0.014	385319	0.000
		19	0.134	0.021	394589	0.000
		20	0.134	0.021	403798	0.000

Figure A.6: Last 20 (of 200) Autocorrelations and Partial Autocorrelations for EUR/USD Durations

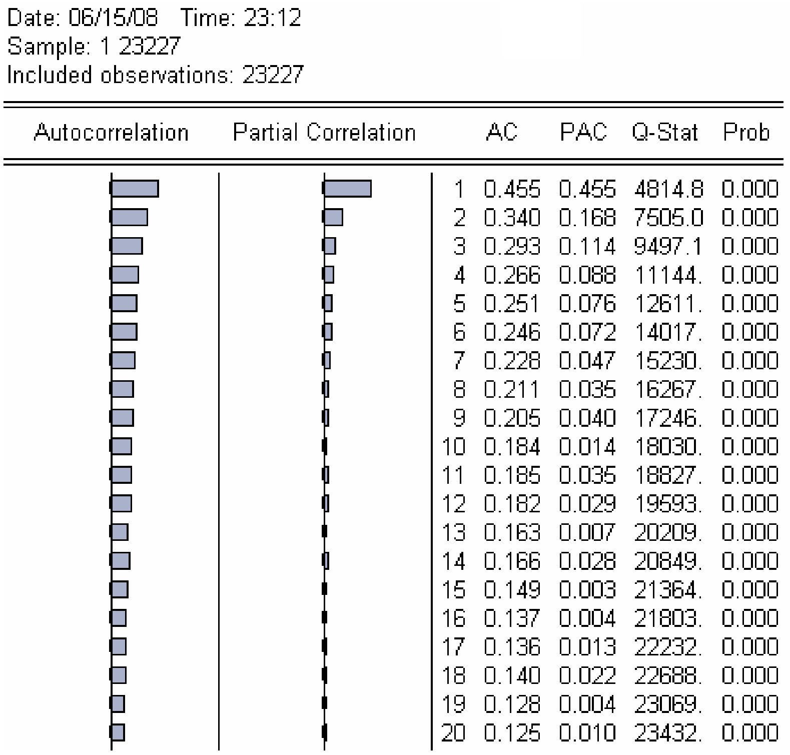


		...	0.007	-0.001	643804	0.000
		...	0.007	-0.000	643830	0.000
		...	0.007	-0.000	643856	0.000
		...	0.007	-0.000	643880	0.000
		...	0.007	-0.000	643905	0.000
		...	0.006	-0.001	643925	0.000
		...	0.006	-0.001	643943	0.000
		...	0.007	0.001	643970	0.000
		...	0.007	0.000	643992	0.000
		...	0.006	-0.001	644010	0.000
		...	0.005	-0.001	644025	0.000
		...	0.005	-0.000	644040	0.000
		...	0.006	0.000	644057	0.000
		...	0.005	-0.001	644068	0.000
		...	0.005	0.000	644082	0.000
		...	0.005	0.001	644098	0.000
		...	0.005	0.000	644112	0.000
		...	0.004	-0.001	644122	0.000
		...	0.005	0.001	644136	0.000
		...	0.005	0.000	644149	0.000

The first and last 20 correlations and associated Q-statistics for IBM 1-minute counts are shown in Figures A.7 and A.8 below. These demonstrate that there is significant correlation of up to order 200, supporting autocorrelation in counts and long memory. Similar results are found for EUR/USD data in Figures A.10 and A.11.

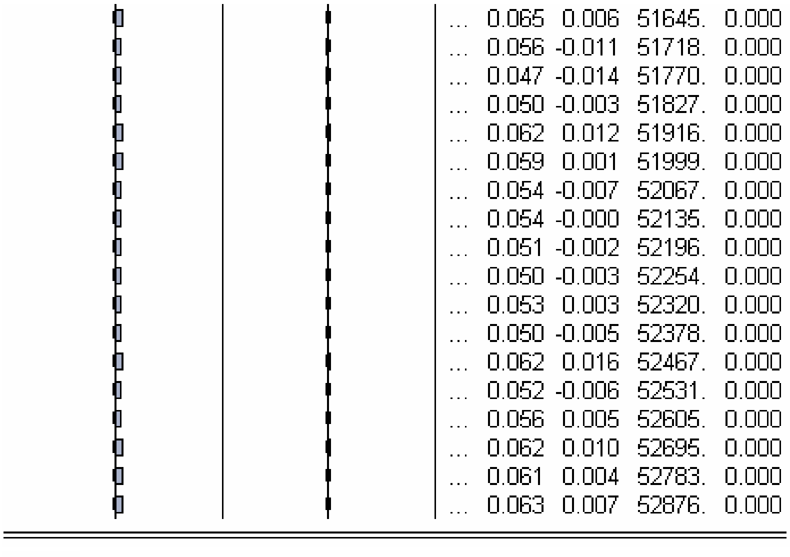
Figure A.9 demonstrates that there is significant lag-1 autocorrelation for 5-day (aggregated) counts, though this is negative. However, only 12 observations existed for the 5-day counts, so it is possible that the sample is not large enough for accurate conclusions. In the larger EUR/USD sample, significant positive lag-1 autocorrelation exists (Figure A.12).

Figure A.7: First 20 (of 200) Autocorrelations and Partial Autocorrelations for IBM 1-Minute Counts



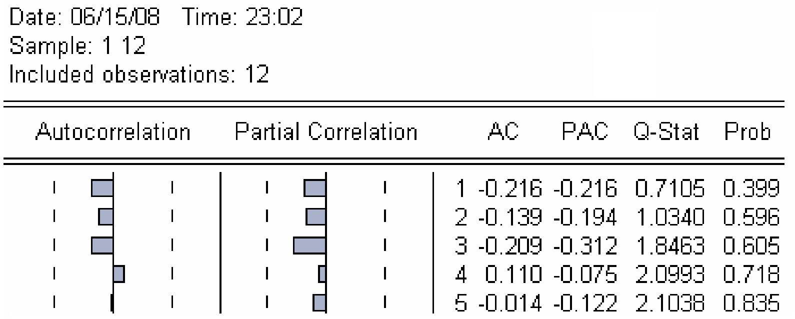
```
Date: 06/15/08   Time: 23:12
Sample: 1 23227
Included observations: 23227

Autocorrelation   Partial Correlation      AC     PAC    Q-Stat   Prob

                                       1  0.455  0.455  4814.8  0.000
                                       2  0.340  0.168  7505.0  0.000
                                       3  0.293  0.114  9497.1  0.000
                                       4  0.266  0.088  11144.  0.000
                                       5  0.251  0.076  12611.  0.000
                                       6  0.246  0.072  14017.  0.000
                                       7  0.228  0.047  15230.  0.000
                                       8  0.211  0.035  16267.  0.000
                                       9  0.205  0.040  17246.  0.000
                                      10  0.184  0.014  18030.  0.000
                                      11  0.185  0.035  18827.  0.000
                                      12  0.182  0.029  19593.  0.000
                                      13  0.163  0.007  20209.  0.000
                                      14  0.166  0.028  20849.  0.000
                                      15  0.149  0.003  21364.  0.000
                                      16  0.137  0.004  21803.  0.000
                                      17  0.136  0.013  22232.  0.000
                                      18  0.140  0.022  22688.  0.000
                                      19  0.128  0.004  23069.  0.000
                                      20  0.125  0.010  23432.  0.000
```

Figure A.8: Last 20 (of 200) Autocorrelations and Partial Autocorrelations for IBM 1-Minute Counts



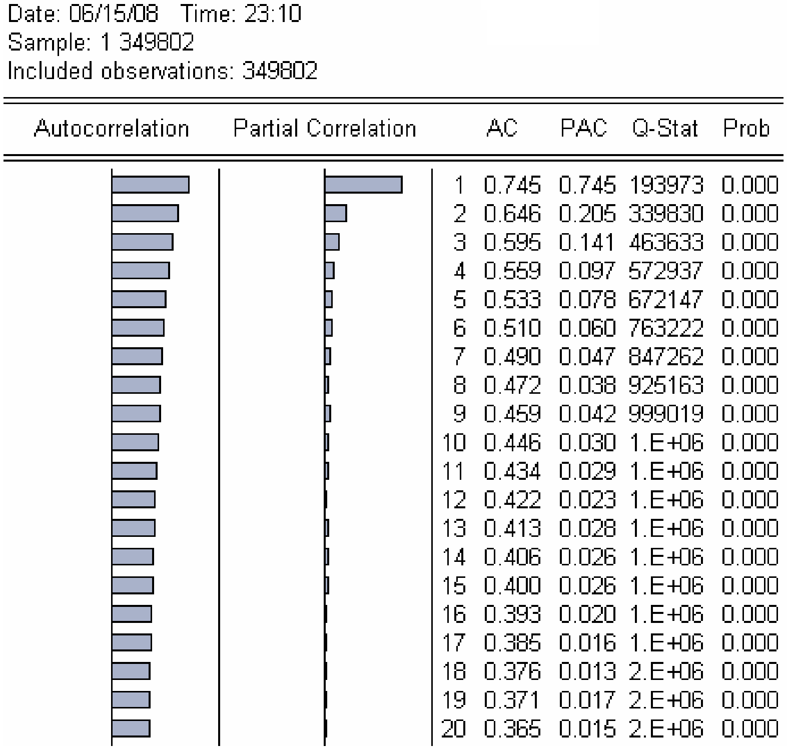
```
... 0.065  0.006 51645. 0.000
... 0.056 -0.011 51718. 0.000
... 0.047 -0.014 51770. 0.000
... 0.050 -0.003 51827. 0.000
... 0.062  0.012 51916. 0.000
... 0.059  0.001 51999. 0.000
... 0.054 -0.007 52067. 0.000
... 0.054 -0.000 52135. 0.000
... 0.051 -0.002 52196. 0.000
... 0.050 -0.003 52254. 0.000
... 0.053  0.003 52320. 0.000
... 0.050 -0.005 52378. 0.000
... 0.062  0.016 52467. 0.000
... 0.052 -0.006 52531. 0.000
... 0.056  0.005 52605. 0.000
... 0.062  0.010 52695. 0.000
... 0.061  0.004 52783. 0.000
... 0.063  0.007 52876. 0.000
```

Figure A.9: First 5 Autocorrelations and Partial Autocorrelations for IBM 5-Day Counts



```
Date: 06/15/08   Time: 23:02
Sample: 1 12
Included observations: 12

Autocorrelation   Partial Correlation      AC     PAC   Q-Stat  Prob

                                       1 -0.216 -0.216 0.7105  0.399
                                       2 -0.139 -0.194 1.0340  0.596
                                       3 -0.209 -0.312 1.8463  0.605
                                       4  0.110 -0.075 2.0993  0.718
                                       5 -0.014 -0.122 2.1038  0.835
```

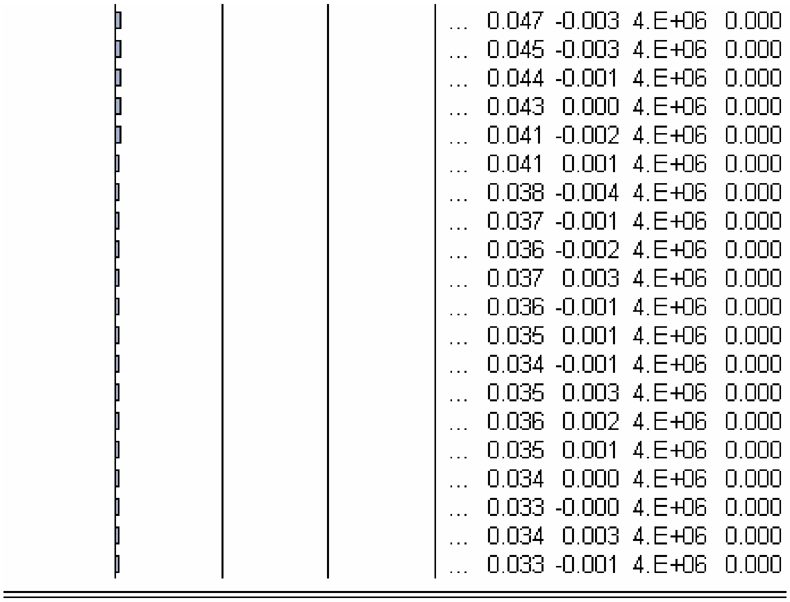
Figure A.10: First 20 (of 200) Autocorrelations and Partial Autocorrelations for EUR/USD 1-Minute Counts



Date: 06/15/08 Time: 23:10
Sample: 1 349802
Included observations: 349802

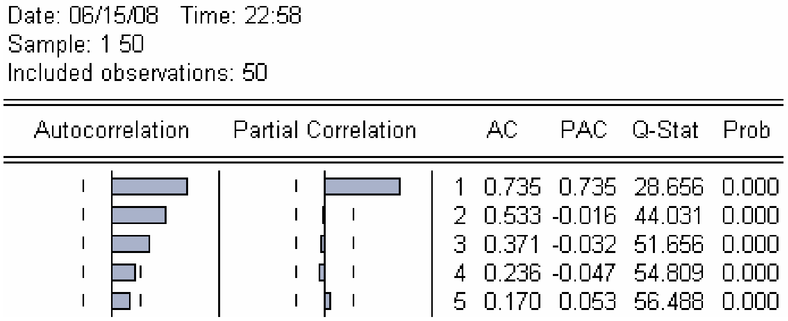
Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.745	0.745	193973	0.000
		2	0.646	0.205	339830	0.000
		3	0.595	0.141	463633	0.000
		4	0.559	0.097	572937	0.000
		5	0.533	0.078	672147	0.000
		6	0.510	0.060	763222	0.000
		7	0.490	0.047	847262	0.000
		8	0.472	0.038	925163	0.000
		9	0.459	0.042	999019	0.000
		10	0.446	0.030	1.E+06	0.000
		11	0.434	0.029	1.E+06	0.000
		12	0.422	0.023	1.E+06	0.000
		13	0.413	0.028	1.E+06	0.000
		14	0.406	0.026	1.E+06	0.000
		15	0.400	0.026	1.E+06	0.000
		16	0.393	0.020	1.E+06	0.000
		17	0.385	0.016	1.E+06	0.000
		18	0.376	0.013	2.E+06	0.000
		19	0.371	0.017	2.E+06	0.000
		20	0.365	0.015	2.E+06	0.000

Figure A.11: Last 20 (of 200) Autocorrelations and Partial Autocorrelations for EUR/USD 1-Minute Counts



```
... 0.047 -0.003 4.E+06 0.000
... 0.045 -0.003 4.E+06 0.000
... 0.044 -0.001 4.E+06 0.000
... 0.043  0.000 4.E+06 0.000
... 0.041 -0.002 4.E+06 0.000
... 0.041  0.001 4.E+06 0.000
... 0.038 -0.004 4.E+06 0.000
... 0.037 -0.001 4.E+06 0.000
... 0.036 -0.002 4.E+06 0.000
... 0.037  0.003 4.E+06 0.000
... 0.036 -0.001 4.E+06 0.000
... 0.035  0.001 4.E+06 0.000
... 0.034 -0.001 4.E+06 0.000
... 0.035  0.003 4.E+06 0.000
... 0.036  0.002 4.E+06 0.000
... 0.035  0.001 4.E+06 0.000
... 0.034  0.000 4.E+06 0.000
... 0.033 -0.000 4.E+06 0.000
... 0.034  0.003 4.E+06 0.000
... 0.033 -0.001 4.E+06 0.000
```

Figure A.12: First 5 Autocorrelations and Partial Autocorrelations for EUR/USD 5-Day Counts

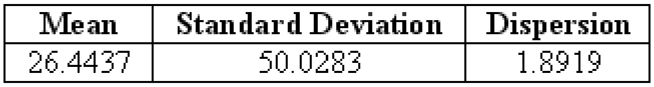


```
Date: 06/15/08   Time: 22:58
Sample: 1 50
Included observations: 50

Autocorrelation   Partial Correlation     AC     PAC    Q-Stat  Prob

                                       1 0.735  0.735  28.656  0.000
                                       2 0.533 -0.016  44.031  0.000
                                       3 0.371 -0.032  51.656  0.000
                                       4 0.236 -0.047  54.809  0.000
                                       5 0.170  0.053  56.488  0.000
```

SF5.1:

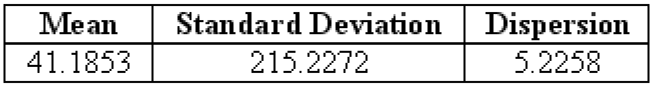
Tables A.3 and A.4 show the dispersion statistics for the IBM and EUR/USD durations. Both are greater than 1, supporting overdispersion.

Table A.3: Dispersion Statistics for IBM Durations



Mean	Standard Deviation	Dispersion
26.4437	50.0283	1.8919

Table A.4: Dispersion Statistics for EUR/USD Durations



Mean	Standard Deviation	Dispersion
41.1853	215.2272	5.2258

CHAPTER 2: MODELLING AND FORECASTING DURATIONS WITH LONG MEMORY MODELS

2.1 Introduction

Note: This is joint work with Dr Filip Žikeš of Imperial College London.

As noted by Pacurar (2006), there is a scarcity of comparisons of duration models. Ideally, we would like to undertake a comparison of all the models she has detailed. However, as noted in Chapter 1, only Long Memory models are able to account for the key stylised feature of SF3.2 - autocorrelation in counts and aggregated counts, as well as SF3.1 - duration clustering. So we would like to compare the empirical performance of the ACD Baseline Duration model with the FIACD and LMSD Long Memory models.

Alternatively, the clustering may be a consequence of regime shifts. The Markov Switching Multifractal (MSM) model of Calvet and Fisher (2004) might also be included in the comparison; Calvet and Fisher have created a regime-switching model which makes use of the scale-invariance or self-similarity property of fractals to model volatility.

In more detail, Calvet and Fisher (2008) note that ”economic shocks have highly heterogeneous degrees of persistence”. Information arrives over varying and random frequencies, translating into multiple frequencies of trading in terms of volatility and volumes. The multifractality in their model is generated by a scale-invariance property which enables the MSM model to fit 1) Long Memory, 2) intermediate frequency volatility dynamics and 3) thick tails in returns. In the volatility setting then, this enables the model to outperform its FIGARCH competitor by having a higher likelihood value and better out of sample forecasts with a lower number of parameters. A comparison with the LMSV model was not conducted, though Calvet and Fisher indicate that one advantage of the MSM model is the availability of a closed-form likelihood function.

We extend this model to the duration setting as has already occurred with GARCH and SV models, naming the resulting model the Markov-Switching Multifractal Duration (MSMD) model. Similarly to Calvet and Fisher (2004), we will compare the performance to the duration analogues of the GARCH model, i.e. the ACD and FIACD models. Furthermore, we will extend the comparison to the latent factor LMSD model. We implement the models using price duration data on the S&P 500 futures contract and the relatively recently introduced VIX futures contract from Tick Data for a period running from January, 2008 till June, 2008. We focus on transactions prices pertaining to the most liquid (front) contract traded on the Chicago Mercantile Exchange (CME) during the main trading hours of 9:30 - 16:15 EST and examine transaction durations.

First we will outline the estimation and forecasting methods for the FIACD and LMSD models, then the MSMD model, and finally we will analyse the results.

2.2 Duration Models

2.2.1 The ACD and FIACD models

Recapping on Chapter 1, Engle and Russell (1998) proposed the Autoregressive Conditional Duration (ACD) model:

$$x_t = \psi_t \varepsilon_t \qquad \varepsilon_t \sim IID(1, \sigma_\varepsilon^2) \tag{2.1}$$

$$\psi_t = \omega + \sum_{i=1}^{p} \alpha_i x_{t-i} + \sum_{i=1}^{q} \beta_i \psi_{t-i} \qquad \varepsilon_t \text{ and } \psi_t \text{ are independent} \tag{2.2}$$

Since the ACF of the ACD model eventually fades geometrically, it cannot exhibit long memory, which is signified by hyperbolic decay (i.e. $\lim\limits_{N\to\infty} \sum\limits_{\tau=-N}^{N} \rho(\tau) = \infty$).

Therefore Jasiak (1999) adapted a FIGARCH specification for durations, creating the Fractionally Integrated ACD (FIACD) model. Now the conditional duration equation (equation (2.2)) is replaced with:

$$\psi_t = \underbrace{[1-\beta(1)]^{-1}\omega}_{:=\varpi} + \underbrace{[1-\beta(L)]^{-1}[1-\beta(L)+[1-\phi(L)](1-L)^d]}_{:=c(L)} x_t$$

$$\psi_t = \varpi + c(L)x_t$$

where $(1-L)^d = \sum\limits_{i=0}^{\infty} \pi_i L^i$ with $\pi_i = F(-d,1;1;1) = \begin{cases} 1 & i = 0 \\ \prod\limits_{k=1}^{i} \frac{k-1-d}{k} & i > 0 \end{cases}$

$F(a,b;c;z)$ is the Hypergeometric function as in e.g. Karanasos et al. (2004)

$c(L)$ is an infinite order lag polynomial; $c(L) = \sum\limits_{i=1}^{\infty} c_i L^i$.

There is now more persistence because the conditional duration equation in (2.2) has changed from an ARMA-type structure to an ARFIMA-type structure.

2.2.2 The LMSD model

Bauwens and Veredas (1999) proposed the the Stochastic Conditional Duration (SCD) model:

$$x_t = \varepsilon_t e^{\psi_t} \qquad \varepsilon_t \sim IID(\mu_\varepsilon, \sigma_\varepsilon^2) \qquad E(\varepsilon_t^r) := g_r \tag{2.3}$$

$$\psi_t = \omega + \beta\psi_{t-1} + u_t \qquad u_t \sim NID(0, \sigma_u^2) \qquad \varepsilon_t \text{ and } u_t \text{ are independent} \tag{2.4}$$

As with the ACD model, the SCD model is only capable of generating geometric decay in the ACF. In order to enable long memory, Deo, Hurvich, Soulier and Wang (2009) and Deo, Hsieh and Hurvich (2010) created the LMSD process, in which the (logged) conditional duration equation in (2.4) is replaced with:

$$\begin{aligned} \psi_t &= \omega + \beta\psi_{t-1} + (1-L)^{-d}u_t \\ \Leftrightarrow \psi_t &= \omega + \sum_{i=0}^{\infty} b_i u_{t-i} \qquad ; b_i \sim Ci^{d-1} \quad (\text{as } i \to \infty) \end{aligned}$$

Here there is more persistence because the logged conditional duration equation has changed from an AR(1) process to an ARFIMA process with slowly decaying coefficients.

Hurvich, Deo, Soulier and Wang (2009) have shown that long memory in durations is required in order to generate SF3.2 - 1) long memory in (short-term) counts (in their Theorem 2), and 2) lag-1 auto-correlation in long-term counts (in their Theorem 4).

2.2.3 Estimation

2.2.3.1 The ACD and FIACD Models

The ACD and FIACD models can be estimated using Maximum Likelihood, given the distribution of the disturbance term. For the ACD model, the parameter vector to be estimated, θ, is $(\omega, \alpha, \beta, \zeta)'$ where $\alpha = (\alpha_1, ..., \alpha_p)'$, $\beta = (\beta_1, ..., \beta_q)'$, and ζ is a vector of extra parameters needed for the distribution of the disturbance term ε_t. For the FIACD model, d, the fractional differencing parameter, also needs to be estimated so $\theta = (\omega, \alpha, \beta, \zeta, d)'$. Then the log-likelihood functions, conditional on ε_t having an Exponential, Weibull, Burr or Generalised Gamma distribution with unit mean, can be shown to be as below. Please refer to Appendix A for more detail on the distributional forms, with notation based on Fernandes and Grammig (2005).

$$\text{Exponential:} \qquad \ell(\theta) = -\sum_{t=1}^{T} \left\{ \ln(\psi_t) + \frac{x_t}{\psi_t} \right\}$$

$$\text{Weibull:} \qquad \ell(\theta) = \sum_{t=1}^{T} \left\{ \ln\left(\frac{\kappa}{x_t}\right) + \kappa \ln\left[\Gamma\left(1 + \frac{1}{\kappa}\right)\frac{x_t}{\psi_t}\right] - \left[\Gamma\left(1 + \frac{1}{\kappa}\right)\frac{x_t}{\psi_t}\right]^{\kappa} \right\}$$

$$\text{Burr:} \qquad \ell(\theta) = \sum_{t=1}^{T} \left\{ \ln\left(\frac{\kappa}{x_t}\right) + \kappa \ln\left(\vartheta_B \frac{x_t}{\psi_t}\right) - \left(1 + \frac{1}{\delta}\right) \ln\left[1 + \delta\left(\vartheta_B \frac{x_t}{\psi_t}\right)^{\kappa}\right] \right\}$$

$$\text{Generalised Gamma:} \qquad \ell(\theta) = \sum_{t=1}^{T} \left\{ \ln\left(\frac{\kappa}{\Gamma(\delta) x_t}\right) + \kappa\delta \ln\left(\vartheta_G \frac{x_t}{\psi_t}\right) - \left(\vartheta_G \frac{x_t}{\psi_t}\right)^{\kappa} \right\}$$

However, since the FIACD involves an infinite series of coefficients, truncation is necessary, which results in bias in the estimates as noted by Baillie, Bollerslev and Mikkelsen (1996).

2.2.3.2 The LMSD Model

Estimation of the LMSD model is less straightforward owing to the unobservable factor. Bauwens and Veredas (1999) advocate employing the Kalman Filter, while Deo, Hsieh and Hurvich (2010) suggest QMLE using the Whittle approximation. The latter approach will be adopted as it expresses the model more compactly. Then the resulting estimators of the parameters are $T^{\frac{1}{2}}$ consistent and asymptotically Normal. Taking logs in the actual duration equation (2.3):

$$\begin{aligned} \ln x_t &= \underbrace{\mathbb{E}(\ln \varepsilon_t)}_{:=m} + \psi_t + \underbrace{\ln \varepsilon_t - \mathbb{E}(\ln \varepsilon_t)}_{:=\xi_t} \\ \ln x_t &= m + \mu + h_t + \xi_t \qquad (2.5) \end{aligned}$$

where $h_t = \psi_t - \mu = b(L)u_t$ i.e. is the de-meaned unobservable factor. Then the parameter vector to be estimated is $\theta = (m, \mu, \omega, \beta, \sigma_u^2, \sigma_\xi^2, d)'$.

We can estimate m and μ jointly using the sample mean of $\ln x_t$, since h_t and ξ_t are zero-mean processes. Then given the distribution of ε_t and the estimate of σ_ξ^2, it is possible to determine the value of m and thereby μ.

For the remaining parameters, the Whittle log-likelihood can be maximised. Define the spectral density function, $f_\theta(\omega_t)$, as:

$$f_\theta(\omega_t) = \frac{\sigma_u^2}{2\pi} \left|1 - \beta e^{-i\omega_t}\right|^{-2} \left|1 - e^{-i\omega_t}\right|^{-2d} + \frac{\sigma_\xi^2}{2\pi}$$

where the angular frequency, $\omega_t = \dfrac{2\pi t}{T}$

Also define the periodogram, $I(\omega_t)$, as:

$$I(\omega_t) = \frac{1}{2\pi T} \sum_{t=1}^{T} \left|(\ln x_t) e^{-it\omega_t}\right|^2$$

Then the Whittle log-likelihood is:

$$\ell(\theta) = -\frac{T}{2}\ln(2\pi) - \frac{1}{2} \sum_{-\frac{T}{2} \le t < \frac{T}{2}} \left\{ \ln f_\theta(\omega_t) + \frac{I(\omega_t)}{f_\theta(\omega_t)} \right\}$$

However, it is sufficient to minimise the following approximate negative log-likelihood, although we revert to the full form when reporting likelihoods in the results section:

$$\ell(\theta) = \sum_{t=1}^{\left[\frac{T-1}{2}\right]} \left\{ \ln f_\theta(\omega_t) + \frac{I(\omega_t)}{f_\theta(\omega_t)} \right\}$$

Note that σ_ξ^2 varies depending on the distribution of ε_t:

$$\text{Exponential:} \qquad \sigma_\xi^2 = \frac{\pi^2}{6}$$

$$\text{Weibull:} \qquad \sigma_\xi^2 = \frac{\pi^2}{6\kappa^2}$$

2.2.4 Model Selection and Specification Testing

In their comparison of the MSM, GARCH and FIGARCH models, Calvet and Fischer (2004) compare the log-likelihood values of the models under consideration, as well as the Bayesian Information Criterion (BIC); BIC = $T^{-1}(-2\ln L(\theta) + NP\ln T)$ where NP is the number of parameters being estimated. We will adopt this approach, although it cannot be applied to the LMSD model directly as its log-likelihood is specified in terms of log durations, whereas the log-likelihood of the other models is specified in terms of (raw) durations. To see this, define $y := \ln(x)$. By Theorem 8.12 in Davidson (1994):

$$\begin{aligned} f_Y(y) &= f_X(x)\left|\frac{dx}{dy}\right| \\ &= f_X(x)\epsilon^y \end{aligned}$$

Then the log-likelihood for log durations is:

$$\begin{aligned} \sum_{t=1}^{T}\ln[f_Y(y_i)] &= \sum_{t=1}^{T}\ln[f_X(x_t)\epsilon^{y_t}] \qquad \text{where } T \text{ is the sample size} \\ &= \sum_{t=1}^{T}\{\ln[f_X(x_t)] + \ln[x_t]\} \\ &\neq \sum_{t=1}^{T}\ln[f_X(x_t)] \qquad \text{which is the log-likelihood for (unlogged) durations} \end{aligned}$$

In addition, we can compare what might be called the “Sum of Relative Deviations” (SRD) - the sum of proportional deviations between fitted and actual durations, i.e. $\sum\limits_{t=1}^{T}\frac{x_t-\hat{\psi}_t}{\hat{\psi}_t}$. For the ACD and FIACD models, the fitted durations are simply the $\{\psi_t(\hat{\theta})\}$, while for the LMSD model, the fitted durations are calculated in the filtering recursions of the Kalman Filter.

Other specification tests have been proposed for duration models. For example, Engle and Russell (1998) proposed tests of the fitted residuals; $\hat{\varepsilon}_t = x_t/\psi_t$ for the ACD and FI-ACD models. In the case of the SCD and LMSD models, $\hat{\varepsilon}_t = x_t/e^{(\hat{\psi}_t)}$, while the extra disturbance term $\hat{u}_t = \hat{\psi}_t - \hat{\omega} - \hat{\beta}(L)\hat{\psi}_t$. These will be tested for dependence using standard Ljung-Box tests.

Note that Li and Yu (2003) have created an alternative diagnostic test based on the residual autocorrelations, although they have only established its properties for the ACD(m) model (as opposed to the ACD($1, 1$) model), so we would not be able to adopt the test for our models. Furthermore, Deo, Hsieh and Hurvich (2010) argue that the autocorrelations of the residuals in the LMSD model do not not behave like those of white noise, even if the model is correctly specified, and admit that there are no completely satisfactory diagnostic tests for the LMSD model. Certainly, this means that there may be no existing method of comparing in-sample fit across models beyond log-likelihood or observation deviation measures (as above), highlighting a potential research gap both in the stochastic volatility and duration literature. However, the ultimate test of a model is its forecast performance as argued by both Friedman (1953) and Pindyck and Rubinfeld (1981); in this context then, the major contribution of this paper is to assess which model forecasts durations best.

Finally, Deo, Hsieh and Hurvich (2010) suggest that the properties of the counting process can be examined to check the properties of the duration process; durations and counts are equivalent means of describing a point process. Principally, as Deo, Hurvich, Soulier and Wang have shown (2009), only a long memory durations process is consistent with long memory in counts. So if the latter exists, this implies that long memory exists in the durations process as well.

2.2.5 Forecasting

2.2.5.1 The ACD and FIACD Models

H-step forecasts of the ACD model are possible by adopting standard GARCH forecasting formulae. In the case of the ACD(1,1) model we have:

$$\mathbb{E}_t(x_{t+1}) = \mathbb{E}_t(\psi_{t+1}) = \omega + \alpha x_t + \beta \psi_t$$

and when $H \geq 2$,

$$\mathbb{E}_t(x_{t+H}) = \mathbb{E}_t(\psi_{t+H}) = \mu + (\alpha+\beta)^{H-1}(\psi_t - \mu) \qquad \text{where } \mu = \tfrac{\omega}{1-\alpha-\beta}$$

A similar multistep expression cannot be derived for the FIACD(1,d,1) model, since this involves the unconditional mean, μ, of x_t, which as Jasiak (1999) points out is not bounded. Instead, we use the infinite MA representation of the FIACD process (as before) to generate forecasts;

$$\begin{aligned} \psi_t &= \varpi + c(L)x_t \\ \mathbb{E}_t(\psi_{t+1}) &= \varpi + c(L)x_{t+1} \qquad \{\text{a function only of } x_t \text{ and previous}\} \\ \Rightarrow \mathbb{E}_t(x_{t+H}) = \mathbb{E}_t(\psi_{t+H}) &= \varpi + \sum_{i=1}^{H-1} c_i \mathbb{E}_t(\psi_{t+H-i}) + \sum_{i=H}^{\infty} c_i x_{t-i} \end{aligned}$$

where the $\mathbb{E}_t(\psi_{t+H-i})$ are calculated recursively given $\mathbb{E}_t(\psi_{t+1})$.

2.2.5.2 The LMSD Model

With respect to the LMSD model, forecasting is possible either through calibration of the best linear predictor, as advocated by Deo, Hurvich and Lu (2006b), or via the Kalman Filter. Following notation as per Brockwell and Davis (1991), let the general state-space form be:

$$\mathbf{X}_t = G_t\mathbf{H}_t + \mathbf{W}_t \qquad \mathbf{R}_t = \mathbb{E}(\mathbf{W}_t\mathbf{W}_t') \qquad \text{(Observation Equation)}$$

$$\mathbf{H}_{t+1} = F_t\mathbf{H}_t + \mathbf{V}_t \qquad \mathbf{Q}_t = \mathbb{E}(\mathbf{V}_t\mathbf{V}_t') \qquad \text{(State Equation)}$$

While the LMSD process contains an infinite series of coefficients, it is still possible to create an approximate / truncated state-space form as observed by Chan and Palma (1998). Demeaning equation (2.5), we have:

$$\underbrace{z_t}_{:=\ln x_t-(m+\mu)} = h_t + \xi_t$$

$$z_t = u_t + b_1u_{t-1} + ... + b_\chi u_{t-\chi} + \xi_t$$

from which it can be seen that the system is similar to a high-order (χ) MA process but with the addition of an extra disturbance term ξ_t. χ is the number of b coefficients we wish to retain before truncating $b(L)$, i.e. $b(L) \approx \sum\limits_{i=0}^{\chi} b_iL^i$. Then a specific state-space form would be:

$$z_t = \begin{pmatrix} b_\chi & b_{\chi-1} & \cdots & b_1 & 1 \end{pmatrix} \begin{pmatrix} b(L)u_{t-\chi} \\ b(L)u_{t-\chi-1} \\ \vdots \\ b(L)u_t \end{pmatrix} + \xi_t$$

$$\begin{pmatrix} b(L)u_{t-\chi} \\ b(L)u_{t-\chi-1} \\ \vdots \\ b(L)u_t \end{pmatrix} = \begin{pmatrix} 0 & I_\chi \\ 0 & 0 \end{pmatrix} \begin{pmatrix} b(L)u_{t-\chi-1} \\ b(L)u_{t-\chi-2} \\ \vdots \\ b(L)u_{t-1} \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} u_{t-1}$$

with covariance matrices:

$$\mathbf{R} = \sigma_\xi^2, \qquad \mathbf{Q} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & 0 \\ 0 & 0 & 0 & \sigma_u^2 \end{pmatrix}$$

and initial conditions:

$$E(X_{1|0}) = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad \Omega_1 = E[(X_{1|0} - E(X_{1|0}))^2] = \sigma_u^2 I_{\chi+1}$$

Then forecasting is possible via the standard prediction formulae/recursions (e.g. Brockwell and Davis, 1991, Chapter 12, page 477).

2.3 The MSMD Model

2.3.1 Overview

Assume that durations, x_t, follow a multiplicative error model given by:

$$x_t = \psi_t \epsilon_t, \quad \epsilon_t \sim IID(1, \sigma_\varepsilon^2),$$

where ψ_t is a Markov switching multifractal process of Calvet and Fisher (2004);

$$\psi_t = \bar{\psi} \left(\prod_{i=1}^{k} M_t^{(i)} \right), \tag{2.6}$$

and ϵ_t is a sequence of independent and identically distributed innovations with mean equal to one. The mean process in (2.6) is determined by k independent unit-mean multipliers, $M_t^{(i)}, i = 1, ..., k$, and a scaling constant, $\bar{\psi}$. At every point in time t, each multiplier $M_t^{(i)}$ assumes, with probability γ_i, a new value drawn from a common distribution M, and remains unchanged with probability $1-\gamma_i$. The transition probabilities are parsimoniously parametrized by

$$\gamma_i = 1 - (1 - \gamma_k)^{(b^{i-k})}, \quad i = 1, ..., k,$$

where $\gamma_k \in [0, 1]$ and $b \in (1, \infty)$. Two specifications for the distribution of the multipliers have been proposed by Calvet and Fisher (2004): binomial and lognormal. In the former specification, each multiplier, if at all, is renewed by drawing the values m_0 and $2-m_0$ with equal probability, ensuring that the mean is equal to one. The latter specification replaces the binomial distribution with a lognormal one, i.e.

$$\log M_t^{(i)} \sim \mathrm{N}(-\lambda, 2\lambda).$$

where the parameter restriction again imposes a unit mean on the multipliers.

Having specified the law governing the multipliers, it remains to choose a distribution for the innovations, ϵ_i. A number of alternatives have been put forward in the literature. Here we experiment with two distributions - the Exponential:

$$f_E(\epsilon; \theta_E) = \exp(-\epsilon)$$

and the Weibull:

$$f_W(\epsilon; \theta_W) = \kappa \vartheta_W^{\kappa} \epsilon^{\kappa-1} \exp(-\vartheta_W^{\kappa} \epsilon^{\kappa}).$$

Other, more flexible alternatives have been proposed, e.g. Burr and Generalized Gamma, but in the interest of parsimony we confine our attention to the Exponential and Weibull cases, which are the standard set of distributions implemented by most authors following full maximum likelihood, particularly Engle and Russell (1998). Moreover, we cannot invoke Quasi-Maximum Likelihood (QML) results for all the competing models as these results do not presently exist for the MSMD model. However, at the very least, the Exponential distribution which we have selected supports QML in the cases of the ACD and FIACD models (Jasiak (1999)) and perhaps the MSMD in future, while the Whittle estimator for the LMSD model enables QML for both Exponential and Weibull distributions (Deo, Hsieh and Hurvich (2010)).

The attractive property of the MSMD model is that its autocorrelation function (ACF) exhibits long-memory like behaviour. It follows directly from Proposition 1 in Calvet & Fisher (2004) that the ACF of MSMD durations decays hyperbolically over a large range of lags before transitioning smoothly into Exponential decay. Formally, take two arbitrary numbers α_1 and α_2 in $(0, 1)$ and let $I_k = \{n : \alpha_1 \log_b(b^k) \leq \log_b n \leq \alpha_2 \log_b(b^k)\}$ denote a set of integers containing a wide range of lags. Then:

$$\sup_{n \in I_k} \left| \frac{\log \mathrm{Corr}(x_t, x_{t+n})}{\log n^{-\delta}} - 1 \right| \to 0$$

for some constant δ as $k \to \infty$. So despite being a short-memory process, the MSMD can mimic the persistence of a genuine long-memory process.

2.3.2 Relation to Counts and Realized Volatility

Deo, Hurvich, Soulier & Wang (2009) recently investigated the propagation of memory of durations to counts and realized volatility. They show that if durations are long (short) memory then under certain conditions the counts are long (short) memory as well.

In more detail, counts and durations are stationary under different measures, since they define the irregularly-spaced event process (”point process”) in terms of different sets of events. These measures are referred to as P and P^0 respectively by Deo et al. However, by making use of equivalence theorems in e.g. Nieuwenhuis (1989), they show that long memory propagates from durations to counts. Furthermore, Deo, Hsieh and Hurvich (2010) show that long memory propagates from counts to squared returns and realized volatility.

With respect to the MSMD process now, the following proposition establishes the conditions under which the short-memory feature of the MSMD translates into short memory in the induced counts.

Proposition 2.3.1. *Let $\{x_i\}$ be a binomial MSMD process with $\mathbb{E}(\epsilon_i^{3+r}) < \infty$, $r > 0$. Then $\{x_i\}$ is Exponential strong mixing and the induced counting process $N(t)$ satisfies* $\operatorname{Var}(N(t)) \sim ct$ *for some $c < \infty$.*

To link the counts and realized volatility, we follow Deo et al. (2009) and employ the simple continuous-time pure-jump model of Oomen (2006). The logarithmic price process p is assumed to have the following dynamics:

$$p(t) = p(0) + \sum_{j=1}^{N(t)} \xi_j, \quad \xi_j \overset{iid}{\sim} \mathrm{D}(0, \sigma_\xi^2),$$

where $N(t)$ counts the number jumps up to time t and ξ_j is the size of the j-th jump. A natural measure of variation in the model is the quadratic variation given by:

$$\langle p \rangle_t = \sum_{j=1}^{N(t)} \xi_j^2.$$

The quadratic variation can be estimated consistently by realized variance. Dividing the time interval $[0, t]$ into n non-overlapping intervals of length $\delta t = t/n$, the realized variance is defined as

$$RV_{t,n} = \sum_{i=1}^{n} (p(i/n) - p((i-1)/n))^2.$$

It follows from Deo et al. (2009) that for the MSMD duration process satisfying the assumptions of Proposition 1, the realized volatility is a short memory process.

It is difficult to derive analytically the autocorrelation function of counts and realized volatility induced by the MSMD process. Nonetheless future research can investigate by simulation how the persistence of MSMD durations affects the persistence of counts and realized volatility.

2.3.3 Estimation and Specification Testing

Collecting the parameters of the binomial and lognormal MSM into vectors $\theta'_B = (m_0, b, \gamma_k, \bar{\psi}, \theta'_\epsilon)$ and $\theta'_L = (\lambda, b, \gamma_k, \bar{\psi}, \theta_\epsilon)$, respectively, we now turn to the problem of their estimation. Two approaches have been proposed in the literature, and they differ quite substantially in terms of their applicability, generality and computational feasibility.

2.3.3.1 Exact Maximum Likelihood Estimation

The binomial MSM with finite k implies a finite number of states of the hidden Markov process and hence can be estimated by exact maximum likelihood (MLE) via Bayesian updating. This has been advocated by Calvet and Fisher (2004) for the binomial MSM model for stochastic volatility, and has been shown to work well for sample sizes typically used for estimating models of time-varying volatility. Moreover, the Bayesian filter allows for estimation of the unobserved state probabilities, which in turn permits optimal forecasting.

Formally, let $\mathcal{F}_{t_i}$ denote the σ-field generated by the duration process up to time t_i, and let $f(x_i|\mathcal{F}_{t_{i-1}}; \theta_B)$ be the conditional density of x_i given $\mathcal{F}_{t_i}$. The Markov process ψ_i takes a finite number of values in $\mathbb{R}$ depending on the state, s_{t_i}, in which the Markov chain resides at time t_i. For a given state $s_{t_i} = j$, the conditional distribution of x_i is given by $\eta_{t_i,j} := f(x_i|s_{t_i} = j, \mathcal{F}_{t_{i-1}}; \theta_\epsilon)$, $j = 1, ..., 2^k$. We write the probability that the Markov chains visits state j in time t_i, conditional on $\mathcal{F}_{t_{i-1}}$, as $\xi_{t_i|t_{i-1},j} := \mathrm{P}(s_{t_i} = j|\mathcal{F}_{t_{i-1}}; \theta_B)$, and, similarly, the probability that the Markov chains resides in state j at time t_i given $\mathcal{F}_{t_i}$ as $\xi_{t_i|t_i,j} := \mathrm{P}(s_{t_i} = j|\mathcal{F}_{t_i}; \theta_B)$. It follows (e.g. Hamilton, 1994, Chapter 22) that the optimal inference and forecast at each point in time t_i can be obtained by iterating on:

$$\begin{aligned}\xi_{t_i|t_i} &= \frac{\xi_{t_i|t_{i-1}} \odot \eta_{t_i}}{\imath'(\xi_{t_i|t_{i-1}} \odot \eta_{t_i})}, \\ \xi_{t_{i+1}|t_i} &= \mathbf{P}\xi_{t_i|t_i},\end{aligned}$$

$i = 1, ..., n$, where $\mathbf{P}$ is the $(2^k \times 2^k)$ transition matrix associated with the Markov chain, $\imath$ is a vector of 1s, and the symbol $\odot$ denotes element-by-element multiplication. The log-likelihood function for a sample of size n is then given by:

$$\ell(\theta_B|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\mathcal{F}_{t_{i-1}}; \theta_B) = \sum_{i=1}^{n} \imath'(\xi_{t_i|t_{i-1}} \odot \eta_{t_i}). \tag{2.7}$$

Standard numerical algorithms can be employed to maximize (2.7) with respect to θ_B.

The disadvantage of the exact maximum likelihood estimator is that it becomes computationally infeasible for $k \geq 10$, since the dimension of the transition matrix grows with 2^k. Computation is similarly intensive for the lognormal MSM, where the state space is infinite; our research found that simulation based estimation methods such as Markov Chain Monte Carlo (MCMC) do not offer an improvement in estimation time over the previous approach in the univariate case, while they are more difficult to implement (however, we also found that in the multivariate case, MCMC would be faster to estimate, so have earmarked it for future research). Such issues have motivated Lux (2008) to develop a generalized method of moments approach (GMM), which works for a wide range of MSM specifications and requires only moderate computational resources. The drawback of the GMM estimator of Lux is its inability to precisely estimate the key dynamic parameters, b and γ_k, that drive the persistence of multipliers. Lux circumvents this problem by setting these parameters to some values that seem to work well for a number of data sets and estimating by GMM the remaining two parameters only. This is quite restrictive, however, especially in our context where no previous evidence exists to suggest possible meaningful values for b and γ_k. We therefore avoid the GMM estimation in this chapter and concentrate on the binomial MSM.

Before taking the model to the data it is worthwhile exploring the finite-sample properties of the maximum likelihood estimator. We run a simple Monte Carlo experiment in which we draw random samples of size $n \in \{1000, 2500, 5000\}$ from the binomial MSMD model with $k = 8$ multipliers, and estimate the parameters by exact maximum likelihood. The true parameter vector is set to $\theta_{\mathbf{B}} = (1.3, 2.0, 0.5, 10, \theta_\epsilon)$. The first three parameters are similar to those in Lux, while the fourth and fifth parameters do not impact the results significantly. The parameter of the Weibull distribution ($\theta_\epsilon = \kappa$) is set equal to 1.45. Due to the computational burden the number of Monte Carlo replications is limited to 500, 250, and 100 replications for $n = 1000, 2500, 5000$, respectively. Table 2.1 (next page) summarizes the simulation results. Clearly, the ML estimator performs in line with asymptotic theory, delivering precise and almost unbiased estimates even in samples of moderate size (e.g. 1,000 observations).

Table 2.1: Monte Carlo Simulation of the MLE of MSMD with $k = 8$ Multipliers.

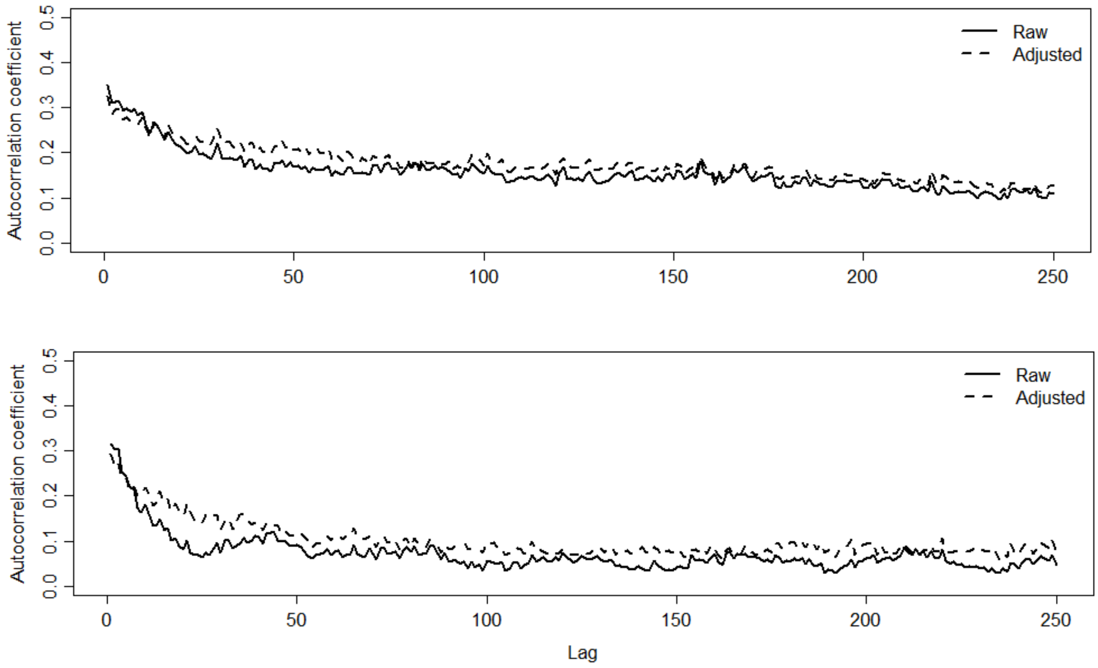
The row labeled "Mean" reports the mean parameter estimates in the simulations, "S.E" denotes standard errors and "RMSE" the root mean-square errors. The true parameter values are: $m_0 = 1.40$, $b = 2.00$, $\gamma_k = 0.5$, $\bar{\psi} = 10.0$, and $\kappa = 1.45$. The simulations for $N = 1,000$, 2,500 and 5,000 observations are based on 500, 200 and 100 replications, respectively.

		$\exp(1)$			$\mathcal{W}(\kappa)$		
		1,000	2,500	5,000	1,000	2,500	5,000
m_0	Mean	1.382	1.393	1.395	1.393	1.3993	1.400
	S.E.	0.035	0.021	0.016	0.036	0.022	0.015
	RMSE	0.039	0.022	0.017	0.037	0.022	0.015
b	Mean	1.832	1.936	1.949	1.982	1.998	2.022
	S.E.	0.346	0.250	0.162	0.438	0.258	0.180
	RMSE	0.384	0.258	0.170	0.439	0.258	0.183
γ_k	Mean	0.501	0.499	0.494	0.506	0.502	0.509
	S.E.	0.164	0.107	0.069	0.160	0.101	0.064
	RMSE	0.164	0.107	0.069	0.160	0.101	0.064
$\bar{\psi}$	Mean	9.853	9.682	9.890	10.29	10.081	10.03
	S.E.	3.475	2.420	1.781	2.740	1.732	1.410
	RMSE	3.478	2.441	1.785	2.755	1.734	1.410
κ	Mean	-	-	-	1.465	1.458	1.453
	S.E.	-	-	-	0.098	0.064	0.037
	RMSE	-	-	-	0.099	0.065	0.037

2.4 Data Description

Our empirical work is based on two data sets. We obtained high-frequency data on the S&P 500 futures contract from Tick Data for a period running from January 1996 until June 2008. We focus on transactions prices pertaining to the most liquid (front) contract traded on the Chicago Mercantile Exchange (CME) during the main U.S. trading hours of 9:30 - 16:00 EST. We also obtained price durations of the VIX (front) futures contract traded on the CBOE futures exchange over the period July 2, 2010 - December 30, 2010. The underlying asset of this futures contract is the model-free implied volatility index, VIX, calculated by the CBOE. We used the first 12,000 price durations of both datasets. The autocorrelation functions for the price durations are plotted in Figure 2.1 below. It confirms the findings many papers cited above in that the durations exhibit highly persistent behavior and the ACF tends to decay very slowly.

Figure 2.1: Autocorrelation Function of S&P 500 Futures Price Durations (top panel) and VIX Futures Price Durations (bottom panel).
The legend indicates the raw and adjusted durations.

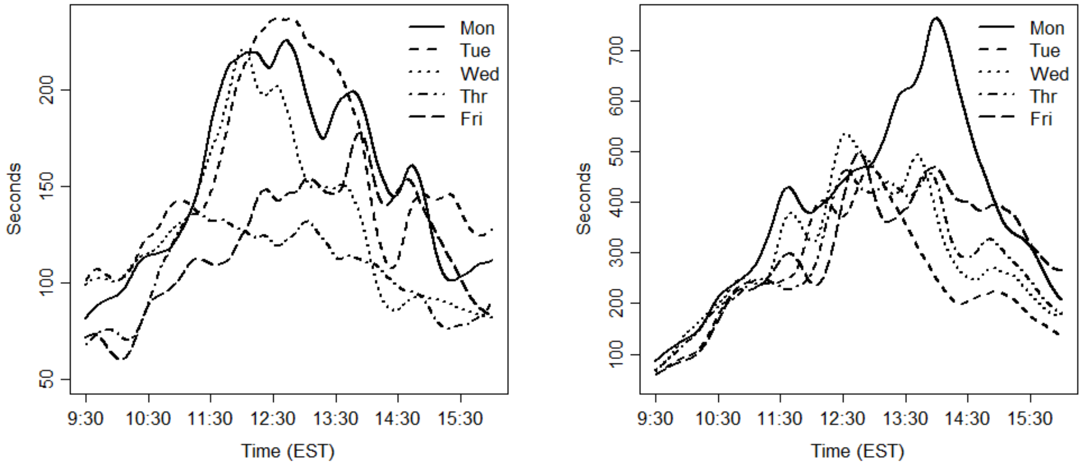


Going into detail on the definition of our price durations, for the S&P data, we selected the time between transactions with an absolute cumulative return of at least 0.05%, while for the less liquid VIX data, we required an absolute cumulative return of at least 0.25%. This approach is more scale-free than using absolute price changes.

As per SF4.1 (diurnal pattern of durations), it is well-known that trading activity in most financial markets varies considerably over the course of the day, see e.g. Engle and Russell (1998) who noted a hump-shaped pattern for transaction durations, with less time between trades at the start and end of the day, and most time during the middle of the day. Consequently, the duration process contains a significant seasonal component that has to be accounted for when estimating a duration model.

There are in principle two ways to achieve this. First, by incorporating seasonality into the duration models directly and estimating the seasonal parameters jointly with the dynamic parameters of the duration process (Rodríguez-Poo, Veredas & Espasa, 2007). Alternatively, and much more widely used, one can first estimate the seasonal component semi- or non-parametrically and fit the duration model to the seasonally-adjusted durations (e.g. Engle & Russell, 1998, and Fernandes & Grammig, 2006). Given the complexity of the duration models we are considering in this chapter, we opt for the latter approach and employ nonparametric regression (the boundary-corrected Nadaraya-Watson estimator with a quartic kernel) to estimate the seasonal component of price duration, separately for each day of the week (Bauwens & Veredas, 2004). The results are as in Figure 2.2 below.

Figure 2.2: Estimated Intraday Pattern for S&P 500 Futures Price Durations (left panel) and VIX Futures Price Durations (right panel).
The legend indicates the differing intraday patterns by day.



Returning to Figure 2.1, it is clear that the persistence in the price durations is not due to the seasonal component for even after filtering, the ACFs exhibit long-memory like behavior. However, we might distinguish between the two datasets by noting that the adjusted ACF for the VIX dataset declines faster and to a lower level than the ACF for the S&P dataset, suggesting that the VIX data exhibits less serial correlation than the S&P data.

Table 2.2: Descriptive Statistics for Price Durations.

The sample period for the S&P futures durations consists of the first 12000 observations in the period August 21, 2007 - November 16, 2007. For the VIX futures durations, the sample period comprises the first 12000 observations of the period July 2, 2010 - December 30, 2010.

	S&P 500 futures		VIX futures	
	raw	adjusted	raw	adjusted
Sample size	12,000	12,000	12,000	12,000
Mean	121.1	0.999	240.9	0.986
Median	68.00	0.599	95.00	0.488
Minimum	1.000	0.006	1.000	0.002
Maximum	3014	22.63	8086	32.87
Overdispersion	1.315	1.215	1.834	1.533

Going further, we have listed some basic summary statistics for the two datasets in Table 2.2 above. The VIX data has a higher mean, median and maximum than the S&P data. Both datasets exhibit overdispersion, confirming SF5.1 - overdispersion of durations. Since the VIX overdispersion is larger, it has a standard deviation which is even larger than in the S&P data.

2.5 Empirical Results

The following sections compare the estimation and forecast performance of the ACD, FIACD, LMSD and MSMD models. At times it may be helpful to classify the ACD, FIACD and LMSD models as *standard duration models* to enable comparison within this subset before comparison with the MSMD models.

2.5.1 Estimation Results

Table 2.3 shows the results from estimating the Exponential and Weibull forms of the ACD, FIACD and LMSD models, including Ljung-Box tests for the residuals $\{\hat{\varepsilon}_t\}$ - while the tests for the second set of residuals for the LMSD model, $\{\hat{u}_t\}$, are in Table 2.4. Plots of the residuals are in Figure 2.3, while Table 2.5 shows the results for the MSMD models. Standard errors are in parentheses, and p-values in square brackets, where applicable. Standard errors were calculated using the second derivative estimate i.e. the square root of the diagonal of $\left[-\frac{\partial^2 L(\theta)}{\partial\theta\partial\theta'}\bigg|_{\theta=\hat{\theta}} \right]^{-1}$. Note that in Table 2.3, the value of σ_u^2 for the Exponential LMSD model with the S&P dataset is actually 2.128E-04, although to 3 decimal places, it appears as 0.000. Similarly, the values of d for the LMSD model are never 0.5, but are close to it (0.499967040639577, 0.499841141572372 and 0.499999999999861). Nevertheless, the closeness of the estimates to the boundary values is a possible indicator of the model not fitting completely. Similarly, in Table 2.5, a standard error of “-” (a hyphen) occurs twice as γ_k had to be fixed at 0.999 in order to avoid hitting the boundary during estimation.

For the standard duration models, most parameters are significant at the 99% significance level except for the estimates of β for the Weibull LMSD model for both datasets. For the MSMD models, all parameters are significant except for b when $k = 2$ (Exponential MSMD) for the S&P dataset. Persistence in the datasets is reflected by the sum of $(\alpha + \beta)$ for the ACD model; as it is close to 1, this indicates autocorrelation possibly signalling long memory, which is picked up by significant values of d in the FIACD model. The relatively lower correlation in the VIX dataset is also reflected by the lower sum ($\alpha + \beta$) and value of d there. Long memory is also confirmed in the LMSD model by the significant values of d found, and partly by the persistence parameter β when the distribution is Exponential. The Weibull shape parameter κ differs from 1, the case where it would imply an Exponential distribution, although it seems to compensate for persistence, resulting in smaller β values when the distribution is Weibull. Finally, the σ_u^2 values are low implying low deviation in the second set of disturbances, $\{u_t\}$, but since σ_u^2 also contributes to the level of the autocorrelation function, the values also indicate relatively lower persistence in the LMSD processes than otherwise.

In general, the MSMD parameters display monotonic trends as the number of multipliers k increases, e.g. the highest switching probability γ_k rises as k rises, across both datasets and distributions, while b, which also governs persistence with the set $\{\gamma_i\}_{i=1}^{k}$ generally falls. This is partly expected; as the number of multipliers increases, the flexibility of the model to fit varying levels of persistence in the data increases, causing the shifts in the parameter values. For example, γ_1, the lowest switching probability (corresponding to the lowest frequency multiplier) falls as k rises, but falls at a higher rate than observed if γ_k or b are constant, rather than rising and falling respectively. The other parameters' trends are not as pronounced; the values of the multipliers, m_0 and $2-m_0$, draw closer, also perhaps a consequence of the greater number of multipliers, while the scaling constant $\bar{\psi}$ rises away from 1. Again, the Weibull shape parameter κ differs from 1 (Exponential case), rising as k rises.

The bottom sections of Tables 2.3 and 2.5 show the log-likelihood values, BIC and Sum of Relative Deviations (SRD) values for the models. Table 2.3 also shows the Ljung-Box

Q-Statistic and associated p-value for the residuals - in the case of the LMSD models, the residuals tested here are the $\{\hat{\varepsilon}_t\}$, which we call "primary residuals", while the residuals $\{\hat{u}_t\}$ in Table 2.4 are termed "secondary residuals".

For the standard duration models, the LMSD models always have the lowest log-likelihood and BIC values, and the absolutely highest SRD values. The FIACD models have the highest BIC values and the highest log-likelihoods for the VIX dataset, although the ACD models have higher log-likelihoods for the S&P dataset and absolutely smaller SRD values throughout. Overall then, the LMSD models have the worst in-sample fit and the FIACD arguably (but weakly) the best. Now comparing with the MSMD models, the Weibull MSMD models have higher log-likelihood and BIC values than all standard duration models, showing that the Weibull MSMD model fits the best in-sample. So despite it not being a true long-memory model, it outperforms the other models in terms of estimation.

A sample of the residual plots (Figure 2.3) is not conclusive on whether the data demonstrates dependence, but the Ljung-Box tests reject the null hypothesis of independently distributed residuals for all models except the ACD models, indicating that dependence persists after fitting the models. An approximate calculation can put the Q-statistics in context; the statistic $Q = N(N+2)\sum_{k=1}^{h}\frac{\hat{\rho}_k^2}{N-k}$; N = the sample size = 10000 $\gg h$ = the lag order = 50 > 2. Also, the maximum value of $\hat{\rho_k}$ is 1. Then $Q_{max} \approx N^2 \times \frac{50\times1}{N} = 50N = 500000$. So even though the statistics are large, they are not infeasible. Our estimation results are potentially weakened, although again, we refer to Deo, Hsieh and Hurvich's (2010) argument that the residuals do not tend to white noise, which possibly explains the large Q-statistics. Finally, we assert that the most important property of the models is their relative forecasting ability. We analyse this in the next subsection.

Table 2.3: Maximum likelihood estimates of the Exponential and Weibull ACD, FIACD and LMSD models for de-seasonalised S&P500 and VIX futures price durations.

Standard errors, calculated using the second derivative estimate, are reported in parentheses, p-values are reported in square brackets. All values are correct to 3 decimal places except for the BIC which is correct to 4 decimal places. The sample period for the S&P futures durations consists of the first 12000 observations in the period August 21, 2007 - November 16, 2007. For the VIX futures durations, the sample period comprises the first 12000 observations of the period July 2, 2010 - December 30, 2010.

	ACD		FIACD		LMSD	
	Exponential	Weibull	Exponential	Weibull	Exponential	Weibull
A. S&P 500 futures						
ω	0.006 (0.001)	0.005 (0.001)	0.018 (0.004)	0.018 (0.003)	–	–
α	0.089 (0.006)	0.085 (0.005)	0.171 (0.039)	0.174 (0.032)	–	–
β	0.906 (0.007)	0.911 (0.005)	0.659 (0.113)	0.652 (0.093)	0.967 (0.012)	-0.015 (0.171)
d	–	–	0.739 (0.078)	0.739 (0.064)	0.500 (0.092)	0.500 (0.031)
κ	–	1.202 (0.008)	–	1.203 (0.008)	–	1.560 (0.060)
σ_u^2	–	–	–	–	0.000 (0.000)	0.130 (0.066)
$m + \mu$	–	–	–	–	-0.537 (0.000)	-0.537 (0.000)
log L	-10184.131	-9858.858	-10186.11	-9860.431	-19840.703	-20428.710
BIC	0.0008	0.0016	0.0016	0.0024	0.0007	0.0015
SRD	-33.677	-72.881	-1146.906	-1160.692	5046.611	4697.879
Ljung-Box Q	66.250 [0.062]	65.768 [0.067]	103.956 [0.000]	106.831 [0.000]	4512.725 [0.000]	701.807 [0.000]
B. VIX futures						
ω	0.012 (0.002)	0.012 (0.002)	0.033 (0.004)	0.032 (0.005)	–	–
α	0.119 (0.007)	0.123 (0.008)	0.254 (0.036)	0.255 (0.040)	–	–
β	0.871 (0.007)	0.867 (0.008)	0.482 (0.033)	0.487 (0.037)	0.620 (0.058)	0.180 (0.105)
d	–	–	0.5583 (0.0502)	0.559 (0.056)	0.500 (0.030)	0.444 (0.020)
κ	–	0.895 (0.006)	–	0.897 (0.006)	–	1.167 (0.033)
σ_u^2	–	–	–	–	0.041 (0.009)	0.300 (0.088)
$m + \mu$	–	–	–	–	-0.889 (0.000)	-0.889 (0.000)
log L	-9416.860	-9279.574	-9384.089	-9251.690	-22651.355	-22687.678
BIC	0.0008	0.0016	0.0016	0.0024	0.0007	0.0015
SRD	105.288	143.104	-1968.862	-1969.404	10490.719	10500.716
Ljung-Box Q	99.924 [0.000]	99.626 [0.000]	208.263 [0.000]	198.408 [0.000]	192.056 [0.000]	134.772 [0.000]

Table 2.4: Ljung-Box Q-Statistics for Secondary Residuals for LMSD models.

Ljung-Box Q-Statistics with p-values in square brackets for the series $\{\hat{u}_t\}$. All values are correct to 3 decimal places. The sample period for the S&P futures durations consists of the first 12000 observations in the period August 21, 2007 - November 16, 2007. For the VIX futures durations, the sample period comprises the first 12000 observations of the period July 2, 2010 - December 30, 2010.

	Exponential	Weibull
S&P 500 Futures	888.348 [0.000]	3061.586 [0.000]
VIX Futures	257.118 [0.000]	1715.965 [0.000]

Figure 2.3: Plots of First 100 Residuals from the Standard Duration Models.

The top-left panel shows the residuals from estimating the Exponential ACD model using the S&P dataset, while the top-right panel shows the residuals for the Weibull FIACD model using the S&P dataset. The bottom panels show the primary residuals ($\{\hat{\varepsilon}_t\}$) and secondary residuals ($\{\hat{u}_t\}$) for the Weibull LMSD model using the VIX dataset.

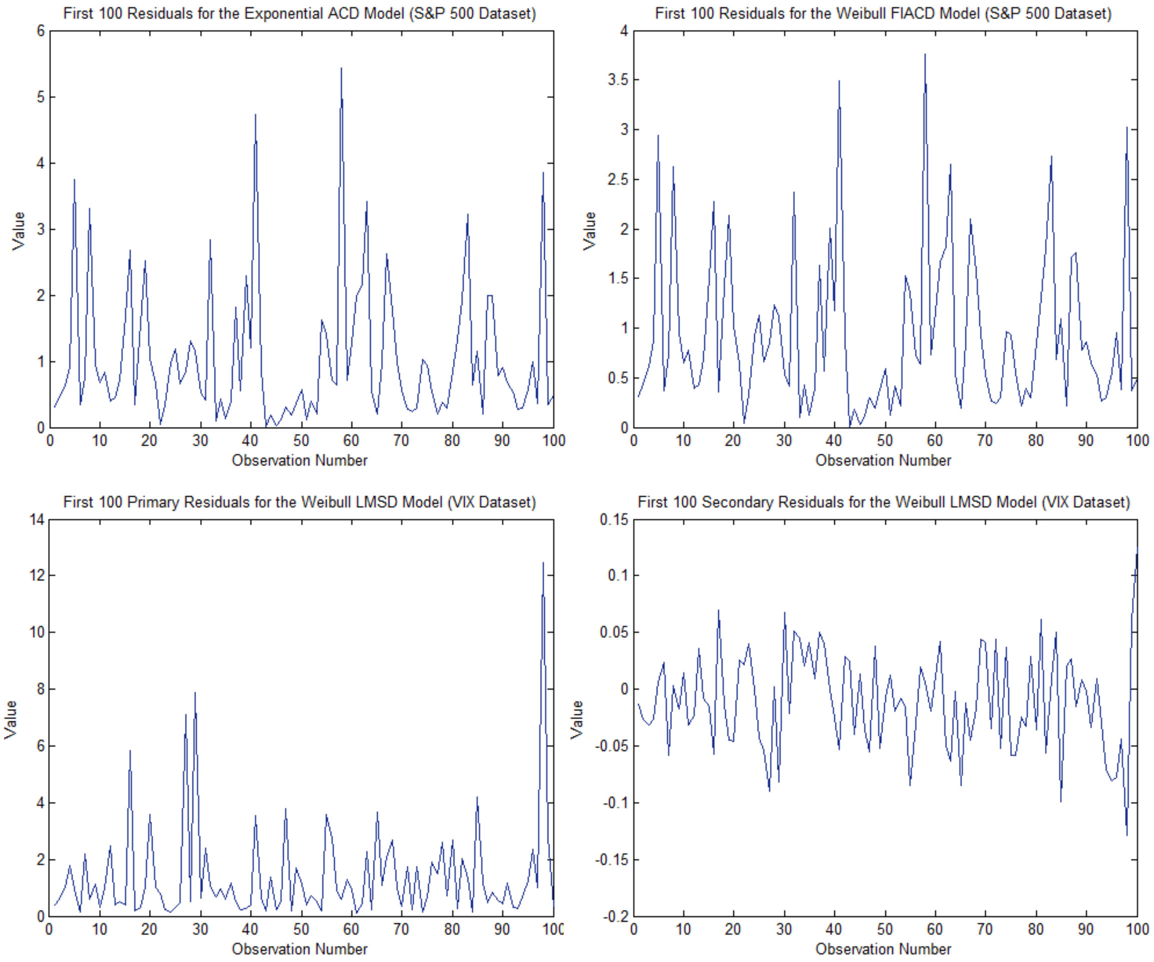


Table 2.5: Maximum likelihood estimates of the Exponential and Weibull MSMD models for de-seasonalised S&P500 and VIX futures price durations.

Standard errors, calculated using the second derivative estimate, are reported in parentheses, p-values are reported in square brackets. The sample period for the S&P futures durations consists of the first 12000 observations in the period August 21, 2007 - November 16, 2007. For the VIX futures durations, the sample period comprises the first 12000 observations of the period July 2, 2010 - December 30, 2010.

	Exponential				Weibull			
	$k=2$	4	6	8	$k=2$	4	6	8
A. S&P 500 futures								
m_0	1.370 (0.009)	1.294 (0.008)	1.244 (0.006)	1.219 (0.008)	1.399 (0.007)	1.374 (0.006)	1.326 (0.006)	1.325 (0.007)
b	10.01 (15.44)	3.487 (0.914)	1.328 (0.363)	1.257 (0.256)	10.02 (2.523)	19.25 (2.467)	11.79 (0.927)	11.46 (1.454)
γ_k	0.013 (0.004)	0.016 (0.003)	0.006 (0.003)	0.006 (0.003)	0.045 (0.007)	0.932 (0.033)	0.999 (−)	0.999 (−)
$\bar{\psi}$	1.050 (0.026)	1.070 (0.037)	1.254 (0.030)	1.283 (0.037)	1.113 (0.022)	1.040 (0.021)	1.056 (0.029)	1.187 (0.039)
κ	−	−	−	−	1.248 (0.010)	1.563 (0.022)	1.612 (0.023)	1.613 (0.024)
log L	-10383.0	-10242.7	-10207.9	-10206.5	-10007.0	-9678.7	-9594.0	-9594.4
BIC	0.0008	0.0016	0.0016	0.0016	0.0024	0.0024	0.0024	0.0024
B. VIX futures								
m_0	1.525 (0.007)	1.396 (0.009)	1.361 (0.008)	1.361 (0.009)	1.534 (0.007)	1.413 (0.009)	1.389 (0.010)	1.370 (0.011)
b	23.15 (5.102)	7.006 (0.873)	5.451 (0.599)	5.461 (0.594)	24.50 (5.055)	7.434 (0.818)	6.407 (0.655)	5.560 (0.588)
γ_k	0.169 (0.014)	0.264 (0.031)	0.308 (0.037)	0.308 (0.037)	0.193 (0.018)	0.409 (0.061)	0.820 (0.074)	0.855 (0.066)
$\bar{\psi}$	0.963 (0.021)	0.876 (0.033)	1.063 (0.039)	1.223 (0.045)	0.970 (0.020)	0.919 (0.037)	1.348 (0.055)	1.661 (0.071)
κ	−	−	−	−	1.024 (0.010)	1.066 (0.013)	1.127 (0.018)	1.126 (0.017)
log L	-9178.5	-9058.6	-9046.0	-9046.6	-9175.4	-9043.5	-9016.4	-9016.3
BIC	0.0008	0.0016	0.0016	0.0016	0.0024	0.0024	0.0024	0.0024

2.5.2 Forecast Results

Tables 2.6-2.9 show the forecast results for the duration models, with standard errors in parentheses below where applicable. The in-sample period comprises the first 10000 observations of our datasets, while the out-of-sample period contains the final 2000 observations. We calculate H-step forecasts with H ranging over 1, 5, 10 and 20, by forecasting from the end of the in-sample dataset and then successively augmenting the dataset by actual out-of-sample duration observations, one at a time. Mincer-Zarnowitz regressions, the Mean Square Error (MSE) and Mean Absolute Deviation (MAD) were then calculated using the forecasts and actual observations.

Mincer-Zarnowitz regressions are a standard means of assessing the forecasting ability of volatility models, even for multiplicative error models such as GARCH, and are employed in Calvet and Fisher (2004) and Andersen, Bollerslev and Meddahi (2004). We therefore adapt them to the duration setting, following Calvet and Fisher by using Heteroskedasticity and Autocorrelation-Consistent (HAC) standard errors. The R^2 values are similar for all models except for the FIACD models, which were substantially lower. In terms of the regression coefficients, the Wald test statistics reject the ideal of $a = 0$ and $b = 1$ for all models, although the p-values in parentheses show that the MSMD models are closest. The standard errors (figures in parentheses for the coefficients) reveal that the a and b values are generally significant, although there are exceptions for all models. The values for the ACD models are overall the closest to the ideal, while the FIACD models are the furthest, and the other two sets of models are approximately equivalent. But while the ACD models' values of a and b generally remain near 0 and 1 respectively, the other models' skew away as the forecast horizon increases, with values of a generally becoming much larger (of order upto 20) and values of b also deviating (order upto 3). This suggests that for the more complex duration models, not only does systematic (intercept) bias increase substantially, but also that forecasts react excessively to changes in the actual observation (slope).

With respect to the precision criteria, the MSE for the S&P data is generally worst for the FIACD models, with all other models being similar until the forecast period increases to 20, when the LMSD models are clearly better. In the VIX dataset, the FIACD models are again poorest, but this time the ACD models emerge as the best, followed by the MSMD and then LMSD models. In terms of the MAD, the FIACD models are again the weakest, though this time, across both datasets, the LMSD is overall the best, followed by ACD and then MSMD models. It is surprising that the FIACD model performs so poorly, but perhaps this is a result of misspecification of the d parameter in the estimation section, where it is approximately 0.74. Such a value would indicate the data series is nonstationary but mean-reverting, which is not evident from the ACF in Figure 2.2.

Table 2.6: Comparison of Out-Of-Sample Forecast Performance: 1-Step-Ahead Forecasts.

In-sample data for the S&P futures durations consists of the first 10000 observations in the period August 21, 2007 - November 16, 2007, while out-of-sample data comprised the next 2000 observations. For the VIX futures durations, the in-sample period comprises the first 10000 observations of the period July 2, 2010 - December 30, 2010, with the out-of-sample period comprising the next 2000 observations. All values are correct to 3 decimal places.

	ACD		FIACD		LMSD		MSMD					
							Exponential			Weibull		
	Exponential	Weibull	Exponential	Weibull	Exponential	Weibull	$k = 4$	6	8	$k = 4$	6	8
A. S&P 500 futures												
a	0.046 (0.046)	0.054 (0.045)	0.642 (0.049)	0.644 (0.050)	−0.149 (0.064)	0.114 (0.040)	0.121 (0.040)	0.116 (0.039)	0.119 (0.039)	0.174 (0.031)	0.143 (0.033)	0.116 (0.033)
b	0.803 (0.076)	0.789 (0.074)	−0.102 (0.065)	−0.104 (0.066)	1.384 (0.132)	0.911 (0.084)	0.733 (0.072)	0.762 (0.073)	0.763 (0.073)	0.653 (0.055)	0.740 (0.063)	0.797 (0.064)
$Wald^{a=0}_{b=1}$	47.233 (0.000)	47.454 (0.000)	367.397 (0.000)	377.741 (0.000)	18.261 (0.000)	49.309 (0.000)	17.367 (0.000)	10.941 (0.004)	10.784 (0.005)	40.073 (0.000)	18.346 (0.000)	12.749 (0.002)
R^2	0.071	0.071	0.002	0.002	0.063	0.060	0.071	0.073	0.074	0.076	0.071	0.078
MSE	0.276	0.276	0.378	0.381	0.274	0.276	0.272	0.270	0.269	0.274	0.270	0.267
MAD	0.390	0.390	0.472	0.476	0.353	0.352	0.377	0.371	0.370	0.376	0.367	0.363
B. VIX futures												
a	−0.198 (0.106)	−0.196 (0.105)	1.006 (0.127)	1.004 (0.126)	−0.141 (0.092)	−0.064 (0.077)	−0.313 (0.119)	−0.102 (0.092)	−0.101 (0.092)	−0.213 (0.107)	−0.067 (0.085)	−0.064 (0.083)
b	1.111 (0.142)	1.111 (0.141)	−0.185 (0.075)	−0.184 (0.074)	2.377 (0.273)	2.132 (0.225)	1.665 (0.227)	1.234 (0.157)	1.233 (0.157)	1.464 (0.200)	1.206 (0.146)	1.205 (0.145)
$Wald^{a=0}_{b=1}$	20.646 (0.000)	20.075 (0.000)	473.902 (0.000)	480.629 (0.000)	151.195 (0.000)	167.624 (0.000)	10.683 (0.005)	4.026 (0.134)	4.016 (0.134)	7.987 (0.018)	6.702 (0.035)	7.034 (0.030)
R^2	0.284	0.285	0.004	0.004	0.311	0.308	0.241	0.291	0.291	0.244	0.304	0.302
MSE	2.273	2.272	3.715	3.720	2.679	2.623	2.535	2.274	2.273	2.473	2.229	2.236
MAD	0.789	0.788	1.039	1.040	0.660	0.656	0.716	0.716	0.716	0.718	0.706	0.706

Table 2.7: Comparison of Out-Of-Sample Forecast Performance: 5-Step-Ahead Forecasts.

In-sample data for the S&P futures durations consists of the first 10000 observations in the period August 21, 2007 - November 16, 2007, while out-of-sample data comprised the next 2000 observations. For the VIX futures durations, the in-sample period comprises the first 10000 observations of the period July 2, 2010 - December 30, 2010, with the out-of-sample period comprising the next 2000 observations. All values are correct to 3 decimal places.

	ACD		FIACD		LMSD		MSMD					
	Exponential	Weibull	Exponential	Weibull	Exponential	Weibull	Exponential			Weibull		
							$k = 4$	6	8	$k = 4$	6	8
A. S&P 500 futures												
a	0.358 (0.212)	0.396 (0.208)	3.229 (0.236)	3.24 (0.242)	−0.655 (0.310)	0.117 (0.215)	0.623 (0.192)	0.629 (0.188)	0.651 (0.185)	0.619 (0.155)	0.422 (0.183)	0.212 (0.174)
b	0.763 (0.070)	0.751 (0.069)	−0.104 (0.061)	−0.105 (0.062)	1.329 (0.124)	1.062 (0.089)	0.718 (0.068)	0.734 (0.069)	0.732 (0.069)	0.722 (0.056)	0.842 (0.070)	0.935 (0.069)
$Wald_{b=1}^{a=0}$	50.129 (0.000)	50.822 (0.000)	449.468 (0.000)	465.231 (0.000)	14.463 (0.001)	25.487 (0.000)	24.28 (0.000)	16.554 (0.000)	16.05 (0.000)	30.202 (0.000)	5.316 (0.070)	2.291 (0.318)
R^2	0.212	0.214	0.007	0.007	0.179	0.182	0.219	0.224	0.226	0.249	0.227	0.257
MSE	1.932	1.936	4.618	4.714	1.856	1.857	1.836	1.784	1.774	1.769	1.700	1.621
MAD	1.100	1.100	1.743	1.769	0.988	0.977	1.052	1.019	1.013	1.015	0.983	0.955
B. VIX futures												
a	−0.756 (0.561)	−0.745 (0.556)	5.108 (0.656)	5.102 (0.652)	−1.108 (0.585)	−0.938 (0.563)	−1.834 (0.655)	−0.534 (0.517)	−0.53 (0.517)	−1.443 (0.615)	−0.447 (0.508)	−0.415 (0.499)
b	1.061 (0.151)	1.061 (0.150)	−0.198 (0.081)	−0.197 (0.080)	2.597 (0.344)	2.5 (0.332)	1.768 (0.252)	1.254 (0.181)	1.253 (0.181)	1.59 (0.229)	1.252 (0.179)	1.248 (0.177)
$Wald_{b=1}^{a=0}$	15.587 (0.000)	15.063 (0.001)	417.816 (0.000)	422.186 (0.000)	139.323 (0.000)	143.912 (0.000)	10.922 (0.004)	3.833 (0.147)	3.825 (0.148)	7.669 (0.022)	5.923 (0.052)	6.318 (0.042)
R^2	0.506	0.507	0.008	0.008	0.537	0.524	0.405	0.462	0.463	0.401	0.485	0.479
MSE	20.269	20.236	53.154	53.268	31.486	31.378	27.772	22.697	22.684	26.896	21.873	22.091
MAD	2.555	2.551	4.429	4.430	2.621	2.617	2.436	2.390	2.389	2.448	2.364	2.365

Table 2.8: Comparison of Out-Of-Sample Forecast Performance: 10-Step-Ahead Forecasts.

In-sample data for the S&P futures durations consists of the first 10000 observations in the period August 21, 2007 - November 16, 2007, while out-of-sample data comprised the next 2000 observations. For the VIX futures durations, the in-sample period comprises the first 10000 observations of the period July 2, 2010 - December 30, 2010, with the out-of-sample period comprising the next 2000 observations. All values are correct to 3 decimal places.

	ACD		FIACD		LMSD		MSMD					
	Exponential	Weibull	Exponential	Weibull	Exponential	Weibull	Exponential			Weibull		
							$k=4$	6	8	$k=4$	6	8
A. S&P 500 futures												
a	1.025 (0.385)	1.088 (0.379)	6.515 (0.450)	6.539 (0.461)	−1.107 (0.586)	−0.391 (0.466)	−0.156 (0.477)	−0.222 (0.491)	0.087 (0.462)	−0.556 (0.422)	0.049 (0.441)	−1.308 (0.43)
b	0.716 (0.064)	0.706 (0.063)	−0.107 (0.057)	−0.108 (0.057)	1.270 (0.114)	1.154 (0.093)	0.831 (0.073)	0.814 (0.073)	0.778 (0.069)	0.961 (0.070)	0.969 (0.083)	1.245 (0.084)
$Wald^{a=0}_{b=1}$	61.35 (0.000)	62.699 (0.000)	573.084 (0.000)	599.742 (0.000)	11.374 (0.003)	15.377 (0.000)	171.826 (0.000)	244.872 (0.000)	221.345 (0.000)	77.069 (0.000)	1.599 (0.449)	9.311 (0.010)
R^2	0.267	0.271	0.011	0.011	0.208	0.221	0.291	0.295	0.297	0.333	0.285	0.339
MSE	5.421	5.437	17.055	17.586	5.009	4.959	6.214	6.884	6.749	4.739	4.390	4.125
MAD	1.832	1.830	3.400	3.466	1.654	1.638	2.058	2.167	2.125	1.743	1.596	1.522
B. VIX futures												
a	−1.08 (1.032)	−1.063 (1.023)	10.241 (1.308)	10.23 (1.300)	−3.487 (1.327)	−3.261 (1.330)	−6.727 (1.683)	−2.522 (1.255)	−2.495 (1.253)	−6.095 (1.632)	−2.475 (1.250)	−2.301 (1.222)
b	1.016 (0.141)	1.016 (0.141)	−0.196 (0.082)	−0.196 (0.081)	2.917 (0.387)	2.857 (0.389)	2.217 (0.308)	1.445 (0.214)	1.441 (0.213)	2.042 (0.288)	1.446 (0.213)	1.437 (0.211)
$Wald^{a=0}_{b=1}$	13.429 (0.001)	12.946 (0.002)	416.916 (0.000)	421.482 (0.000)	138.773 (0.000)	140.075 (0.000)	15.999 (0.000)	4.358 (0.113)	4.302 (0.116)	14.405 (0.001)	4.531 (0.104)	4.738 (0.094)
R^2	0.552	0.553	0.009	0.009	0.587	0.573	0.440	0.467	0.467	0.437	0.489	0.483
MSE	62.114	61.926	182.411	182.833	109.153	109.48	96.839	79.504	79.442	93.951	76.797	77.531
MAD	4.521	4.511	8.364	8.366	4.876	4.876	4.388	4.328	4.328	4.407	4.289	4.278

Table 2.9: Comparison of Out-Of-Sample Forecast Performance: 20-Step-Ahead Forecasts.

In-sample data for the S&P futures durations consists of the first 10000 observations in the period August 21, 2007 - November 16, 2007, while out-of-sample data comprised the next 2000 observations. For the VIX futures durations, the in-sample period comprises the first 10000 observations of the period July 2, 2010 - December 30, 2010, with the out-of-sample period comprising the next 2000 observations. All values are correct to 3 decimal places.

	ACD		FIACD		LMSD		MSMD					
	Exponential	Weibull	Exponential	Weibull	Exponential	Weibull	Exponential			Weibull		
							$k = 4$	6	8	$k = 4$	6	8
A. S&P 500 futures												
a	2.576 (0.674)	2.681 (0.663)	13.194 (0.845)	13.252 (0.866)	-2.854 (1.160)	-2.985 (1.042)	-17.321 (2.13)	-19.128 (2.391)	-18.654 (2.347)	-18.349 (1.929)	-12.759 (1.847)	-21.481 (1.931)
b	0.678 (0.055)	0.669 (0.054)	-0.107 (0.051)	-0.108 (0.051)	1.290 (0.106)	1.316 (0.098)	1.566 (0.119)	1.553 (0.125)	1.474 (0.118)	1.797 (0.12)	1.667 (0.133)	2.256 (0.137)
$Wald^{a=0}_{b=1}$	85.509 (0.000)	88.043 (0.000)	861.833 (0.000)	919.237 (0.000)	9.270 (0.010)	13.28 (0.001)	1778.245 (0.000)	2685.583 (0.000)	3146.172 (0.000)	1229.635 (0.000)	478.008 (0.000)	725.403 (0.000)
R^2	0.326	0.331	0.015	0.015	0.224	0.256	0.361	0.369	0.373	0.410	0.34	0.415
MSE	16.131	16.193	73.511	77.15	14.312	13.892	60.469	80.254	92.649	38.603	22.413	22.872
MAD	3.146	3.142	7.137	7.326	2.885	2.818	7.154	8.361	9.055	5.607	4.087	4.151
B. VIX futures												
a	-0.007 (1.939)	0.017 (1.930)	20.125 (2.691)	20.106 (2.674)	-10.936 (3.250)	-10.417 (3.206)	-42.391 (7.251)	-19.572 (4.826)	-18.993 (4.810)	-39.504 (7.016)	-27.452 (5.932)	-23.663 (5.465)
b	0.900 (0.137)	0.901 (0.137)	-0.173 (0.086)	-0.172 (0.085)	3.382 (0.465)	3.320 (0.461)	4.089 (0.557)	2.245 (0.345)	2.150 (0.335)	3.776 (0.525)	2.327 (0.352)	2.176 (0.335)
$Wald^{a=0}_{b=1}$	6.952 (0.031)	6.628 (0.036)	402.077 (0.000)	406.175 (0.000)	121.998 (0.000)	121.628 (0.000)	65.330 (0.000)	31.855 (0.000)	35.216 (0.000)	65.586 (0.000)	118.496 (0.000)	89.304 (0.000)
R^2	0.493	0.494	0.007	0.007	0.539	0.526	0.460	0.412	0.405	0.454	0.437	0.422
MSE	251.129	250.349	639.22	640.667	423.191	424.42	391.896	345.182	342.922	384.923	345.591	341.962
MAD	8.876	8.862	15.965	15.966	9.461	9.455	9.266	9.453	9.628	9.374	11.044	10.707

In terms of the relative performance of the different distributions, with respect to the log-likelihood and BIC values, the estimates for the ACD and FIACD models are better with the Weibull distribution, though the Exponential is better for the LMSD model. These results are exactly reversed with respect to the SRD values, except for the LMSD models with the VIX dataset. Overall, this indicates the distributions are approximately equivalent. For the forecasts, the distributions are again approximately equivalent, although the Weibull forecasts are arguably slightly better overall. The Weibull MSMD models generally outperform the Exponential MSMD models in terms of estimation and forecasting too. However, the result that Exponential estimates can outperform Weibull at times is surprising with respect to estimation, since the Weibull specifications encompass the Exponential. Given the truncations implicit in implementing the models and also the (Matlab) optimisation routines, it is perhaps possible that a simpler specification may optimise with better properties than a richer one.

Note that we trialled different values of χ, the number of coefficients to include before truncation / the size of the Kalman Filter system used to generate forecasts for the LMSD models. As Table 2.10 shows, we found that using a value of 50 was best in terms of minimising the MSE for the Weibull version applied to S&P data, so used this value for all our forecasting results. This value also dramatically improved the speed of forecasting, a pertinent issue when contrasting with the more time-consuming MSMD process.

Table 2.10: Sensitivity of Forecast Results to Choice of Truncation Parameter χ

As χ increases, the mean MSE rises from a minimum at $\chi = 50$ to a maximum at $\chi = 200$ before falling again.

χ	Mean MSE across all H
50	51.998
100	54.723
150	56.189
200	60.389
250	57.685
300	58.254

2.6 Conclusion

It was the aim of this chapter to make an empirical comparison of durations models which can exhibit high order serial correlation via Long Memory or Regime Shifts, in order to be able to fit SF3.1-2 (duration clustering and autocorrelation in counts). This analysis was undertaken using the two prominent alternatives within the durations literature, the FIACD and LMSD models, and by adapting a new volatility model to the durations literature, the MSMD model. Different distributions for the innovations term driving the models were considered - Exponential and Weibull.

The MSMD models fit the data best in-sample, with higher log-likelihood and BIC values than the FIACD, ACD and LMSD models. Out of sample, the conclusion is not as definite. Within the standard duration models, the ACD model performs the best in terms of the Mincer-Zarnowitz regression. Arguably though, the LMSD is best with respect to MSE and MAD measures - it is only dominated by the ACD and MSMD models in the VIX dataset, although this data features less high order serial correlation so may be in less need of a long memory model.

Overall, these findings suggest that the LMSD specification is more powerful than the others for forecasting, while the MSMD is best for in-sample fit. However, the standard ACD model does reasonably well throughout and its relative parsimony and higher adoption rate give it clear advantages over the other models. Interestingly, the divergence between the LMSD estimation and forecast performance contrasts with the observation that SV models in the volatility literature are less good at forecasting than estimating. In terms of the distribution of model innovations, the richer specification of the Weibull distribution (which nests the Exponential distribution), did not enable substantially better estimates.

The partially superior estimation performance of the MSMD model shows that Regime Shifts may also be a useful modelling approach. However, the computational burden required to estimate the MSMD model with a high number of multipliers is a major hurdle. As Table 2.11 below shows, the standard duration models can be estimated much more quickly:

Table 2.11: Estimation Times for Duration Models

Estimation times for the more computationally intensive Weibull versions of models are shown below.

WACD	WFIACD	WLMSD	WMSMD		
			$k = 8$	9	10
10s	1min	5mins	1hr	6hrs	75hrs

Overall, we conclude that the ACD model is the best model for modelling durations when serial correlation is not large, since its advantages in terms of estimation performance and speed sufficiently outweigh any disadvantages in terms of estimation and forecast performance with respect to the MSMD and LMSD models. However, the LMSD and MSMD models are notable competitors, for their forecasting and estimation superiority respectively, especially for data which exhibits high order serial correlation, such as that of the S&P dataset. Since forecasting is key though, and the LMSD is so much faster to estimate than the MSMD, we conclude that the LMSD is the best model in the presence of serial correlation. The MSMD may become increasingly attractive as computing power continues to improve, or if simplifications can be made to its estimation process, perhaps via Whittle estimation as is used for the LMSD model (in fact, forthcoming work by Žikeš includes the Whittle form of the MSMD model).

Further work would concentrate on expanding the analysis to include more distributions, namely the Burr and Generalised Gamma distributions, and to explore the findings over multiple datasets. It would also be useful to compare the various models' performance in forecasting realized volatility and to investigate whether long memory in durations generated by the MSMD model can propagate to long memory in realized volatility, as shown for the LMSD model by Deo, Hurvich, Soulier and Wang (2009).

APPENDIX A: DISTRIBUTIONS OF INNOVATION ϵ_t

A.0.1 Exponential

Density:

$$f_E(\epsilon; \theta_E) = \exp(-\epsilon),$$

Distribution function:

$$F_E(\epsilon; \theta_E) = 1 - \exp(-\epsilon)$$

A.0.2 Weibull

Density:

$$f_W(\epsilon; \theta_W) = \kappa \vartheta_W^{\kappa} \epsilon^{\kappa-1} \exp(-\vartheta_W^{\kappa} \epsilon^{\kappa}),$$

Distribution function:

$$F_W(\epsilon; \theta_W) = 1 - \exp(-\vartheta_W^{\kappa} \epsilon^{\kappa}),$$

where

$$\vartheta_W = \Gamma(1 + 1/\kappa)$$

A.0.3 Burr

Density:

$$f_B(\epsilon; \theta_B) = \frac{\kappa \vartheta_B^{\kappa} \epsilon^{\kappa-1}}{(1 + \delta \vartheta_B^{\kappa} \epsilon^{\kappa})^{1+1/\delta}},$$

Distribution function:

$$F_B(\epsilon; \theta_B) = 1 - (1 + \delta \vartheta_B^{\kappa} \epsilon^{\kappa})^{-1/\delta},$$

where

$$\vartheta_B = \frac{\Gamma(1+1/\kappa)\Gamma(1/\delta-1/\kappa)}{\delta^{1+1/\kappa}\Gamma(1+1/\delta)}$$

A.0.4 Generalized Gamma

Density:

$$f_G(\epsilon;\theta_G) = \frac{\vartheta_G^{\kappa\delta}\kappa\epsilon^{\kappa\delta-1}}{\Gamma(\delta)}\exp(-\vartheta_G^{\kappa}\epsilon^{\kappa}),$$

where

$$\vartheta_G = \Gamma(\delta+1/\kappa)/\Gamma(\delta)$$

Distribution function: not available in closed form.

Appendix B: Moments of the MSM model

The unconditional moments of x_t are given by:

$$\begin{aligned}
\mathrm{Var}(x_t) &= \bar{\psi}^2(\mathbb{E}(\psi_t^2)\mathbb{E}(\epsilon_t^2) - \bar{\psi}^2), \\
\mathrm{Cov}(x_t, x_{t-h}) &= \bar{\psi}^2(\mathbb{E}(\psi_t\psi_{t-h}) - 1)
\end{aligned}$$

Lux (2008) shows that

$$\mathbb{E}(\psi_i^2) = \begin{cases} (0.5(m_0^2 + (2-m_0)^2))^k & \text{binomial MSM} \\ \\ \exp(2\lambda k) & \text{lognormal MSM} \end{cases}$$

$$\mathbb{E}(\psi_i\psi_{i-h}) = \begin{cases} \prod_{i=1}^k\{0.5(1-(1-\gamma_i)^h)m_0(2-m_0)+ & \\ \\ ((1-\gamma_i)^h + 0.5(1-(1-\gamma_i)^h))(0.5(m_0^2+(2-m_0)^2))\} & \text{binomial MSM} \\ \\ \prod_{i=1}^k\{(1-(1-\gamma_i)^h) + (1-\gamma_i)^h\exp(2\lambda)\} & \text{lognormal MSM} \end{cases}$$

Chapter 3: An Exploration of Volumes across Sampling Intervals

3.1 Introduction

In this chapter, we wish to gain a thorough understanding of volume dynamics. We wish to determine whether features in the data are purely a consequence of aggregation from higher time frequencies, or if instead these same features are intrinsic in volume data at all time frequencies. For example, Gallant, Rossi and Tauchen (1992) found that daily volume data has a quadratic time trend, we wish to investigate whether the same trend exists across all frequencies or whether it is purely a consequence of aggregation of linear time trends from a higher frequency (e.g. a time trend at a high frequency might be represented by $t \in \mathbb{Z}^+$, a linear trend, but then at a lower frequency it would aggregate as: $\sum\limits_{t=1}^{T} t = \frac{1}{2}(T^2 + T)$, which is a quadratic trend). Modelling and forecasting volume is important as it enables strategies which trade closer to the Volume Weighted Average Price (VWAP) benchmark and so can reduce execution risk when trading.

In studies such as those by Bollerslev and Jubinsky (1999) and Lobato and Velasco (2000), volumes are investigated in reference to share return volatility, usually in line with the Mixture of Distributions Hypothesis put forward by Press (1967), among others. However, new work by Darolles and le Fol (2003) and Bialkowski, Darolles and le Fol (2008) investigates volume on its own. We will follow this latter approach in conducting a thorough exploration of volume data in its own right. A better univariate understanding might also better inform joint modelling with other variables in the long run.

Darolles and le Fol, and Bialkowski, Darolles and le Fol suggest that conventional time series models such as ARMA and Self-Exciting Threshold AutoRegressive (SETAR) can be applied to the data, while Manganelli (2002) has specified volumes as a Multiplicative Error Model. Manchaldore, Palit and Soloviev (2010) in turn model the volume process directly in a continuous time framework, but we will concentrate on discrete-time modelling. It is well known that daily volumes exhibit the property of long memory (as found by Bollerslev and Jubinsky, and Lobato and Velasco), so we will extend to AutoRegressive Fractionally Integrated Moving Average (ARFIMA) as opposed to ARMA models.

Bialkowski, Darolles and le Fol investigated the data using 20-minute intervals, while Gallant, Rossi and Tauchen, Bollerslev and Jubinsky, and Lobato and Velasco conducted their volume analyses using daily data. However, we argue that there may be information embedded in the different frequencies of data, so will explore data over multiple frequencies, ranging from half-minute to daily data, in order to understand where certain features of the data may be coming from i.e. does a trend at the daily frequency also exist at higher frequencies? Alternatively, if the same features exist across all data frequencies, our findings will be robust to the data frequency. However, since the processing of the dataset over multiple frequencies is computationally intensive (as will be detailed in the next section), we will concentrate on application of simple time series methods to the dataset. Essentially, we will explore volume data using the Classical Decomposition method as outlined in Brockwell and Davis (1991).

Linking to another strand of literature, literature exists on how time series processes aggregate. The principle result we reference is that of Chambers (1998), who showed that stochastic processes with an order of fractional integration, d, such that $-0.5 < d < 0.5$, aggregate to have the same order of fractional integration no matter what the level of aggregation. This theoretical result holds for continuous and discrete-time processes, and also for both stock and flow processes.

However, he tested this result empirically using only quarterly and annual flow data, so any support we find for his work over a larger range of frequencies is new, and we know of no other research which has been conducted at as high a maximal frequency of data as ours (most studies have a maximal frequency of daily data, whereas ours is 30 seconds). Furthermore, Chambers' paper also highlights a divergence which can occur between theory and empirics; despite his theoretical result, he found that the estimates of d can vary substantially across the time frequencies he investigated (for example, his estimates of d for the UK Investment Expenditure time series during the period 1955 to 1992 are 0.4439 for quarterly data and 0.9217 for annual data). He ascribes this to a combination of sampling variation and possible misspecification in the estimation of the orders p and q. He also cites Granger (1980) as saying : "researchers should not completely believe their identified models, and so should not be surprised if the results from aggregation theory do not work perfectly with estimated models".

Further results exist with respect to ARFIMA specifications after temporal aggregation. Tschernig (1994) states that Baillie, Nijman and Tschernig (1994) found that the class of ARFIMA processes is not closed under temporal aggregation; an ARFIMA(p, d, q) process aggregates to an ARFIMA(p, d, ∞) process, although the aggregated process can be approximated by an ARFIMA(p, d, q^*) model with a low value of q^*. Man and Tiao (2006) explored approximating processes further with respect to ARFIMA(0, d, 0) processes where d is now any positive real number, and showed that as the level of aggregation tends to infinity, the aggregate process tends to an ARFIMA(0, d, $\lfloor d+1 \rfloor$) where $\lfloor . \rfloor$ denotes the smallest integer smaller than its argument. Finally, Tsai and Chan (2005) found that an ARFIMA (p, d, q) process aggregates to a process with the same autocorrelation properties as a Continuous-Time ARFIMA (CARFIMA) model with a similar level of integration.

Note that we will investigate the properties of volume over different time frequencies in order to understand aggregation effects better. So we will investigate 11 separate series. An alternative approach is to assume that the time series at different frequencies all have the same underlying series, and then investigate this underlying series using tests which jointly incorporate all time frequencies, principally the modified Geweke-Porter-Hudak (GPH) test of Ohanissian, Russell and Tsay (2008). However, we feel we may overlook features of the data if we do not investigate it over separate frequencies, moreover, we do not know of corresponding (joint multiple frequency) tests for stochastic trendedness, serial correlation and heteroskedasticity, and given Chambers' aforementioned noting of possible divergence between theory and empirics, we prefer to investigate the data ourselves.

Finally, we acknowledge that long memory is a modelling feature with an alternative in regime-switching; both features can explain high order serial correlation. However, as pointed out by Diebold and Inoue (2001) and Fleming and Kirby (2001), long memory may be a convenient description regardless of which feature is correct. Diebold and Inoue further state that there may be no "correct" approach: "structural change and long memory are effectively different labels for the same phenomenon, in which case attempts to label one as true and the other as spurious may be of dubious value". Furthermore, Fleming and Kirby point out that forecasting regime switches is difficult so long memory forecasts are easier to generate. Henceforth, regardless of which approach might be better, we refer to the phenomenon generating high order serial correlation as "long memory".

Overall then, we aim to answer the following questions:

Q1. (Data Exploration - Sections 3.2 and 3.3) Are there any trends or features of the data over the different sampling intervals? E.g. long memory and trendedness at lower frequencies may be a result of aggregation of data at higher frequencies.

Q2. (Model Simulation - Sections 3.4 and 3.5) If we simulate a process for a given frequency and aggregate it to produce data at lower frequencies, do we obtain similar features to those in the data? We will explore this principally in terms of how the estimated memory parameter behaves, but also perform diagnostics to check that the aggregated models are still valid.

This chapter addresses these questions through four sections. The first section, 3.2, introduces the dataset and the computing system created to conduct the analysis. In a sense, it is the crucial section as it underpins all else, and itself required much research, even though it is not a result. Section 3.4 outlines the empirics of the dataset with regard to stationarity and correlation properties, yielding two patterns in the memory of data as the time frequency reduces / sampling interval increases. We attempt to determine if ARFIMA processes might underlie and aggregate in a consistent way over the time frequencies in Section 3.4, and finally investigate whether the deseasonalisation procedure might be producing one of the patterns in the memory of data in Section 3.5. We end with a conclusion and suggest possible future work.

3.1.1 Terminology

In what follows we make use of certain non-standard terms to communicate the key concepts. First, we prefer to use the term *(sampling) intervals* to the *frequency* of data, as we wish to reserve frequency to refer to the angular frequencies used in the computation of the periodogram of the volume process. So analysis of data at "high frequency" is synonomous with analysis over "small sampling intervals" (e.g. 1 minute), while "low frequency" is equivalent to a "large sampling interval" (e.g. 1 day).

As stated before, *long memory* corresponds to the memory parameter, $d \in (0, 0.5)$, and *short memory* corresponds to $d = 0$. As Brockwell and Davis (1991) observe, some authors further specify $d \in (-0.5, 0)$ as *intermediate memory*. We also refer to *memory patterns*; these are patterns in the memory parameter of volume across sampling intervals. E.g. as the sampling interval increases from 30s to 1 day, the memory parameter may rise through intermediate memory, short memory and then long memory, producing an *increasing memory pattern*.

Overall, we are investigating the data over different intervals in the expectation of a memory pattern. If this memory pattern is what is predicted as a result of theoretical or simulated aggregation from smaller sampling intervals, then we say that the memory pattern has *aggregational consistency*. A related concept is that of an ARFIMA specification being *closed under temporal aggregation / closed to time aggregation*, meaning that an ARFIMA process maintains its values of p, d, and q as it is aggregated temporally. Our notion of aggregational consistency relates more to an observed memory pattern associated with d, whether stable or changing, and if we can fit this.

Finally, to enable easier identification of patterns in the data, we frequently make use of *colour scales* in tables: gradients in the colour of cells which demonstrate which cells contain relatively large or small values. For example, a typical colour scale for memory parameter tables is red for the most negative value (say -0.5), transitioning to white at 0, and then transitioning to green at the most positive value (say 0.5).

3.2 Data and Computing Design

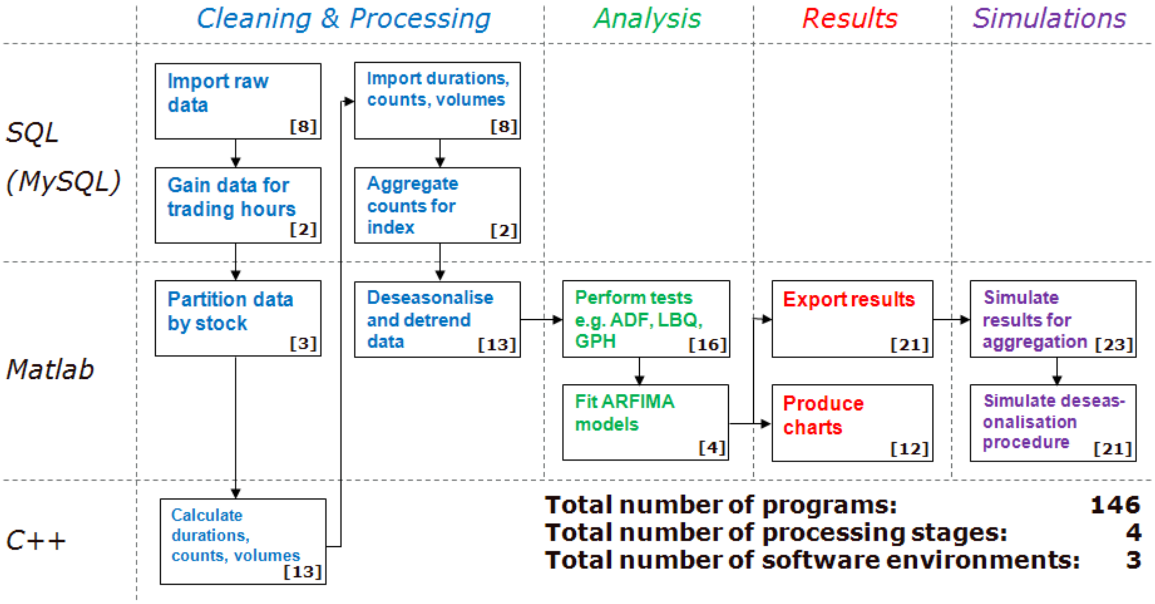
High-frequency data was obtained on FTSE shares via Ionic Information from the LSE. This dataset contains the timestamp for every trade and quote related to a stock and the associated traded price and volume, and bid and ask prices respectively. The period runs from the beginning of April, 2009 until the end of November, 2009 (8 months). The top 45 stocks of the FTSE 100 by traded volume were selected to ensure that the shares were liquid enough that data was not sparse. The full list of stocks is presented in Appendix A, Table A.2. Finally, the data was aggregated to create an index, which should have similar features to the FTSE 100 index, although more manageable computationally (45 as opposed to 100 stocks).

Altogether, this dataset comprises 1,095,469,908 (1 bn) items of data / 91,289,159 (91 m) records. Methods for processing such a large quantity of data were investigated. Pre-existing software did not exist for this, requiring custom components to be built. A design of the necessary computing architecture was created, as per Figure 3.1 on the next page. Each box represents a component of the computing effort, with a count of the number of functions used to implement the component in the lower right-hand corner. The boxes are arranged in rows and columns enclosed by dashed lines - each row or column is called a “lane”.

Overall, the data passed through 4 stages of processing and 3 different software environments, making use of 146 programs (built and tested for accuracy myself, additional programs were used from Matlab and from Kanzler (1998)). This explains the significant amount of time spent in completing the work and the relative simplicity of analysis. However, it will now be faster to process new datasets or extend the architecture to enable deeper analysis in further research.

Figure 3.1: Data Processing Flow

Data was processed in the sequence shown by the arrows below. The number of custom programs created for each software component (box) is shown in square brackets in the lower right corner. Vertical lanes (e.g. "Cleaning and Processing") represent processing stages, horizontal lanes represent software environments.


Cleaning & Processing
Analysis
Results
Simulations
SQL (MySQL)
Matlab
C++
Import raw data [8]
Gain data for trading hours [2]
Import durations, counts, volumes [8]
Aggregate counts for index [2]
Partition data by stock [3]
Deseasonalise and detrend data [13]
Perform tests e.g. ADF, LBQ, GPH [16]
Fit ARFIMA models [4]
Export results [21]
Produce charts [12]
Simulate results for aggregation [23]
Simulate deseas-onalisation procedure [21]
Calculate durations, counts, volumes [13]
Total number of programs: 146
Total number of processing stages: 4
Total number of software environments: 3

Further detail will now be given on the data processing stages (vertical lanes of the diagram) and the software environments (horizontal lanes).

3.2.1 Data Processing Stages (Vertical Lanes)

Cleaning and Processing

Transactional datasets are very large and contain errors in some records, so significant preprocessing was required to clean the data and calculate the variable measures required. First, data was imported. Anomalous data, such as negative bids, were removed and the data for the trading hours (08:00 - 16:30) was selected. Then midquotes, durations, counts, squared returns, realized volatilies and absolute trading volumes were calculated for the individual stocks and aggregated for the index proxy. Finally, data was deseasonalised and detrended as will be outlined in Section 3.3.

Tests

ACFs and PACFs were calculated for the data over any sampling interval. Tests of nonstationarity, serial correlation and long memory were also conducted, and ARFIMA specifications were fitted.

Results

Spreadsheets and charts based on the data were generated automatically using Matlab.

Simulations

Aggregation behaviour was analysed by simulating over the sampling intervals and aggregating to compare with actual specifications over larger intervals. Data was also simulated to check the effect of the deseasonalisation procedure on data with and without seasonal components.

3.2.2 Software Environments (Horizontal Lanes)

The vast quantities of data required fast software to be used. SQL is a 4th generation programming language designed for storage and fast selection and manipulation of records, so was mainly used for cleaning and aggregating the data, in a MySQL database (freely scaleable).

C++ is powerful and industry standard for 3rd generation languages so was used to perform the basic calculations of quantities such as counts and squared returns.

However, as observed by Chan (2009), the gain in processing speed from using C++ can be outweighed by its debugging and testing time when compared to Matlab, whose visual environment makes prototyping much easier. Therefore Matlab was used for more complex procedures such as those required for deseasonalising the data and performing tests. Matlab is also able to connect directly to the MySQL database unlike C++.

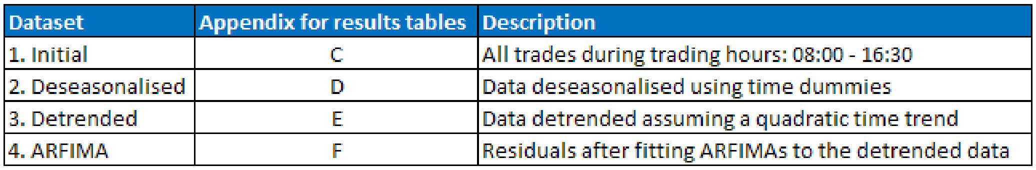
We acknowledge that alternative computing designs could have been developed - one prominent alternative is kdb+, a type of database software from Kx Systems. However, at most, kdb+ would have been able to replace the SQL and C++ software environments; it does not contain as extensive an array of prebuilt time series functions as Matlab, nor can it automatically generate custom Excel spreadsheets as were used to create the results tables in Appendices C-F. Overall then, Matlab enabled the majority of processing (113 out of the 146 functions built were in Matlab) so was an irreplaceable software environment, at least to an academic researcher with fewer resources than in industry.

3.3 Data Exploration

We proceeded to investigate the data for patterns, following a Classical Decomposition procedure similar to that outlined in Brockwell and Davis (1991); we checked for and removed seasonal and trend components from the time series before fitting ARFIMAs. Using the computing setup above, we were able to aggregate the high frequency data into the following fixed / regularly-spaced sampling intervals: 30s, 60s, 120s, 300s, 600s, 1200s, 1800s, 3600s, 7200s, 14400s and 30600s, or alternatively $\frac{1}{2}$ min, 1 min, 2 mins, 5 mins, 10 mins, 20 mins, $\frac{1}{2}$ hr, 1 hr, 2 hrs, 4 hrs and 1 trading day (7.5 hours). Our investigation produced 4 successive datasets, as detailed in Table 3.1 below. For ease of display, we show data for only 15 of the stocks which were selected randomly. The list is given in Appendix A, Table A.1.

Table 3.1: Datasets and Descriptions

A guide to the datasets - the "Description" column details each dataset, while the middle column shows which appendix the tables are in.



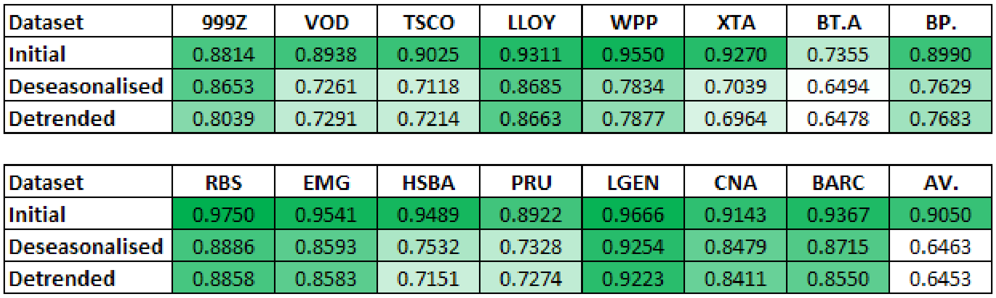
Dataset	Appendix for results tables	Description
1. Initial	C	All trades during trading hours: 08:00 - 16:30
2. Deseasonalised	D	Data deseasonalised using time dummies
3. Detrended	E	Data detrended assuming a quadratic time trend
4. ARFIMA	F	Residuals after fitting ARFIMAs to the detrended data

For each dataset, we followed the same diagnostic procedure: 1) examine the series, 2) perform Augmented Dickey-Fuller (ADF) tests, 3) perform Ljung-Box Q (LBQ) tests of order 50 (altering the order did not affect the results), and 4) estimate the memory parameter using the GPH estimator. Table 3.1 also indicates which Appendix contains the results for each dataset. A key to the results is in Appendix B.

As a check of the reliability of the computing design, we compared our data at each stage of processing with daily Datastream data over the same period (note that fitting ARFIMAs caused a reduction in the number of observations so this dataset could not be compared). Table 3.2 shows the correlation figures for the 15 stocks and the aggregate/index analogue ("999Z") - the table has been split to fit on the page. It clearly shows generally high correlation especially for the Initial dataset, with correlation reducing but still high over the various stages of processing. This confirms the integrity of our datasets.

Table 3.2: Datasets' Correlation with Datastream Daily Data

Correlation between the data we use from Ionic Information and Datastream daily data for 15 stocks and the aggregate/index analogue ("999Z"). Colour scale: The darker the green in the table, the closer the correlation is to 1.



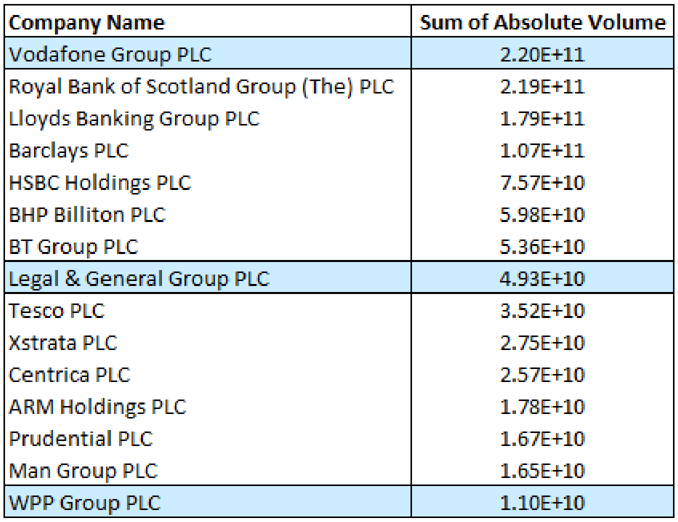
Dataset	999Z	VOD	TSCO	LLOY	WPP	XTA	BT.A	BP.
Initial	0.8814	0.8938	0.9025	0.9311	0.9550	0.9270	0.7355	0.8990
Deseasonalised	0.8653	0.7261	0.7118	0.8685	0.7834	0.7039	0.6494	0.7629
Detrended	0.8039	0.7291	0.7214	0.8663	0.7877	0.6964	0.6478	0.7683

Dataset	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV.
Initial	0.9750	0.9541	0.9489	0.8922	0.9666	0.9143	0.9367	0.9050
Deseasonalised	0.8886	0.8593	0.7532	0.7328	0.9254	0.8479	0.8715	0.6463
Detrended	0.8858	0.8583	0.7151	0.7274	0.9223	0.8411	0.8550	0.6453

We will now proceed to discuss the results from our investigation. The diagnostics for all 15 companies' volumes are shown in Appendices C-F, but with reference to basic statistics and charts during the following analysis, we will only display these for 3 companies, which span the range of liquidity as measured by the sum of absolute volume over the data period (i.e. the total number of contracts traded for the stocks). There are highlighted in Table 3.3 below - they are Vodafone Group PLC (highest liquidity), Legal and General Group PLC (median liquidity) and WPP Group PLC (lowest liquidity). Henceforth we will abbreviate these companies' names to Vodafone, Legal and General, and WPP.

Table 3.3: Liquidity of Stocks

Liquidity of Stocks as measured by the Total Number of Contracts traded during the period April 2009 to November 2009 (inclusive). Blue highlighting shows the highest, median and lowest liquidity.



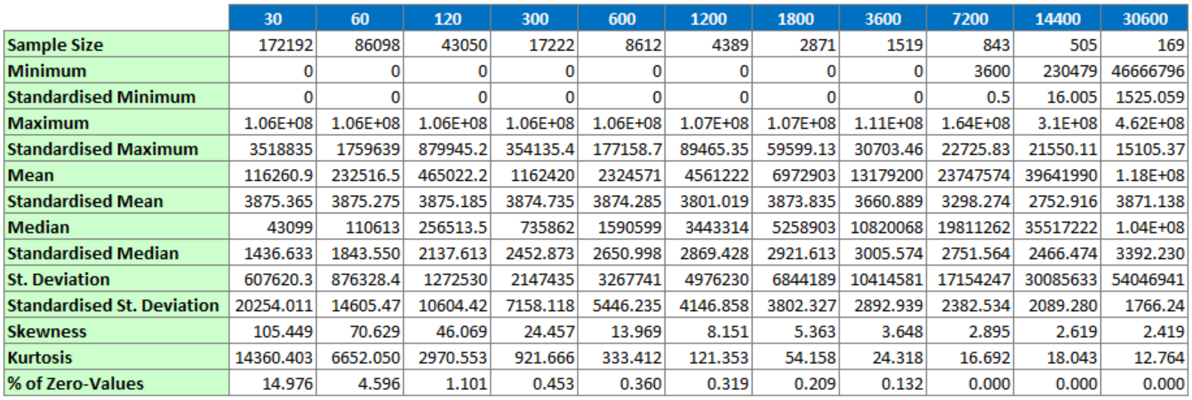
Company Name	Sum of Absolute Volume
Vodafone Group PLC	2.20E+11
Royal Bank of Scotland Group (The) PLC	2.19E+11
Lloyds Banking Group PLC	1.79E+11
Barclays PLC	1.07E+11
HSBC Holdings PLC	7.57E+10
BHP Billiton PLC	5.98E+10
BT Group PLC	5.36E+10
Legal & General Group PLC	4.93E+10
Tesco PLC	3.52E+10
Xstrata PLC	2.75E+10
Centrica PLC	2.57E+10
ARM Holdings PLC	1.78E+10
Prudential PLC	1.67E+10
Man Group PLC	1.65E+10
WPP Group PLC	1.10E+10

3.3.1 Initial Data

Tables 3.4-3.6 below show summary statistics for Vodafone, Legal and General, and WPP respectively. In order to ensure comparability across sampling intervals for the minimum, maximum, mean, median and standard deviation, we have produced versions of these statistics divided by the number of seconds in the sampling interval, e.g. the mean for Vodafone at the 30s sampling interval is 116260.9, but the standardised mean below it is 3875.365.

Table 3.4: Summary Statistics for Absolute Volume for Vodafone across Sampling Intervals

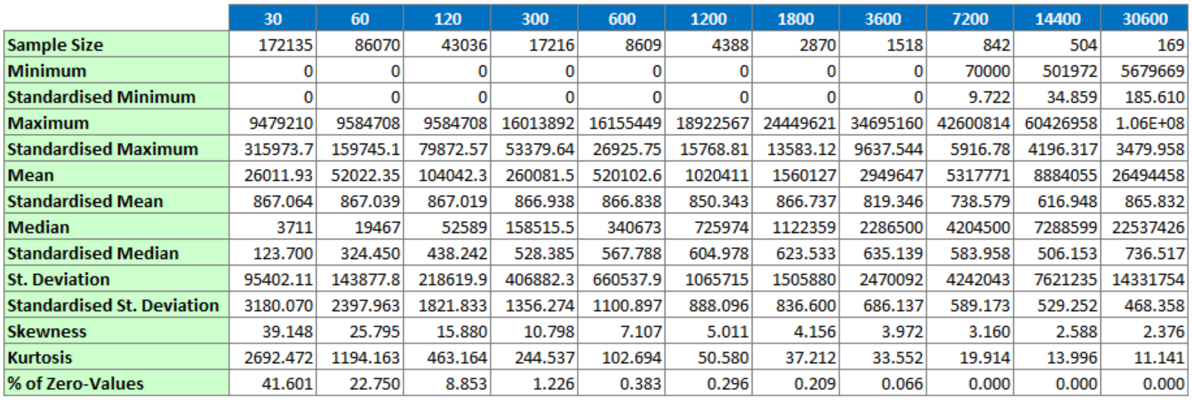
Each column shows statistics for a different sampling interval, ranging from 30s to 30600s (1 day). Statistics with ”Standardised” in their name, e.g. ”Standardised Mean”, have been divided by the number of seconds in the sampling interval to enable comparability across columns.



	30	60	120	300	600	1200	1800	3600	7200	14400	30600
Sample Size	172192	86098	43050	17222	8612	4389	2871	1519	843	505	169
Minimum	0	0	0	0	0	0	0	0	3600	230479	46666796
Standardised Minimum	0	0	0	0	0	0	0	0	0.5	16.005	1525.059
Maximum	1.06E+08	1.06E+08	1.06E+08	1.06E+08	1.06E+08	1.07E+08	1.07E+08	1.11E+08	1.64E+08	3.1E+08	4.62E+08
Standardised Maximum	3518835	1759639	879945.2	354135.4	177158.7	89465.35	59599.13	30703.46	22725.83	21550.11	15105.37
Mean	116260.9	232516.5	465022.2	1162420	2324571	4561222	6972903	13179200	23747574	39641990	1.18E+08
Standardised Mean	3875.365	3875.275	3875.185	3874.735	3874.285	3801.019	3873.835	3660.889	3298.274	2752.916	3871.138
Median	43099	110613	256513.5	735862	1590599	3443314	5258903	10820068	19811262	35517222	1.04E+08
Standardised Median	1436.633	1843.550	2137.613	2452.873	2650.998	2869.428	2921.613	3005.574	2751.564	2466.474	3392.230
St. Deviation	607620.3	876328.4	1272530	2147435	3267741	4976230	6844189	10414581	17154247	30085633	54046941
Standardised St. Deviation	20254.011	14605.47	10604.42	7158.118	5446.235	4146.858	3802.327	2892.939	2382.534	2089.280	1766.24
Skewness	105.449	70.629	46.069	24.457	13.969	8.151	5.363	3.648	2.895	2.619	2.419
Kurtosis	14360.403	6652.050	2970.553	921.666	333.412	121.353	54.158	24.318	16.692	18.043	12.764
% of Zero-Values	14.976	4.596	1.101	0.453	0.360	0.319	0.209	0.132	0.000	0.000	0.000

Table 3.5: Summary Statistics for Absolute Volume for Legal and General across Sampling Intervals

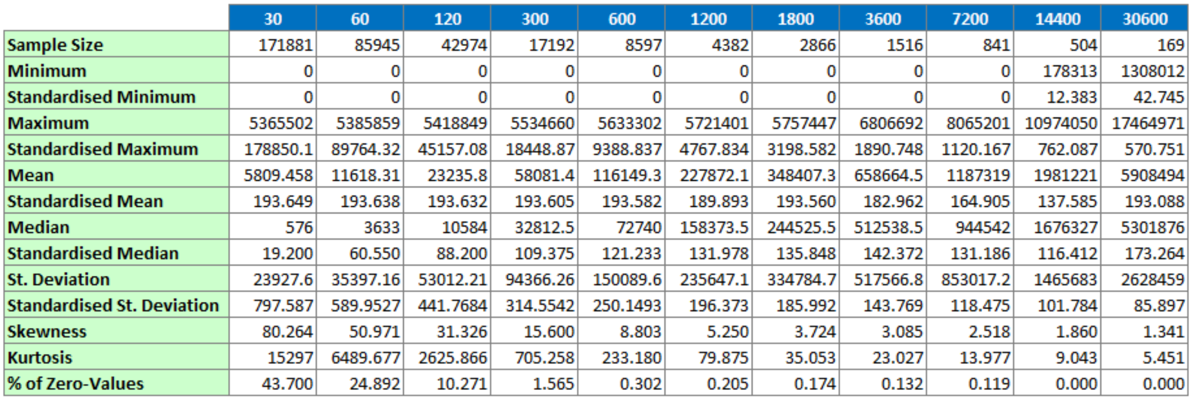
Each column shows statistics for a different sampling interval, ranging from 30s to 30600s (1 day). Statistics with ”Standardised” in their name, e.g. ”Standardised Mean”, have been divided by the number of seconds in the sampling interval to enable comparability across columns.



	30	60	120	300	600	1200	1800	3600	7200	14400	30600
Sample Size	172135	86070	43036	17216	8609	4388	2870	1518	842	504	169
Minimum	0	0	0	0	0	0	0	0	70000	501972	5679669
Standardised Minimum	0	0	0	0	0	0	0	0	9.722	34.859	185.610
Maximum	9479210	9584708	9584708	16013892	16155449	18922567	24449621	34695160	42600814	60426958	1.06E+08
Standardised Maximum	315973.7	159745.1	79872.57	53379.64	26925.75	15768.81	13583.12	9637.544	5916.78	4196.317	3479.958
Mean	26011.93	52022.35	104042.3	260081.5	520102.6	1020411	1560127	2949647	5317771	8884055	26494458
Standardised Mean	867.064	867.039	867.019	866.938	866.838	850.343	866.737	819.346	738.579	616.948	865.832
Median	3711	19467	52589	158515.5	340673	725974	1122359	2286500	4204500	7288599	22537426
Standardised Median	123.700	324.450	438.242	528.385	567.788	604.978	623.533	635.139	583.958	506.153	736.517
St. Deviation	95402.11	143877.8	218619.9	406882.3	660537.9	1065715	1505880	2470092	4242043	7621235	14331754
Standardised St. Deviation	3180.070	2397.963	1821.833	1356.274	1100.897	888.096	836.600	686.137	589.173	529.252	468.358
Skewness	39.148	25.795	15.880	10.798	7.107	5.011	4.156	3.972	3.160	2.588	2.376
Kurtosis	2692.472	1194.163	463.164	244.537	102.694	50.580	37.212	33.552	19.914	13.996	11.141
% of Zero-Values	41.601	22.750	8.853	1.226	0.383	0.296	0.209	0.066	0.000	0.000	0.000

Table 3.6: Summary Statistics for Absolute Volume for WPP across Sampling Intervals

Each column shows statistics for a different sampling interval, ranging from 30s to 30600s (1 day). Statistics with ”Standardised” in their name, e.g. ”Standardised Mean”, have been divided by the number of seconds in the sampling interval to enable comparability across columns.



	30	60	120	300	600	1200	1800	3600	7200	14400	30600
Sample Size	171881	85945	42974	17192	8597	4382	2866	1516	841	504	169
Minimum	0	0	0	0	0	0	0	0	0	178313	1308012
Standardised Minimum	0	0	0	0	0	0	0	0	0	12.383	42.745
Maximum	5365502	5385859	5418849	5534660	5633302	5721401	5757447	6806692	8065201	10974050	17464971
Standardised Maximum	178850.1	89764.32	45157.08	18448.87	9388.837	4767.834	3198.582	1890.748	1120.167	762.087	570.751
Mean	5809.458	11618.31	23235.8	58081.4	116149.3	227872.1	348407.3	658664.5	1187319	1981221	5908494
Standardised Mean	193.649	193.638	193.632	193.605	193.582	189.893	193.560	182.962	164.905	137.585	193.088
Median	576	3633	10584	32812.5	72740	158373.5	244525.5	512538.5	944542	1676327	5301876
Standardised Median	19.200	60.550	88.200	109.375	121.233	131.978	135.848	142.372	131.186	116.412	173.264
St. Deviation	23927.6	35397.16	53012.21	94366.26	150089.6	235647.1	334784.7	517566.8	853017.2	1465683	2628459
Standardised St. Deviation	797.587	589.9527	441.7684	314.5542	250.1493	196.373	185.992	143.769	118.475	101.784	85.897
Skewness	80.264	50.971	31.326	15.600	8.803	5.250	3.724	3.085	2.518	1.860	1.341
Kurtosis	15297	6489.677	2625.866	705.258	233.180	79.875	35.053	23.027	13.977	9.043	5.451
% of Zero-Values	43.700	24.892	10.271	1.565	0.302	0.205	0.174	0.132	0.119	0.000	0.000

The sample sizes naturally reduce as the sampling interval (top row, in seconds) increases, while the minima and maxima (and standardised minima and maxima) increase as the volume processes for all three stocks aggregate. The minima are zero as the variable examined for each stock is absolute volume. The bottom row for each table shows the proportion of

0-value observations at each sampling interval; these arise for trading periods in which no trading occurs, and naturally reduce in proportion as the sampling interval grows. They are also fewest for Vodafone (the most liquid stock) and most numerous for WPP (the least liquid stock).

The means, medians and standard deviations increase as the volume processes aggregate and there are fewer 0-value observations. In contrast, the standardised standard deviations fall as the sampling interval grows, implying less relative variation with aggregation, while the standardised means remain relatively constant and the standardised medians rise. The patterns in the standardised means and medians experience dips for the 3600s, 7200s and 14400s sampling intervals. These dips arise because the sampling intervals do not divide the daily sampling interval (30600s) an integer number of times, so part of the intervals for some observations contain no trades.

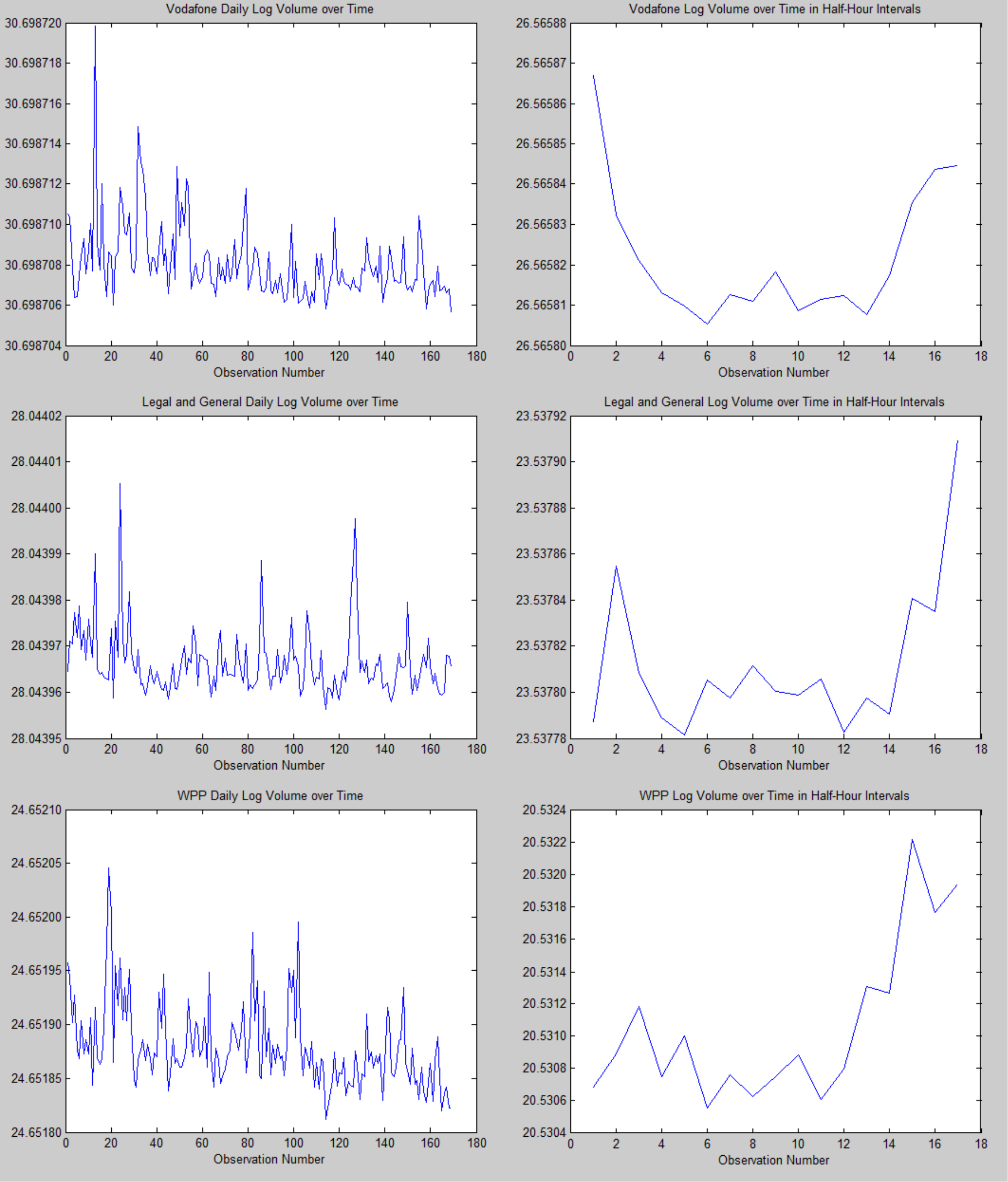
In terms of the higher moments, there is substantial positive skewness and (excess) kurtosis, but these reduce with the sampling interval, reflecting less tail observations with aggregation. Overall, absolute volume is a positive random variable and is not close to Gaussianity at small intervals. While we would like to align with Bialkowski et al.'s (2008) and Darolles et al.'s treatment (2003), where they processed absolute volumes, a transformation of absolute volume is necessary. Ishida and Watanabe (2009) suggest taking square-roots or logs, adding that logs transform the series most suitably. However, the log transformation encounters an issue when absolute volume is zero; using a high negative value to approximate negative infinity would cause the rest of the data points to be dominated. We therefore follow Bollerslev and Wright (2001) in applying what might be termed the "Fuller Transformation", which dampens the effects of 0-value observations; if V_t denotes the absolute volume process and V_t^* is the transformed volume process, then:

$$V_t^* = \log(V_t + \tau s^2) - \frac{\tau s^2}{V_t + \tau s^2} \qquad \text{where } s^2 \text{ is the sample variance of } V_t \text{ and } \tau = 0.02$$

We also adopt Bollerslev and Wright's convention of referring to V_t^* as "log volume" as opposed to "Fuller-Transformed volume". We now inspect plots of this log volume.

The left-hand side of Figure 3.2 below shows plots of the volume data series for the daily sampling interval over the whole data period, demonstrating that daily data seems relatively random. The right-hand side shows plots at the 1800s interval, for the first day of the sample period. This data shows the typical U-shape highlighted by Engle and Russell (1998), confirming seasonality in volumes (SF4.2).

Figure 3.2: Log Volumes for Vodafone, Legal and General, and WPP (Initial Dataset). Left Column: Daily Log Volumes over the whole dataset, Right Column: Half-Hourly Log Volumes on 01/04/2009 (first day of the dataset)



The ADF tests in Appendix C (Table C.1) show that over all sampling intervals, all volume data rejects the null hypothesis of nonstationarity. However, as can be seen in the Ljung-Box Q tests in Table C.2, there is persistent autocorrelation upto orders of 50. This suggests that there is (non-explosive) long memory in the data. GPH estimates of the memory parameter, $\hat{d}$, were obtained through log-periodogram regression in the frequency domain as in Barkoulas and Baum (1996):

$$\ln\{I(\omega_t)\} = \beta_0 - \hat{d}\ln\left\{sin^2\left(\frac{\omega_t}{2}\right)\right\} + \eta_t; \qquad \eta_t \sim NID(0, \sigma^2_\eta)$$

where $\{x_t\}$ is the data series,

T is the sample size,

$L << T$ is the number of ordinates included in the spectral regression,

$t = 1, ..., L$,

the ordinate, $\omega_t = \dfrac{2\pi t}{T}$,

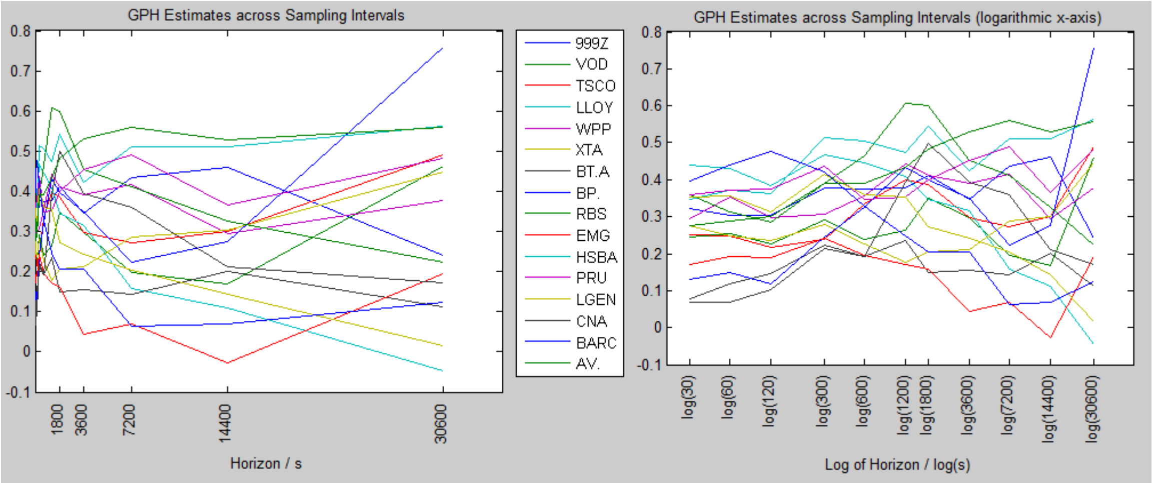
the periodogram, $I(\omega_t) = \dfrac{1}{2\pi T}\displaystyle\sum_{t=1}^{T} |x_t e^{-it\omega_t}|^2$

Expanding on the requirement that $L << T$ in the spectral regression (”trimming of frequencies”), Kanzler (1998) does not exclude any lower frequencies, but does exclude higher frequencies beyond the $\sqrt{T}$-th, as advocated by Künsch (1986). However, this means that there is not always a large number of observations for the larger sampling intervals; at the smallest sampling interval (30s), we have 696 observations, while for the largest sampling interval (daily) we have only 13 observations. Regardless, as a check we experimented with various ranges of frequencies to include and exclude and found that Kanzler’s choice was the best.

The GPH estimates and their p-values are shown below the Ljung-Box Q tests in Tables

C.3-C.4. Plots of the GPH estimates are also shown in Figure 3.3 below; the plot on the right has a logarithmic x-axis so that points appear more evenly spaced horizontally. The estimates by and large indicate long memory which remains stable or increases weakly as the sampling interval increases. The corresponding p-values in the table below the estimates indicate general significance at the 5% significance, although significance falls as the sampling interval increases. These findings corroborate the general finding of long memory in daily volume data, as found by Bollerslev and Jubinsky (1999) and Lobato and Velasco (2000).

Figure 3.3: Charts of GPH Values for Stocks across Sampling Intervals (Initial Dataset)
Left Panel: Normal X-Axis, Middle Panel: Legend, Right Panel: Logarithmic X-Axis

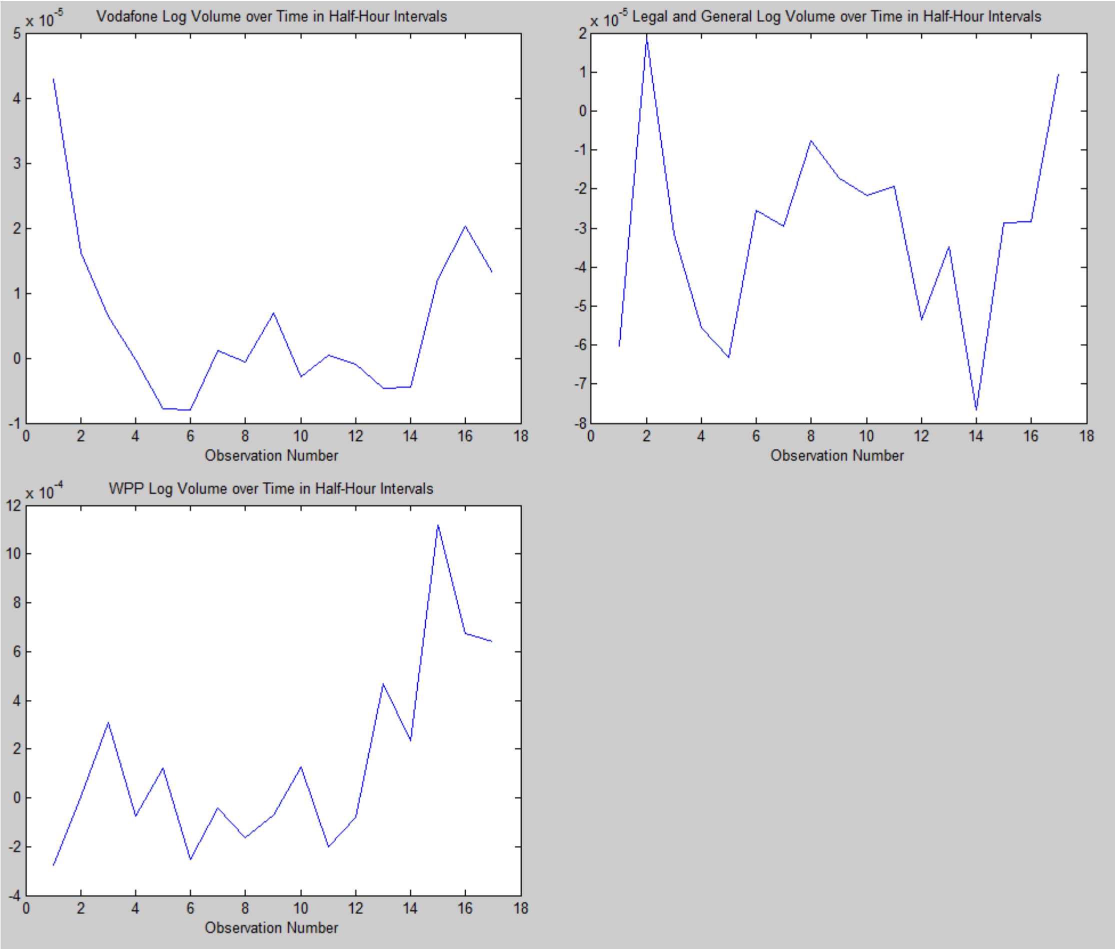


However, the seasonality in the data may be obscuring other inference, so we propose to deseasonalise the data. This yields the results in the next subsection.

3.3.2 Deseasonalised Data

The plots in Figure 3.4 below show the results of deseasonalising the data. We adapted the deseasonlisation and detrending process of Gallant, Rossi and Tauchen (1992) who removed systematic calendar effects and long-run trends in daily volumes using dummy variables (year, month, week, day and number of days since preceding trading day) and a quadratic trend. However, we split the procedure into two stages - deseasonalisation first and then detrending (next subsection). For sampling intervals of less than an hour, we also extended their deseasonalisation procedure to include time-of-day dummies, specifically one dummy for every half-hour in the trading day. Starting from 8:00 and ending at 16:30, this makes 17 extra dummies. This is also partly in line with Engel and Russell (2005), who used half-hour nodes in their cubic spline deseasonalisation procedure. At the same time, since all our data is from the same year, we do not have yearly dummies.

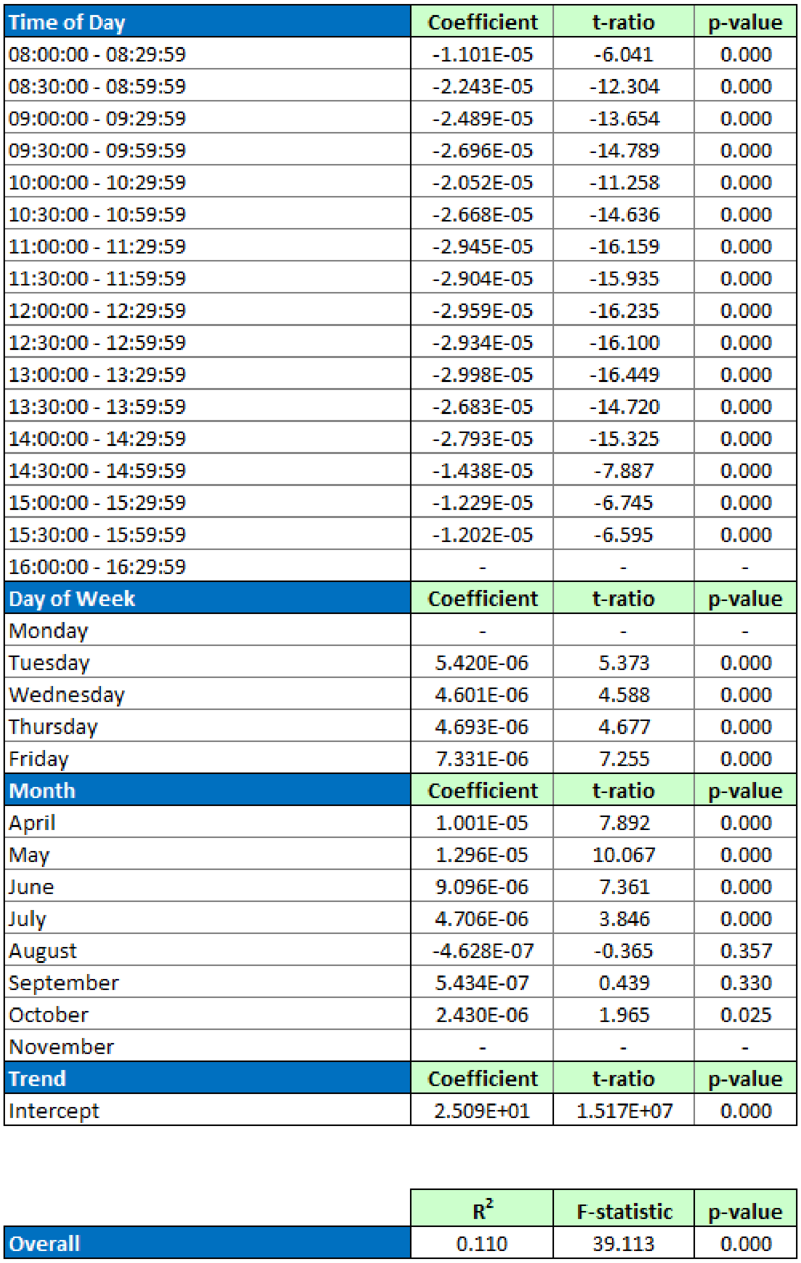
Figure 3.4: Log Volumes over the course of a day for Vodafone, Legal and General and WPP at the $\frac{1}{2}$ hour Sampling Interval (Deseasonalised Dataset).



The dummies are shown with regression coefficients at the 600s sampling interval in Tables 3.7-3.9. The coefficients are small and generally of the magnitude of 1e-5. Where coefficients are marked ”-”, the corresponding dummies have been omitted from estimation to avoid the dummy variable trap which leads to multicollinearity. Multicollinearity was still encountered initially due to the ”GAP” dummies - these show the number of days since the preceding trading day, so will take the value of 1 for the first trade on a day following a previous trading day, a value of 3 for the first trade on a Monday (after the weekend) etc. Since our smaller sampling intervals contain much more data points than the daily sampling interval, the number of trades for which GAP > 0 was less than 1%. If we call the dummy variable matrix δ, this issue resulted in a very large $\delta'\delta$. Due to a limit to machine precision, the Matlab engine gave warnings that $\delta'\delta$ was effectively singular, leading to very large standard errors, zero-value t-ratios and no effective p-values. Further research (beyond this chapter) therefore demands more advanced or precise computing systems. However, dropping the GAP dummies led to the elimination of the Matlab warnings and to the provision of correct t-ratios and p-values, with little change to the coefficients. Only the Month dummies for September are generally insignificant at the 5% significance level for Vodafone, Legal and General, and WPP.

Table 3.7: Vodafone Dummy Regression at the 600s Sampling Interval without GAP Dummies

For each dummy listed in the leftmost column, regression coefficients, t-ratios and p-values are shown in the next three columns. ”-” indicates that the dummy has not been estimated to avoid the dummy variable trap. A summary R^2 for the regression, along with the F-statistic and p-value are given at the bottom of the table.

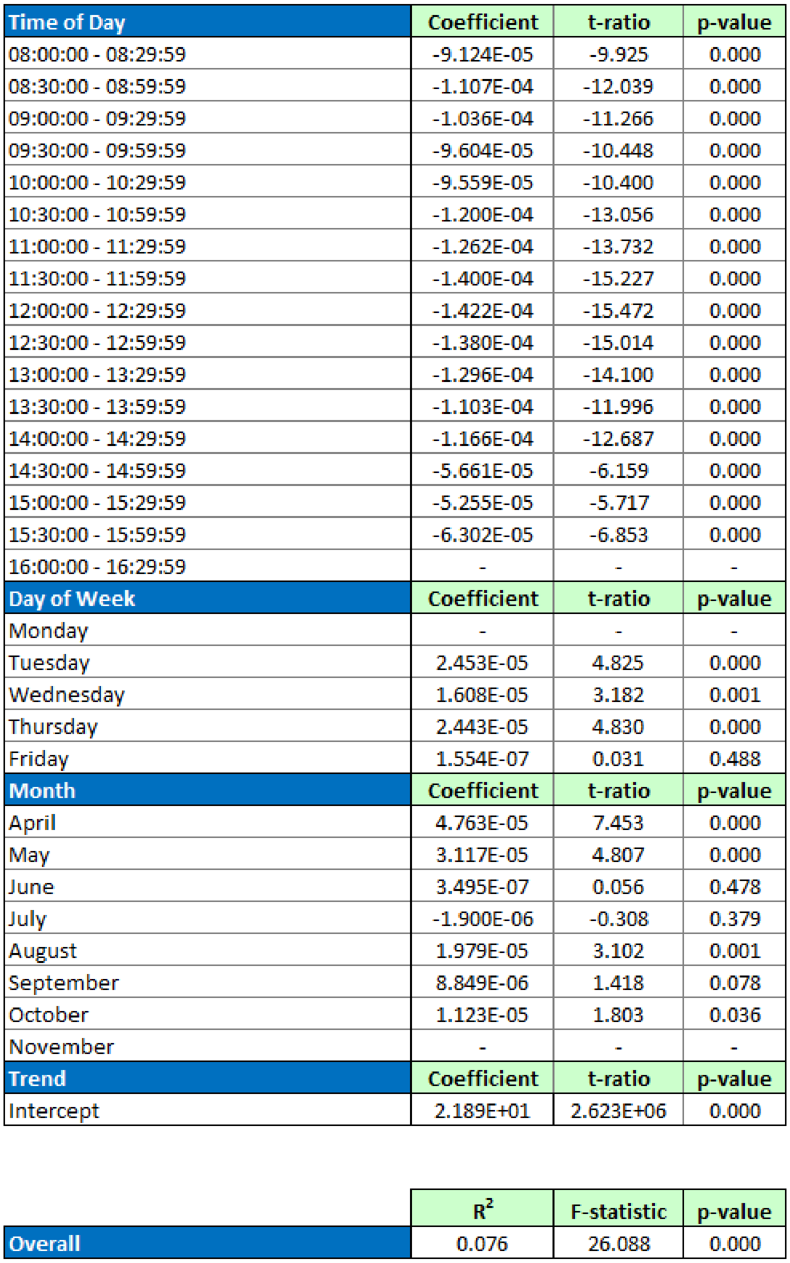


Time of Day	Coefficient	t-ratio	p-value
08:00:00 - 08:29:59	-1.101E-05	-6.041	0.000
08:30:00 - 08:59:59	-2.243E-05	-12.304	0.000
09:00:00 - 09:29:59	-2.489E-05	-13.654	0.000
09:30:00 - 09:59:59	-2.696E-05	-14.789	0.000
10:00:00 - 10:29:59	-2.052E-05	-11.258	0.000
10:30:00 - 10:59:59	-2.668E-05	-14.636	0.000
11:00:00 - 11:29:59	-2.945E-05	-16.159	0.000
11:30:00 - 11:59:59	-2.904E-05	-15.935	0.000
12:00:00 - 12:29:59	-2.959E-05	-16.235	0.000
12:30:00 - 12:59:59	-2.934E-05	-16.100	0.000
13:00:00 - 13:29:59	-2.998E-05	-16.449	0.000
13:30:00 - 13:59:59	-2.683E-05	-14.720	0.000
14:00:00 - 14:29:59	-2.793E-05	-15.325	0.000
14:30:00 - 14:59:59	-1.438E-05	-7.887	0.000
15:00:00 - 15:29:59	-1.229E-05	-6.745	0.000
15:30:00 - 15:59:59	-1.202E-05	-6.595	0.000
16:00:00 - 16:29:59	-	-	-
Day of Week	**Coefficient**	**t-ratio**	**p-value**
Monday	-	-	-
Tuesday	5.420E-06	5.373	0.000
Wednesday	4.601E-06	4.588	0.000
Thursday	4.693E-06	4.677	0.000
Friday	7.331E-06	7.255	0.000
Month	**Coefficient**	**t-ratio**	**p-value**
April	1.001E-05	7.892	0.000
May	1.296E-05	10.067	0.000
June	9.096E-06	7.361	0.000
July	4.706E-06	3.846	0.000
August	-4.628E-07	-0.365	0.357
September	5.434E-07	0.439	0.330
October	2.430E-06	1.965	0.025
November	-	-	-
Trend	**Coefficient**	**t-ratio**	**p-value**
Intercept	2.509E+01	1.517E+07	0.000

	R^2	F-statistic	p-value
Overall	0.110	39.113	0.000

Table 3.8: Legal and General Dummy Regression at the 600s Sampling Interval without GAP Dummies

For each dummy listed in the leftmost column, regression coefficients, t-ratios and p-values are shown in the next three columns. ”-” indicates that the dummy has not been estimated to avoid the dummy variable trap. A summary R^2 for the regression, along with the F-statistic and p-value are given at the bottom of the table.

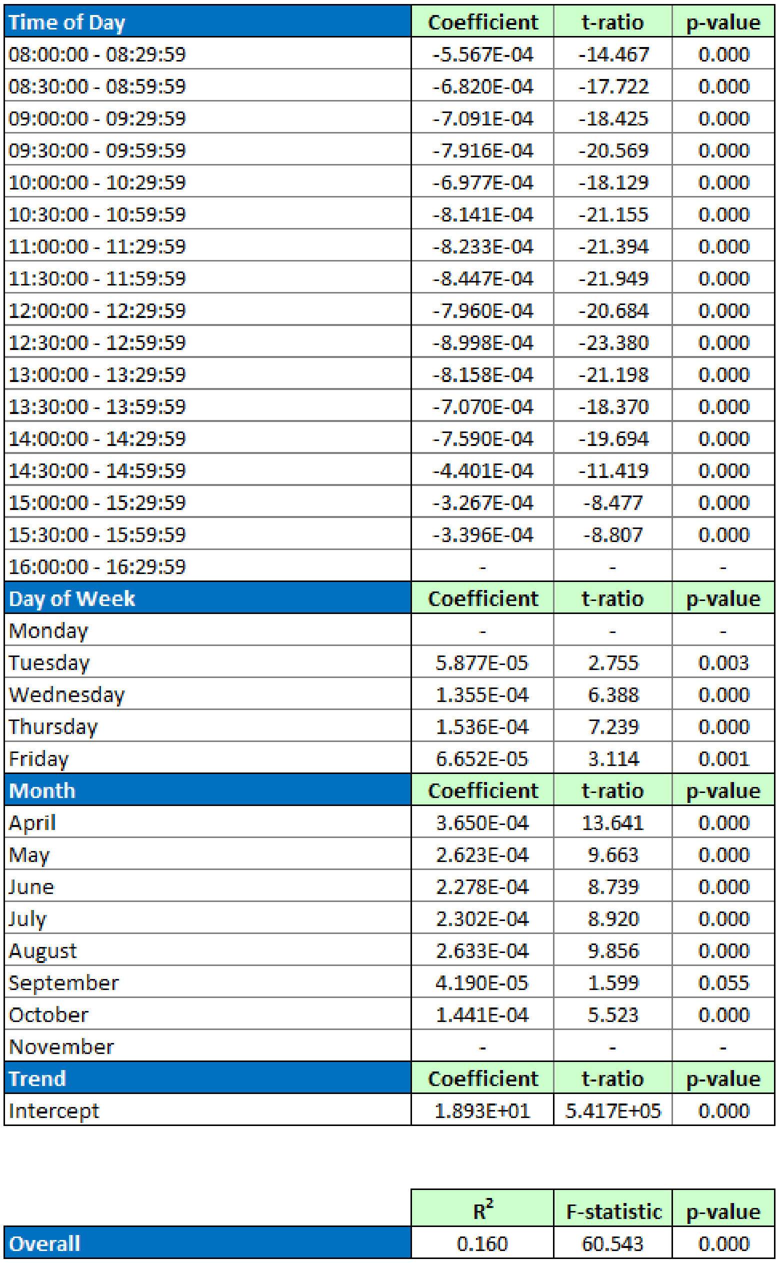


Time of Day	Coefficient	t-ratio	p-value
08:00:00 - 08:29:59	-9.124E-05	-9.925	0.000
08:30:00 - 08:59:59	-1.107E-04	-12.039	0.000
09:00:00 - 09:29:59	-1.036E-04	-11.266	0.000
09:30:00 - 09:59:59	-9.604E-05	-10.448	0.000
10:00:00 - 10:29:59	-9.559E-05	-10.400	0.000
10:30:00 - 10:59:59	-1.200E-04	-13.056	0.000
11:00:00 - 11:29:59	-1.262E-04	-13.732	0.000
11:30:00 - 11:59:59	-1.400E-04	-15.227	0.000
12:00:00 - 12:29:59	-1.422E-04	-15.472	0.000
12:30:00 - 12:59:59	-1.380E-04	-15.014	0.000
13:00:00 - 13:29:59	-1.296E-04	-14.100	0.000
13:30:00 - 13:59:59	-1.103E-04	-11.996	0.000
14:00:00 - 14:29:59	-1.166E-04	-12.687	0.000
14:30:00 - 14:59:59	-5.661E-05	-6.159	0.000
15:00:00 - 15:29:59	-5.255E-05	-5.717	0.000
15:30:00 - 15:59:59	-6.302E-05	-6.853	0.000
16:00:00 - 16:29:59	-	-	-
Day of Week	**Coefficient**	**t-ratio**	**p-value**
Monday	-	-	-
Tuesday	2.453E-05	4.825	0.000
Wednesday	1.608E-05	3.182	0.001
Thursday	2.443E-05	4.830	0.000
Friday	1.554E-07	0.031	0.488
Month	**Coefficient**	**t-ratio**	**p-value**
April	4.763E-05	7.453	0.000
May	3.117E-05	4.807	0.000
June	3.495E-07	0.056	0.478
July	-1.900E-06	-0.308	0.379
August	1.979E-05	3.102	0.001
September	8.849E-06	1.418	0.078
October	1.123E-05	1.803	0.036
November	-	-	-
Trend	**Coefficient**	**t-ratio**	**p-value**
Intercept	2.189E+01	2.623E+06	0.000

	R^2	F-statistic	p-value
Overall	0.076	26.088	0.000

Table 3.9: WPP Dummy Regression at the 600s Sampling Interval without GAP Dummies

For each dummy listed in the leftmost column, regression coefficients, t-ratios and p-values are shown in the next three columns. ”-” indicates that the dummy has not been estimated to avoid the dummy variable trap. A summary R^2 for the regression, along with the F-statistic and p-value are given at the bottom of the table.



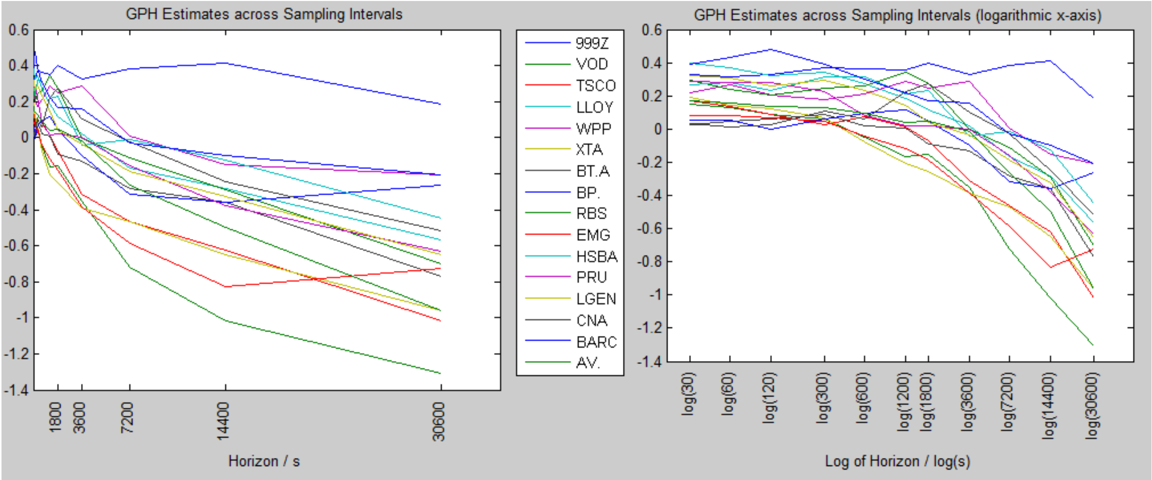
Time of Day	Coefficient	t-ratio	p-value
08:00:00 - 08:29:59	-5.567E-04	-14.467	0.000
08:30:00 - 08:59:59	-6.820E-04	-17.722	0.000
09:00:00 - 09:29:59	-7.091E-04	-18.425	0.000
09:30:00 - 09:59:59	-7.916E-04	-20.569	0.000
10:00:00 - 10:29:59	-6.977E-04	-18.129	0.000
10:30:00 - 10:59:59	-8.141E-04	-21.155	0.000
11:00:00 - 11:29:59	-8.233E-04	-21.394	0.000
11:30:00 - 11:59:59	-8.447E-04	-21.949	0.000
12:00:00 - 12:29:59	-7.960E-04	-20.684	0.000
12:30:00 - 12:59:59	-8.998E-04	-23.380	0.000
13:00:00 - 13:29:59	-8.158E-04	-21.198	0.000
13:30:00 - 13:59:59	-7.070E-04	-18.370	0.000
14:00:00 - 14:29:59	-7.590E-04	-19.694	0.000
14:30:00 - 14:59:59	-4.401E-04	-11.419	0.000
15:00:00 - 15:29:59	-3.267E-04	-8.477	0.000
15:30:00 - 15:59:59	-3.396E-04	-8.807	0.000
16:00:00 - 16:29:59	-	-	-
Day of Week	**Coefficient**	**t-ratio**	**p-value**
Monday	-	-	-
Tuesday	5.877E-05	2.755	0.003
Wednesday	1.355E-04	6.388	0.000
Thursday	1.536E-04	7.239	0.000
Friday	6.652E-05	3.114	0.001
Month	**Coefficient**	**t-ratio**	**p-value**
April	3.650E-04	13.641	0.000
May	2.623E-04	9.663	0.000
June	2.278E-04	8.739	0.000
July	2.302E-04	8.920	0.000
August	2.633E-04	9.856	0.000
September	4.190E-05	1.599	0.055
October	1.441E-04	5.523	0.000
November	-	-	-
Trend	**Coefficient**	**t-ratio**	**p-value**
Intercept	1.893E+01	5.417E+05	0.000

	R^2	F-statistic	p-value
Overall	0.160	60.543	0.000

Overall, the difference in Figures 3.2 and 3.4 indicates that deseasonalisation has dampened the data in the sense that the peaks are now smaller. Performing the usual sets of tests, it can be seen again from Tables D.1-D.4 in Appendix D that over all sampling intervals, all stocks fail to have a unit root, though there is persistent serial correlation upto orders of 50. The majority of the GPH estimates are significant at the 5% significance level. As can also be seen in Figure 3.5 below, there is a temporal trend of long memory at the small intervals which aggregates to intermediate memory over the larger intervals. This trend bears further investigation. Note that there is no discernible cross-sectional trend, although analysis of the stocks in sectors may yield such a pattern. We defer this to a further paper.

Figure 3.5: Charts of GPH Values for Stocks across Sampling Intervals (Deseasonalised Dataset)

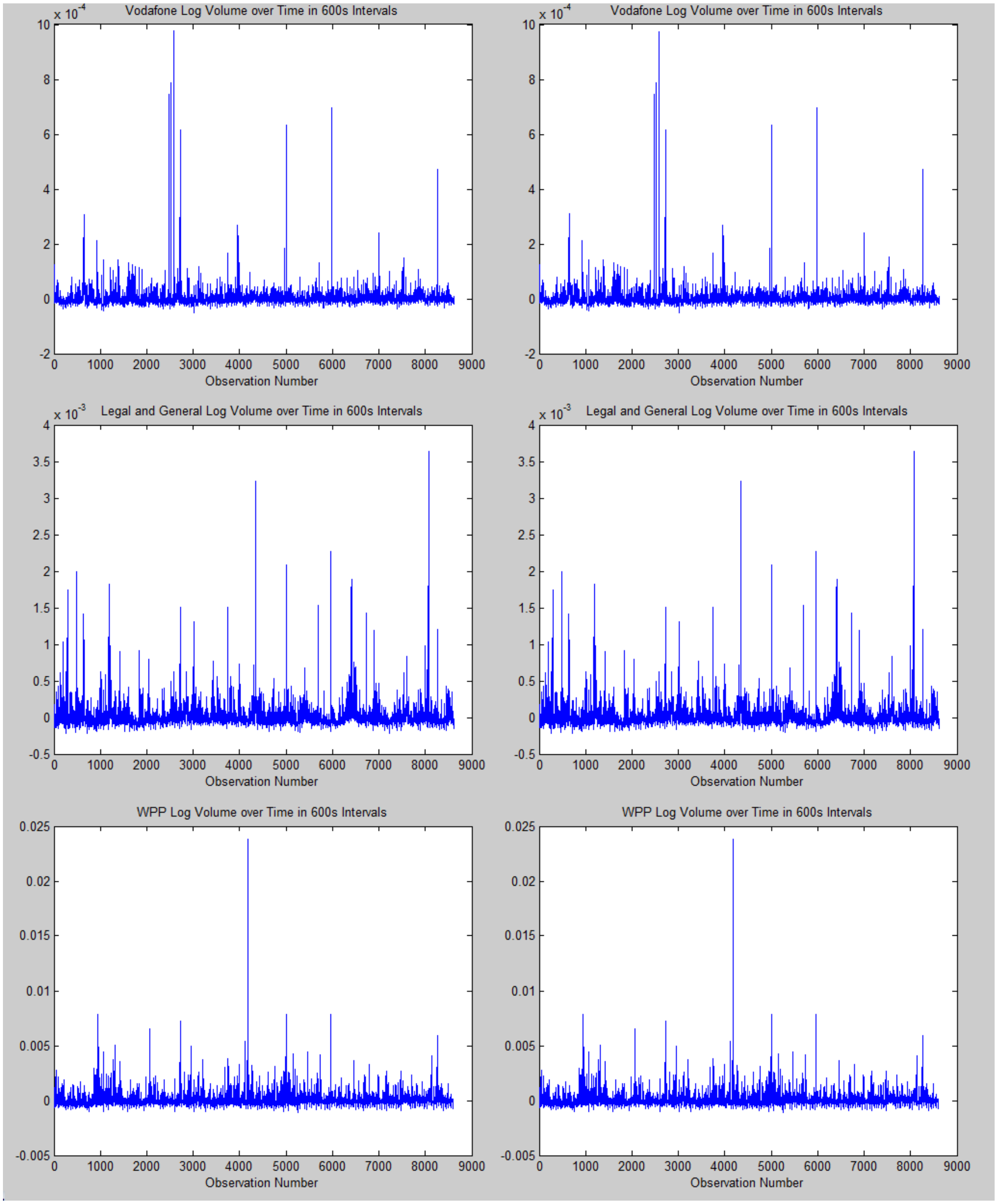
Left Panel: Normal X-Axis, Middle Panel: Legend, Right Panel: Logarithmic X-Axis



3.3.3 Deseasonalised, then Detrended Data

Figure 3.6: Log Volumes for Vodafone, Legal and General, and WPP at the 600s Sampling Interval (Deseasonalised, then Detrended Dataset).

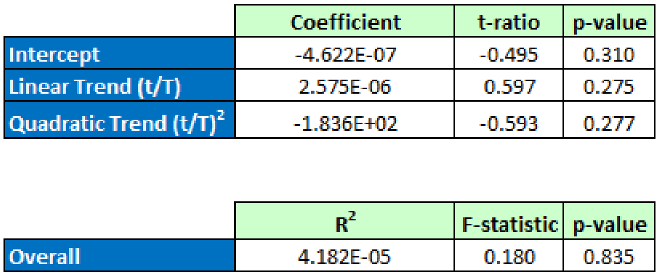
Left Column: Data before Detrending, Right Column: Data after Detrending



We now detrend the series using a quadratic trend to complete the Gallant, Rossi and Tauchen (1992) procedure, producing the plots in Figure 3.6 above; the left-hand side shows the data before detrending, the right-hand side shows the data after detrending. They do not look particularly different from each other, indicating the detrending procedure is possibly redundant. Example regressions for Vodafone, Legal and General and WPP at the 600s sampling interval are shown in Tables 3.10-3.12 below.

Table 3.10: Vodafone Trend Regression at the 600s Sampling Interval

For each regressor listed in the leftmost column, regression coefficients, t-ratios and p-values are shown in the next three columns. A summary R^2 for the regression, along with the F-statistic and p-value are given at the bottom of the table.

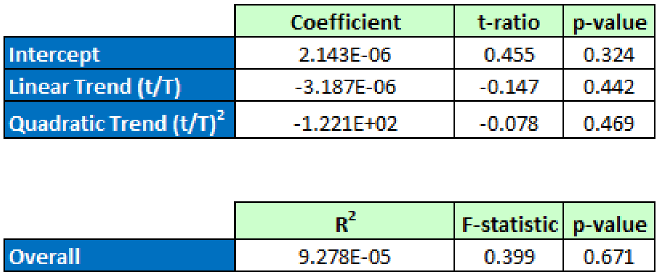


	Coefficient	t-ratio	p-value
Intercept	-4.622E-07	-0.495	0.310
Linear Trend (t/T)	2.575E-06	0.597	0.275
Quadratic Trend $(t/T)^2$	-1.836E+02	-0.593	0.277

	R^2	F-statistic	p-value
Overall	4.182E-05	0.180	0.835

Table 3.11: Legal and General Trend Regression at the 600s Sampling Interval

For each regressor listed in the leftmost column, regression coefficients, t-ratios and p-values are shown in the next three columns. A summary R^2 for the regression, along with the F-statistic and p-value are given at the bottom of the table.

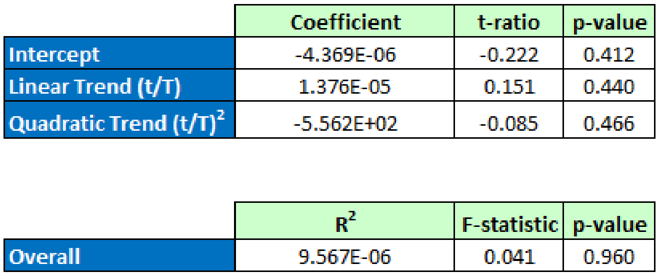


	Coefficient	t-ratio	p-value
Intercept	2.143E-06	0.455	0.324
Linear Trend (t/T)	-3.187E-06	-0.147	0.442
Quadratic Trend $(t/T)^2$	-1.221E+02	-0.078	0.469

	R^2	F-statistic	p-value
Overall	9.278E-05	0.399	0.671

Table 3.12: WPP Trend Regression at the 600s Sampling Interval

For each regressor listed in the leftmost column, regression coefficients, t-ratios and p-values are shown in the next three columns. A summary R^2 for the regression, along with the F-statistic and p-value are given at the bottom of the table.



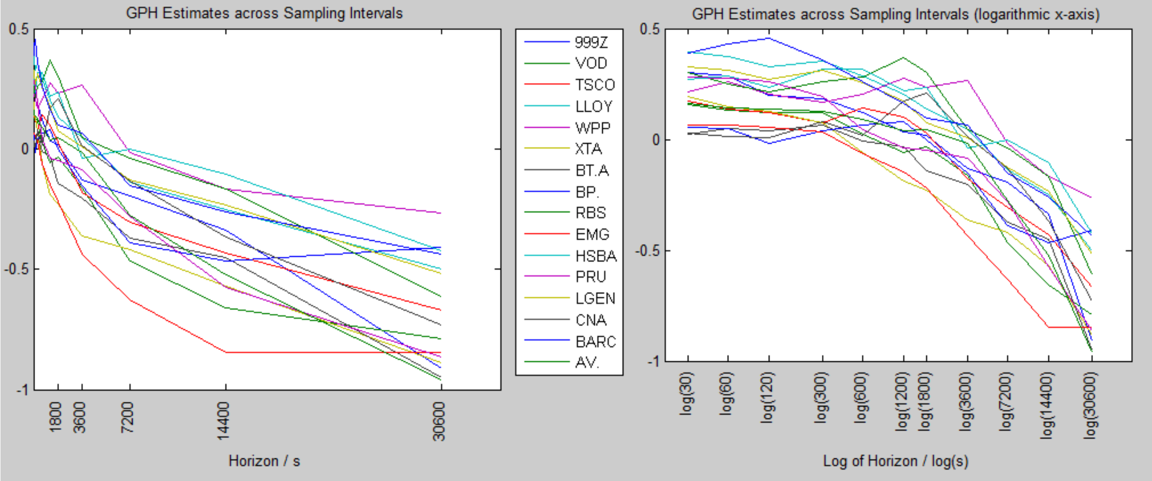
	Coefficient	t-ratio	p-value
Intercept	-4.369E-06	-0.222	0.412
Linear Trend (t/T)	1.376E-05	0.151	0.440
Quadratic Trend $(t/T)^2$	-5.562E+02	-0.085	0.466

	R^2	F-statistic	p-value
Overall	9.567E-06	0.041	0.960

In general, the intercept, linear trend and quadratic trend were found to be insignificant at the 5% significance level. Note that this conflicts with SF 6.2 (trendedness in volumes); however, these results may instead reflect the lack of trendedness over relatively short sampling intervals; our largest sampling intervals are daily, but trends may arise over months or years. ADF tests, LBQ tests and GPH estimates and their p-values for the detrended data are presented in Appendix E (Tables E.1-E.4); since the detrending procedure was not significant, these statistics are largely the same as before detrending as in Appendix D. The decreasing memory pattern remains, as in Figure 3.7 below. We now investigate whether valid ARFIMA models can be fitted to the data.

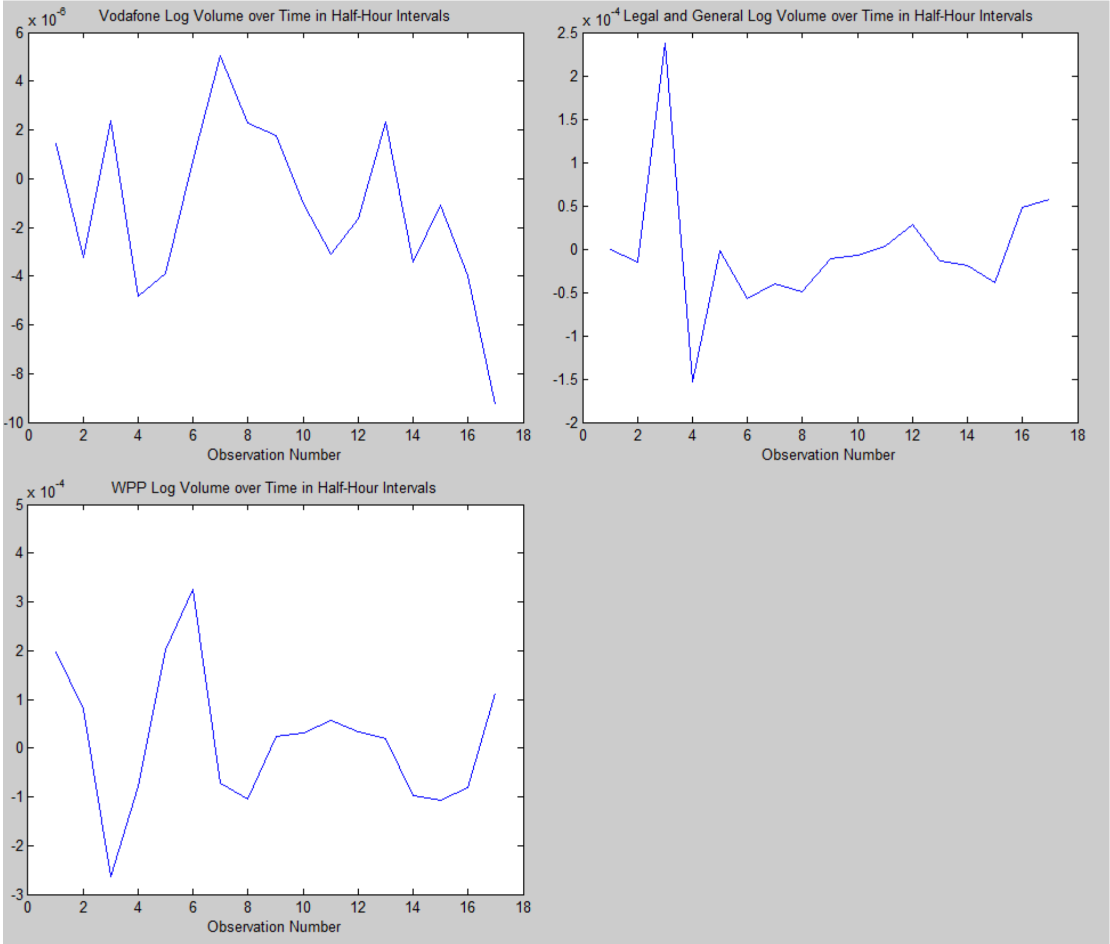
Figure 3.7: Charts of GPH Values for Stocks across Sampling Intervals (Deseasonalised, then Detrended Dataset)

Left Panel: Normal X-Axis, Middle Panel: Legend, Right Panel: Logarithmic X-Axis



3.3.4 ARFIMA Specifications Fitted

Figure 3.8: Log Volumes over the course of a day for Vodafone, Legal and General, and WPP, at the $\frac{1}{2}$ hour Sampling Interval (ARFIMA Dataset).

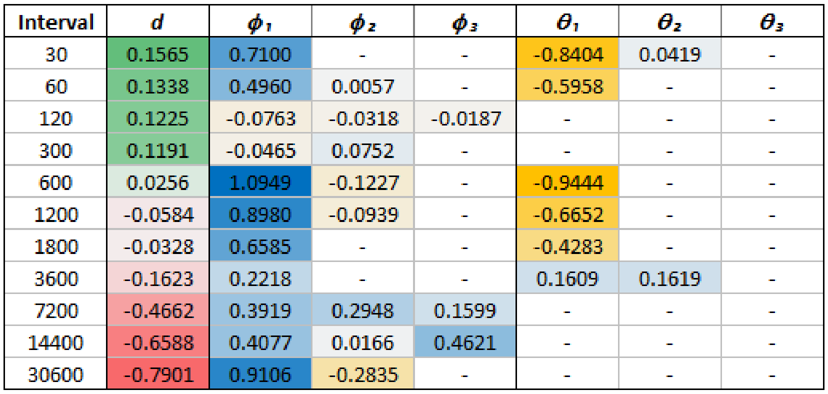


The residuals after fitting ARFIMAs to the data are shown in the plots in Figure 3.8 above, and these plots are markedly different from Figures 3.2 and 3.4, arguably exhibiting more randomness. In order to fit ARFIMA models to the data, we used a two-step procedure; we started by fractionally differencing each data series using the corresponding estimates of d from Appendix E, and then trialled different orders of p and q upto a maximum of 3 (i.e. $p + q \leq 3$), selecting the specification with the lowest AIC. Baillie and Kapetanios

(2008) showed this is a possible way of estimating ARFIMAs albeit with a local Whittle estimator rather than the GPH estimator. Below are the specifications for Vodafone, Legal and General, and WPP (Tables 3.13-3.15). The second leftmost column reproduces the estimates of d with the same colour scale of red to green, while the next columns display the ϕ and θ coefficients - here the colour scale is blue for more positive, white for zero and yellow for more negative.

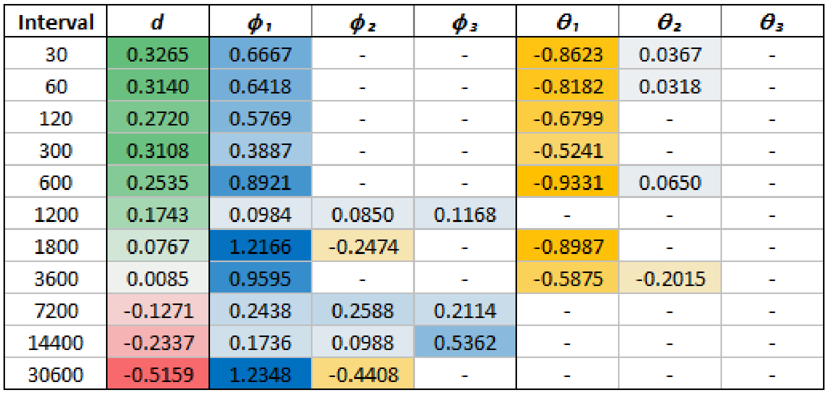
Table 3.13: Vodafone ARFIMA Specifications across Sampling Intervals

Each row shows the ARFIMA specification corresponding to the sampling interval in the first column. d estimates are from Table E.3 in Appendix E. ”-” indicates that the ARFIMA specification does not contain this parameter.



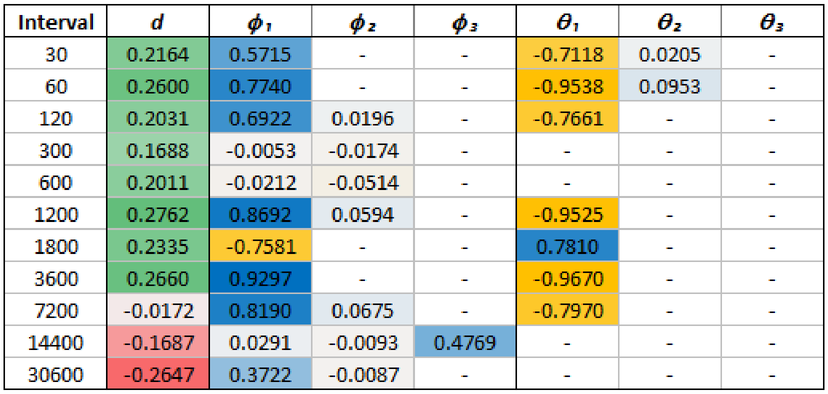
Interval	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	θ_3
30	0.1565	0.7100	-	-	-0.8404	0.0419	-
60	0.1338	0.4960	0.0057	-	-0.5958	-	-
120	0.1225	-0.0763	-0.0318	-0.0187	-	-	-
300	0.1191	-0.0465	0.0752	-	-	-	-
600	0.0256	1.0949	-0.1227	-	-0.9444	-	-
1200	-0.0584	0.8980	-0.0939	-	-0.6652	-	-
1800	-0.0328	0.6585	-	-	-0.4283	-	-
3600	-0.1623	0.2218	-	-	0.1609	0.1619	-
7200	-0.4662	0.3919	0.2948	0.1599	-	-	-
14400	-0.6588	0.4077	0.0166	0.4621	-	-	-
30600	-0.7901	0.9106	-0.2835	-	-	-	-

Table 3.14: Legal and General ARFIMA Specifications across Sampling Intervals
Each row shows the ARFIMA specification corresponding to the sampling interval in the first column. d estimates are from Table E.3 in Appendix E. ”-” indicates that the ARFIMA specification does not contain this parameter.



Interval	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	θ_3
30	0.3265	0.6667	-	-	-0.8623	0.0367	-
60	0.3140	0.6418	-	-	-0.8182	0.0318	-
120	0.2720	0.5769	-	-	-0.6799	-	-
300	0.3108	0.3887	-	-	-0.5241	-	-
600	0.2535	0.8921	-	-	-0.9331	0.0650	-
1200	0.1743	0.0984	0.0850	0.1168	-	-	-
1800	0.0767	1.2166	-0.2474	-	-0.8987	-	-
3600	0.0085	0.9595	-	-	-0.5875	-0.2015	-
7200	-0.1271	0.2438	0.2588	0.2114	-	-	-
14400	-0.2337	0.1736	0.0988	0.5362	-	-	-
30600	-0.5159	1.2348	-0.4408	-	-	-	-

Table 3.15: WPP ARFIMA Specifications across Sampling Intervals
Each row shows the ARFIMA specification corresponding to the sampling interval in the first column. d estimates are from Table E.3 in Appendix E. ”-” indicates that the ARFIMA specification does not contain this parameter.

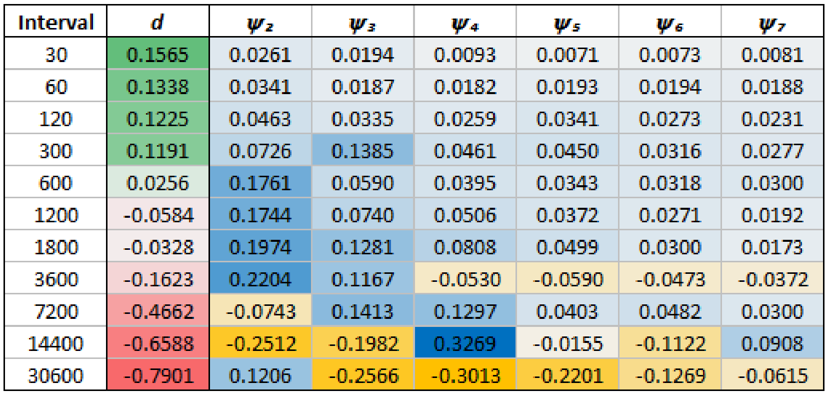


Interval	d	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	θ_3
30	0.2164	0.5715	-	-	-0.7118	0.0205	-
60	0.2600	0.7740	-	-	-0.9538	0.0953	-
120	0.2031	0.6922	0.0196	-	-0.7661	-	-
300	0.1688	-0.0053	-0.0174	-	-	-	-
600	0.2011	-0.0212	-0.0514	-	-	-	-
1200	0.2762	0.8692	0.0594	-	-0.9525	-	-
1800	0.2335	-0.7581	-	-	0.7810	-	-
3600	0.2660	0.9297	-	-	-0.9670	-	-
7200	-0.0172	0.8190	0.0675	-	-0.7970	-	-
14400	-0.1687	0.0291	-0.0093	0.4769	-	-	-
30600	-0.2647	0.3722	-0.0087	-	-	-	-

It may be easier to compare the specifications across intervals by analysing the equivalent infinite order MAs - these are shown in Tables 3.16-3.18 below, although truncated after 6 lags. Since $\psi_1 = 1$, we show the second to seventh coefficients. Plots of these coefficients are below the tables, in Figure 3.9. Due to differences in the memory parameter, the tables and plots show that the moving average coefficients vary substantially across sampling intervals; when d is positive, the MA coefficients decline slowly to zero as consistent with long memory, whereas when d is negative, the MA coefficients are sometimes negative, which is consistent with intermediate memory (also known as “antipersistence”).

Table 3.16: Vodafone Infinite MA Specifications across Sampling Intervals - Coefficients 2-7

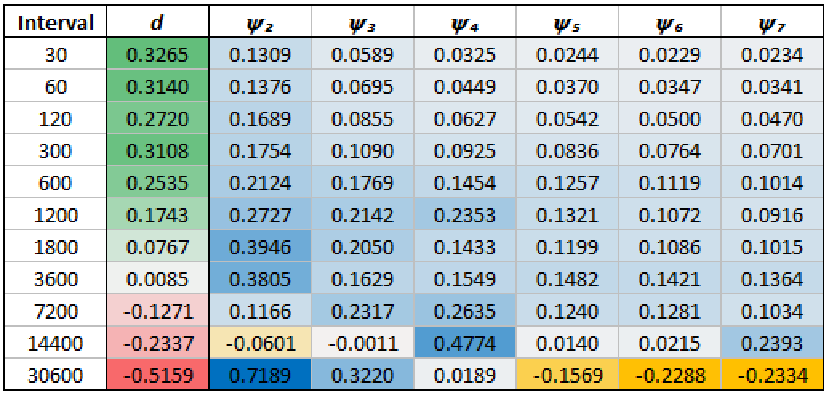
Each row shows the infinite MA specification corresponding to the sampling interval in the first column. d estimates are from Table E.3 in Appendix E.



Interval	d	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
30	0.1565	0.0261	0.0194	0.0093	0.0071	0.0073	0.0081
60	0.1338	0.0341	0.0187	0.0182	0.0193	0.0194	0.0188
120	0.1225	0.0463	0.0335	0.0259	0.0341	0.0273	0.0231
300	0.1191	0.0726	0.1385	0.0461	0.0450	0.0316	0.0277
600	0.0256	0.1761	0.0590	0.0395	0.0343	0.0318	0.0300
1200	-0.0584	0.1744	0.0740	0.0506	0.0372	0.0271	0.0192
1800	-0.0328	0.1974	0.1281	0.0808	0.0499	0.0300	0.0173
3600	-0.1623	0.2204	0.1167	-0.0530	-0.0590	-0.0473	-0.0372
7200	-0.4662	-0.0743	0.1413	0.1297	0.0403	0.0482	0.0300
14400	-0.6588	-0.2512	-0.1982	0.3269	-0.0155	-0.1122	0.0908
30600	-0.7901	0.1206	-0.2566	-0.3013	-0.2201	-0.1269	-0.0615

Table 3.17: Legal and General Infinite MA Specifications across Sampling Intervals - Coefficients 2-7

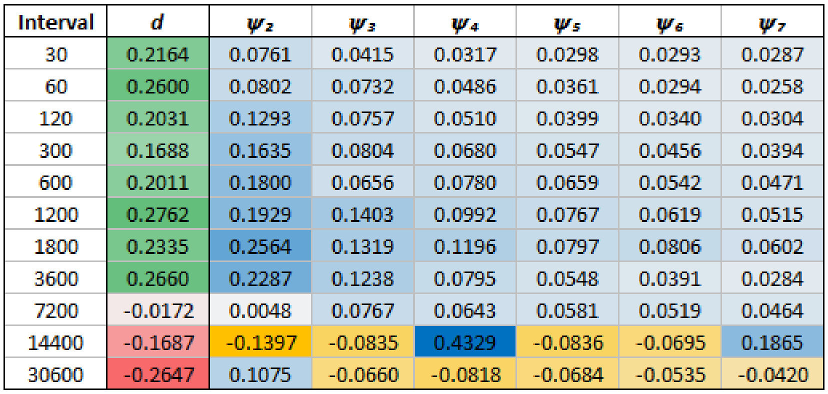
Each row shows the infinite MA specification corresponding to the sampling interval in the first column. d estimates are from Table E.3 in Appendix E.



Interval	d	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
30	0.3265	0.1309	0.0589	0.0325	0.0244	0.0229	0.0234
60	0.3140	0.1376	0.0695	0.0449	0.0370	0.0347	0.0341
120	0.2720	0.1689	0.0855	0.0627	0.0542	0.0500	0.0470
300	0.3108	0.1754	0.1090	0.0925	0.0836	0.0764	0.0701
600	0.2535	0.2124	0.1769	0.1454	0.1257	0.1119	0.1014
1200	0.1743	0.2727	0.2142	0.2353	0.1321	0.1072	0.0916
1800	0.0767	0.3946	0.2050	0.1433	0.1199	0.1086	0.1015
3600	0.0085	0.3805	0.1629	0.1549	0.1482	0.1421	0.1364
7200	-0.1271	0.1166	0.2317	0.2635	0.1240	0.1281	0.1034
14400	-0.2337	-0.0601	-0.0011	0.4774	0.0140	0.0215	0.2393
30600	-0.5159	0.7189	0.3220	0.0189	-0.1569	-0.2288	-0.2334

Table 3.18: WPP Infinite MA Specifications across Sampling Intervals - Coefficients 2-7

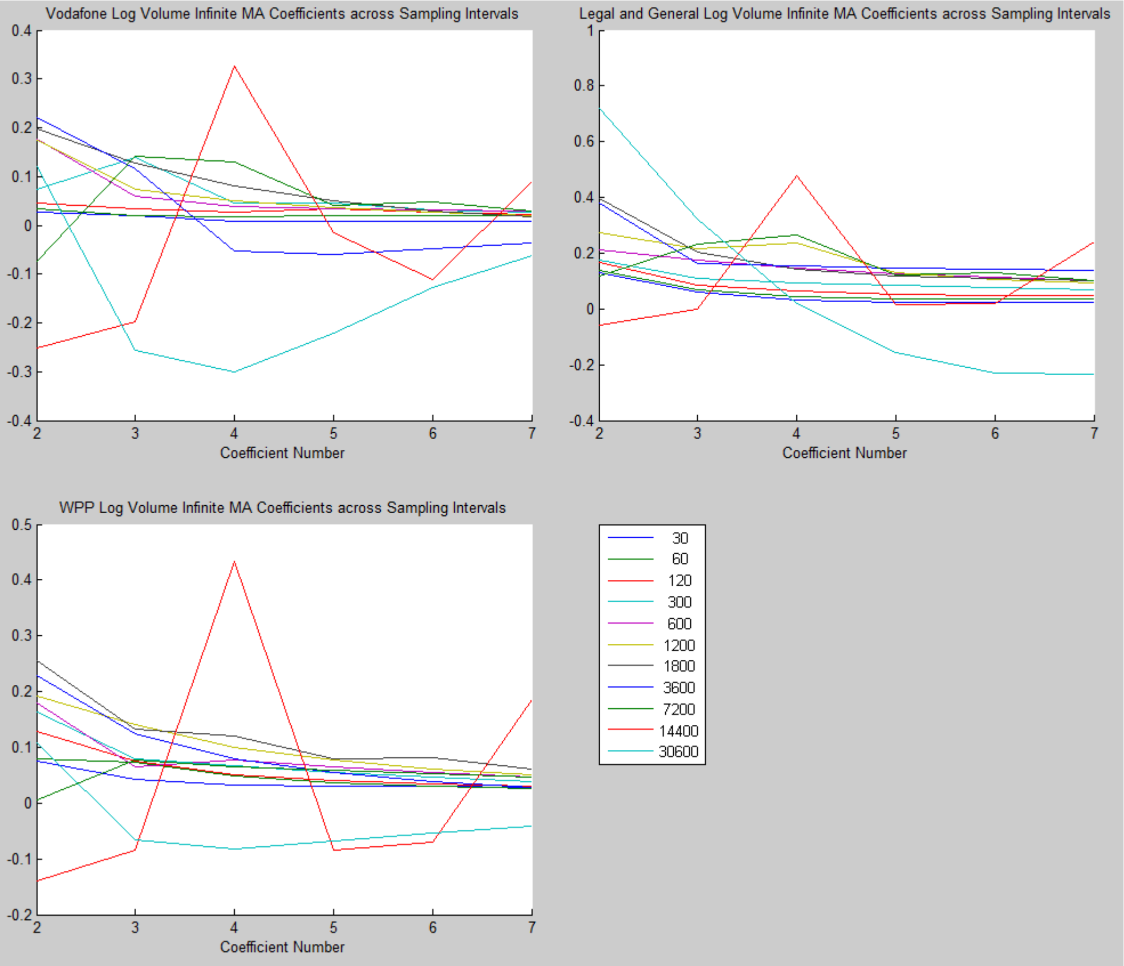
Each row shows the infinite MA specification corresponding to the sampling interval in the first column. d estimates are from Table E.3 in Appendix E.



Interval	d	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_7
30	0.2164	0.0761	0.0415	0.0317	0.0298	0.0293	0.0287
60	0.2600	0.0802	0.0732	0.0486	0.0361	0.0294	0.0258
120	0.2031	0.1293	0.0757	0.0510	0.0399	0.0340	0.0304
300	0.1688	0.1635	0.0804	0.0680	0.0547	0.0456	0.0394
600	0.2011	0.1800	0.0656	0.0780	0.0659	0.0542	0.0471
1200	0.2762	0.1929	0.1403	0.0992	0.0767	0.0619	0.0515
1800	0.2335	0.2564	0.1319	0.1196	0.0797	0.0806	0.0602
3600	0.2660	0.2287	0.1238	0.0795	0.0548	0.0391	0.0284
7200	-0.0172	0.0048	0.0767	0.0643	0.0581	0.0519	0.0464
14400	-0.1687	-0.1397	-0.0835	0.4329	-0.0836	-0.0695	0.1865
30600	-0.2647	0.1075	-0.0660	-0.0818	-0.0684	-0.0535	-0.0420

Figure 3.9: Vodafone, Legal and General, and WPP Infinite MA Specifications across Sampling Intervals - Plot

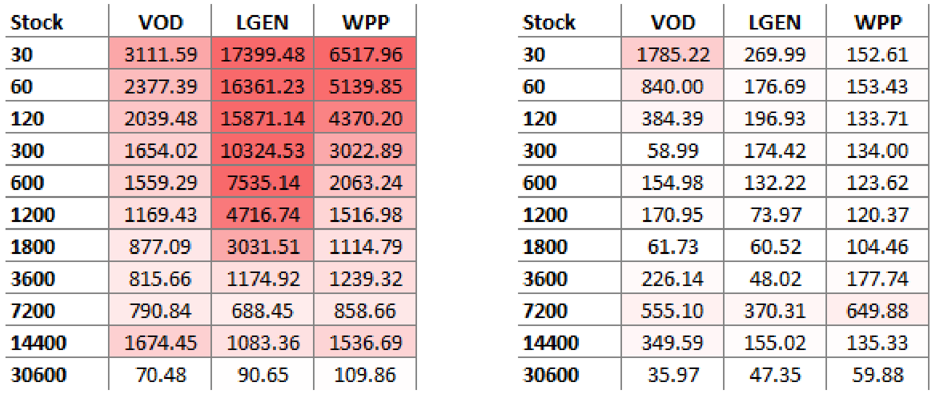
Each line plots the infinite MA specification corresponding to the sampling interval in the legend. *d* estimates are from Table E.3. in Appendix E



After fitting the ARFIMAs, the residuals were again tested for nonstationarity, serial correlation and long memory - the results are in Appendix F. The ADF results (Table F.1) are as before, indicating rejection of nonstationarity, and there is less widespread serial correlation as shown by the LBQ tests (Table F.2), now only 65.9% of the LBQ tests reject the null hypothesis of no serial correlation upto lag order 50 (the proportion in the previous subsection was 84.3%). This can be further seen by a large reduction in the Q-statistics used in the LBQ tests, as shown for Vodafone, Legal and General, and WPP in the tables below (3.19):

Table 3.19: Tables of Q-Statistics before and after fitting ARFIMAs

Left Panel: Q-Statistics before fitting ARFIMAs, Right Panel: Q-Statistics after fitting ARFIMAs, Colour Scale: Values become redder with magnitude



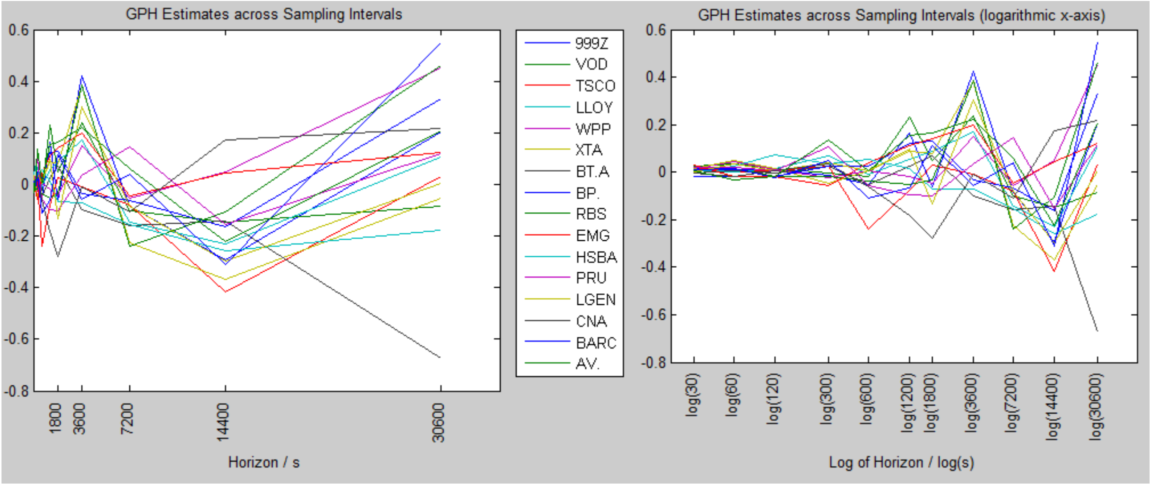
Stock	VOD	LGEN	WPP
30	3111.59	17399.48	6517.96
60	2377.39	16361.23	5139.85
120	2039.48	15871.14	4370.20
300	1654.02	10324.53	3022.89
600	1559.29	7535.14	2063.24
1200	1169.43	4716.74	1516.98
1800	877.09	3031.51	1114.79
3600	815.66	1174.92	1239.32
7200	790.84	688.45	858.66
14400	1674.45	1083.36	1536.69
30600	70.48	90.65	109.86

Stock	VOD	LGEN	WPP
30	1785.22	269.99	152.61
60	840.00	176.69	153.43
120	384.39	196.93	133.71
300	58.99	174.42	134.00
600	154.98	132.22	123.62
1200	170.95	73.97	120.37
1800	61.73	60.52	104.46
3600	226.14	48.02	177.74
7200	555.10	370.31	649.88
14400	349.59	155.02	135.33
30600	35.97	47.35	59.88

We also performed tests for AutoRegressive Conditional Heteroskedasticity (ARCH) tests of order 1 (Table F.3), finding that 34.7% of the tests rejected the null of homoskedastic residuals, with rejection tending to cluster for certain companies, most notably Lloyds Banking Group. Finally, there is no longer a clear memory pattern as evident in Tables F.4-F.5 and Figure 3.10 below (also the large *d* estimates are insignificant).

Figure 3.10: Charts of GPH Values for Stocks across sampling intervals (ARFIMA Dataset)

Left Panel: Normal X-Axis, Middle Panel: Legend, Right Panel: Logarithmic X-Axis



This suggests that the ARFIMA specifications have explained some but not all of the variation in the data. Further research could analyse ways of improving model performance with respect to the diagnostics, e.g. through modelling the ARFIMA residuals as an ARCH process. We defer this to a further paper as the computing effort was already significant (so far, we have generated 2992 hypothesis tests, 352 regressions and 704 ARFIMA specifications). We end this section with our findings with respect to Q1 and log volume:

1. There are significant seasonal effects across sampling intervals

2. Detrending is not found to be significant

3. After Classical Decomposition, ARFIMA specifications can fit the data to some extent. Different specifications for different sampling intervals show that empirically, ARFIMA specifications are not closed under temporal aggregation after deseasonalisation. The fit of the ARFIMA specifications will be explored further in the next section.

4. Concentrating on the memory parameter, there is a relatively stable memory pat-

tern with temporal aggregation before deseasonalisation, and a decreasing memory pattern with temporal aggregation after deseasonalisation

In the next section, we investigate Q2 - whether simulated ARFIMA models are aggregationally consistent with the observed memory patterns.

3.4 Simulated Results on Aggregation

In this section we simulate models in order to explain the behaviour noted in the datasets. In summary, for the initial dataset, we have a relatively stable memory pattern, while for the deseasonalised dataset, we have a decreasing memory pattern. Theoretically, the former pattern is consistent with Chambers (1998), in which it is shown that temporal aggregation of time series should lead to no change in the memory parameter of the aggregate series. While the latter pattern contradicts this, it may be a result of the data procedures we conducted - deseasonalising or detrending. Since there was no change to the pattern of decreasing memory after detrending, we will investigate the deseasonalisation procedure more.

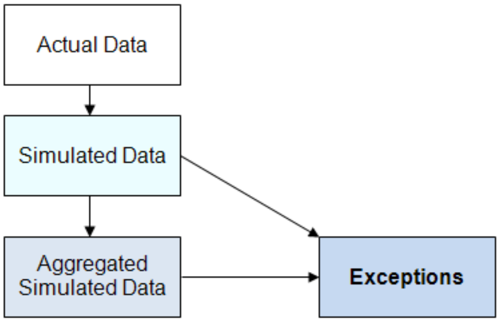
Overall, we seek processes which can explain these two patterns. Classical Decomposition of time series shows that any time series, once a stochastic trend has been removed, can be decomposed into a stationary time series and deterministic trend and seasonal components. The ADF tests show that there is no stochastic trend, and since the dataset only spans 8 months of data, we suggest that the deterministic trend should not be significant. So at most, the Initial dataset should be representable as a stationary time series and seasonal component, while the deseasonalised dataset should be purely representable by a stationary time series.

As Brockwell and Davis (1991) state, for any time series with covariance function fading to zero, an ARMA process can be found to model the series (by replicating the covariance function). Extending to allow for long memory, ARFIMA processes (or equivalently, infinite-order ARMA processes) should be able to model the volume series.

We attempt to discover the data generating process behind the Initial and Deseasonalised data as shown in the process flow below in Figure 3.11, which we call ”Memory Exception Analysis”. Essentially, we estimate ARFIMA models on the raw data (“Actual Data”), then use the estimated ARFIMA parameters to simulate much more data (“Simulated Data”), which we aggregate over all intervals (creating ”Aggregated Simulated Data”). We are primarily interested in the memory parameter of the data, so we check whether the Aggregated Simulated Data at a given interval have a similar memory parameter to the Actual Data at the same interval. If it does, then the ARFIMA specification at the smallest interval underlies all the data reasonably.

Figure 3.11: Memory Exception Analysis Overview

Actual Data is analysed to gain ARFIMA specifications. These are simulated and then aggregated. Exceptions (where d estimates from Simulated or Aggregated Simulated data do not agree with Actual Data) are highlighted in bold.



Our results will be presented in tables with the colour code below:

Actual Data - White: First we show the GPH estimates of the memory parameter for the actual data. Since they are t-distributed, 95% confidence bounds can be constructed and are displayed in smaller numbers in the ’+’ and ’-’ columns.

Simulated Data - Light Blue: Next, we estimate the best ARFIMA on the real data (by selecting the model with the highest AIC). Using the ARFIMA parameters for p, d and q, we simulate series 50 times and calculate the average d and confidence bounds.

Aggregated Simulated Data - Dark Blue: Finally, we take the simulated series from the previous stage and aggregate to create simulated series over longer sampling intervals, and estimate d and its associated confidence bounds.

Exceptions - Bold Data: Wherever the Simulated Data or Aggregated Simulated confidence bounds fail to overlap with the confidence bounds for the Actual Data, they are emboldened to show that the memory parameter they exhibit is not consistent with that of the real data.

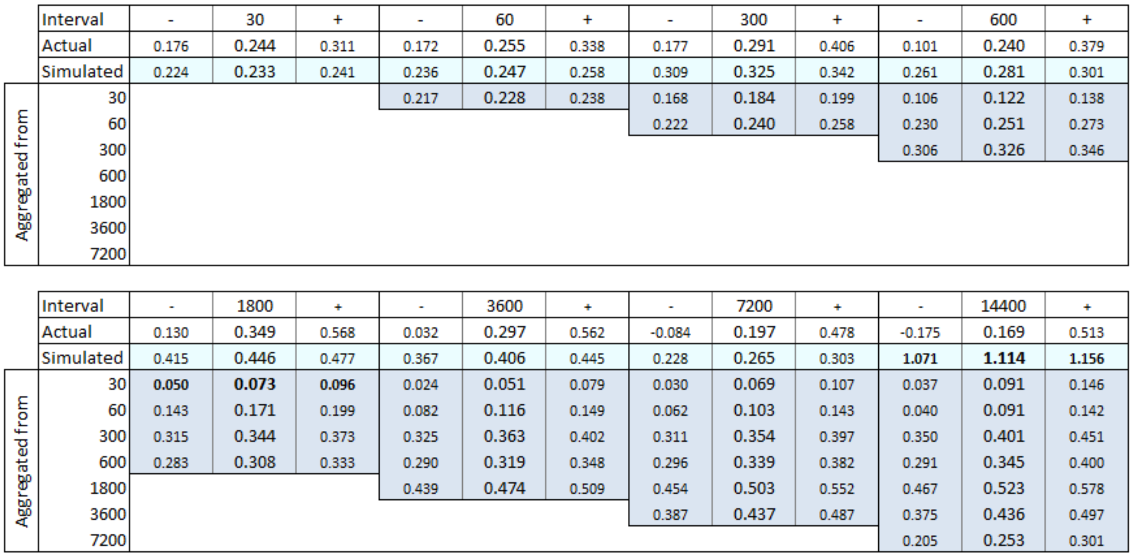
Note that for the Actual Data, the confidence interval calculation is $d \pm 1.96\frac{s}{\sqrt{T}}$ where s is the sample standard deviation and T is the number of observations. In contrast, the d estimates for the Simulated and Aggregated Simulated Data are calculated as an estimate over many simulation loops, so the confidence interval calculation is now $\overline{d} \pm 1.96\frac{s}{\sqrt{N}}$, where N is the number of loops.

3.4.1 Initial Dataset - Stable Memory with Aggregation

Tables 3.20-3.22 below show the results for Vodafone, Legal and General and WPP. Note that the intervals in our standard set of sampling intervals ({30s, 60s, 120s, 300s, 600s, 1200s, 1800s, 3600s, 7200s, 14400s, 30600s}) do not always aggregate a whole number of times to higher sampling intervals, e.g. 1200s does not divide 1800s an integer number of times. So to maintain integer aggregation/division, we now restrict the set of sampling intervals to {30s, 60s, 120s, 300s, 600s, 1800s, 3600s, 7200s, 14400s}. Unfortunately we lose the daily data (30600s), leaving us with a maximum sampling interval of 4 hours (14400s).

Table 3.20: Vodafone Memory Exception Analysis - Initial Dataset

Confidence bounds for d for Actual data (white cells), Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).

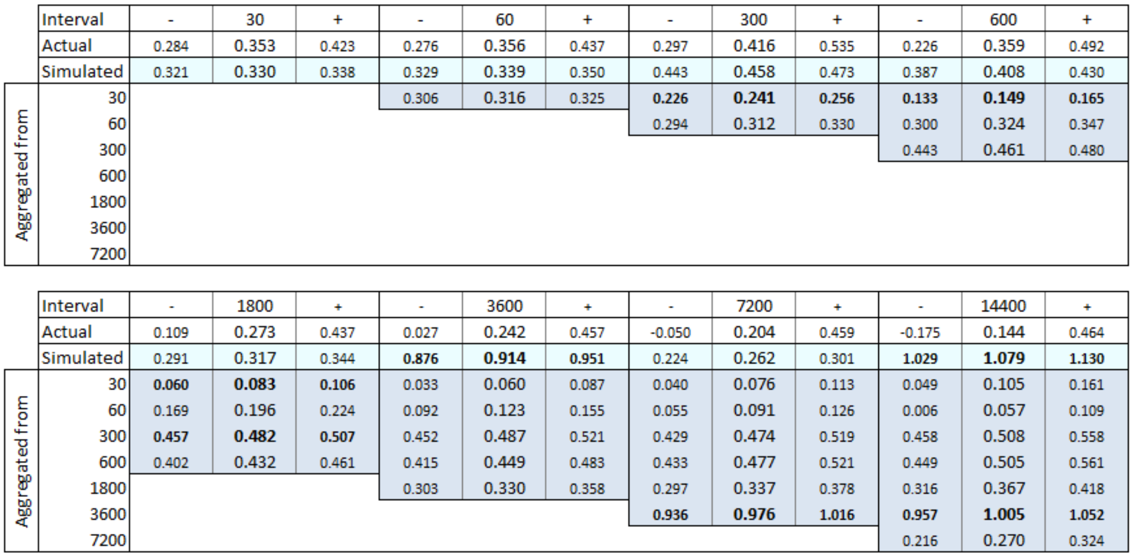


	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Actual	0.176	0.244	0.311	0.172	0.255	0.338	0.177	0.291	0.406	0.101	0.240	0.379
	Simulated	0.224	0.233	0.241	0.236	0.247	0.258	0.309	0.325	0.342	0.261	0.281	0.301
Aggregated from	30				0.217	0.228	0.238	0.168	0.184	0.199	0.106	0.122	0.138
	60							0.222	0.240	0.258	0.230	0.251	0.273
	300										0.306	0.326	0.346
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Actual	0.130	0.349	0.568	0.032	0.297	0.562	-0.084	0.197	0.478	-0.175	0.169	0.513
	Simulated	0.415	0.446	0.477	0.367	0.406	0.445	0.228	0.265	0.303	**1.071**	**1.114**	**1.156**
Aggregated from	30	**0.050**	**0.073**	**0.096**	0.024	0.051	0.079	0.030	0.069	0.107	0.037	0.091	0.146
	60	0.143	0.171	0.199	0.082	0.116	0.149	0.062	0.103	0.143	0.040	0.091	0.142
	300	0.315	0.344	0.373	0.325	0.363	0.402	0.311	0.354	0.397	0.350	0.401	0.451
	600	0.283	0.308	0.333	0.290	0.319	0.348	0.296	0.339	0.382	0.291	0.345	0.400
	1800				0.439	0.474	0.509	0.454	0.503	0.552	0.467	0.523	0.578
	3600							0.387	0.437	0.487	0.375	0.436	0.497
	7200										0.205	0.253	0.301

Table 3.21: Legal and General Memory Exception Analysis - Initial Dataset

Confidence bounds for d for Actual data (white cells), Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).

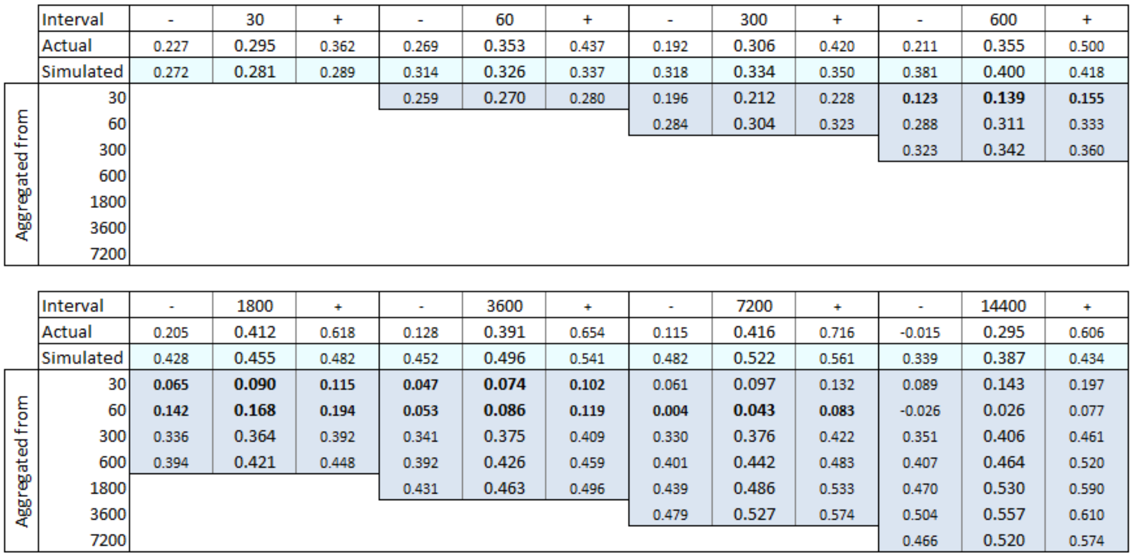


	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Actual	0.284	0.353	0.423	0.276	0.356	0.437	0.297	0.416	0.535	0.226	0.359	0.492
	Simulated	0.321	0.330	0.338	0.329	0.339	0.350	0.443	0.458	0.473	0.387	0.408	0.430
Aggregated from	30				0.306	0.316	0.325	**0.226**	**0.241**	**0.256**	**0.133**	**0.149**	**0.165**
	60							0.294	0.312	0.330	0.300	0.324	0.347
	300										0.443	0.461	0.480
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Actual	0.109	0.273	0.437	0.027	0.242	0.457	-0.050	0.204	0.459	-0.175	0.144	0.464
	Simulated	0.291	0.317	0.344	**0.876**	**0.914**	**0.951**	0.224	0.262	0.301	**1.029**	**1.079**	**1.130**
Aggregated from	30	**0.060**	**0.083**	**0.106**	0.033	0.060	0.087	0.040	0.076	0.113	0.049	0.105	0.161
	60	0.169	0.196	0.224	0.092	0.123	0.155	0.055	0.091	0.126	0.006	0.057	0.109
	300	**0.457**	**0.482**	**0.507**	0.452	0.487	0.521	0.429	0.474	0.519	0.458	0.508	0.558
	600	0.402	0.432	0.461	0.415	0.449	0.483	0.433	0.477	0.521	0.449	0.505	0.561
	1800				0.303	0.330	0.358	0.297	0.337	0.378	0.316	0.367	0.418
	3600							**0.936**	**0.976**	**1.016**	**0.957**	**1.005**	**1.052**
	7200										0.216	0.270	0.324

Table 3.22: WPP Memory Exception Analysis - Initial Dataset

Confidence bounds for d for Actual data (white cells), Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).



	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Actual	0.227	0.295	0.362	0.269	0.353	0.437	0.192	0.306	0.420	0.211	0.355	0.500
	Simulated	0.272	0.281	0.289	0.314	0.326	0.337	0.318	0.334	0.350	0.381	0.400	0.418
Aggregated from	30				0.259	0.270	0.280	0.196	0.212	0.228	**0.123**	**0.139**	**0.155**
	60							0.284	0.304	0.323	0.288	0.311	0.333
	300										0.323	0.342	0.360
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Actual	0.205	0.412	0.618	0.128	0.391	0.654	0.115	0.416	0.716	-0.015	0.295	0.606
	Simulated	0.428	0.455	0.482	0.452	0.496	0.541	0.482	0.522	0.561	0.339	0.387	0.434
Aggregated from	30	**0.065**	**0.090**	**0.115**	**0.047**	**0.074**	**0.102**	0.061	0.097	0.132	0.089	0.143	0.197
	60	**0.142**	**0.168**	**0.194**	**0.053**	**0.086**	**0.119**	**0.004**	**0.043**	**0.083**	-0.026	0.026	0.077
	300	0.336	0.364	0.392	0.341	0.375	0.409	0.330	0.376	0.422	0.351	0.406	0.461
	600	0.394	0.421	0.448	0.392	0.426	0.459	0.401	0.442	0.483	0.407	0.464	0.520
	1800				0.431	0.463	0.496	0.439	0.486	0.533	0.470	0.530	0.590
	3600							0.479	0.527	0.574	0.504	0.557	0.610
	7200										0.466	0.520	0.574

In general, the memory parameter confidence bounds for simulated and aggregated simulated data overlap with the confidence bounds for the Actual Data. Elaborating further on Table 3.20, the first row ('Interval') shows the sampling intervals over which estimates were calculated. Moving down one row, Actual Data is shown, e.g. over the 30s sampling interval (first triplet of data), the memory parameter has a point estimate of 0.244 with outer bounds on either side: 0.176 and 0.311. Moving down one more row, the ARFIMA model generated on this data yielded simulated data with average d being 0.233. The next rows down demonstrate the memory parameter estimates created by aggregating data with smaller intervals - since 30s is the smallest interval, this is blank for the first column of triplets, but moving to the 60s triplet, we can see that simulated 30s data was aggregated to create 60s data with an estimate for d of 0.228. The table has been split into two in order to fit all the data, hence the duplication of the headers halfway down.

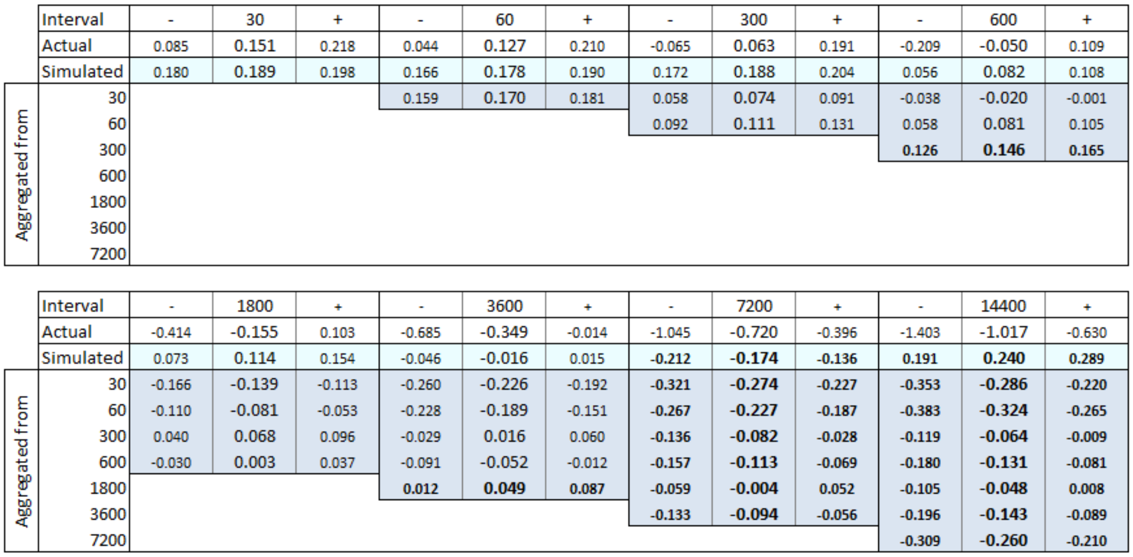
Concentrating on Vodafone, in terms of exceptions then, only the simulated data over the 14400s interval, and the Aggregated Data for the 1800s interval created from 30s data are in bold, demonstrating that these series do not conform to the Actual data in terms of memory properties. However, the relative paucity of exceptions shows that in general, ARFIMA processes fit the data across sampling intervals and aggregate in a way consistent with the data. This is confirmed by LBQ and ARCH tests conducted for the first loop of data - Tables G.1 and G.4 in Appendix G. These show that by and large, the simulated models produce residuals without significant serial correlation or heteroskedasticity, even if the Actual data do exhibit these. Moving onto the other stocks, there are 5 more exceptions in the Aggregated Data for both Legal and General, and WPP, and there is 1 more exception in the Simulated data for Legal and General. Again, the diagnostics in Tables G.2-G.3 and G.5-G.6 show general conformity to residuals without serial correlation or heteroskedasticity, indicating that ARFIMAs fit the data reasonably.

3.4.2 Deseasonalised Dataset - Decreasing Memory with Aggregation

As Table 3.23 shows below, ARFIMA models fit the data at well at all intervals except the largest for Vodafone, for which they almost uniformly fail to capture the correct memory. However, they fit Legal and General (Table 3.24), and WPP data (Table 3.25) well over all the sampling intervals. Nevertheless, all three tables show a decreasing memory pattern as the sampling interval increases; this decreasing memory may be a consequence of the deseasonalisation process as it did not exist for the initial data.

Table 3.23: Vodafone Memory Exception Analysis - Deseasonalised Dataset

Confidence bounds for d for Actual data (white cells), Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).

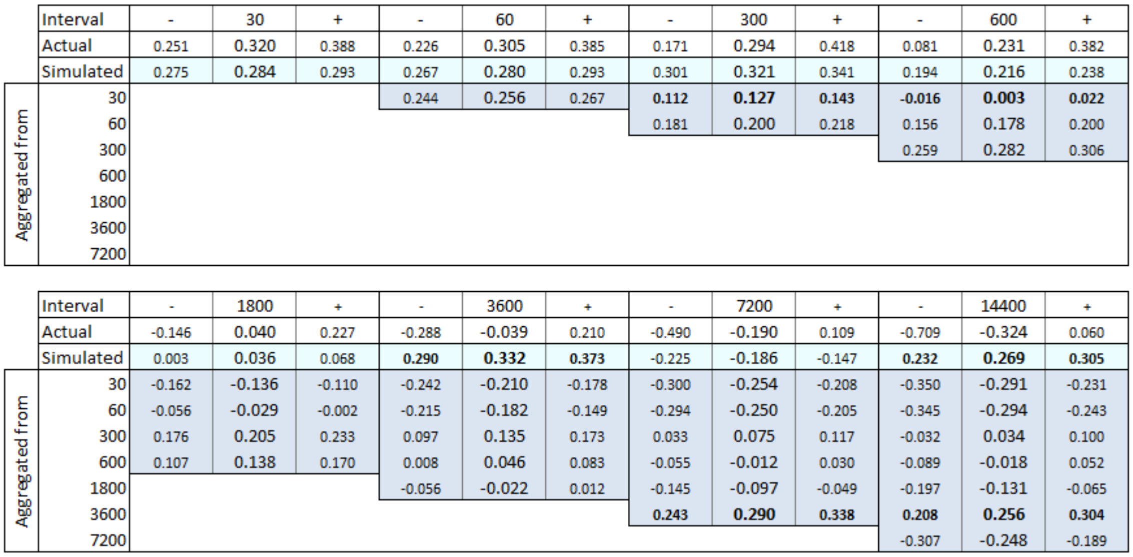


	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Actual	0.085	0.151	0.218	0.044	0.127	0.210	-0.065	0.063	0.191	-0.209	-0.050	0.109
	Simulated	0.180	0.189	0.198	0.166	0.178	0.190	0.172	0.188	0.204	0.056	0.082	0.108
Aggregated from	30				0.159	0.170	0.181	0.058	0.074	0.091	-0.038	-0.020	-0.001
	60							0.092	0.111	0.131	0.058	0.081	0.105
	300										0.126	0.146	0.165
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Actual	-0.414	-0.155	0.103	-0.685	-0.349	-0.014	-1.045	-0.720	-0.396	-1.403	-1.017	-0.630
	Simulated	0.073	0.114	0.154	-0.046	-0.016	0.015	-0.212	-0.174	-0.136	0.191	0.240	0.289
Aggregated from	30	-0.166	-0.139	-0.113	-0.260	-0.226	-0.192	-0.321	-0.274	-0.227	-0.353	-0.286	-0.220
	60	-0.110	-0.081	-0.053	-0.228	-0.189	-0.151	-0.267	-0.227	-0.187	-0.383	-0.324	-0.265
	300	0.040	0.068	0.096	-0.029	0.016	0.060	-0.136	-0.082	-0.028	-0.119	-0.064	-0.009
	600	-0.030	0.003	0.037	-0.091	-0.052	-0.012	-0.157	-0.113	-0.069	-0.180	-0.131	-0.081
	1800				0.012	0.049	0.087	-0.059	-0.004	0.052	-0.105	-0.048	0.008
	3600							-0.133	-0.094	-0.056	-0.196	-0.143	-0.089
	7200										-0.309	-0.260	-0.210

Table 3.24: Legal and General Memory Exception Analysis - Deseasonalised Dataset

Confidence bounds for d for Actual data (white cells), Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).

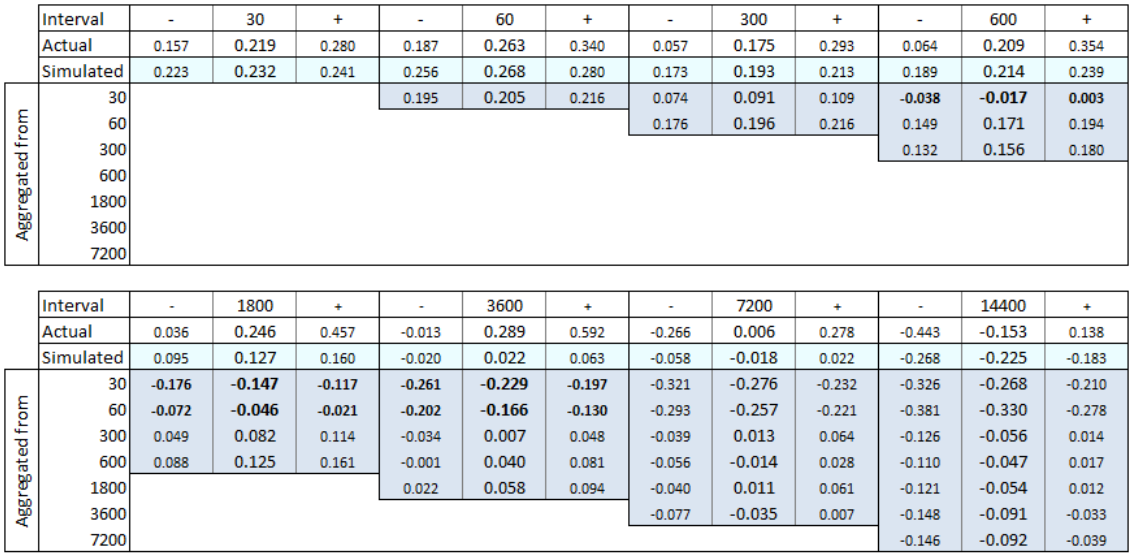


	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Actual	0.251	0.320	0.388	0.226	0.305	0.385	0.171	0.294	0.418	0.081	0.231	0.382
	Simulated	0.275	0.284	0.293	0.267	0.280	0.293	0.301	0.321	0.341	0.194	0.216	0.238
Aggregated from	30				0.244	0.256	0.267	0.112	0.127	0.143	-0.016	0.003	0.022
	60							0.181	0.200	0.218	0.156	0.178	0.200
	300										0.259	0.282	0.306
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Actual	-0.146	0.040	0.227	-0.288	-0.039	0.210	-0.490	-0.190	0.109	-0.709	-0.324	0.060
	Simulated	0.003	0.036	0.068	0.290	0.332	0.373	-0.225	-0.186	-0.147	0.232	0.269	0.305
Aggregated from	30	-0.162	-0.136	-0.110	-0.242	-0.210	-0.178	-0.300	-0.254	-0.208	-0.350	-0.291	-0.231
	60	-0.056	-0.029	-0.002	-0.215	-0.182	-0.149	-0.294	-0.250	-0.205	-0.345	-0.294	-0.243
	300	0.176	0.205	0.233	0.097	0.135	0.173	0.033	0.075	0.117	-0.032	0.034	0.100
	600	0.107	0.138	0.170	0.008	0.046	0.083	-0.055	-0.012	0.030	-0.089	-0.018	0.052
	1800				-0.056	-0.022	0.012	-0.145	-0.097	-0.049	-0.197	-0.131	-0.065
	3600							0.243	0.290	0.338	0.208	0.256	0.304
	7200										-0.307	-0.248	-0.189

Table 3.25: WPP Memory Exception Analysis - Deseasonalised Dataset

Confidence bounds for d for Actual data (white cells), Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).



	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Actual	0.157	0.219	0.280	0.187	0.263	0.340	0.057	0.175	0.293	0.064	0.209	0.354
	Simulated	0.223	0.232	0.241	0.256	0.268	0.280	0.173	0.193	0.213	0.189	0.214	0.239
Aggregated from	30				0.195	0.205	0.216	0.074	0.091	0.109	**-0.038**	**-0.017**	**0.003**
	60							0.176	0.196	0.216	0.149	0.171	0.194
	300										0.132	0.156	0.180
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Actual	0.036	0.246	0.457	-0.013	0.289	0.592	-0.266	0.006	0.278	-0.443	-0.153	0.138
	Simulated	0.095	0.127	0.160	-0.020	0.022	0.063	-0.058	-0.018	0.022	-0.268	-0.225	-0.183
Aggregated from	30	**-0.176**	**-0.147**	**-0.117**	**-0.261**	**-0.229**	**-0.197**	-0.321	-0.276	-0.232	-0.326	-0.268	-0.210
	60	**-0.072**	**-0.046**	**-0.021**	**-0.202**	**-0.166**	**-0.130**	-0.293	-0.257	-0.221	-0.381	-0.330	-0.278
	300	0.049	0.082	0.114	-0.034	0.007	0.048	-0.039	0.013	0.064	-0.126	-0.056	0.014
	600	0.088	0.125	0.161	-0.001	0.040	0.081	-0.056	-0.014	0.028	-0.110	-0.047	0.017
	1800				0.022	0.058	0.094	-0.040	0.011	0.061	-0.121	-0.054	0.012
	3600							-0.077	-0.035	0.007	-0.148	-0.091	-0.033
	7200										-0.146	-0.092	-0.039

3.4.3 Extension to Other Stocks

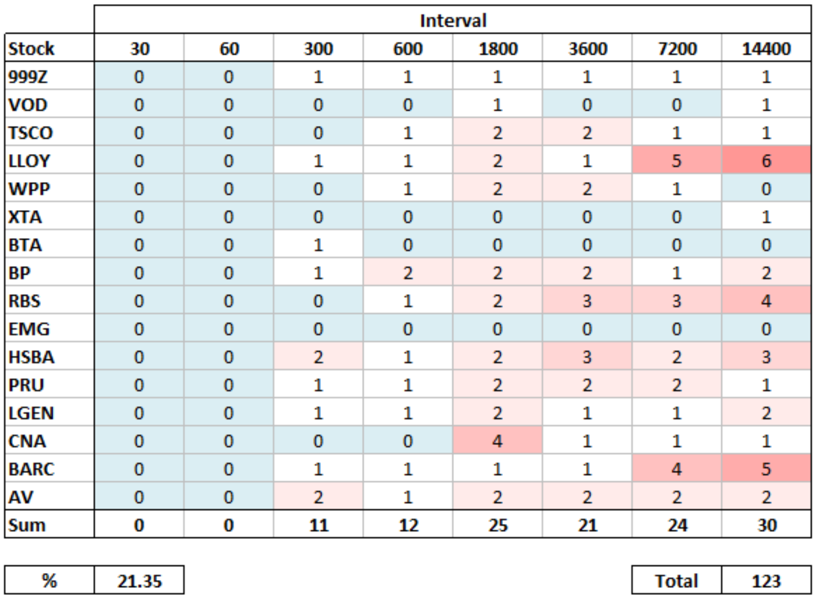
In terms of the other stocks, we created similar tables for them and then counted the number of times the estimates of d failed to lie within base confidence bounds. Tables 3.26 and 3.27 are below. A colour scale of blue to increasing red has been used, with blue corresponding to zero exceptions and increasing red to progressively higher numbers of exceptions.

As can be seen, the number of exceptions rises as the level of aggregation rises, which fits intuitively - bands of uncertainty / forecast variance rises as time increases. Overall though, the tables show that approximately 20% of the time, the d estimates are not in line, with more exceptions arising as the time interval increases. Since this is a new approach to investigation, it is hard to establish whether this is a large number, so further research may yield greater context.

ARFIMAs fit the initial dataset and deseasonalised dataset equally well in terms of memory exceptions, while the deseasonalised dataset features one fewer instance of ARCH(1) effects in the residuals. This implies there is no significant gain to deseasonalising, at least with the data observed. However, the non-standard decreasing memory pattern indicates that 1) this may be artificially induced by the deseasonalisation procedure, and 2) researchers who conduct their research based on one interval may observe certain artificial features of the data but since they lack the other intervals, may not be able to see that the pattern is artificial. We therefore propose to complete this chapter with a simulation study into how the deseasonalisation procedure affects datasets, incorporating suitable control sets.

Table 3.26: Memory Exception Analysis by Stock across Sampling Intervals - Initial Dataset

For each stock (1st column), each row shows the number of exceptions as the sampling interval rises. Colour scale: blue to red as exceptions increase.

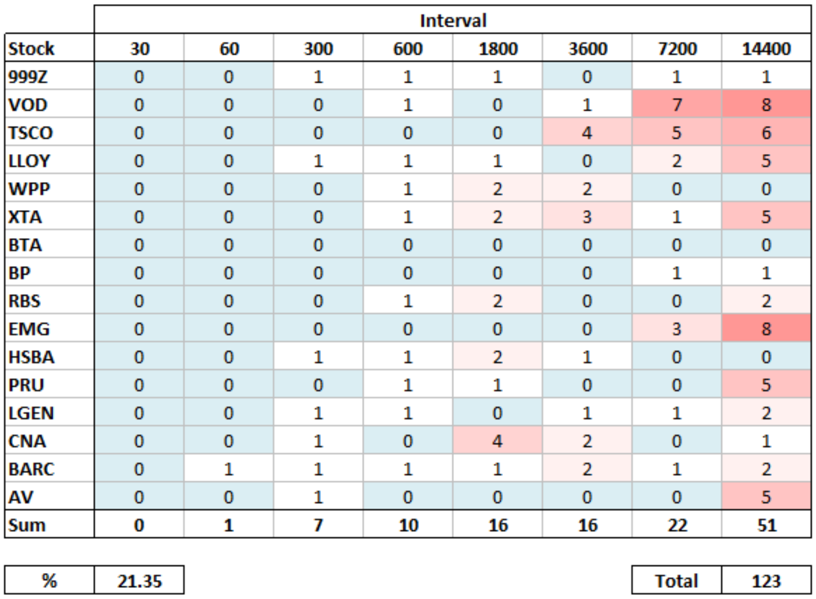


	Interval							
Stock	30	60	300	600	1800	3600	7200	14400
999Z	0	0	1	1	1	1	1	1
VOD	0	0	0	0	1	0	0	1
TSCO	0	0	0	1	2	2	1	1
LLOY	0	0	1	1	2	1	5	6
WPP	0	0	0	1	2	2	1	0
XTA	0	0	0	0	0	0	0	1
BTA	0	0	1	0	0	0	0	0
BP	0	0	1	2	2	2	1	2
RBS	0	0	0	1	2	3	3	4
EMG	0	0	0	0	0	0	0	0
HSBA	0	0	2	1	2	3	2	3
PRU	0	0	1	1	2	2	2	1
LGEN	0	0	1	1	2	1	1	2
CNA	0	0	0	0	4	1	1	1
BARC	0	0	1	1	1	1	4	5
AV	0	0	2	1	2	2	2	2
Sum	0	0	11	12	25	21	24	30

%	21.35	Total	123

Table 3.27: Memory Exception Analysis by Stock across Sampling Intervals - Deseasonalised Dataset

For each stock (1st column), each row shows the number of exceptions as the sampling interval rises. Colour scale: blue to red as exceptions increase.



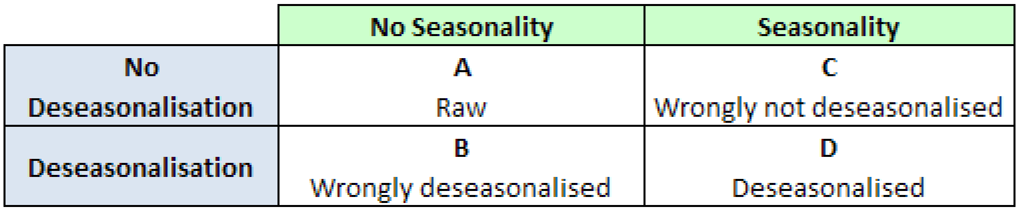
	Interval							
Stock	30	60	300	600	1800	3600	7200	14400
999Z	0	0	1	1	1	0	1	1
VOD	0	0	0	1	0	1	7	8
TSCO	0	0	0	0	0	4	5	6
LLOY	0	0	1	1	1	0	2	5
WPP	0	0	0	1	2	2	0	0
XTA	0	0	0	1	2	3	1	5
BTA	0	0	0	0	0	0	0	0
BP	0	0	0	0	0	0	1	1
RBS	0	0	0	1	2	0	0	2
EMG	0	0	0	0	0	0	3	8
HSBA	0	0	1	1	2	1	0	0
PRU	0	0	0	1	1	0	0	5
LGEN	0	0	1	1	0	1	1	2
CNA	0	0	1	0	4	2	0	1
BARC	0	1	1	1	1	2	1	2
AV	0	0	1	0	0	0	0	5
Sum	0	1	7	10	16	16	22	51

%	21.35	Total	123

3.5 Simulations for the Deseasonalisation Procedure

We now conduct simulations to assess whether the deaseasonalisation procedure of Gallant, Rossi and Tauchen (1992) procedure is neutral. As shown in previous sections, a decreasing memory pattern was observed. It is possible that this is a consequence of the deaseasonalisation procedure. Theoretically, one would expect that implementation of the deaseasonalisation procedure when it is unnecessary would not have any adverse effects on the data. However, we conducted simulations to determine the effects of deseasonalising, yielding 4 sets of results (A-D) as shown in Table 3.28 below.

Table 3.28: Results Sets for Deseasonalisation Simulations
Left Column: results without seasonality, Right Column: results with seasonality
Top Row: results without deseasonalisation, Bottom Row: results with deseasonalisation



	No Seasonality	Seasonality
No Deseasonalisation	A Raw	C Wrongly not deseasonalised
Deseasonalisation	B Wrongly deseasonalised	D Deseasonalised

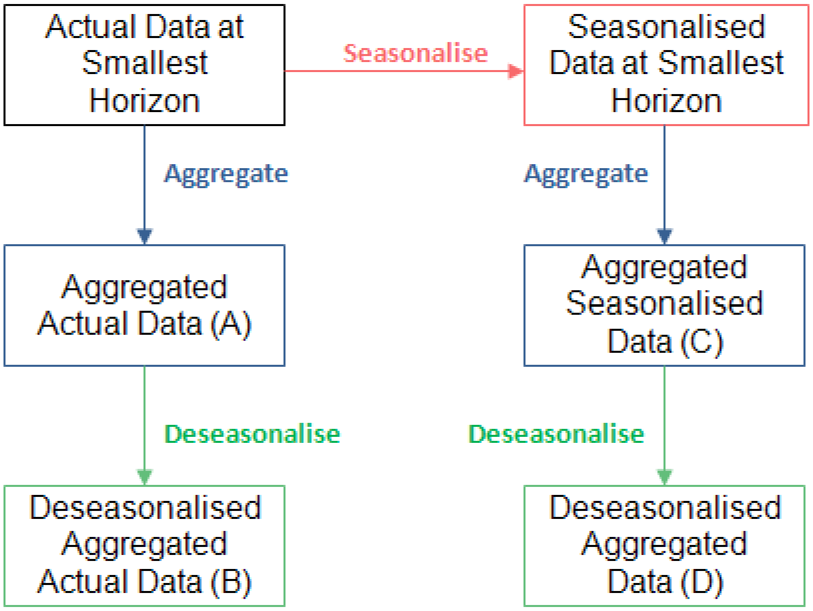
We simulated two sets of series as shown in the figure above - 1) pure ARFIMA (first column) and 2) pure ARFIMA with seasonal components added (second column). We then applied the deseasonalisation procedure to both series and subsequently estimated the memory parameter, producing results sets B and D. As controls, we also estimated the memory parameter without deseasonalisation, producing results sets A and C. The process flow in Figure 3.12 below goes into more detail on the order of operations - note that we aggregate the smallest interval data before deseasonalising (as opposed to aggregating after deseasonalising). We feel this is more realistic - usually researchers take the sampling interval as given and then deseasonalise, rather than deseasonalise and then aggregate over all intervals.

Figure 3.12: Order of Operations for Deseasonalisation Simulations

Top Row: Seasonality is added to the Actual Data

Middle Row: Both sets of data are aggregated to all possible larger sampling intervals

Bottom Row: Both sets of data, over all sampling intervals, are deseasonalised


Actual Data at Smallest Horizon
Seasonalise
Seasonalised Data at Smallest Horizon
Aggregate
Aggregate
Aggregated Actual Data (A)
Aggregated Seasonalised Data (C)
Deseasonalise
Deseasonalise
Deseasonalised Aggregated Actual Data (B)
Deseasonalised Aggregated Data (D)

We can obtain different sets of results based on the scale of the seasonal component with respect to the ARFIMA component; the series, $V_t = s_t + f\varepsilon_t$ where s_t is the seasonal component, ε_t is the pure ARFIMA component and f is some scaling factor. So while changes in the value of f will not change any estimates of d for results sets A and B (since there is no seasonal component and hence the GPH estimator will simply work on a pure ARFIMA with a different scale), they will change results for sets C and D. We take $f = 1e - 9$, as is approximately the case for the real data. However, unlike for the real data, we do not recreate the seasonal component, s_t, to mimic that found, as the seasonal component could effectively take any form. Therefore we have chosen a simple form: $s_t = \delta b$ where δ is the set of all dummies and b is the column vector of means of each column of δ; $b_j = 1/n \sum_{i=1}^{n} \delta_{i,j}$.

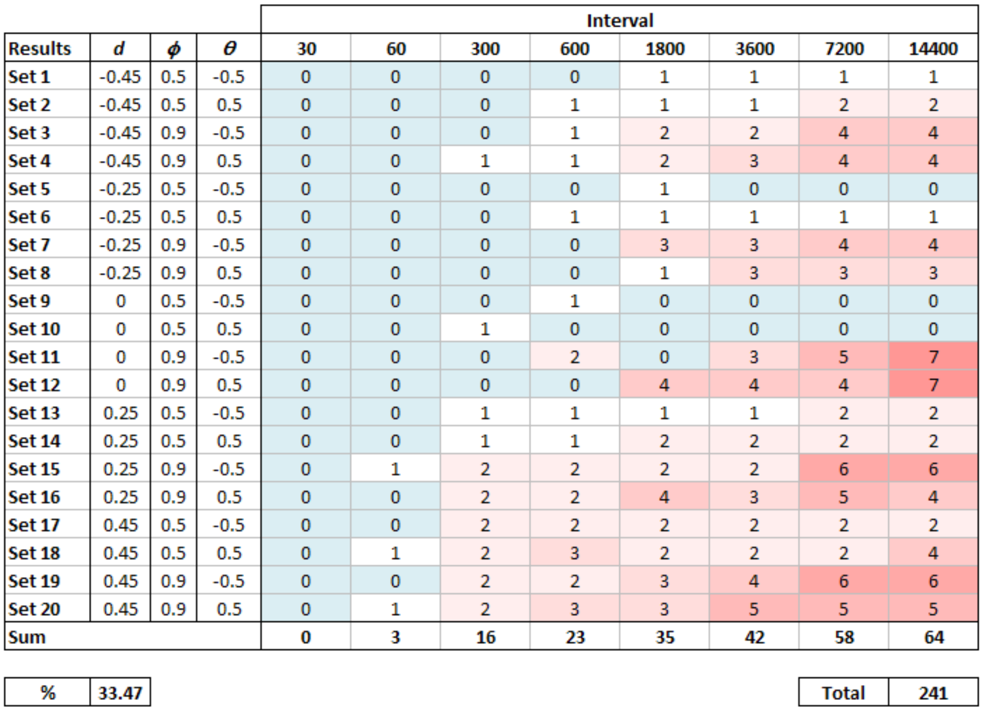
Hence we do not recreate exactly the same pattern as in the real data, but it is useful to see whether similar patterns arise for differents sets of parameters within the parameter space. Clearly further investigation is required over different scaling factors and seasonal components, though this lies beyond the scope and resources of the current investigation.

We conducted the simulation investigation assuming an ARFIMA(1,d,1) process over different sets of values for the parameters: $\phi \in \{0.5, 0.9\}, d \in \{-0.45, -0.25, 0, 0.25, 0.45\}$ and $\theta \in \{-0.5, 0.5\}$. We chose these values in order to span the parameter space as best we could, and to see if there was any conflict between high values of ϕ and the value of d; as ϕ nears 1, a time series becomes closer to nonstationarity, which is detected in periodograms at the low frequencies, although these also pick up long memory. As the tables below in Tables 3.29-3.32 show, exceptions arise approximately 40% of the time on average, with most occurring for the larger sampling intervals though without a specific pattern for the values of d or ϕ (there is a weak increase in the number of exceptions with a high ϕ in Table 3.29, but this only occurs for larger sampling intervals and is not found in the other tables). So no significant conflict arose between a high ϕ and d.

Observe that we have the fewest exceptions in the wrongly deseasonalised data (set B) even though the deseasonalisation procedure induces an artificial memory pattern. Meanwhile, we have most exceptions when there is seasonality and the deseasonalistion procedure is implemented (set D). Since the wrongly not-deseasonalised data (set C) represents the worst case for deseasonalisation, and it actually has less exceptions that the correctly deseasonalised data (set D), it suggests there is not much loss if the deseasonalisation procedure is not implemented at all; sets A and C do not have the most number of exceptions.

Table 3.29: Memory Exception Analysis by Results Set across sampling intervals - A / Raw (no Seasonality, no Deseasonalisation)

For each results set (1st column), each row shows the number of exceptions as the sampling interval rises. Colour scale: blue to red as exceptions increase.



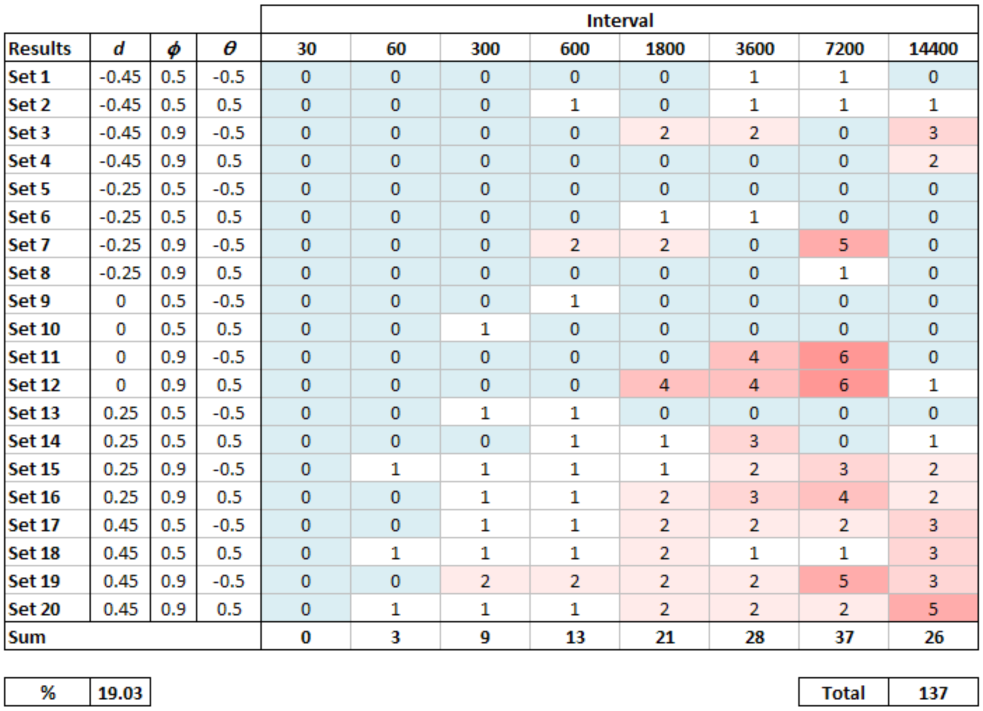
				Interval							
Results	***d***	***ϕ***	***θ***	**30**	**60**	**300**	**600**	**1800**	**3600**	**7200**	**14400**
Set 1	-0.45	0.5	-0.5	0	0	0	0	1	1	1	1
Set 2	-0.45	0.5	0.5	0	0	0	1	1	1	2	2
Set 3	-0.45	0.9	-0.5	0	0	0	1	2	2	4	4
Set 4	-0.45	0.9	0.5	0	0	1	1	2	3	4	4
Set 5	-0.25	0.5	-0.5	0	0	0	0	1	0	0	0
Set 6	-0.25	0.5	0.5	0	0	0	1	1	1	1	1
Set 7	-0.25	0.9	-0.5	0	0	0	0	3	3	4	4
Set 8	-0.25	0.9	0.5	0	0	0	0	1	3	3	3
Set 9	0	0.5	-0.5	0	0	0	1	0	0	0	0
Set 10	0	0.5	0.5	0	0	1	0	0	0	0	0
Set 11	0	0.9	-0.5	0	0	0	2	0	3	5	7
Set 12	0	0.9	0.5	0	0	0	0	4	4	4	7
Set 13	0.25	0.5	-0.5	0	0	1	1	1	1	2	2
Set 14	0.25	0.5	0.5	0	0	1	1	2	2	2	2
Set 15	0.25	0.9	-0.5	0	1	2	2	2	2	6	6
Set 16	0.25	0.9	0.5	0	0	2	2	4	3	5	4
Set 17	0.45	0.5	-0.5	0	0	2	2	2	2	2	2
Set 18	0.45	0.5	0.5	0	1	2	3	2	2	2	4
Set 19	0.45	0.9	-0.5	0	0	2	2	3	4	6	6
Set 20	0.45	0.9	0.5	0	1	2	3	3	5	5	5
Sum				**0**	**3**	**16**	**23**	**35**	**42**	**58**	**64**

%	**33.47**

Total	**241**

Table 3.30: Memory Exception Analysis by Results Set across sampling intervals - B / Wrongly Deseasonalised (no Seasonality, but Deseasonalisation)

For each results set (1st column), each row shows the number of exceptions as the sampling interval rises. Colour scale: blue to red as exceptions increase.



				Interval							
Results	*d*	*ϕ*	*θ*	30	60	300	600	1800	3600	7200	14400
Set 1	-0.45	0.5	-0.5	0	0	0	0	0	1	1	0
Set 2	-0.45	0.5	0.5	0	0	0	1	0	1	1	1
Set 3	-0.45	0.9	-0.5	0	0	0	0	2	2	0	3
Set 4	-0.45	0.9	0.5	0	0	0	0	0	0	0	2
Set 5	-0.25	0.5	-0.5	0	0	0	0	0	0	0	0
Set 6	-0.25	0.5	0.5	0	0	0	0	1	1	0	0
Set 7	-0.25	0.9	-0.5	0	0	0	2	2	0	5	0
Set 8	-0.25	0.9	0.5	0	0	0	0	0	0	1	0
Set 9	0	0.5	-0.5	0	0	0	1	0	0	0	0
Set 10	0	0.5	0.5	0	0	1	0	0	0	0	0
Set 11	0	0.9	-0.5	0	0	0	0	0	4	6	0
Set 12	0	0.9	0.5	0	0	0	0	4	4	6	1
Set 13	0.25	0.5	-0.5	0	0	1	1	0	0	0	0
Set 14	0.25	0.5	0.5	0	0	0	1	1	3	0	1
Set 15	0.25	0.9	-0.5	0	1	1	1	1	2	3	2
Set 16	0.25	0.9	0.5	0	0	1	1	2	3	4	2
Set 17	0.45	0.5	-0.5	0	0	1	1	2	2	2	3
Set 18	0.45	0.5	0.5	0	1	1	1	2	1	1	3
Set 19	0.45	0.9	-0.5	0	0	2	2	2	2	5	3
Set 20	0.45	0.9	0.5	0	1	1	1	2	2	2	5
Sum				0	3	9	13	21	28	37	26

%	19.03

Total	137

Table 3.31: Memory Exception Analysis by Results Set across sampling intervals - C / Wrongly NOT Deseasonalised (Seasonality, but no Deseasonalisation)

For each results set (1st column), each row shows the number of exceptions as the sampling interval rises. Colour scale: blue to red as exceptions increase.

				Interval							
Results	***d***	***ϕ***	***θ***	**30**	**60**	**300**	**600**	**1800**	**3600**	**7200**	**14400**
Set 1	-0.45	0.5	-0.5	0	1	2	3	4	5	6	7
Set 2	-0.45	0.5	0.5	0	0	2	3	4	3	5	6
Set 3	-0.45	0.9	-0.5	0	0	2	0	4	2	5	6
Set 4	-0.45	0.9	0.5	0	0	2	1	2	2	5	7
Set 5	-0.25	0.5	-0.5	0	1	2	1	4	5	6	5
Set 6	-0.25	0.5	0.5	0	0	2	0	3	2	5	6
Set 7	-0.25	0.9	-0.5	0	0	2	0	2	2	5	6
Set 8	-0.25	0.9	0.5	0	0	2	0	0	0	5	7
Set 9	0	0.5	-0.5	0	0	2	1	3	2	5	6
Set 10	0	0.5	0.5	0	0	2	0	2	2	5	7
Set 11	0	0.9	-0.5	0	0	2	0	2	3	5	6
Set 12	0	0.9	0.5	0	0	2	0	1	0	5	6
Set 13	0.25	0.5	-0.5	0	0	2	0	4	2	5	6
Set 14	0.25	0.5	0.5	0	0	2	0	1	2	5	7
Set 15	0.25	0.9	-0.5	0	0	1	2	1	0	4	6
Set 16	0.25	0.9	0.5	0	1	0	0	0	0	3	4
Set 17	0.45	0.5	-0.5	0	0	1	0	1	5	5	6
Set 18	0.45	0.5	0.5	0	0	0	0	0	0	4	7
Set 19	0.45	0.9	-0.5	0	0	0	0	0	0	0	0
Set 20	0.45	0.9	0.5	0	1	0	1	4	4	0	2
Sum				**0**	**4**	**30**	**12**	**42**	**41**	**88**	**113**

%	**45.83**

Total	**330**

Table 3.32: Memory Exception Analysis by Results Set across sampling intervals - D / Deseasonalised (Seasonality and Deseasonalisation)

For each results set (1st column), each row shows the number of exceptions as the sampling interval rises. Colour scale: blue to red as exceptions increase.

				Interval							
Results	***d***	***ϕ***	***θ***	**30**	**60**	**300**	**600**	**1800**	**3600**	**7200**	**14400**
Set 1	-0.45	0.5	-0.5	0	1	2	3	4	5	6	7
Set 2	-0.45	0.5	0.5	0	1	2	3	4	5	6	7
Set 3	-0.45	0.9	-0.5	0	1	2	3	4	5	1	5
Set 4	-0.45	0.9	0.5	0	1	2	2	4	5	1	1
Set 5	-0.25	0.5	-0.5	0	1	2	2	4	5	1	7
Set 6	-0.25	0.5	0.5	0	1	2	2	4	5	1	6
Set 7	-0.25	0.9	-0.5	0	1	2	1	2	5	6	1
Set 8	-0.25	0.9	0.5	0	1	2	2	3	5	4	3
Set 9	0	0.5	-0.5	0	0	2	2	2	5	1	7
Set 10	0	0.5	0.5	0	1	2	0	2	5	1	3
Set 11	0	0.9	-0.5	0	0	2	3	2	0	6	1
Set 12	0	0.9	0.5	0	0	1	0	3	0	6	1
Set 13	0.25	0.5	-0.5	0	1	2	3	2	0	6	5
Set 14	0.25	0.5	0.5	0	1	2	3	2	3	2	1
Set 15	0.25	0.9	-0.5	0	1	2	2	2	3	5	6
Set 16	0.25	0.9	0.5	0	0	2	2	2	5	5	4
Set 17	0.45	0.5	-0.5	0	1	2	3	3	5	6	2
Set 18	0.45	0.5	0.5	0	1	2	2	2	5	6	1
Set 19	0.45	0.9	-0.5	0	0	2	2	3	5	5	5
Set 20	0.45	0.9	0.5	0	1	2	1	2	5	4	5
Sum				**0**	**15**	**39**	**41**	**56**	**81**	**79**	**78**

%	54.03

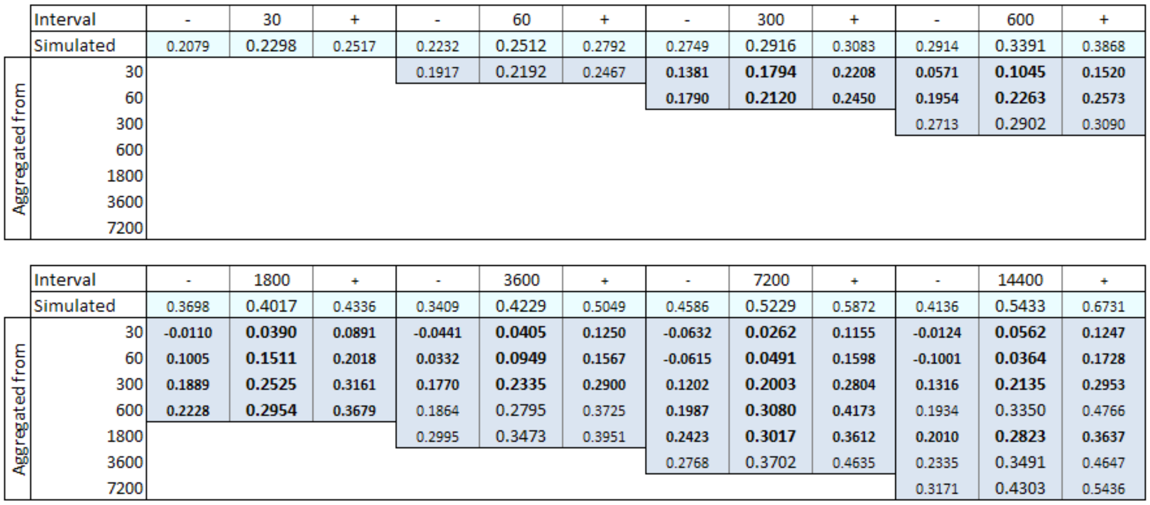
Total	389

Looking at specific data for one of the sets of parameters with most exceptions, Set 16, we see in Tables 3.33-3.36 that the two non-deseasonalised datasets (sets A and C in Tables 3.33 and 3.35) feature increasing or fluctuating memory with aggregation, while we again have (weakly) decreasing memory for sets B and D (Tables 3.34 and 3.36). Most importantly, this shows that it is only as an effect of the deseasonalisation procedure that we gain a decreasing memory pattern, and that this outcome results regardless of whether seasonality exists in the data.

Note also that Sets B and D have almost identical results for the highest row of the data. This indicates that the deseasonalisation procedure tends to dominate the data it acts on, at least at small intervals. There is a divergence thereafter because the data is aggregated differently; the smallest interval data used for producing the top row of Set B is aggregated and then deseasonalised, whereas the data used for set D is aggregated from the seasonalised data (as was shown in the "Order of Operations Diagram" - Figure 3.12). Overall, we conclude that the deseasonalisation procedure is possibly redundant and at worst harmful in the sense that it may generate artificial memory patterns.

Table 3.33: Set 16 Memory Exception Analysis - A / Raw (no Seasonality, no Deseasonalisation)

Confidence bounds for d for Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).

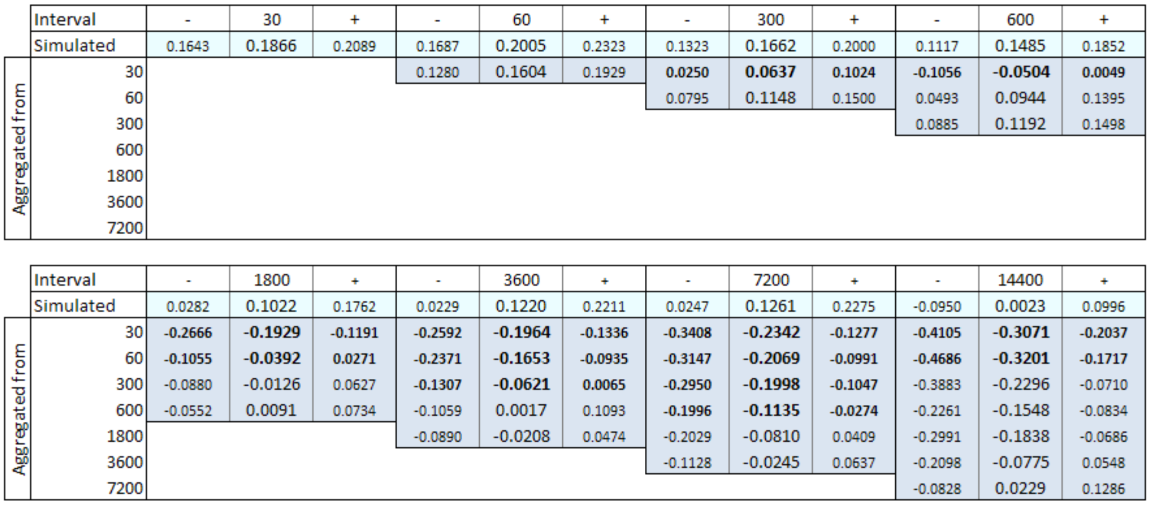


	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Simulated	0.2079	0.2298	0.2517	0.2232	0.2512	0.2792	0.2749	0.2916	0.3083	0.2914	0.3391	0.3868
Aggregated from	30				0.1917	0.2192	0.2467	**0.1381**	**0.1794**	**0.2208**	**0.0571**	**0.1045**	**0.1520**
	60							**0.1790**	**0.2120**	**0.2450**	**0.1954**	**0.2263**	**0.2573**
	300										0.2713	0.2902	0.3090
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Simulated	0.3698	0.4017	0.4336	0.3409	0.4229	0.5049	0.4586	0.5229	0.5872	0.4136	0.5433	0.6731
Aggregated from	30	**-0.0110**	**0.0390**	**0.0891**	**-0.0441**	**0.0405**	**0.1250**	**-0.0632**	**0.0262**	**0.1155**	**-0.0124**	**0.0562**	**0.1247**
	60	**0.1005**	**0.1511**	**0.2018**	**0.0332**	**0.0949**	**0.1567**	**-0.0615**	**0.0491**	**0.1598**	**-0.1001**	**0.0364**	**0.1728**
	300	**0.1889**	**0.2525**	**0.3161**	**0.1770**	**0.2335**	**0.2900**	**0.1202**	**0.2003**	**0.2804**	**0.1316**	**0.2135**	**0.2953**
	600	**0.2228**	**0.2954**	**0.3679**	0.1864	0.2795	0.3725	**0.1987**	**0.3080**	**0.4173**	0.1934	0.3350	0.4766
	1800				0.2995	0.3473	0.3951	**0.2423**	**0.3017**	**0.3612**	**0.2010**	**0.2823**	**0.3637**
	3600							0.2768	0.3702	0.4635	0.2335	0.3491	0.4647
	7200										0.3171	0.4303	0.5436

Table 3.34: Set 16 Memory Exception Analysis - B / Wrongly Deseasonalised (no Seasonality, but Deseasonalisation)

Confidence bounds for d for Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).

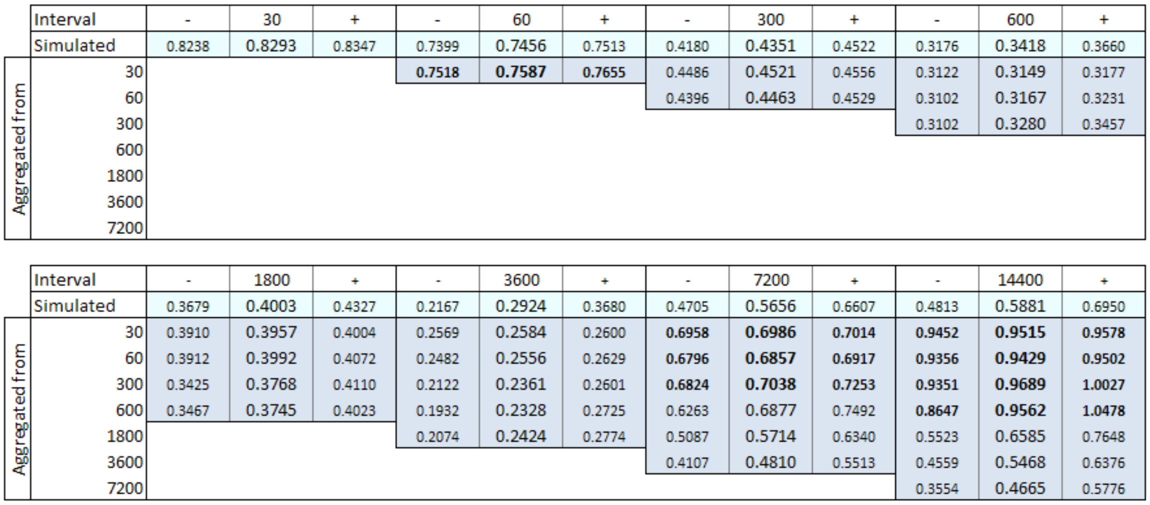


	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Simulated	0.1643	0.1866	0.2089	0.1687	0.2005	0.2323	0.1323	0.1662	0.2000	0.1117	0.1485	0.1852
Aggregated from	30				0.1280	0.1604	0.1929	**0.0250**	**0.0637**	**0.1024**	**-0.1056**	**-0.0504**	**0.0049**
	60							0.0795	0.1148	0.1500	0.0493	0.0944	0.1395
	300										0.0885	0.1192	0.1498
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Simulated	0.0282	0.1022	0.1762	0.0229	0.1220	0.2211	0.0247	0.1261	0.2275	-0.0950	0.0023	0.0996
Aggregated from	30	**-0.2666**	**-0.1929**	**-0.1191**	**-0.2592**	**-0.1964**	**-0.1336**	**-0.3408**	**-0.2342**	**-0.1277**	**-0.4105**	**-0.3071**	**-0.2037**
	60	**-0.1055**	**-0.0392**	**0.0271**	**-0.2371**	**-0.1653**	**-0.0935**	**-0.3147**	**-0.2069**	**-0.0991**	**-0.4686**	**-0.3201**	**-0.1717**
	300	-0.0880	-0.0126	0.0627	**-0.1307**	**-0.0621**	**0.0065**	**-0.2950**	**-0.1998**	**-0.1047**	-0.3883	-0.2296	-0.0710
	600	-0.0552	0.0091	0.0734	-0.1059	0.0017	0.1093	**-0.1996**	**-0.1135**	**-0.0274**	-0.2261	-0.1548	-0.0834
	1800				-0.0890	-0.0208	0.0474	-0.2029	-0.0810	0.0409	-0.2991	-0.1838	-0.0686
	3600							-0.1128	-0.0245	0.0637	-0.2098	-0.0775	0.0548
	7200										-0.0828	0.0229	0.1286

Table 3.35: Set 16 Memory Exception Analysis - C / Wrongly NOT Deseasonalised (Seasonality, but no Deseasonalisation)

Confidence bounds for d for Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).

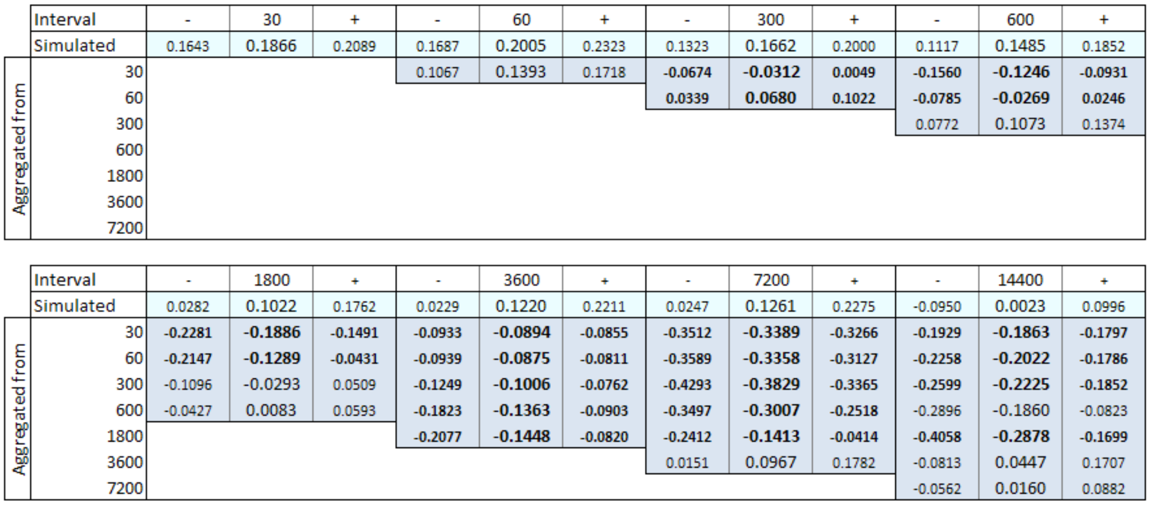


	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Simulated	0.8238	0.8293	0.8347	0.7399	0.7456	0.7513	0.4180	0.4351	0.4522	0.3176	0.3418	0.3660
Aggregated from	30				**0.7518**	**0.7587**	**0.7655**	0.4486	0.4521	0.4556	0.3122	0.3149	0.3177
	60							0.4396	0.4463	0.4529	0.3102	0.3167	0.3231
	300										0.3102	0.3280	0.3457
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Simulated	0.3679	0.4003	0.4327	0.2167	0.2924	0.3680	0.4705	0.5656	0.6607	0.4813	0.5881	0.6950
Aggregated from	30	0.3910	0.3957	0.4004	0.2569	0.2584	0.2600	**0.6958**	**0.6986**	**0.7014**	**0.9452**	**0.9515**	**0.9578**
	60	0.3912	0.3992	0.4072	0.2482	0.2556	0.2629	**0.6796**	**0.6857**	**0.6917**	**0.9356**	**0.9429**	**0.9502**
	300	0.3425	0.3768	0.4110	0.2122	0.2361	0.2601	**0.6824**	**0.7038**	**0.7253**	**0.9351**	**0.9689**	**1.0027**
	600	0.3467	0.3745	0.4023	0.1932	0.2328	0.2725	0.6263	0.6877	0.7492	**0.8647**	**0.9562**	**1.0478**
	1800				0.2074	0.2424	0.2774	0.5087	0.5714	0.6340	0.5523	0.6585	0.7648
	3600							0.4107	0.4810	0.5513	0.4559	0.5468	0.6376
	7200										0.3554	0.4665	0.5776

Table 3.36: Set 16 Memory Exception Analysis - D / Deseasonalised (Seasonality and Deseasonalisation)

Confidence bounds for d for Simulated data (light blue cells) and Aggregated Simulated data (dark blue cells).



	Interval	-	30	+	-	60	+	-	300	+	-	600	+
	Simulated	0.1643	0.1866	0.2089	0.1687	0.2005	0.2323	0.1323	0.1662	0.2000	0.1117	0.1485	0.1852
Aggregated from	30				0.1067	0.1393	0.1718	**-0.0674**	**-0.0312**	**0.0049**	**-0.1560**	**-0.1246**	**-0.0931**
	60							**0.0339**	**0.0680**	**0.1022**	**-0.0785**	**-0.0269**	**0.0246**
	300										0.0772	0.1073	0.1374
	600												
	1800												
	3600												
	7200												

	Interval	-	1800	+	-	3600	+	-	7200	+	-	14400	+
	Simulated	0.0282	0.1022	0.1762	0.0229	0.1220	0.2211	0.0247	0.1261	0.2275	-0.0950	0.0023	0.0996
Aggregated from	30	**-0.2281**	**-0.1886**	**-0.1491**	**-0.0933**	**-0.0894**	**-0.0855**	**-0.3512**	**-0.3389**	**-0.3266**	**-0.1929**	**-0.1863**	**-0.1797**
	60	**-0.2147**	**-0.1289**	**-0.0431**	**-0.0939**	**-0.0875**	**-0.0811**	**-0.3589**	**-0.3358**	**-0.3127**	**-0.2258**	**-0.2022**	**-0.1786**
	300	-0.1096	-0.0293	0.0509	**-0.1249**	**-0.1006**	**-0.0762**	**-0.4293**	**-0.3829**	**-0.3365**	**-0.2599**	**-0.2225**	**-0.1852**
	600	-0.0427	0.0083	0.0593	**-0.1823**	**-0.1363**	**-0.0903**	**-0.3497**	**-0.3007**	**-0.2518**	-0.2896	-0.1860	-0.0823
	1800				**-0.2077**	**-0.1448**	**-0.0820**	**-0.2412**	**-0.1413**	**-0.0414**	**-0.4058**	**-0.2878**	**-0.1699**
	3600							0.0151	0.0967	0.1782	-0.0813	0.0447	0.1707
	7200										-0.0562	0.0160	0.0882

3.6 Conclusion

High frequency data is highly computationally intensive. It was the aim of this chapter to conduct a basic investigation of volume time series for a number of stocks on the London Stock Exchange. We have created a computing architecture and a set of techniques for processing and analysing the data in our use of colour scales and Memory Exception Analysis. This is our first contribution.

After processing the data, an initial dataset was produced which was deseasonalised to create a deseasonalised dataset as per Gallant, Rossi and Tauchen's (1992) procedures. Our other contributions deal with addressing the questions posed at the beginning of the chapter.

Q1 (Data Features): We found that there was significant seasonality in the data, so the deseasonalisation procedure was needed. In contrast, detrending the data was not found to be significant. For the initial/raw dataset, we found a pattern of stable long memory over all intervals, which is theoretically consistent with the work of Chambers (1998). Empirically, long memory at daily intervals agrees with findings by Bollerslev and Jubinsky (1999) and Lobato and Velasco (2000). However, as mentioned by Engle (2000), intradaily data is subject to seasonality. With deseasonalised data, we found a pattern of decreasing memory with temporal aggregation. This is a new empirical finding, although it clashes with the theoretical result of Chambers (1998). However, it is possible that this decreasing memory pattern was artificially induced by the deseasonalisation process, as shown in simulations to address Q2. Finally, ARFIMA models could fit the Classically Decomposed data to some extent.

Since different ARFIMA orders and values for the memory parameter are produced over different intervals after Classical Decomposition, the same models are not equally applicable across all intervals. This suggests that there are effects from aggregation and processing which require different models. Further research may also yield more sophisticated modelling features which may be required such as correcting for more relative discreteness at small intervals (as per SF1.2 - discreteness of marks) as well as increased microstructure noise which may cancel or fade as the interval increases. A rationale for the existence of the microstructure noise may be that at small intervals, informed high-frequency traders make transactions to exploit price differences, whereas over larger intervals, their moves may cancel out.

Q2 (Model Simulations): Using simulated data for all the sampling intervals, we found similar but not identical results to both memory patterns can be replicated. Using confidence intervals for the GPH estimates, we found that for the Initial dataset, the stable memory pattern was almost always consistent with aggregated simulated data.

The decreasing memory pattern was also replicated, but using simulations for the deseasonalisation procedure, we have shown that the deseasonalisation procedure is not neutral in the sense that it induces an artificial memory pattern regardless of whether seasonality exists. Moreover, whether or not it is used, a similar number of exceptions arose, indicating the possible superfluousness of the procedure (even for the induced pattern).

Overall we have created new research which shows features of the data over a range of previously unexplored sampling intervals. We also confirmed that the raw data series has a stable memory pattern across sampling intervals as theoretically predicted by Chambers (1998) while deseasonalising data may induce a decreasing memory pattern. In addition, we have also implemented new ways of analysing data over different intervals, in terms of colour scales and ”Memory Exception Analysis”, which can be refined and built on in further research.

It would be interesting to conduct further research along these lines over more high-frequency datasets, as well as extend the simulation section by increasing the number of simulation loops, extending the parameter grid, and trialling more values of the scale factor and seasonal components. In addition, alternatives to the GPH estimator could be trialled, such as Lo’s R/S statistic. Such work would likely need an expanded version of the Computing Design (Figure 3.1), possibly with more powerful sets of computers which would also alleviate the problem of multicollinearity in deseasonalising the data when including GAP dummies. Further decomposition of the data could also be conducted by extending the work of Bialkowski et al. (2008) to split the random volume component into systematic and idiosyncratic components as for the CAPM. Finally, we have explored temporal aggregation in this chapter, so it would be interesting to explore if effects exist cross-sectionally. We suggest that such research be conducted in teams though, owing to its highly computationally intensive requirements.

A theoretical explanation for decreasing memory after deseasonalisation could also be pursued to explain the deseasonalisation procedure’s effects on the data. For example[1], if we model volumes, V_t, as a simple AR(1) process plus seasonal effects for the morning and afternoon, say M_t and A_t respectively, with δ_1 and δ_2 denoting the size of these effects, then we would have the following specification:

$$V_t = \phi V_{t-1} + \delta_1 M_t + \delta_2 A_t$$

Now aggregating to a larger sample interval, say double the length of the original, we

[1] We thank Professor Marcelo Fernandes of Queen Mary University of London for calling our attention to this

obtain:

$$V_t + V_{t+1} = \phi(V_t + V_{t-1}) + \delta_1(M_{t+1} + M_t) + \delta_2(A_{t+1} + A_t) + \varepsilon_{t+1} + \varepsilon_t$$

$$\begin{aligned} &= \phi^2(V_{t-1} + V_{t-2}) \\ &+ \delta_1 M_{t+1} + (\phi + 1)\delta_1(M_t) + \phi\delta_1 M_{t-1} \\ &+ \delta_2 A_{t+1} + (\phi + 1)\delta_2(A_t) + \phi\delta_2 A_{t-1} \\ &+ \varepsilon_{t+1} + (\phi + 1)\varepsilon_t + \phi\varepsilon_{t-1} \end{aligned}$$

As can be seen from the coefficients of $\{M_t\}$ and $\{A_t\}$, the seasonal pattern at the larger sampling interval is different from the pattern at the smaller sampling interval. This is not an issue if the smaller interval (e.g. 7200s) divides the larger (e.g. 14400s); the first 14400 interval will have different properties in terms of the mean and variance from the next, but deseasonalisation within this first interval at the 7200s frequency will still yield valid coefficients for the $\{M_t\}$ and $\{A_t\}$. However, if the larger interval (e.g. 30600s) is not divided an integer number of times by the smaller, then eventually a 7200s interval may span two 30600s intervals with different means, potentially causing a different seasonal pattern, and structural breaks in the data. If these are large enough, they may generate spurious persistence properties. Further research could expand on this idea with more general models to explore the level of spurious persistence which might arise.

APPENDIX A: LISTS OF STOCKS

Table A.1: List of Stocks used for Analysis

EPIC / Symbol	Name / Description
999Z	Index - sum of other volumes
VOD	Vodafone Group PLC
TSCO	Tesco PLC
LLOY	Lloyds Banking Group PLC
WPP	WPP Group PLC
XTA	Xstrata PLC
BTA	BT Group PLC
BP	BHP Billiton PLC
RBS	Royal Bank of Scotland Group (The) PLC
EMG	Man Group PLC
HSBA	HSBC Holdings PLC
PRU	Prudential PLC
LGEN	Legal & General Group PLC
CNA	Centrica PLC
BARC	Barclays PLC
AV	ARM Holdings PLC

Table A.2: List of Stocks used for Creation of Index (“999Z”)

EPIC / Symbol	Name / Description
AAL	Anglo American PLC
ARM	ARM Holdings PLC
AV.	Aviva PLC
BA.	BAE Systems PLC
BARC	Barclays PLC
BAY	British Airways PLC
BG.	BG Group PLC
BLND	British Land Co PLC
BLT	BHP Billiton PLC
BP.	BP PLC
BT.A	BT Group PLC
CNA	Centrica PLC
CNE	Cairn Energy PLC
COB	Cobham PLC
DGE	Diageo PLC
EMG	Man Group PLC
GFS	G4S PLC
GSK	GlaxoSmithKline PLC
HOME	Home Retail Group PLC
HSBA	HSBC Holdings PLC
IPR	International Power PLC
KGF	Kingfisher PLC
LAND	Land Securities Group PLC
LGEN	Legal & General Group PLC
LLOY	Lloyds Banking Group PLC
MKS	Marks & Spencer Group PLC
MRW	Morrison (Wm) Supermarkets PLC
NG.	National Grid PLC
OML	Old Mutual PLC
PRU	Prudential PLC
RBS	Royal Bank of Scotland Group (The) PLC
RDSB	Royal Dutch Shell PLC
REL	Reed Elsevier PLC
REX	Rexam PLC
RIO	Rio Tinto PLC
RR.	Rolls-Royce Group PLC
RSA	RSA Insurance Group PLC
SBRY	Sainsbury (J) PLC
SGE	Sage Group (The) PLC
STAN	Standard Chartered PLC
TCG	Thomas Cook Group PLC
TSCO	Tesco PLC
VOD	Vodafone Group PLC
WPP	WPP Group PLC
XTA	Xstrata PLC

Appendix B: Colour Scales for Appendix Results Tables

Table B.1: Colour Scales for Results Tables

Different colour scales are used according to the type of test conducted, with extremes as detailed below. For example, for the GPH Estimates tables, the darker the red, the lower the estimate of d, while the darker the green, the higher the estimate. Values close to zero will be closer to white in colour.

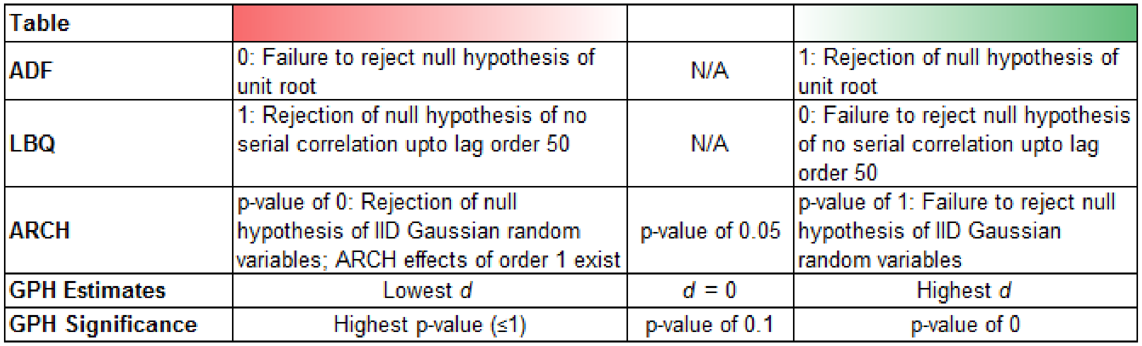
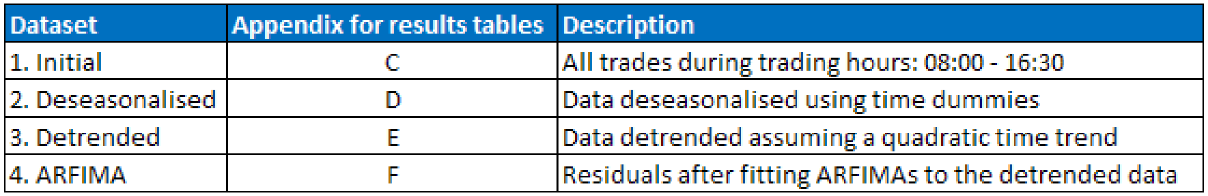


Table	🟥		🟩
ADF	0: Failure to reject null hypothesis of unit root	N/A	1: Rejection of null hypothesis of unit root
LBQ	1: Rejection of null hypothesis of no serial correlation upto lag order 50	N/A	0: Failure to reject null hypothesis of no serial correlation upto lag order 50
ARCH	p-value of 0: Rejection of null hypothesis of IID Gaussian random variables; ARCH effects of order 1 exist	p-value of 0.05	p-value of 1: Failure to reject null hypothesis of IID Gaussian random variables
GPH Estimates	Lowest d	$d = 0$	Highest d
GPH Significance	Highest p-value (≤1)	p-value of 0.1	p-value of 0

The guide to the different datasets is reproduced below:

Table B.2: Datasets and Descriptions

A guide to the datasets - the “Description” column details each dataset, while the middle column shows which appendix the tables are in.



Dataset	Appendix for results tables	Description
1. Initial	C	All trades during trading hours: 08:00 - 16:30
2. Deseasonalised	D	Data deseasonalised using time dummies
3. Detrended	E	Data detrended assuming a quadratic time trend
4. ARFIMA	F	Residuals after fitting ARFIMAs to the detrended data

Appendix C: Initial Tests

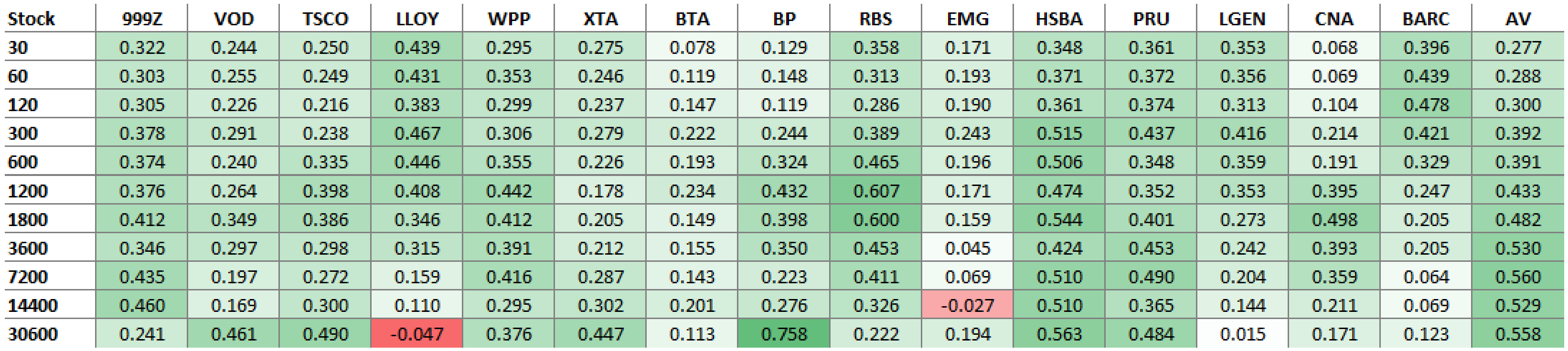
Table C.1: ADF Tests (Initial Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
120	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
300	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1800	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table C.2: Ljung-Box Q Tests with 50 Lags (Initial Dataset)

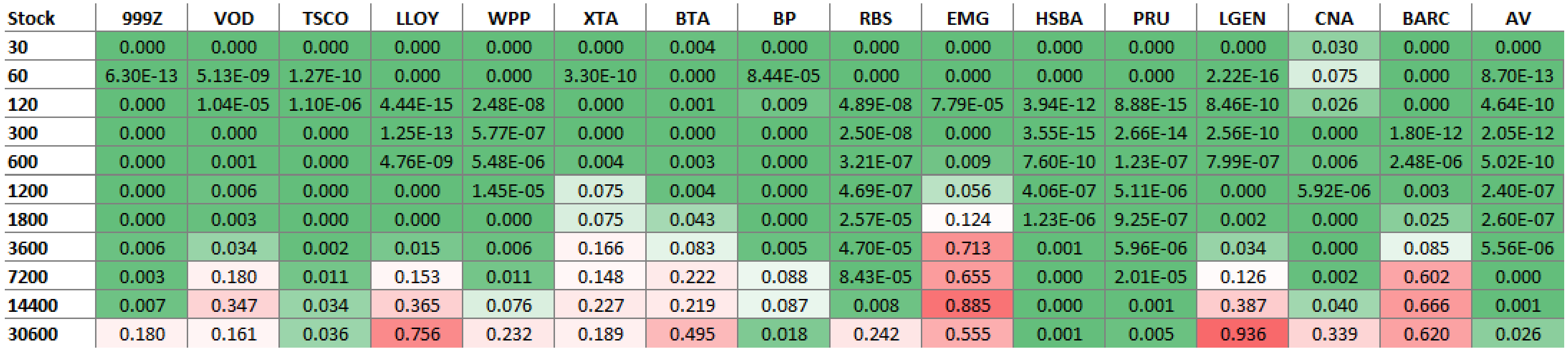
Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
120	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
300	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1800	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	1	1	1	1	1	1	0	1	1	0	1	1	1	0	1	1

Table C.3: GPH Estimates of the Memory Parameter (Initial Dataset)



Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.322	0.244	0.250	0.439	0.295	0.275	0.078	0.129	0.358	0.171	0.348	0.361	0.353	0.068	0.396	0.277
60	0.303	0.255	0.249	0.431	0.353	0.246	0.119	0.148	0.313	0.193	0.371	0.372	0.356	0.069	0.439	0.288
120	0.305	0.226	0.216	0.383	0.299	0.237	0.147	0.119	0.286	0.190	0.361	0.374	0.313	0.104	0.478	0.300
300	0.378	0.291	0.238	0.467	0.306	0.279	0.222	0.244	0.389	0.243	0.515	0.437	0.416	0.214	0.421	0.392
600	0.374	0.240	0.335	0.446	0.355	0.226	0.193	0.324	0.465	0.196	0.506	0.348	0.359	0.191	0.329	0.391
1200	0.376	0.264	0.398	0.408	0.442	0.178	0.234	0.432	0.607	0.171	0.474	0.352	0.353	0.395	0.247	0.433
1800	0.412	0.349	0.386	0.346	0.412	0.205	0.149	0.398	0.600	0.159	0.544	0.401	0.273	0.498	0.205	0.482
3600	0.346	0.297	0.298	0.315	0.391	0.212	0.155	0.350	0.453	0.045	0.424	0.453	0.242	0.393	0.205	0.530
7200	0.435	0.197	0.272	0.159	0.416	0.287	0.143	0.223	0.411	0.069	0.510	0.490	0.204	0.359	0.064	0.560
14400	0.460	0.169	0.300	0.110	0.295	0.302	0.201	0.276	0.326	-0.027	0.510	0.365	0.144	0.211	0.069	0.529
30600	0.241	0.461	0.490	-0.047	0.376	0.447	0.113	0.758	0.222	0.194	0.563	0.484	0.015	0.171	0.123	0.558

Table C.4: p-values for the Estimates (Initial Dataset)



Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.000	0.000
60	6.30E-13	5.13E-09	1.27E-10	0.000	0.000	3.30E-10	0.000	8.44E-05	0.000	0.000	0.000	0.000	2.22E-16	0.075	0.000	8.70E-13
120	0.000	1.04E-05	1.10E-06	4.44E-15	2.48E-08	0.000	0.001	0.009	4.89E-08	7.79E-05	3.94E-12	8.88E-15	8.46E-10	0.026	0.000	4.64E-10
300	0.000	0.000	0.000	1.25E-13	5.77E-07	0.000	0.000	0.000	2.50E-08	0.000	3.55E-15	2.66E-14	2.56E-10	0.000	1.80E-12	2.05E-12
600	0.000	0.001	0.000	4.76E-09	5.48E-06	0.004	0.003	0.000	3.21E-07	0.009	7.60E-10	1.23E-07	7.99E-07	0.006	2.48E-06	5.02E-10
1200	0.000	0.006	0.000	0.000	1.45E-05	0.075	0.004	0.000	4.69E-07	0.056	4.06E-07	5.11E-06	0.000	5.92E-06	0.003	2.40E-07
1800	0.000	0.003	0.000	0.000	0.000	0.075	0.043	0.000	2.57E-05	0.124	1.23E-06	9.25E-07	0.002	0.000	0.025	2.60E-07
3600	0.006	0.034	0.002	0.015	0.006	0.166	0.083	0.005	4.70E-05	0.713	0.001	5.96E-06	0.034	0.000	0.085	5.56E-06
7200	0.003	0.180	0.011	0.153	0.011	0.148	0.222	0.088	8.43E-05	0.655	0.000	2.01E-05	0.126	0.002	0.602	0.000
14400	0.007	0.347	0.034	0.365	0.076	0.227	0.219	0.087	0.008	0.885	0.000	0.001	0.387	0.040	0.666	0.001
30600	0.180	0.161	0.036	0.756	0.232	0.189	0.495	0.018	0.242	0.555	0.001	0.005	0.936	0.339	0.620	0.026

Appendix D: Tests after Deseasonalising

Table D.1: ADF Tests (Deseasonalised Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
120	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
300	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1800	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table D.2: Ljung-Box Q Tests with 50 Lags (Deseasonalised Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
120	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
300	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1200	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
1800	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
3600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	1	1	0	1	1	1	0	0	1	1	1	0	1	0	0	1

Table D.3: GPH Estimates of the Memory Parameter (Deseasonalised Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.328	0.151	0.173	0.395	0.219	0.193	0.025	0.056	0.297	0.078	0.269	0.284	0.320	0.030	0.390	0.174
60	0.317	0.127	0.137	0.371	0.263	0.144	0.050	0.051	0.240	0.082	0.284	0.279	0.305	0.011	0.433	0.154
120	0.330	0.085	0.090	0.322	0.207	0.121	0.057	-0.003	0.202	0.065	0.234	0.279	0.260	0.028	0.477	0.139
300	0.367	0.063	0.027	0.346	0.175	0.065	0.089	0.062	0.244	0.050	0.315	0.225	0.294	0.112	0.391	0.131
600	0.363	-0.050	0.077	0.274	0.209	-0.074	0.022	0.094	0.262	-0.040	0.315	0.083	0.231	0.058	0.303	0.092
1200	0.353	-0.163	0.014	0.187	0.287	-0.208	0.009	0.117	0.343	-0.119	0.215	0.017	0.145	0.225	0.220	0.039
1800	0.398	-0.155	-0.073	0.113	0.246	-0.253	-0.093	0.044	0.272	-0.184	0.231	0.018	0.040	0.275	0.167	0.048
3600	0.325	-0.349	-0.313	0.019	0.289	-0.389	-0.132	-0.097	-0.002	-0.385	-0.045	0.001	-0.039	0.105	0.158	-0.015
7200	0.383	-0.720	-0.464	-0.168	0.006	-0.467	-0.286	-0.316	-0.112	-0.585	-0.012	-0.157	-0.190	-0.026	-0.030	-0.265
14400	0.411	-1.017	-0.623	-0.280	-0.153	-0.651	-0.356	-0.361	-0.290	-0.830	-0.128	-0.378	-0.324	-0.246	-0.096	-0.497
30600	0.183	-1.305	-1.015	-0.565	-0.210	-0.961	-0.770	-0.264	-0.700	-0.724	-0.446	-0.632	-0.651	-0.515	-0.208	-0.962

Table D.4: p-values for the Estimates (Deseasonalised Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.000	0.000	0.000	0.000	0.000	0.000	0.328	0.064	0.000	0.023	0.000	0.000	0.000	0.332	0.000	0.000
60	1.29E-14	0.003	0.000	0.000	0.000	0.000	0.117	0.170	2.57E-08	0.046724	1.84E-10	0.000	6.34E-13	0.77901	0.000	6.16E-05
120	0.000	0.103	0.056	2.05E-11	0.000	0.022	0.168	0.953	0.000	0.197	9.45E-07	1.72E-08	2.35E-07	0.52993	0.000	0.002
300	0.000	0.336	0.667	0.000	0.004	0.367	0.123	0.264	0.001	0.479	0.000	0.000	0.000	0.063	0.000	0.022
600	0.000	0.537	0.392	0.000	0.006	0.335	0.734	0.183	0.006	0.645	0.000	0.209	0.003	0.417	0.000	0.189
1200	0.000	0.129	0.881	0.041	0.004	0.033	0.917	0.201	0.008	0.280	0.028	0.829	0.107	0.013	0.006	0.676
1800	0.000	0.243	0.473	0.276	0.026	0.024	0.265	0.686	0.071	0.149	0.055	0.837	0.673	0.011	0.061	0.672
3600	0.007	0.048	0.005	0.888	0.069	0.006	0.206	0.440	0.986	0.014	0.689	0.993	0.761	0.346	0.173	0.920
7200	0.005	0.000	0.001	0.228	0.966	0.008	0.033	0.020	0.454	0.001	0.930	0.231	0.223	0.831	0.763	0.103
14400	0.013	0.000	0.000	0.065	0.314	0.003	0.049	0.036	0.146	0.000	0.430	0.009	0.113	0.037	0.358	0.009
30600	0.320	0.001	0.000	0.029	0.435	0.003	0.002	0.421	0.023	0.033	0.013	0.004	0.020	0.015	0.160	0.001

Appendix E: Tests after Deseasonalising, then Detrending

Table E.1: ADF Tests (Deseasonalised, then Detrended Dataset Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
120	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
300	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1800	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table E.2: Ljung-Box Q Tests with 50 Lags (Deseasonalised, then Detrended Dataset Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
120	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
300	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1200	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
1800	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1
3600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	0	1	0	1	1	1	0	0	1	1	1	0	1	0	0	1

Table E.3: GPH Estimates of the Memory Parameter (Deseasonalised, then Detrended Dataset Dataset)

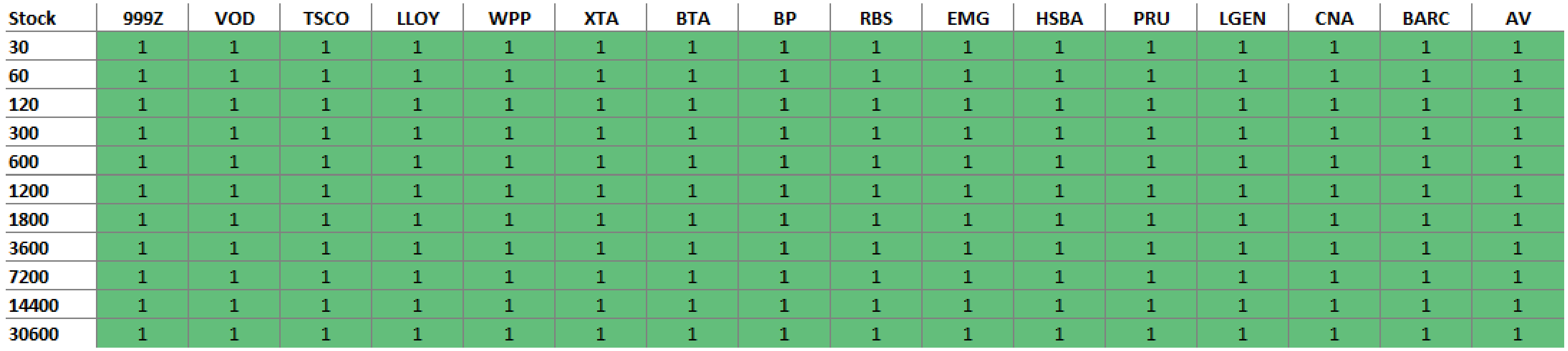
Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.302	0.156	0.174	0.396	0.216	0.194	0.025	0.057	0.303	0.064	0.272	0.283	0.327	0.030	0.389	0.163
60	0.286	0.134	0.137	0.372	0.260	0.146	0.050	0.051	0.249	0.064	0.288	0.278	0.314	0.011	0.431	0.140
120	0.200	0.123	0.123	0.329	0.203	0.127	0.041	-0.017	0.212	0.054	0.234	0.259	0.272	0.008	0.455	0.137
300	0.182	0.119	0.075	0.356	0.169	0.076	0.065	0.040	0.258	0.033	0.316	0.194	0.311	0.082	0.358	0.128
600	0.122	0.026	0.142	0.287	0.201	-0.060	-0.010	0.065	0.282	-0.062	0.316	0.042	0.254	0.017	0.259	0.090
1200	0.033	-0.058	0.100	0.204	0.276	-0.188	-0.034	0.080	0.369	-0.148	0.217	-0.038	0.174	0.171	0.162	0.037
1800	0.017	-0.033	0.028	0.134	0.233	-0.231	-0.144	-0.002	0.303	-0.219	0.235	-0.048	0.077	0.210	0.097	0.046
3600	-0.131	-0.162	-0.179	0.045	0.266	-0.361	-0.205	-0.153	0.045	-0.437	-0.040	-0.084	0.008	0.015	0.065	-0.016
7200	-0.195	-0.466	-0.304	-0.139	-0.017	-0.419	-0.372	-0.391	-0.037	-0.629	-0.001	-0.282	-0.127	-0.132	-0.151	-0.275
14400	-0.337	-0.659	-0.433	-0.251	-0.169	-0.569	-0.453	-0.464	-0.166	-0.845	-0.106	-0.577	-0.234	-0.368	-0.260	-0.521
30600	-0.911	-0.790	-0.670	-0.501	-0.265	-0.887	-0.950	-0.409	-0.611	-0.847	-0.424	-0.865	-0.516	-0.731	-0.437	-0.959

Table E.4: p-values for the Estimates (Deseasonalised, then Detrended Dataset Dataset)

Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.000	0.000	0.000	0.000	0.000	0.000	0.331	0.061	0.000	0.066	0.000	0.000	0.000	0.331	0.000	0.000
60	9.56E-12	0.002	0.000	0.000	1.57E-10	0.000	0.120	0.163	5.77E-09	0.131	8.85E-11	0.000	9.46E-14	0.776	0.000	0.000
120	0.000	0.014	0.007	5.71E-12	2.99E-05	0.015	0.32978	0.701	4.03E-05	0.288	8.5E-07	2.83E-07	5.2E-08	0.866	0.000	0.002
300	0.004	0.052	0.209	0.000	0.006	0.288	0.269	0.472	0.000	0.639	0.000	0.001	0.000	0.184	0.000	0.024
600	0.147	0.729	0.097	0.000	0.008	0.430	0.886	0.365	0.003	0.486	0.000	0.547	0.001	0.816	0.000	0.199
1200	0.746	0.541	0.234	0.023	0.006	0.048	0.698	0.398	0.004	0.190	0.025	0.654	0.046	0.070	0.050	0.694
1800	0.889	0.782	0.749	0.184	0.036	0.035	0.105	0.987	0.041	0.094	0.050	0.615	0.409	0.065	0.292	0.687
3600	0.378	0.280	0.034	0.729	0.100	0.008	0.072	0.241	0.695	0.007	0.721	0.513	0.945	0.899	0.593	0.912
7200	0.281	0.001	0.003	0.301	0.904	0.013	0.011	0.006	0.780	0.001	0.994	0.058	0.397	0.336	0.165	0.093
14400	0.158	0.000	0.001	0.081	0.274	0.005	0.020	0.012	0.327	0.000	0.502	0.001	0.236	0.009	0.036	0.008
30600	0.002	0.006	0.000	0.038	0.342	0.003	0.000	0.233	0.030	0.018	0.015	0.001	0.048	0.003	0.015	0.001

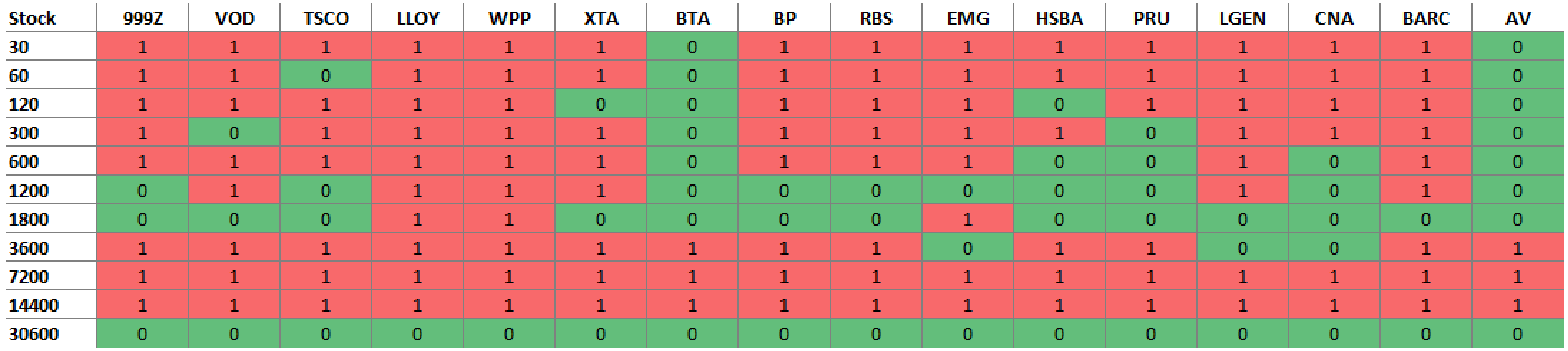
Appendix F: Tests after Fitting ARFIMAs

Table F.1: ADF Tests (ARFIMA Dataset)



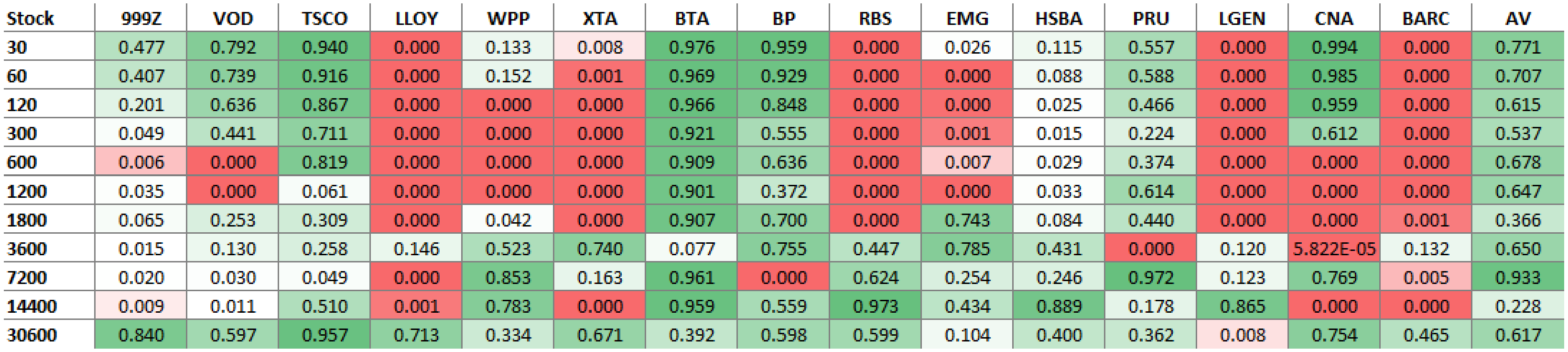
Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
60	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
120	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
300	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1800	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table F.2: Ljung-Box Q Tests with 50 Lags (ARFIMA Dataset)



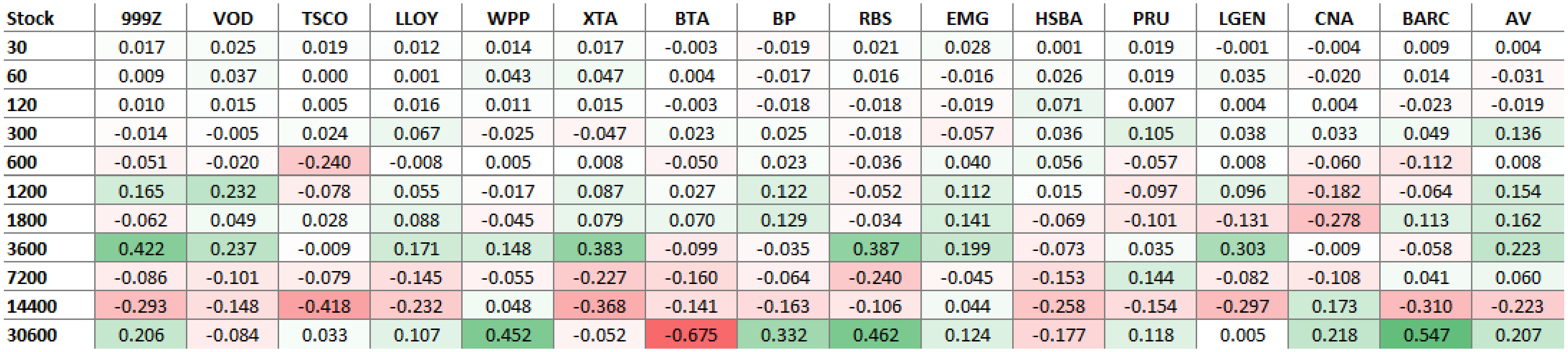
Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0
60	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	0
120	1	1	1	1	1	0	0	1	1	1	0	1	1	1	1	0
300	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	0
600	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	0
1200	0	1	0	1	1	1	0	0	0	0	0	0	1	0	1	0
1800	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
3600	1	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1
7200	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
14400	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
30600	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table F.3: ARCH(1) Tests (ARFIMA Dataset)



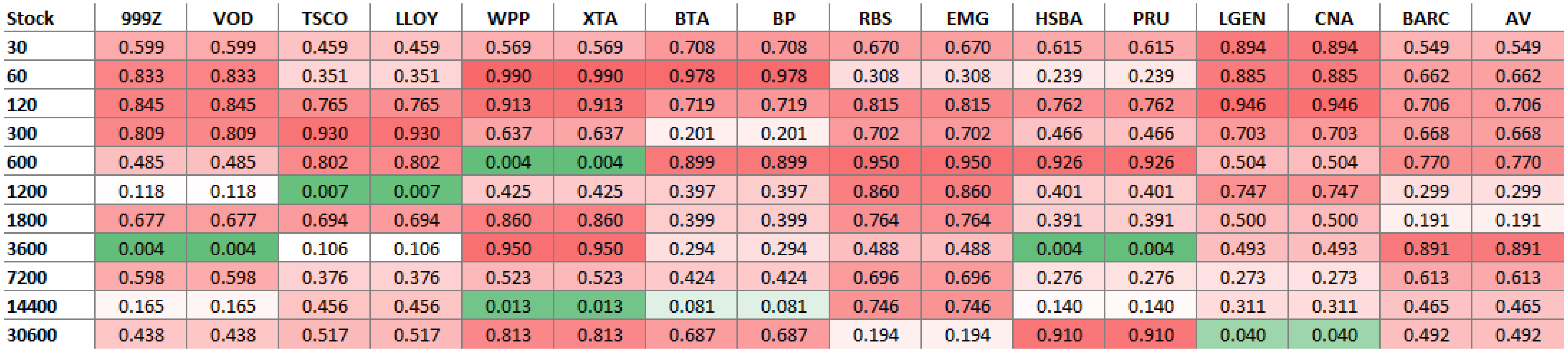
Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.477	0.792	0.940	0.000	0.133	0.008	0.976	0.959	0.000	0.026	0.115	0.557	0.000	0.994	0.000	0.771
60	0.407	0.739	0.916	0.000	0.152	0.001	0.969	0.929	0.000	0.000	0.088	0.588	0.000	0.985	0.000	0.707
120	0.201	0.636	0.867	0.000	0.000	0.000	0.966	0.848	0.000	0.000	0.025	0.466	0.000	0.959	0.000	0.615
300	0.049	0.441	0.711	0.000	0.000	0.000	0.921	0.555	0.000	0.001	0.015	0.224	0.000	0.612	0.000	0.537
600	0.006	0.000	0.819	0.000	0.000	0.000	0.909	0.636	0.000	0.007	0.029	0.374	0.000	0.000	0.000	0.678
1200	0.035	0.000	0.061	0.000	0.000	0.000	0.901	0.372	0.000	0.000	0.033	0.614	0.000	0.000	0.000	0.647
1800	0.065	0.253	0.309	0.000	0.042	0.000	0.907	0.700	0.000	0.743	0.084	0.440	0.000	0.000	0.001	0.366
3600	0.015	0.130	0.258	0.146	0.523	0.740	0.077	0.755	0.447	0.785	0.431	0.000	0.120	5.822E-05	0.132	0.650
7200	0.020	0.030	0.049	0.000	0.853	0.163	0.961	0.000	0.624	0.254	0.246	0.972	0.123	0.769	0.005	0.933
14400	0.009	0.011	0.510	0.001	0.783	0.000	0.959	0.559	0.973	0.434	0.889	0.178	0.865	0.000	0.000	0.228
30600	0.840	0.597	0.957	0.713	0.334	0.671	0.392	0.598	0.599	0.104	0.400	0.362	0.008	0.754	0.465	0.617

Table F.4: GPH Estimates of the Memory Parameter (ARFIMA Dataset)



Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.017	0.025	0.019	0.012	0.014	0.017	-0.003	-0.019	0.021	0.028	0.001	0.019	-0.001	-0.004	0.009	0.004
60	0.009	0.037	0.000	0.001	0.043	0.047	0.004	-0.017	0.016	-0.016	0.026	0.019	0.035	-0.020	0.014	-0.031
120	0.010	0.015	0.005	0.016	0.011	0.015	-0.003	-0.018	-0.018	-0.019	0.071	0.007	0.004	0.004	-0.023	-0.019
300	-0.014	-0.005	0.024	0.067	-0.025	-0.047	0.023	0.025	-0.018	-0.057	0.036	0.105	0.038	0.033	0.049	0.136
600	-0.051	-0.020	-0.240	-0.008	0.005	0.008	-0.050	0.023	-0.036	0.040	0.056	-0.057	0.008	-0.060	-0.112	0.008
1200	0.165	0.232	-0.078	0.055	-0.017	0.087	0.027	0.122	-0.052	0.112	0.015	-0.097	0.096	-0.182	-0.064	0.154
1800	-0.062	0.049	0.028	0.088	-0.045	0.079	0.070	0.129	-0.034	0.141	-0.069	-0.101	-0.131	-0.278	0.113	0.162
3600	0.422	0.237	-0.009	0.171	0.148	0.383	-0.099	-0.035	0.387	0.199	-0.073	0.035	0.303	-0.009	-0.058	0.223
7200	-0.086	-0.101	-0.079	-0.145	-0.055	-0.227	-0.160	-0.064	-0.240	-0.045	-0.153	0.144	-0.082	-0.108	0.041	0.060
14400	-0.293	-0.148	-0.418	-0.232	0.048	-0.368	-0.141	-0.163	-0.106	0.044	-0.258	-0.154	-0.297	0.173	-0.310	-0.223
30600	0.206	-0.084	0.033	0.107	0.452	-0.052	-0.675	0.332	0.462	0.124	-0.177	0.118	0.005	0.218	0.547	0.207

Table F.5: p-values for the Estimates (ARFIMA Dataset)



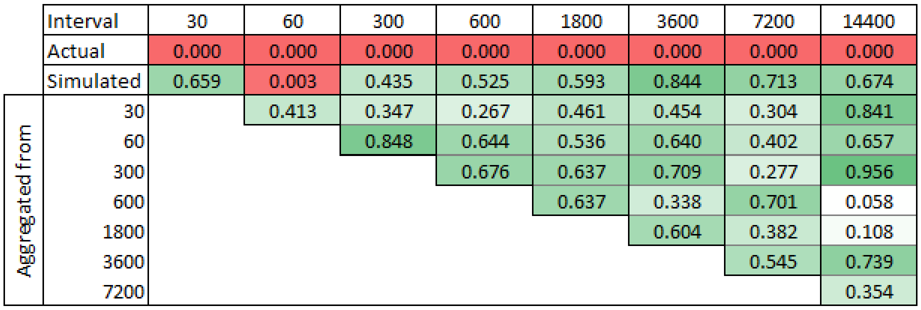
Stock	999Z	VOD	TSCO	LLOY	WPP	XTA	BTA	BP	RBS	EMG	HSBA	PRU	LGEN	CNA	BARC	AV
30	0.599	0.599	0.459	0.459	0.569	0.569	0.708	0.708	0.670	0.670	0.615	0.615	0.894	0.894	0.549	0.549
60	0.833	0.833	0.351	0.351	0.990	0.990	0.978	0.978	0.308	0.308	0.239	0.239	0.885	0.885	0.662	0.662
120	0.845	0.845	0.765	0.765	0.913	0.913	0.719	0.719	0.815	0.815	0.762	0.762	0.946	0.946	0.706	0.706
300	0.809	0.809	0.930	0.930	0.637	0.637	0.201	0.201	0.702	0.702	0.466	0.466	0.703	0.703	0.668	0.668
600	0.485	0.485	0.802	0.802	0.004	0.004	0.899	0.899	0.950	0.950	0.926	0.926	0.504	0.504	0.770	0.770
1200	0.118	0.118	0.007	0.007	0.425	0.425	0.397	0.397	0.860	0.860	0.401	0.401	0.747	0.747	0.299	0.299
1800	0.677	0.677	0.694	0.694	0.860	0.860	0.399	0.399	0.764	0.764	0.391	0.391	0.500	0.500	0.191	0.191
3600	0.004	0.004	0.106	0.106	0.950	0.950	0.294	0.294	0.488	0.488	0.004	0.004	0.493	0.493	0.891	0.891
7200	0.598	0.598	0.376	0.376	0.523	0.523	0.424	0.424	0.696	0.696	0.276	0.276	0.273	0.273	0.613	0.613
14400	0.165	0.165	0.456	0.456	0.013	0.013	0.081	0.081	0.746	0.746	0.140	0.140	0.311	0.311	0.465	0.465
30600	0.438	0.438	0.517	0.517	0.813	0.813	0.687	0.687	0.194	0.194	0.910	0.910	0.040	0.040	0.492	0.492

Appendix G: Tests for Simulations

G.1 Initial Data

Table G.1: Vodafone Ljung-Box Q Tests with 50 Lags (Initial Dataset)

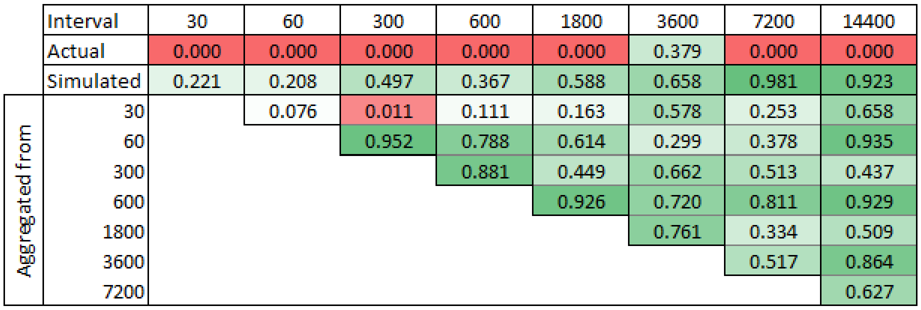
p-values of Ljung-Box Q tests with lag order upto 50 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of no serial correlation, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Simulated	0.659	0.003	0.435	0.525	0.593	0.844	0.713	0.674
Aggregated from	30		0.413	0.347	0.267	0.461	0.454	0.304	0.841
	60			0.848	0.644	0.536	0.640	0.402	0.657
	300				0.676	0.637	0.709	0.277	0.956
	600					0.637	0.338	0.701	0.058
	1800						0.604	0.382	0.108
	3600							0.545	0.739
	7200								0.354

Table G.2: Legal and General Ljung-Box Q Tests with 50 Lags (Initial Dataset)

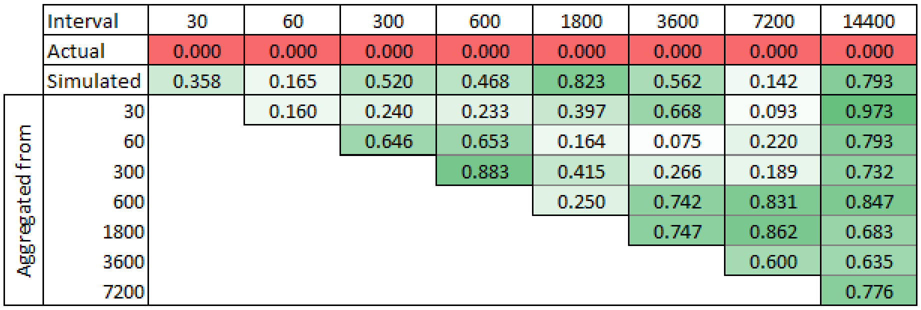
p-values of Ljung-Box Q tests with lag order upto 50 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of no serial correlation, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.000	0.379	0.000	0.000
	Simulated	0.221	0.208	0.497	0.367	0.588	0.658	0.981	0.923
Aggregated from	30		0.076	0.011	0.111	0.163	0.578	0.253	0.658
	60			0.952	0.788	0.614	0.299	0.378	0.935
	300				0.881	0.449	0.662	0.513	0.437
	600					0.926	0.720	0.811	0.929
	1800						0.761	0.334	0.509
	3600							0.517	0.864
	7200								0.627

Table G.3: WPP Ljung-Box Q Tests with 50 Lags (Initial Dataset)

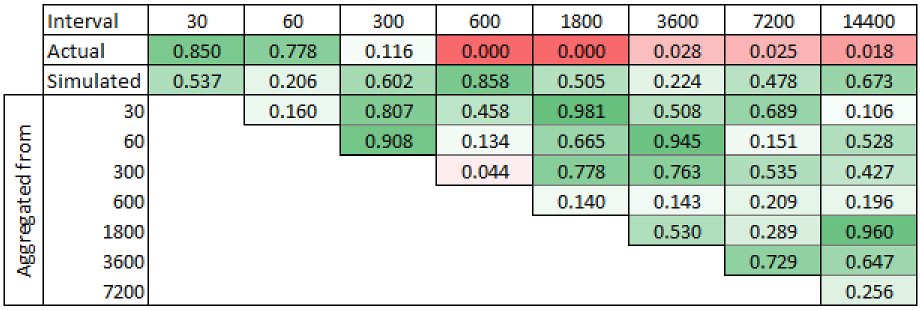
p-values of Ljung-Box Q tests with lag order upto 50 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of no serial correlation, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Simulated	0.358	0.165	0.520	0.468	0.823	0.562	0.142	0.793
Aggregated from	30		0.160	0.240	0.233	0.397	0.668	0.093	0.973
	60			0.646	0.653	0.164	0.075	0.220	0.793
	300				0.883	0.415	0.266	0.189	0.732
	600					0.250	0.742	0.831	0.847
	1800						0.747	0.862	0.683
	3600							0.600	0.635
	7200								0.776

Table G.4: Vodafone ARCH(1) Tests (Initial Dataset)

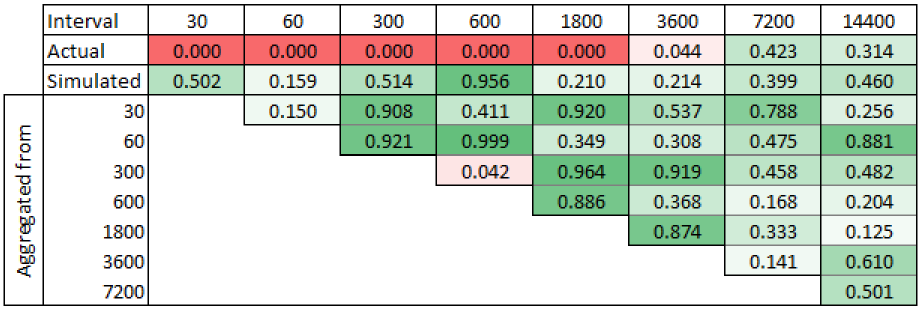
p-values of ARCH tests of order 1 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of IID Gaussian random variables, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.850	0.778	0.116	0.000	0.000	0.028	0.025	0.018
	Simulated	0.537	0.206	0.602	0.858	0.505	0.224	0.478	0.673
Aggregated from	30		0.160	0.807	0.458	0.981	0.508	0.689	0.106
	60			0.908	0.134	0.665	0.945	0.151	0.528
	300				0.044	0.778	0.763	0.535	0.427
	600					0.140	0.143	0.209	0.196
	1800						0.530	0.289	0.960
	3600							0.729	0.647
	7200								0.256

Table G.5: Legal and General ARCH(1) Tests (Initial Dataset)

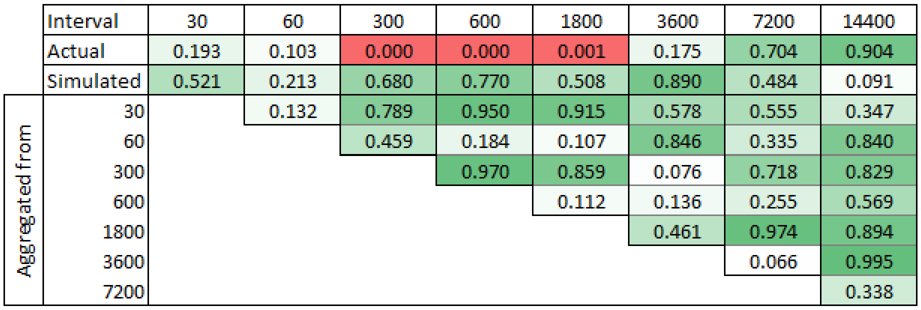
p-values of ARCH tests of order 1 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of IID Gaussian random variables, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.000	0.044	0.423	0.314
	Simulated	0.502	0.159	0.514	0.956	0.210	0.214	0.399	0.460
Aggregated from	30		0.150	0.908	0.411	0.920	0.537	0.788	0.256
	60			0.921	0.999	0.349	0.308	0.475	0.881
	300				0.042	0.964	0.919	0.458	0.482
	600					0.886	0.368	0.168	0.204
	1800						0.874	0.333	0.125
	3600							0.141	0.610
	7200								0.501

Table G.6: WPP ARCH(1) Tests (Initial Dataset)

p-values of ARCH tests of order 1 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of IID Gaussian random variables, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.193	0.103	0.000	0.000	0.001	0.175	0.704	0.904
	Simulated	0.521	0.213	0.680	0.770	0.508	0.890	0.484	0.091
Aggregated from	30		0.132	0.789	0.950	0.915	0.578	0.555	0.347
	60			0.459	0.184	0.107	0.846	0.335	0.840
	300				0.970	0.859	0.076	0.718	0.829
	600					0.112	0.136	0.255	0.569
	1800						0.461	0.974	0.894
	3600							0.066	0.995
	7200								0.338

G.2 Deseasonalised Data

Table G.7: Vodafone Ljung-Box Q Tests with 50 Lags (Deseasonalised Dataset)

p-values of Ljung-Box Q tests with lag order upto 50 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of no serial correlation, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.

	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.176	0.000	0.000	0.000
	Simulated	0.270	0.258	0.333	0.380	0.508	0.862	0.703	0.019
Aggregated from	30		0.319	0.344	0.085	0.965	0.775	0.789	0.850
	60			0.688	0.823	0.968	0.295	0.672	0.884
	300				0.817	0.810	0.881	0.445	0.846
	600					0.693	0.506	0.669	0.061
	1800						0.601	0.320	0.145
	3600							0.409	0.543
	7200								0.372

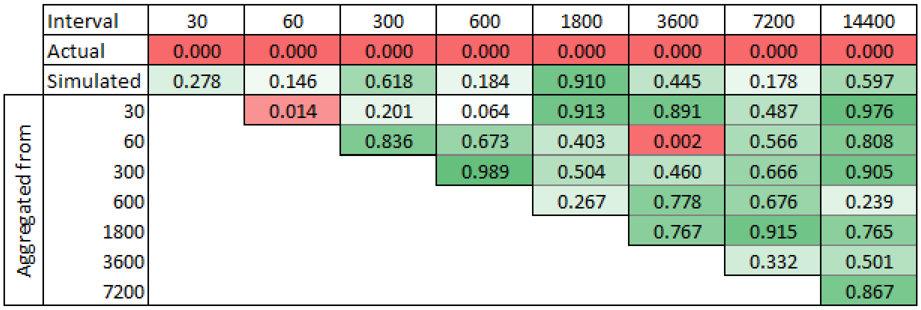
Table G.8: Legal and General Ljung-Box Q Tests with 50 Lags (Deseasonalised Dataset)

p-values of Ljung-Box Q tests with lag order upto 50 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of no serial correlation, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.

	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.136	0.563	0.000	0.000
	Simulated	0.076	0.171	0.565	0.443	0.656	0.639	0.960	0.925
Aggregated from	30		0.012	0.115	0.064	0.844	0.955	0.804	0.863
	60			0.907	0.643	0.575	0.162	0.282	0.951
	300				0.894	0.453	0.894	0.720	0.506
	600					0.908	0.725	0.858	0.890
	1800						0.880	0.413	0.598
	3600							0.592	0.290
	7200								0.689

Table G.9: WPP Ljung-Box Q Tests with 50 Lags (Deseasonalised Dataset)

p-values of Ljung-Box Q tests with lag order upto 50 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of no serial correlation, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Simulated	0.278	0.146	0.618	0.184	0.910	0.445	0.178	0.597
Aggregated from	30		0.014	0.201	0.064	0.913	0.891	0.487	0.976
	60			0.836	0.673	0.403	0.002	0.566	0.808
	300				0.989	0.504	0.460	0.666	0.905
	600					0.267	0.778	0.676	0.239
	1800						0.767	0.915	0.765
	3600							0.332	0.501
	7200								0.867

Table G.10: Vodafone ARCH(1) Tests (Deseasonalised Dataset)

p-values of ARCH tests of order 1 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of IID Gaussian random variables, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.

	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.790	0.715	0.395	0.000	0.232	0.174	0.021	0.039
	Simulated	0.479	0.159	0.604	0.847	0.500	0.328	0.706	0.696
Aggregated from	30		0.158	0.913	0.232	0.934	0.800	0.373	0.127
	60			0.909	0.237	0.424	0.744	0.384	0.960
	300				0.058	0.701	0.746	0.656	0.333
	600					0.208	0.119	0.297	0.106
	1800						0.588	0.628	0.807
	3600							0.552	0.353
	7200								0.412

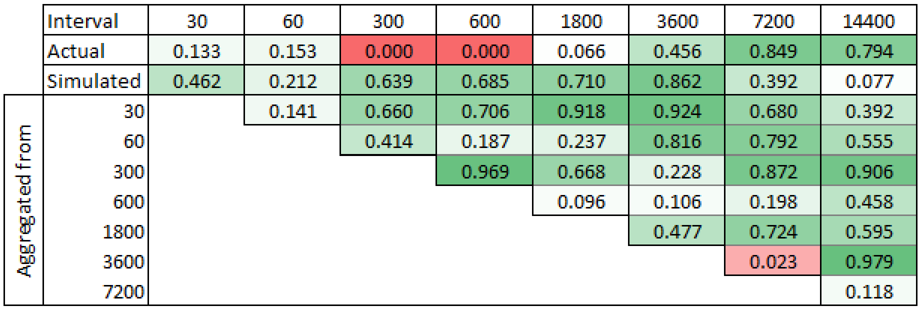
Table G.11: Legal and General ARCH(1) Tests (Deseasonalised Dataset)

p-values of ARCH tests of order 1 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of IID Gaussian random variables, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.

	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.000	0.000	0.000	0.000	0.000	0.131	0.114	0.945
	Simulated	0.428	0.183	0.505	0.724	0.243	0.269	0.503	0.033
Aggregated from	30		0.161	0.971	0.240	0.808	0.573	0.352	0.098
	60			0.956	0.859	0.502	0.354	0.772	0.359
	300				0.036	0.900	0.774	0.394	0.639
	600					0.984	0.374	0.195	0.239
	1800						0.755	0.564	0.136
	3600							0.000	0.000
	7200								0.147

Table G.12: WPP ARCH(1) Tests (Deseasonalised Dataset)

p-values of ARCH tests of order 1 for actual and simulated data, and simulated data aggregated from smaller sampling intervals. Colour scale: red for a p-value of 0, rejecting the null hypothesis of IID Gaussian random variables, white for a p-value of 0.05, and green for a p-value of 1, retaining the null hypothesis.



	Interval	30	60	300	600	1800	3600	7200	14400
	Actual	0.133	0.153	0.000	0.000	0.066	0.456	0.849	0.794
	Simulated	0.462	0.212	0.639	0.685	0.710	0.862	0.392	0.077
Aggregated from	30		0.141	0.660	0.706	0.918	0.924	0.680	0.392
	60			0.414	0.187	0.237	0.816	0.792	0.555
	300				0.969	0.668	0.228	0.872	0.906
	600					0.096	0.106	0.198	0.458
	1800						0.477	0.724	0.595
	3600							0.023	0.979
	7200								0.118

Chapter 4: Long Memory and Cointegration in High-Frequency Prices and Volumes

4.1 Introduction

As shown empirically in Chapter 2, the LMSD model is arguably the best model for forecasting durations in the presence of high order serial correlation, while Chapter 3 showed that volumes do exhibit long memory at a range of high frequencies (upto daily). This chapter builds on theoretical work in Deo, Hurvich, Soulier and Wang (2009), Deo, Hsieh and Hurvich (2010), and Hurvich and Wang (2010) to develop a baseline model which connects prices and volumes through irregular time. The analysis makes further use of the theory of point processes to make four propositions on the behaviour of prices and volumes, as well as a statistical lemma on dependence between point processes. A general introduction to the theory of point processes is given by Daley and Vere-Jones (2003) and by Cox and Isham (1980).

The analysis extends consideration to 2 assets. The first section of this chapter highlights some properties of the LMSD process and a linear combination of prices based on the LMSD process which features cointegration (we term this combination a "cointegrated system"), as derived by Hurvich and Wang (2010). These will be exploited in the following sections in which we make the following propositions on prices and volumes:

1. (Prices) Long Memory exists in the squared returns and (hence) in realized volatility for the fractionally cointegrated system of Hurvich and Wang.

2. (Volumes) Volume from a similar system is at least I(1) and at most just less than I(2).

3. (Volumes) Volume from a similar system can feature cointegration if the microstructure component has intermediate memory.

4. (Prices and volumes) Positive correlation at tick level between the series generates positive unconditional correlation between returns and volumes, and positive time-varying conditional covariance.

For ease of reference, let us label and abbreviate the variable specifications we detail below as follows:

$$\begin{aligned} \text{TM\#} &= \text{Time Model \#} \\ \text{PM\#} &= \text{Price Model \#} \\ \text{VM\#} &= \text{Volume Model \#} \\ \text{JM\#} &= \text{Price-Volume Model \#} \end{aligned}$$

4.2 Times

To restate, the LMSD process is:

TM1:

$$x_{m,t} = e^{\psi_{m,t}}\varepsilon_{m,t} \qquad \varepsilon_{m,t} \sim IID(1, \sigma_{m,\varepsilon^2}) \text{ with all moments finite}$$

$$\psi_{m,t} = \sum_{i=0}^{\infty} b_{m,i}\upsilon_{m,t-i} \qquad \upsilon_{m,t} \sim NID(0, \sigma^2_{m,\upsilon}) \qquad b_{m,i} \sim Ci^{d_{x_m}-1}$$

where d_{x_m} is the memory parameter for the asset m durations process; different memory parameters may exist in the price and volume processes.

This generates a count process, $N(t)$, with the same memory parameter d_x, as shown by Theorem 2 in Deo, Hurvich, Soulier and Wang (2009). Let μ be the unconditional mean of the durations. Then $\lambda = \frac{1}{\mu}$ represents the intensity or instantaneous expected number of counts, so that the unconditional mean number of counts in a given time interval Δt is $\lambda\Delta t$.

If the nth moment of durations generated by the LMSD model is finite, then by Lemmas 3 and 4 in Hurvich and Wang, the nth moment of $N(t)$ is bounded by $[K_n(\Delta t)]^n$ where K_n is a constant. In Proposition 1 of this chapter, the highest order moment we need is 8, while in Propositions 2 and 3, the highest order we need is 2.

4.3 Prices

We wish to analyse the dynamics of prices over irregular time. Oomen (2006) follows Andersen, Bollerslev, Diebold and Labys (2003) in representing the logarithmic price process as follows:

PM1: $\ln P_t = A_t + C_t + D_t$ where A is a finite variation predictable mean component,

C and D are local martingales;

C is a continuous sample-path process,

D is a compensated pure jump process.

Whereas previous authors, such as Andersen et al., have focussed on C and assumed $D = 0$ in analyzing realized variance, Oomen takes the opposite route by focussing on D and assuming $C = 0$. This enables analysis of prices using irregularly-spaced events. Oomen also argues that C is the limit of D as the time between jumps tends to zero.

In particular, Oomen adapts the Compound Poisson Process (CPP) of Press (1967) to form what he terms a CPP-MA(q) model. Then:

PM2: $\ln P_t = \ln P_0 + \sum_{i=1}^{N(t)} (e_i + \eta_i)$ where $\eta_i = \Delta v_{i-1} + \rho_2 \Delta v_{i-2} + ... + \rho_q \Delta v_{i-q+1}$,

$v_i \sim NID(0, \sigma_v^2)$ and $e_i \sim NID(\mu_e, \sigma_e^2)$,

$\{v_i\}$ and $\{e_i\}$ are mutually independent,

$N(t)$ is a Poisson process independent of $\{e_i\}$ and $\{\eta_i\}$

Overall, log prices are a Poisson sum of efficient prices shocks e_i and microstructure noise η_i. Without the microstructure noise component, the log price process would be a martingale.

Deo, Hurvich, Soulier and Wang (2009) extended this model by respecifying the counting process as being generated by a LMSD process in the durations, thereby incorporating long memory in time. They also weakened the assumptions on the disturbances:

PM3: $\ln P_t = \ln P_0 + \sum_{i=1}^{N(t)} (e_i + \eta_i)$ where $e_i \sim IID(0, \sigma_e^2)$ and $\eta_i \sim IID(0, \sigma_\eta^2)$
both with finite eighth moment,
$\{e_i\}$ and $\{\eta_i\}$ are mutually independent,
$N(t)$ is generated by a LMSD process
independent of $\{e_i\}$ and $\{\eta_i\}$

By adopting this specification, Deo, Hurvich, Soulier and Wang are able to show (in their Theorem 5) that long memory propagates from the durations process via the counting process to squared returns. They also remark that the realized volatility series constructed from squared returns over any fixed interval Δt will inherit long memory with the same memory parameter.

Hurvich and Wang (2010) extended this model further to allow for cointegration between different assets' price series even if their trades are asynchronous; if both assets have some degree of memory, there is the possibility that they share a (possibly fractional) cointegrating relationship if there is reason to believe that there is an underlying economic relationship between prices. E.g. Hurvich and Wang suggest that spot and future prices may feature such a relationship.

Specifically, they specify processes for the log prices of two assets, denoted with subscript $m \in \{1, 2\}$, which have feedback effects between them:

PM4:

$$\ln P_{1,t} = \sum_{i=1}^{N_1(t)} (e_{1,i} + \eta_{1,i}) + \sum_{i=1}^{N_2(t_1,N_1(t))} (\theta e_{2,i} + g_{21}\eta_{2,i}) \qquad e_{m,i} \sim NID(0, \sigma^2_{m,e})$$

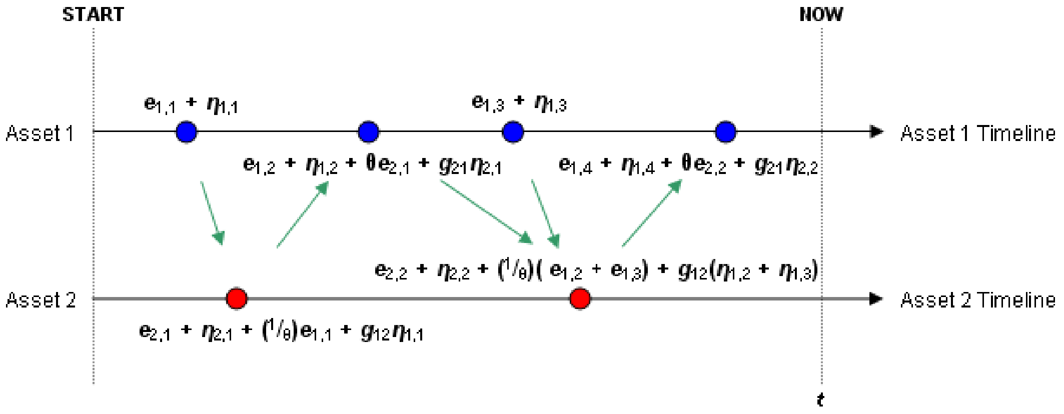
$$\ln P_{2,t} = \sum_{i=1}^{N_2(t)} (e_{2,i} + \eta_{2,i}) + \sum_{i=1}^{N_1(t_2,N_2(t))} (\frac{1}{\theta} e_{1,i} + g_{12}\eta_{1,i}) \qquad \eta_{m,i} \sim FGN : d_{\eta_m} \in (-\frac{1}{2}, 0)$$

where FGN denotes Fractional Gaussian Noise, the increment of a Fractional Brownian Motion. Then $\eta_{m,i}$ has mean zero, variance $\sigma^2_{m,\eta}$ and covariance $\gamma_m(\tau) = \frac{\sigma^2_{m,\eta}}{2}[(\tau+1)^{2d_{\eta_m}+1} - (2\tau)^{2d_{\eta_m}+1} + (\tau-1)^{2d_{\eta_m}+1}]$. More detail is available in Beran (1994).

Note that two counting processes exist now, $N_1(t)$ and $N_2(t)$, corresponding to the events for both assets. In both cases, the second summation term represents weighted feedback from one asset to the other, and incorporates the other counting process. E.g. for asset 1, the feedback from asset 2 is a sum of weighted shocks over asset 2's timeline which occur before the last trade which occurs for asset 1: this is the meaning of $N_2(t_1, N_1(t))$.

The arrows in Figure 4.1 below show this feedback of shocks from one asset to the other more clearly. The first event on asset 1's timeline is the first event for both assets so incorporates no feedback. The first event on asset 2's timeline occurs after this though, so contains feedback from asset 1 as $\frac{1}{\theta}e_{1,1} + g_{12}\eta_{1,1}$. Similarly, the second event for asset 1 contains feedback from the first event for asset 2. However, the third event for asset 2 contains no feedback since there are no intervening asset 2 events before it, while the second event for asset 2 has two intervening asset 1 events so contains two sets of feedback shocks from asset 1.

Figure 4.1: Feedback between Assets


START
NOW
$e_{1,1} + \eta_{1,1}$
$e_{1,3} + \eta_{1,3}$
Asset 1
Asset 1 Timeline
$e_{1,2} + \eta_{1,2} + \theta e_{2,1} + g_{21}\eta_{2,1}$
$e_{1,4} + \eta_{1,4} + \theta e_{2,2} + g_{21}\eta_{2,2}$
$e_{2,2} + \eta_{2,2} + (^1/_\theta)(e_{1,2} + e_{1,3}) + g_{12}(\eta_{1,2} + \eta_{1,3})$
Asset 2
Asset 2 Timeline
$e_{2,1} + \eta_{2,1} + (^1/_\theta)e_{1,1} + g_{12}\eta_{1,1}$
t

Attractively, Hurvich and Wang show that these log price processes act as martingales in the long run / for large Δt (in Theorem 1), while returns formed from them display decreasing correlation at all orders as the sampling interval increases (Theorem 2) - note that the latter point enables correlation at high frequency, which fades to zero for large Δt. The martingale property enables prices to be trended as per SF6.1 (prices are I(1)).

Also note that the η_j terms have been modified to a Fractional Gaussian Noise process with intermediate memory. This term mimics the MA process previously used while enabling fractional cointegration. By forming the cointegrating relationship $\ln P_{1,t} - \theta \ln P_{2,t}$, Hurvich and Wang show by Theorem 3 that the combination reduces in memory from 1 to $1 + \max(d_{\eta_1}, d_{\eta_2})$.

Hurvich and Wang have not yet shown that PM4 yields long memory in the squared returns as before (for PM3). For completeness, I have attempted to prove this by extending the logic of Theorem 5 in Deo, Hurvich, Soulier and Wang (2009). This is done by 1) incorporating PM4's Fractional Gaussian Noise term instead of PM3's MA term, and 2) incorporating the feedback effects from another asset.

Proposition 1: Long memory exists in the squared returns of PM4, with a common long memory parameter across both assets.

Proof: See Section A.1 (Appendix).

4.4 Volumes

In this subsection we wish to add volume to the price-time framework we already have. Tauchen & Pitts (1983) suggested a joint model for daily prices and volumes along these lines. They specified the process as:

JM1:

$$\Delta P_t = \sum_{i=1}^{N(t)} \Delta P_i, \qquad \Delta P_i \sim N(0, \sigma_1^2)$$

$$V = \sum_{i=1}^{N(t)} V_i, \qquad V_i \sim N(\mu_2, \sigma_2^2)$$

So the daily price changes and volumes are mixtures of independent normal distributions with $N(t)$ as the common mixing variable.

4.4.1 Volumes over time

Let us concentrate on the implications for volume on its own first. If Tauchen and Pitts' specification is extended for random times following a LMSD process and microstructure effects, the following may result:

$$\mathbf{VM1}: \qquad V = \sum_{i=1}^{N(t)} u_i + \nu_i, \qquad u_i \sim N(\mu_2, \sigma_2^2)$$
$$\nu_i \sim FGN : d_\nu \in (-\frac{1}{2}, 0)$$

Overall though, this specification may result in negative volumes, which should not be possible. A better specification may be in terms of logs in volume, then we obtain the following instead:

$$\mathbf{VM2}: \qquad \ln V = \sum_{i=1}^{N(t)} u_i + \nu_i, \qquad u_i \sim N(\mu_2, \sigma_2^2)$$
$$\nu_i \sim FGN : d_\nu \in (-\frac{1}{2}, \frac{1}{2})$$

Note that the u_i term will cause volume to have a trend, which is consistent with Bollerslev and Jubinsky's paper (1999). Also, we wish to promote long memory in volume, so we have modified the microstructure effect to enable the case where the memory parameter $d_\nu \in (-\frac{1}{2}, \frac{1}{2})$. However, long memory in the microstructure shock may cause a dominating stochastic trend, as will be seen in Proposition 2.

We can further extend the model to incorporate feedback from a second asset as in the framework of Hurvich and Wang (2010). However, rather than constraining the feedback coefficients of the efficient volume shocks $\{u_{m,i}\}$ to the values $\{\rho, \frac{1}{\rho}\}$ where ρ is some constant like θ in PM4, let us simply label the coefficients for $\{u_{m,i}\}$ as $\{a_1, a_2\}$, and those for $\{\nu_{m,i}\}$ as $\{b_1, b_2\}$. Then:

VM3:

$$\ln V_{1,t} = \sum_{i=1}^{N_1(t)} (u_{1,i} + \nu_{1,i}) + \sum_{i=1}^{N_2(t_1,N_1(t))} (a_1 u_{2,i} + b_1 \nu_{2,i}) \qquad u_{m,i} \sim NID(0, \sigma_{m,u}^2)$$

$$\ln V_{2,t} = \sum_{i=1}^{N_2(t)} (u_{2,i} + \nu_{2,i}) + \sum_{i=1}^{N_1(t_2,N_2(t))} (a_2 u_{1,i} + b_2 \nu_{1,i}) \qquad \nu_{m,i} \sim FGN : d_{\nu_m} \in (-\frac{1}{2}, \frac{1}{2})$$

As per SF6.2 (volumes are I $\geq$ 1), this specification generates volume as at least I(1) as required. This is shown in the following extension of Theorem 1 in Hurvich and Wang:

Proposition 2: Volume as described by VM3 is at least I(1) and at most just less than I(2).

Proof: See Section A.2, distinguishing two cases: 1) $d_{\nu_m} \in (-\frac{1}{2}, 0)$ and 2) $d_{\nu_m} \in (0, \frac{1}{2})$.

For Case 1, we can show how fractional cointegration might occur in the volumes by applying Theorem 3 in Hurvich and Wang. In Case 2 though, even with relatively weaker conditions, we cannot show fractional cointegration yet.

Proposition 3: Cointegration can exist between volume processes as defined in VM3 if the microstucture component has intermediate memory ($d_{\nu_m} \in (-\frac{1}{2}, 0)$).

Proof: See Section A.3.

4.4.2 Volumes and prices

In this section, we suggest two different statistical relationships arising from the MDH which may exist between prices and volumes, and analyse whether the models suggested in previous sections can accomodate the first. Essentially, integer powers of absolute returns and volumes may 1) have positive contemporaneous correlation, or 2) display positive, asymptotically hyperbolic cross-correlation with increasing lag orders (if we extend the findings in Bollerslev and Jubinsky (1999)).

Concentrating on 1), Karpoff (1987) contains a review of some of the suggested relationships dealing with contemporaneous correlation between various measures of price change and volume. They are summarised as follows:

1. No price-volume correlation exists
2. Positive correlation exists between volume and the absolute change in price
3. A positive or negative correlation exists between volume and the signed change in price
4. Volume is higher when prices increase than when prices decrease

1 is a case which does not necessitate any further analysis of prices and volumes. Karpoff shows that a positive relationship along the lines of 3 may exist because of short-selling constraints, but let us simplify by abstracting from such constraints. Out of 2 and 4, we choose to concentrate on 2, since this is the relationship suggested by the MDH. To generate 2, it is enough to assume positive correlation in the microstructure components of PM4 and VM3.

Proposition 4: Positive correlation between the microstructure components of returns and volumes generates positive unconditional correlation between returns and volumes, and positive time-varying conditional covariance.

Proof: See Section A.4.

As well as being consistent with the MDH, this result implies that if time-varying correlation is observed between returns and volumes over fixed intervals, it may actually be generated by the action of the irregular time process on a fixed covariance between assets at the transaction level. Then if we knew the fixed transaction-level covariance and could forecast the time process, we would be able to forecast the correlation over time.

4.5 Conclusion

It was the aim of this chapter to suggest a link between prices and volumes at irregular times, allowing for cointegration and long memory. We have produced one possible system by adapting the original work of Deo, Hurvich, Soulier and Wang (2009), Deo, Hsieh and Hurvich (2010), and Hurvich and Wang (2010). The strength of their framework enabled the system to support a variety of propositions which fit the stylised features.

This system features propagation of long memory from two individual LMSD processes (with feedback effects) to the squared returns of the cointegrated system (Proposition 1). The higher long memory parameter (from the two assets) eventually dominates the cointegrated system's squared returns and hence realized volatility. The system also shows that volumes will be between I(1) and I(2) (Proposition 2), which agrees with SF6.2 (Volumes are I($\geq$ 1)), and shows cointegration is possible between such volumes if the microstructure components of the assets have intermediate memory (Proposition 3). Finally, the system enables correlation between prices and volumes, demonstrating that a fixed tick-level correlation can lead to time-varying conditional correlation over regularly-spaced time intervals and that positive tick-level correlation leads to positive correlation at regularly-spaced time intervals (Proposition 4). More generally though, Proposition 4 can be applied to enable prediction of correlation over time between any variables in regularly-spaced time intervals if the tick-level correlation is known and the time process is forecasted. Alternatively, it can explain time-varying correlation over fixed time intervals as purely an irregular time effect.

This work is original in terms of its extension to new propositions. It also makes use of theory on point processes, and we suggest that Lemma 1, which concerns the long run behaviour of the covariance of two counting processes, is an original contribution in this respect.

Overall, the system can explain the propagation of long memory from the durations process to realized volaility in two assets and can be extended to incorporate volumes. This baseline system can be built upon to model the other stylised features (essentially distributional properties), requiring work at the post-doctoral level. It would also be useful to explore whether Proposition 3 can be extended to enable cointegration in volumes with long memory in the microstructure components.

The model used so far is mainly statistical rather than economic. One extension of the model would be to determine whether economic relations can explain the statistical relations. Specific parameterisations of the price-volume relationship as suggested by Tauchen and Pitts (1983) and Liesenfeld (2001) may yield more economic meaning.

Finally, alternative specifications to the framework of Hurvich et al. used so far might be investigated to determine whether they can generate the same results as easily or with greater extensibility. As mentioned before (Chapter 1), Time Deformation models can also generate irregular spacing of events through random time changes, albeit in a continuous time setting. Similarly, point processes can be specified in terms of an intensity function as opposed to a duration process, so the Autoregressive Conditional Intensity model of Russell (2001) or the Stochastic Conditional Intensity model of Bauwens and Hautsch (2006) might also be investigated, as it can be easier to conduct multivariate analysis of point processes using intensity functions. Alternatively, models based on Markov chains, such the MSMD model in Chapter 2, adapted from Calvet and Fisher (2004), might be explored further, since it can generate duration clustering similar to Long Memory. Engle and Russell (2005) have also created a new model called the ACM model which enables analysis of multivariate point processes taking values in a discrete number of states via Markov Chains.

Appendix A: Proofs of Propositions

Here we present our proofs of the various propositions.

It may be helpful to distinguish the usage of subscripts:

t : point in time: $t \in (0, \infty)$

i, j and l : transaction-level subscripts, e.g. $j \in \mathbb{Z}$

k : number of fixed interval of size Δt, then $k \in \mathbb{Z}^+$

m : asset number : $m \in \{1, 2\}$

τ : lag number: $\tau \in \mathbb{Z}$

A.1 Proposition 1

Long memory exists in the squared returns of PM4, with a common long memory parameter across both assets.

To start, let us simplify PM4 by grouping terms.

$$\ln P_{1,t} = \ln P_{1,0} + \sum_{i=1}^{N_1(t)} \underbrace{(e_{1,i} + \eta_{1,i})}_{:=\xi_{1,i}} + \sum_{i=1}^{N_2(t_{1,N_1(t)})} \underbrace{(\theta e_{2,i} + g_{21}\eta_{2,i})}_{:=\xi_{2,i}}$$

$$= \ln P_{1,0} + \sum_{i=1}^{N_1(t)} \xi_{1,i} + \sum_{i=1}^{N_2(t_{1,N_1(t)})} \xi_{2,i}$$

Define the *k*th return over the fixed time interval, Δt, as:

$$\begin{aligned} r_{1,k} &:= \ln P_{1,k\Delta t} - \ln P_{1,(k-1)\Delta t} \\ &= \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \xi_{1,i} + \sum_{i=N_2[t_{1,N_1[(k-1)\Delta t)}]+1}^{N_2(t_{1,N_1(k\Delta t)})} \xi_{2,i} \end{aligned}$$

We need to calculate $Cov(r_{1,k}^2, r_{1,k+\tau}^2)$, which equals $\mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2) - \mathbb{E}(r_{1,k}^2)\mathbb{E}(r_{1,k+\tau}^2)$. To calculate $\mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2)$, it is helpful to split it into two terms using an indicator function $\mathbf{1}_{A_m}$;

Let: $A_m = \{\{N_m[(k+\tau-1)\Delta t] - N_m(k\Delta t)\} > 1\}$

$=$ the set of events such that the number of asset *m* events

between $k\Delta t$ and $(k+\tau-1)\Delta t$ is greater than 1

$\mathcal{F}_m = \sigma\{N_m[(k-1)\Delta t], N_m(k\Delta t), N_m[(k+\tau-1)\Delta t], N_m[(k+\tau)\Delta t]\}$

$=$ the σ-field generated by the counts upto $(k-1)\Delta t$, $k\Delta t$, $(k+\tau-1)\Delta t$ and $(k+\tau)\Delta t$

Then $\mathbb{E}(r^2_{1,k}r^2_{1,k+\tau}) = \underbrace{\mathbb{E}(r^2_{1,k}r^2_{1,k+\tau}\mathbf{1}_{A_1})}_{(1)} + \mathbb{E}(r^2_{1,k}r^2_{1,k+\tau}\mathbf{1}_{A_1^c})$

Expanding term (1): $\mathbb{E}(r^2_{1,k}r^2_{1,k+\tau}\mathbf{1}_{A_1}) = \mathbb{E}[\mathbb{E}(r^2_{1,k}r^2_{1,k+\tau}\mathbf{1}_{A_1}|\mathcal{F}_1)]$

On the set A_1, the squared returns $r^2_{1,k}$ and $r^2_{1,k+\tau}$ are generated in fixed intervals with at least one intervening asset price movement. So the price levels associated with the squared returns are independent (in particular, $\ln P_{1,k}$, the log price at the end of the period corresponding to $r^2_{1,k}$, does not necessarily equal $\ln P_{1,k+\tau-1}$, the log price at the beginning of the period corresponding to $r^2_{1,k+\tau}$). So the squared returns are independent and hence:

$$\mathbb{E}[\mathbb{E}(r^2_{1,k}r^2_{1,k+\tau}\mathbf{1}_{A_1}|\mathcal{F}_1)] = \mathbb{E}[\mathbb{E}(r^2_{1,k}\mathbf{1}_{A_1}|\mathcal{F}_1)\mathbb{E}(r^2_{1,k+\tau}\mathbf{1}_{A_1}|\mathcal{F}_1)]$$

But since $A_1 \subset \mathcal{F}_1$,

$$\mathbb{E}[\mathbb{E}(r_{1,k}^2\mathbf{1}_{A_1}|\mathcal{F}_1)\mathbb{E}(r_{1,k+\tau}^2\mathbf{1}_{A_1}|\mathcal{F}_1)] = \mathbb{E}[\mathbf{1}_{A_1}\underbrace{\mathbb{E}(r_{1,k}^2|\mathcal{F}_1)}_{(2)}\mathbb{E}(r_{1,k+\tau}^2|\mathcal{F}_1)]$$

Expanding term (2):

$$\begin{aligned}\mathbb{E}(r^2_{1,k}|\mathcal{F}_1) = \mathbb{E}\Bigg[& \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(\Delta t)} \sum_{j=N_1[(k-1)\Delta t]+1}^{N_1(\Delta t)} \xi_{1,i}\xi_{1,j} \\ & + 2 \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(\Delta t)} \sum_{j=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \xi_{1,i}\xi_{2,j} \\ & + \sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \sum_{j=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \xi_{2,i}\xi_{2,j} \Bigg| \mathcal{F}_1 \Bigg]\end{aligned}$$

$$\begin{aligned} & = \underbrace{\mathbb{E}\left[\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(\Delta t)} \sum_{j=N_1[(k-1)\Delta t]+1}^{N_1(\Delta t)} \xi_{1,i}\xi_{1,j} \middle| \mathcal{F}_1 \right]}_{(3)} \\ & + \underbrace{\mathbb{E}\left[\sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \sum_{j=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \xi_{2,i}\xi_{2,j} \middle| \mathcal{F}_1 \right]}_{(4)}\end{aligned}$$

Expanding term (3):

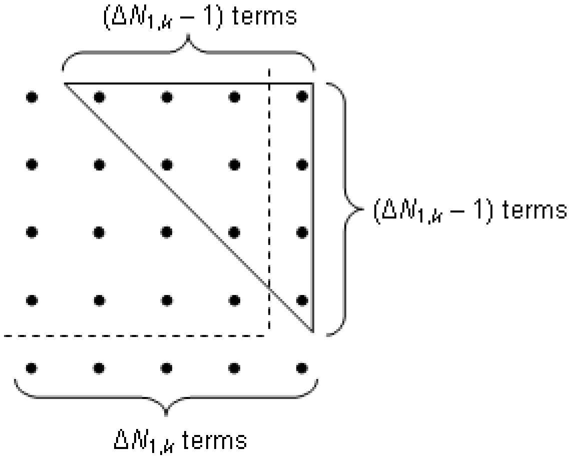
$$\mathbb{E}\left(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \sum_{j=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \xi_{1,i}\xi_{1,j} \middle| \mathcal{F}_1 \right)$$

$$= \{\underbrace{N_1(k\Delta t) - N_1[(k-1)\Delta t]}_{:=\Delta N_{1,k}}\}Var(\xi_{1,i}) + \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \sum_{j\neq i}^{N_1(k\Delta t)} \mathbb{E}(\xi_{1,i}\xi_{1,j})$$

$$= \Delta N_{1,k}Var(\xi_{1,i}) + 2\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \sum_{j>i}^{N_1(k\Delta t)} \mathbb{E}(\xi_{1,i}\xi_{1,j}) \qquad \text{(A.1)}$$

Figure A.1: Upper Off-Diagonal Cross Products

In total, there are $\Delta N_{1,k} \times \Delta N_{1,k}$ cross products. So the upper off-diagonal cross products form a triangle with width / length $\Delta N_{1,k} - 1$.

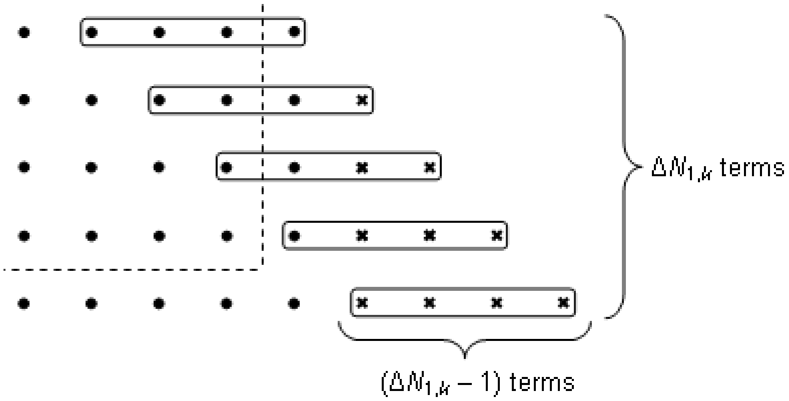


We can consider all the cross-products of $\xi_{1,i}$ with $\xi_{1,j}$ as points on a square $\Delta N_{1,k}$ points wide as in Figure A.1 above (the dashed line indicates that the square may actually be wider than 5 points). Then $\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \sum_{j>i}^{N_1(k\Delta t)} \mathbb{E}(\xi_{1,i}\xi_{1,j})$, the sum of all the upper off-diagonal cross-products, is the sum of all the points in the triangle.

If we now calculate $\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\xi_{1,i}\xi_{1,i+j}) = \Delta N_{1,k} \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\xi_{1,1}\xi_{1,1+j})$, then we calculate the sum of the terms in the rectangles in Figure A.2 below. This overestimates the sum of the terms in the triangle of the previous diagram by all the terms marked as crosses.

Figure A.2: Overestimate of Upper Off-Diagonal Cross Products

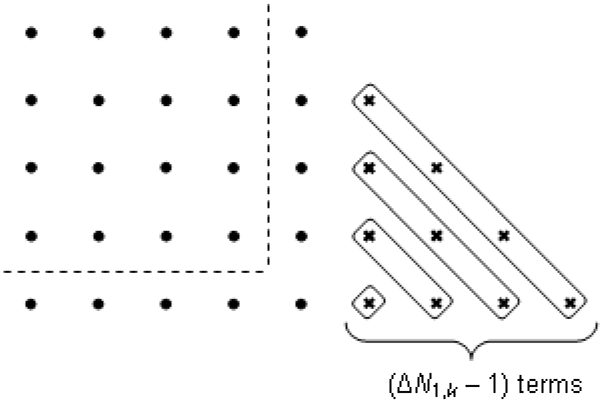
To estimate the sum of the upper off-diagonal cross products, we deliberately overestimate by summing $\Delta N_{1,k}$ rectangles with length $\Delta N_{1,k} - 1$. The products marked as crosses are the excess terms.



The sum of the terms with crosses (as in Figure A.3 below) is $\sum_{j=1}^{\Delta N_{1,k}-1} j\mathbb{E}(\xi_{1,1}\xi_{1,1+j})$

Figure A.3: Upper Off-Diagonal Cross Products

Finally, we subtract the overestimate as the sum of the excess terms (crosses).



Overall then, $\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \sum_{j>i}^{N_1(k\Delta t)} \mathbb{E}(\xi_{1,i}\xi_{1,j}) = \Delta N_{1,k} \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\xi_{1,i}\xi_{1,i+j}) - \sum_{j=1}^{\Delta N_{1,k}-1} j\mathbb{E}(\xi_{1,1}\xi_{1,1+j})$

So returning to equation (A.1),

$$\Delta N_{1,k} Var(\xi_{1,i}) + 2 \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \sum_{j>i}^{N_1(k\Delta t)} \mathbb{E}(\xi_{1,i}\xi_{1,j})$$
$$= \Delta N_{1,k} Var(\xi_{1,i}) + 2\Delta N_{1,k} \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\xi_{1,i}\xi_{1,i+j}) - 2 \sum_{j=1}^{\Delta N_{1,k}-1} j\mathbb{E}(\xi_{1,1}\xi_{1,1+j})$$

$$= \Delta N_{1,k} \underbrace{\left[Var(\xi_{1,i}) + 2 \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\xi_{1,1}\xi_{1,1+j}) \right]}_{:=C} \underbrace{-2 \sum_{j=1}^{\Delta N_{1,k}-1} j\mathbb{E}(\xi_{1,1}\xi_{1,1+j})}_{:=D}$$

$$= \Delta N_{1,k} C + D$$

We can say that C is bounded. For its first term, $0 < Var(\xi_{1,i}) = Var(e_{1,k}) + Var(\eta_{1,k}) = \sigma^2_{1,e} + \sigma^2_{1,\eta} < \infty$

For its second term, $\sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\xi_{1,1}\xi_{1,1+j}) = \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\eta_{1,1}\eta_{1,1+j})$

By the properties of intermediate memory processes (see e.g. Beran (1994)):

$$\sum_{j=1}^{\infty} \mathbb{E}(\eta_{1,1}\eta_{1,1+j}) < \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\eta_{1,1}\eta_{1,1+j}) < \sum_{j=-\infty}^{\infty} \mathbb{E}(\eta_{1,1}\eta_{1,1+j})$$
$$-0.5\sigma^2_{\eta_{1,1}} < \sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\eta_{1,1}\eta_{1,1+j}) < 0$$

We can also say that D is bounded; as supported by Figures A.1 and A.2, it contains a smaller number of cross-products than C, so:

$$\sum_{j=1}^{\Delta N_{1,k}-1} \mathbb{E}(\eta_{1,1}\eta_{1,1+j}) < \sum_{j=1}^{\Delta N_{1,k}-1} j\mathbb{E}(\xi_{1,1}\xi_{1,1+j}) < \sum_{j=-\infty}^{\infty} \mathbb{E}(\eta_{1,1}\eta_{1,1+j})$$

$$-0.5\sigma^2_{\eta_{1,1}} < \sum_{j=1}^{\Delta N_{1,k}-1} j\mathbb{E}(\xi_{1,1}\xi_{1,1+j}) < 0$$

Returning to term (4): $\mathbb{E}\left[\sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \sum_{j=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \xi_{2,i}\xi_{2,j} \middle| \mathcal{F}_1 \right]$

$$= \{\underbrace{N_2(t_{1,N_1(k\Delta t)}) - N_2[t_{1,N_1[(k-1)\Delta t]}]}_{:=\Delta N^*_{2,k}}\} Var(\xi_{2,i})$$

$$+ \sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \sum_{j\neq i} \xi_{2,i}\xi_{2,j}$$

$$= \Delta N^*_{2,k} Var(\xi_{2,i}) + 2 \sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \sum_{j>i}^{N_2[t_{1,N_1(k\Delta t)}]} \mathbb{E}(\xi_{2,i}\xi_{2,j})$$

$$= \Delta N^*_{2,k} Var(\xi_{2,i}) + 2 \sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2[t_{1,N_1(k\Delta t)}]} \left\{ \sum_{j=1}^{\Delta N^*_{2,k}-1} \mathbb{E}(\xi_{2,i}\xi_{2,i+j}) \right.$$
$$\left. - 2 \sum_{j=1}^{\Delta N^*_{2,k}-1} j\mathbb{E}(\xi_{2,1}\xi_{2,1+j}) \right\}$$

$$= \Delta N^*_{2,k} \underbrace{\left[Var(\xi_{2,i}) + 2 \sum_{j=1}^{\Delta N^*_{2,k}-1} \mathbb{E}(\xi_{2,1}\xi_{2,1+j}) \right]}_{:=F} \underbrace{-2 \sum_{j=1}^{\Delta N^*_{2,k}-1} j\mathbb{E}(\xi_{2,1}\xi_{2,1+j})}_{:=G}$$

$$= \Delta N^*_{2,k} F + G$$

where using similar arguments to before, F and G are bounded.

$$\mathbb{E}(r_{1,k}^2|\mathcal{F}_1) = C\Delta N_{1,k} + F\Delta N_{2,k}^* + (D+G)$$

$$\mathbb{E}(r_{1,k+\tau}^2|\mathcal{F}_1) = C\Delta N_{1,k+\tau} + F\Delta N_{2,k+\tau}^* + (D+G)$$

$$\mathbb{E}(r_{1,k}^2) = \mathbb{E}(r_{1,k+\tau}^2) = C\mathbb{E}(\Delta N_{1,k}) + F\mathbb{E}(\Delta N_{2,k}^*) + (D+G) \text{ by stationarity}$$

$$
\begin{aligned}
\mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2 \mathbf{1}_{A_1}) &= \mathbb{E}[\mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2 \mathbf{1}_{A_1} | \mathcal{F}_1)] \\
&= \mathbb{E}\{\mathbb{E}[(C\Delta N_{1,k} + F\Delta N_{2,k}^* + (D+G))(C\Delta N_{1,k+\tau} + F\Delta N_{2,k+\tau}^* + (D+G))\mathbf{1}_{A_1} | \mathcal{F}_1]\} \\
&= \mathbb{E}\{\mathbb{E}[(C^2\Delta N_{1,k}\Delta N_{1,k+\tau} + CF\Delta N_{1,k}\Delta N_{2,k+\tau}^* + CF\Delta N_{1,k+\tau}\Delta N_{2,k}^* \\
&\quad + F^2\Delta N_{2,k}^*\Delta N_{2,k+\tau}^* + C(D+G)\Delta N_{1,k} + C(D+G)\Delta N_{1,k+\tau} \\
&\quad + F(D+G)\Delta N_{2,k}^* + F(D+G)\Delta N_{2,k+\tau}^* + (D+G)^2)\mathbf{1}_{A_1} | \mathcal{F}_1]\} \\
&= \mathbb{E}\{\mathbf{1}_{A_1}\mathbb{E}[(C^2\Delta N_{1,k}\Delta N_{1,k+\tau} + CF\Delta N_{1,k}\Delta N_{2,k+\tau}^* + CF\Delta N_{1,k+\tau}\Delta N_{2,k}^* \\
&\quad + F^2\Delta N_{2,k}^*\Delta N_{2,k+\tau}^* + C(D+G)\Delta N_{1,k} + C(D+G)\Delta N_{1,k+\tau} \\
&\quad + F(D+G)\Delta N_{2,k}^* + F(D+G)\Delta N_{2,k+\tau}^* + (D+G)^2) | \mathcal{F}_1]\} \\
&= \mathbb{E}\{\mathbf{1}_{A_1}[(C^2\mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}) + CF\mathbb{E}(\Delta N_{1,k}\Delta N_{2,k+\tau}^*) + CF\mathbb{E}(\Delta N_{1,k+\tau}\Delta N_{2,k}^*) \\
&\quad + F^2\mathbb{E}(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*) + C(D+G)\mathbb{E}(\Delta N_{1,k}) + C(D+G)\mathbb{E}(\Delta N_{1,k+\tau}) \\
&\quad + F(D+G)\mathbb{E}(\Delta N_{2,k}^*) + F(D+G)\mathbb{E}(\Delta N_{2,k+\tau}^*) + (D+G)^2) | \mathcal{F}_1]\} \\
&= C^2[\mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}) - \mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}\mathbf{1}_{A_1^c})] \\
&\quad + F^2[\mathbb{E}(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*) - \mathbb{E}(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*\mathbf{1}_{A_1^c})] \\
&\quad + CF[\mathbb{E}(\Delta N_{1,k}\Delta N_{2,k+\tau}^*) - \mathbb{E}(\Delta N_{1,k}\Delta N_{2,k+\tau}^*\mathbf{1}_{A_1^c})] \\
&\quad + CF[\mathbb{E}(\Delta N_{1,k+\tau}\Delta N_{2,k}^*) - \mathbb{E}(\Delta N_{1,k+\tau}\Delta N_{2,k}^*\mathbf{1}_{A_1^c})] \\
&\quad + 2C(D+G)[\mathbb{E}(\Delta N_{1,k}) - \mathbb{E}(\Delta N_{1,k}\mathbf{1}_{A_1^c})] \\
&\quad + 2F(D+G)[\mathbb{E}(\Delta N_{2,k}^*) - \mathbb{E}(\Delta N_{2,k}^*\mathbf{1}_{A_1^c})] \\
&\quad + (D+G)^2[1 - \mathbb{E}(\mathbf{1}_{A_1^c})]
\end{aligned}
$$

$$Cov(r_{1,k}^2, r_{1,k+\tau}^2) = \mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2) - \mathbb{E}(r_{1,k}^2)\mathbb{E}(r_{1,k+\tau}^2)$$

$$= \mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2 \mathbf{1}_{A_1}) + \mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2 \mathbf{1}_{A_1^c}) - \mathbb{E}(r_{1,k}^2)\mathbb{E}(r_{1,k+\tau}^2)$$

$$\begin{aligned}
&= C^2\mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}) - C^2\mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}\mathbf{1}_{A_1^c}) \\
&\quad + F^2\mathbb{E}(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*) - F^2\mathbb{E}(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*\mathbf{1}_{A_1^c}) \\
&\quad + CF\mathbb{E}(\Delta N_{1,k}\Delta N_{2,k+\tau}^*) - CF\mathbb{E}(\Delta N_{1,k}\Delta N_{2,k+\tau}^*\mathbf{1}_{A_1^c}) \\
&\quad + CF\mathbb{E}(\Delta N_{1,k+\tau}\Delta N_{2,k}^*) - CF\mathbb{E}(\Delta N_{1,k+\tau}\Delta N_{2,k}^*\mathbf{1}_{A_1^c}) \\
&\quad + 2C(D+G)\mathbb{E}(\Delta N_{1,k}) - 2C(D+G)\mathbb{E}(\Delta N_{1,k}\mathbf{1}_{A_1^c}) \\
&\quad + 2F(D+G)\mathbb{E}(\Delta N_{2,k}^*) - 2F(D+G)\mathbb{E}(\Delta N_{2,k}^*\mathbf{1}_{A_1^c}) \\
&\quad + (D+G)^2 - (D+G)^2\mathbb{E}(\mathbf{1}_{A_1^c}) + \mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2\mathbf{1}_{A_1^c}) \\
&\quad - C^2\mathbb{E}(\Delta N_{1,k})\mathbb{E}(\Delta N_{1,k+\tau}) - 2C(D+G)\mathbb{E}(\Delta N_{1,k}) \\
&\quad - CF\mathbb{E}(\Delta N_{1,k})\mathbb{E}(\Delta N_{2,k+\tau}^*) - CF\mathbb{E}(\Delta N_{1,k+\tau})\mathbb{E}(\Delta N_{2,k}^*) - (D+G)^2 \\
&\quad - 2F(D+G)\mathbb{E}(\Delta N_{2,k}^*) - F^2\mathbb{E}(\Delta N_{2,k}^*)\mathbb{E}(\Delta N_{2,k+\tau}^*)
\end{aligned}$$

$$\begin{aligned}
&= C^2Cov(\Delta N_{1,k}, \Delta N_{1,k+\tau}) + F^2Cov(\Delta N_{2,k}^*, \Delta N_{2,k+\tau}^*) \\
&\quad + CFCov(\Delta N_{1,k}, \Delta N_{2,k+\tau}^*) + CFCov(\Delta N_{1,k+\tau}, \Delta N_{2,k}^*) \\
&\quad - C^2\mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}\mathbf{1}_{A_1^c}) - F^2\mathbb{E}(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*\mathbf{1}_{A_1^c}) \\
&\quad - CF\mathbb{E}(\Delta N_{1,k}\Delta N_{2,k+\tau}^*\mathbf{1}_{A_1^c}) - CF\mathbb{E}(\Delta N_{1,k+\tau}\Delta N_{2,k}^*\mathbf{1}_{A_1^c}) \\
&\quad - 2C(D+G)\mathbb{E}(\Delta N_{1,k}\mathbf{1}_{A_1^c}) - 2F(D+G)\mathbb{E}(\Delta N_{2,k}^*\mathbf{1}_{A_1^c}) \\
&\quad - (D+G)^2\mathbb{E}(\mathbf{1}_{A_1^c}) + \mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2\mathbf{1}_{A_1^c})
\end{aligned}$$

Of the first four terms, $C^2Cov(\Delta N_{1,k}, \Delta N_{1,k+\tau})$ and $F^2Cov(\Delta N^*_{2,k}, \Delta N^*_{2,k+\tau})$ dominate the others. Note that $Cov(\Delta N^*_{2,k}, \Delta N^*_{2,k+\tau}) \neq Cov(\Delta N_{2,k}, \Delta N_{2,k+\tau})$ since the former quantity is not defined over fixed time intervals - however, we provide a proof that it is of the same asymptotic order as $Cov(\Delta N_{2,k}, \Delta N_{2,k+\tau})$ in Lemma 1 (Section A.5). In this case, if $N_1(t)$ has a higher memory parameter than $N_2(t)$, then $C^2Cov(\Delta N_{1,k}, \Delta N_{1,k+\tau})$ dominates the other term, and vice versa.

Of the terms involving $\mathbf{1}_{A_1^c}$, three terms dominate the others: $C^2\mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}\mathbf{1}_{A_1^c})$, $F^2\mathbb{E}(\Delta N^*_{2,k}\Delta N^*_{2,k+\tau}\mathbf{1}_{A_1^c})$ and $\mathbb{E}(r^2_{1,k}r^2_{1,k+\tau}\mathbf{1}_{A_1^c})$.

By the Cauchy-Schwartz inequality, $\mathbb{E}(r_{1,k}^2 r_{1,k+\tau}^2 \mathbf{1}_{A_1^c}) \leq \sqrt{\mathbb{E}(r_{1,k}^4 r_{1,k+\tau}^4)\mathbb{E}(\mathbf{1}_{A_1^c}^2)}$

$$= \sqrt{\mathbb{E}(r_{1,k}^4 r_{1,k+\tau}^4)\mathbb{E}(\mathbf{1}_{A_1^c})}$$

$$= \sqrt{\mathbb{E}(r_{1,k}^4 r_{1,k+\tau}^4)\mathbb{P}(\mathbf{1}_{A_1^c})}$$

{for a stationary process $\{Z_k\}$, $\mathbb{E}(Z_k, Z_{k+\tau}) \leq \mathbb{E}(Z_k^2)\}$ $\leq \sqrt{\mathbb{E}(r_{1,k}^8)\mathbb{P}(\mathbf{1}_{A_1^c})}$

Similarly, $\mathbb{E}(\Delta N_{1,k}\Delta N_{1,k+\tau}\mathbf{1}_{A_1^c}) \leq \sqrt{\mathbb{E}[(\Delta N_{1,k}\Delta N_{1,k+\tau})^2]\mathbb{P}(\mathbf{1}_{A_1^c})}$

$$\leq \sqrt{\mathbb{E}[(\Delta N_{1,k})^4]\mathbb{P}(\mathbf{1}_{A_1^c})}$$

and $\mathbb{E}(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*\mathbf{1}_{A_1^c}) \leq \sqrt{\mathbb{E}[(\Delta N_{2,k}^*\Delta N_{2,k+\tau}^*)^2]\mathbb{P}(\mathbf{1}_{A_1^c})}$

$$\leq \sqrt{\mathbb{E}[(\Delta N_{2,k}^*)^4]\mathbb{P}(\mathbf{1}_{A_1^c})}$$

As per the assumptions at the beginning of Theorem 5 in Deo, Hurvich, Soulier and Wang (2009), since $e_{m,i}$ and $\eta_{m,i}$ have finite 8th moment, $\xi_{m,i}$ has finite 8th moment, so $r_{m,i}$, which is a finite sum of $\xi_{m,i}$ terms, has finite 8th moment. They also assume that $\mathbb{E}\{N_m(t) - \mathbb{E}[N_m(t)]\}^8$ is finite for finite t, so $\{[\mathbb{E}(N_m(t))]\}^8$ is also finite for finite t. This is bigger than both $\mathbb{E}[(\Delta N_m)^4]$ and $\mathbb{E}[(\Delta N_m^*)^4]$. So it suffices to show that $\mathbb{P}(1_{A_1^c}) = o(\tau^{4d_{x_1}-2})$ to show that these terms are dominated by $C^2 Cov(\Delta N_{1,k}, \Delta N_{1,k+\tau})$ and $F^2 Cov(\Delta N_{2,k}^*, \Delta N_{2,k+\tau}^*)$, which is done by Deo, Hsieh and Hurvich (2006a). Overall then,

$Cov(r_{1,k}^2, r_{1,k+\tau}^2) \sim K\tau^{2\max(d_{x_1}, d_{x_2})-1}$ where K is a constant (and finite).

Applying all the logic up to this point to the other asset leads to the conclusion that the memory component from one asset dominates that of the other, so that both assets have the same long memory parameter, as we set out to prove. $\Box$

A.2 Proposition 2

Volume as described by VM3 is at least I(1) and at most just less than I(2).

Case 1: $d_{\nu_m} \in (-\frac{1}{2}, 0)$

$$\begin{aligned} Var(\ln V_{1,t}) &= Var(\sum_{i=1}^{N_1(t)} u_{1,i}) + Var(\sum_{i=1}^{N_1(t)} \nu_{1,i}) \\ &\quad + a_1^2 Var(\sum_{i=1}^{N_2(t_{1,N_1(t)})} u_{2,i}) + b_1^2 Var(\sum_{i=1}^{N_2(t_{1,N_1(t)})} \nu_{2,i}) \end{aligned}$$

$$\begin{aligned} Var(\sum_{i=1}^{N_1(t)} u_{1,i}) &= \mathbb{E}[Var(\sum_{i=1}^{N_1(t)} u_{1,i}|N_1(.))] + Var[\mathbb{E}(\sum_{i=1}^{N_1(t)} u_{1,i}|N_1(.))] \\ &= \mathbb{E}\left(Var(\sum_{i=1}^{N_1(t)} u_{1,i}|N_1(t))\right) + 0 \\ &= \mathbb{E}\left(N_1(t)\sigma_{1,u}^2\right) \\ &= \sigma_{1,u}^2 \mathbb{E}\left(N_1(t)\right) \\ &= \sigma_{1,u}^2 \lambda_1 t = O(t) \end{aligned}$$

Similarly, $Var(\sum_{i=1}^{N_2(t_{1,N_1(t)})} u_{2,i}) = \sigma_{2,u}^2 \mathbb{E}[N_2(t_{1,N_1(t)})]$

$$= \sigma_{2,u}^2 \mathbb{E}[N_2(t)] - \tilde{C}_2)]$$

where $\tilde{C}_2$ represents the Backwards Recurrence Time for asset 2

$$\sim \sigma_{2,u}^2 \lambda_2 t = O(t)$$

$Var(\sum_{i=1}^{N_1(t)} \nu_{1,i}) \sim \sigma_{1,u}^2(\lambda_1 t)^{2d_{\nu_1}+1} = o(t)$ by Lemma 1 in Hurvich and Wang (2010).

$$\text{Similarly, } Var(\sum_{i=1}^{N_2(t_{1,N_1(t)})} \nu_{2,i}) \sim \sigma_{1,u}^2 \mathbb{E}[N_2(t_{1,N_1(t)})]^{2d_{\nu_2}+1}$$

$$\leq \sigma_{1,u}^2[\mathbb{E}(N_2(t_{1,N_1(t)}))]^{2d_{\nu_2}+1} \qquad \text{by Jensen's Inequality}$$

$$= \sigma_{1,u}^2[\mathbb{E}(N_2(t)) - \tilde{C}_2]^{2d_{\nu_2}+1} = o(t)$$

So overall, $Var(\ln V_{1,t}) = O(t)$ so $\ln V_{1,t}$ is I(1).

Case 2: $d_{\nu_m} \in (0, \frac{1}{2})$

The component variances are similar to before, except that Jensen's Inequality can no longer be applied to bound the asymptotic order of $Var\left(\sum_{i=1}^{N_2(t_{1,N_1(t)})} \nu_{2,i}\right)$. In fact, it implies that the order is greater than or equal to $t^{2d_{\nu_1}+1}$. Instead, Lemmas 3 and 4 of Hurvich and Wang can be applied, which imply that $\mathbb{E}\{[N_2(t)]^2\} = O(t^2)$ if the LMSD process has finite moments of order 2 for the durations.

Then overall, the best we can say is that $Var(\ln V_{1,t}) = o(t^2)$ so $\ln V_{1,t}$ is at most just less than I(2). □

A.3 Proposition 3

Cointegration can exist between volume processes as defined in VM3 if the microstucture component has intermediate memory ($d_{\nu_m} \in (-\frac{1}{2}, 0)$).

Case 1: $d_{\nu_m} \in (-\frac{1}{2}, 0)$

Take $a_1 = \frac{1}{a_2}$, as in Hurvich and Wang (2010).

$$\ln V_{1,t} - a_1 \ln V_{2,t} = \sum_{i=1}^{N_1(t)} u_{1,i} - \sum_{i=1}^{N_1(t_{2,N_2(t)})} u_{1,i} + \sum_{i=1}^{N_1(t)} \nu_{1,i} - a_1 b_2 \sum_{i=1}^{N_1(t_{2,N_2(t)})} \nu_{1,i}$$

$$+ a_1 \sum_{i=1}^{N_2(t_{1,N_1(t)})} u_{2,i} - a_1 \sum_{i=1}^{N_2(t)} u_{2,i} + b_1 \sum_{i=1}^{N_2(t_{1,N_1(t)})} \nu_{2,i} - a_1 \sum_{i=1}^{N_2(t)} \nu_{2,i}$$

$$= \underbrace{\sum_{i=N_1(t_{2,N_2(t)})+1}^{N_1(t)} u_{1,i}}_{T_1} - a_1 \underbrace{\sum_{i=N_2(t_{1,N_1(t)})+1}^{N_2(t)} u_{2,i}}_{T_2}$$

$$+ (1 - a_1 b_2) \underbrace{\sum_{i=1}^{N_1(t)} \nu_{1,i}}_{T_3} + a_1 b_2 \underbrace{\sum_{i=N_1(t_{2,N_2(t)})+1}^{N_1(t)} \nu_{1,i}}_{T_4}$$

$$+ (b_1 - a_1) \underbrace{\sum_{i=1}^{N_2(t)} \nu_{2,i}}_{T_5} - b_1 \underbrace{\sum_{i=N_2(t_{1,N_1(t)})+1}^{N_2(t)} \nu_{2,i}}_{T_6}$$

By Theorem 5 of Deo, Hurvich, Soulier and Wang (2009), $Var(T_3) \sim (\sigma^2_{1,\nu} \lambda_1^{2d_{\nu_1}+1}) t^{2d_{\nu_1}+1}$

Similarly, $Var(T_5) \sim (\sigma_{2,\nu}^2 \lambda_2^{2d_{\nu_2}+1}) t^{2d_{\nu_2}+1}$

$$Var(T_4) = Var\{B_{H_1}[N_1(t) + 1] - B_{H_1}[N_1(t_{2,N_2(t)})]\}$$

$$= \mathbb{E}\{\sigma_{1,\nu}^2 [N_1(t) + 1 - N_1(t_{2,N_2(t)}) - 1]^{2d_{\nu_1}+1}\}$$

$$= \sigma_{1,\nu}^2 \mathbb{E}\{[N_1(t) - N_1(t_{2,N_2(t)})]^{2d_{\nu_1}+1}\}$$

$$\leq \sigma_{1,\nu}^2 \{\mathbb{E}[N_1(t) - N_1(t_{2,N_2(t)})]\}^{2d_{\nu_1}+1}$$

(by Jensen's inequality)

$$= \sigma_{1,\nu}^2 \tilde{C}_1^{2d_{\nu_1}+1}$$

Similarly, $Var(T_6) = \sigma_{2,\nu}^2 \tilde{C}_2^{2d_{\nu_2}+1}$

$$Var(T_1) = Var(u_{1,i})\mathbb{E}[N_1(t) - N_1(t_{2,N_2(t)})] = \sigma_{1,e}^2 \tilde{C}_1$$

Similarly, $Var(T_2) = \sigma_{2,e}^2 \tilde{C}_2$

By Cauchy - Schwartz, $|Cov(T_3, T_4)| \leq \sqrt{Var(T_3)Var(T_4)}$

$$\leq \sqrt{\sigma_{1,u}^2 \lambda_1^{2d_{\nu_1}+1})t^{2d_{\nu_1}+1}\sigma_{1,\nu}^2 \tilde{C}_1^{2d_{\nu_1}+1}}$$

(by Jensen's inequality)

$$= \sigma_{1,\nu}^2 (\lambda_1 \tilde{C}_1)^{d_{\nu_1}+\frac{1}{2}} t^{d_{\nu_1}+\frac{1}{2}}$$

$$= o(t^{2d_{\nu_1}+1})$$

Similarly, $|Cov(T_5, T_6)| \leq \sigma_{2,\nu}^2 (\lambda_2 \tilde{C}_2)^{d_{\nu_2}+\frac{1}{2}} t^{d_{\nu_2}+\frac{1}{2}}$

$$= o(t^{2d_{\nu_2}+1})$$

So overall, $Var(T_3)$ and $Var(T_5)$ dominate the other terms

$$\begin{aligned}\text{So } Var(\ln V_{1,t} - (a_1 \ln V_{2,t})) &\sim \underbrace{(1 - a_1 b_2)^2 \sigma_{1,\nu}^2 \lambda_1^{2d_{\nu_1}+1}}_{C_1} t^{2d_{\nu_1}+1} \\ &+ \underbrace{(b_1 - a_1)^2 \sigma_{2,\nu}^2 \lambda_2^{2d_{\nu_2}+1}}_{C_2} t^{2d_{\nu_2}+1} \\ &\sim Ct^{1+2\max(d_{\nu_1}, d_{\nu_1})}\end{aligned}$$

$$\text{where } C = \begin{cases} C_1 & \text{if } d_{\nu_1} > d_{\nu_2} \\ C_2 & \text{if } d_{\nu_1} < d_{\nu_2} \\ C_1 + C_2 & \text{if } d_{\nu_1} = d_{\nu_2} \end{cases}$$

Overall, the cointegrating vector is $\left(\begin{smallmatrix} 1 \\ -a_1 \end{smallmatrix}\right)$ and the memory parameter decreases from 1 for both $\ln V_{m,t}$ to $1 + \max(d_{\nu_1}, d_{\nu_2})$

Case 2: $d_{\nu_m} \in (0, \frac{1}{2})$

Take $a_1 = b_1 = \frac{1}{a_2} = \frac{1}{b_2}$. We are now assuming that one parameter drives all the feedback.

$$\ln V_{1,t} - a_1 \ln V_{2,t} = \sum_{i=1}^{N_1(t)} u_{1,i} - \sum_{i=1}^{N_1(t_{2,N_2(t)})} u_{1,i} + \sum_{i=1}^{N_1(t)} \nu_{1,i} - \sum_{i=1}^{N_1(t_{2,N_2(t)})} \nu_{1,i}$$

$$+ a_1 \sum_{i=1}^{N_2(t_{1,N_1(t)})} u_{2,i} - a_1 \sum_{i=1}^{N_2(t)} u_{2,i} + b_1 \sum_{i=1}^{N_2(t_{1,N_1(t)})} \nu_{2,i} - a_1 \sum_{i=1}^{N_2(t)} \nu_{2,i}$$

$$= \underbrace{\sum_{i=N_1(t_{2,N_2(t)})+1}^{N_1(t)} u_{1,i}}_{T_1} - a_1 \underbrace{\sum_{i=N_2(t_{1,N_1(t)})+1}^{N_2(t)} u_{2,i}}_{T_2}$$

$$+ \underbrace{\sum_{i=N_1(t_{2,N_2(t)})+1}^{N_1(t)} \nu_{1,i}}_{T_4} - a_1 \underbrace{\sum_{i=N_2(t_{1,N_1(t)})+1}^{N_2(t)} \nu_{2,i}}_{T_6}$$

$$Var(T_4) = Var\{B_{H_1}[N_1(t)+1] - B_{H_1}[N_1(t_{2,N_2(t)})]\}$$

$$= \mathbb{E}\{\sigma_{1,\nu}^2 [N_1(t) + 1 - N_1(t_{2,N_2(t)}) - 1]^{2d_{\nu_1}+1}\}$$

$$= \sigma_{1,\nu}^2 \mathbb{E}\{[N_1(t) - N_1(t_{2,N_2(t)})]^{2d_{\nu_1}+1}\}$$

$$\leq \sigma_{1,\nu}^2 \{\mathbb{E}[N_1(t)]\}^{2d_{\nu_1}+1}$$

$$= o(t^2) \text{ by Lemmas 3 and 4 of Hurvich and Wang}$$

Similarly, $Var(T_6) = o(t^2)$

As before, $Var(T_1) = \sigma^2_{1,e}\tilde{C}_1$

$Var(T_2) = \sigma^2_{2,e}\tilde{C}_2$

So overall, $Var(T_4)$ and $Var(T_6)$ now dominate the other terms, and we have not achieved a reduction in the memory parameter which is still between $1 + 2\max(d_{\nu_1}, d_{\nu_2})$ and 2. $\square$

A.4 Proposition 4

Positive correlation between the microstructure components of returns and volumes generates positive unconditional correlation between returns and volumes, and positive time-varying conditional covariance.

Define the *k*th logged volume over the fixed time interval, Δt, as:

$$v_{1,k} := \ln V_{1,0} + \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \underbrace{(u_{1,i} + \nu_{1,i})}_{:=\zeta_{1,i}} + \sum_{i=N_2[(k-1)\Delta t]+1}^{N_2(t_{1,N_1(k\Delta t)})} \underbrace{(a_1 u_{2,i} + b_1 \nu_{2,i})}_{:=\zeta_{2,i}}$$

$$= \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t)} \zeta_{1,i} + \sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]}]+1}^{N_2(t_{1,N_1(k\Delta t)})} \zeta_{2,i}$$

$$Cov(r_{1,k}^2, v_{1,k}|\mathcal{F}_1) = \underbrace{Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \xi_{1,i}\xi_{1,j}, \sum_{l=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \zeta_{1,l}|\mathcal{F}_1)}_{(5)}$$

$$+ \underbrace{Cov(\sum_{i=N_2[t_{1,N_1[(k-1)\Delta t]+1}]}^{N_2[t_{1,N_1(k\Delta t)}]} \sum_{j=N_2[t_{1,N_1[(k-1)\Delta t]+1}]}^{N_2[t_{1,N_1(k\Delta t)}]} \xi_{2,i}\xi_{2,j} \sum_{l=N_2[t_{1,N_1[(k-1)\Delta t]+1}]}^{N_2[t_{1,N_1(k\Delta t)}]} \zeta_{2,l}|\mathcal{F}_1)}_{(6)}$$

Expanding (5): $Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \xi_{1,i}\xi_{1,j}, \sum_{l=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \zeta_{1,l}|\mathcal{F}_1)$

$$\begin{aligned}
&= Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j=i} \xi_{1,i}\xi_{1,j}, \sum_{l=i} \zeta_{1,l}|\mathcal{F}_1) \\
&\quad + Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j=i} \xi_{1,i}\xi_{1,j}, \sum_{l\neq i} \zeta_{1,l}|\mathcal{F}_1) \\
&\quad + Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j\neq i} \xi_{1,i}\xi_{1,j}, \sum_{l=i} \zeta_{1,l}|\mathcal{F}_1) \\
&\quad + Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j\neq i} \xi_{1,i}\xi_{1,j}, \sum_{l\neq i\cap l=j} \zeta_{1,l}|\mathcal{F}_1) \\
&\quad + Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j\neq i} \xi_{1,i}\xi_{1,j}, \sum_{l\neq i\cap l\neq j} \zeta_{1,l}|\mathcal{F}_1)
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} Cov(\xi_{1,i}^2, \zeta_{1,i}|\mathcal{F}_1) + Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \xi_{1,i}^2, \sum_{l\neq i} \zeta_{1,l}|\mathcal{F}_1) \\
&\quad + \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} Cov(\sum_{j\neq i} \xi_{1,i}\xi_{1,j},\zeta_{1,i}|\mathcal{F}_1) \\
&\quad + Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j\neq i} \xi_{1,i}\xi_{1,j}, \zeta_{1,j}|\mathcal{F}_1) \\
&\quad + Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j\neq i} \xi_{1,i}\xi_{1,j}, \sum_{l\neq i\cap l\neq j} \zeta_{1,l}|\mathcal{F}_1)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} Cov(\xi_{1,i}^2, \zeta_{1,i}) + 2Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \xi_{1,i}^2, \sum_{l>i} \zeta_{1,l}|\mathcal{F}_1) \\
&\quad + 2 \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} Cov(\sum_{j>i} \xi_{1,i}\xi_{1,j},\zeta_{1,i}|\mathcal{F}_1) \\
&\quad + 2Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j>i} \xi_{1,i}\xi_{1,j}, \zeta_{1,j}|\mathcal{F}_1) \\
&\quad + 2Cov(\sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} \sum_{j>i} \xi_{1,i}\xi_{1,j}, \sum_{l>j>i} \zeta_{1,l} + \sum_{j>l>i} \zeta_{1,l} + \sum_{j>i>l} \zeta_{1,l}|\mathcal{F}_1)
\end{aligned}
$$

If there is only positive contemporaneous covariance in the microstructure components, equal to κ_1, then we have that all terms apart from the first become zero. So:

$$
(5) = \sum_{i=N_1[(k-1)\Delta t]+1}^{N_1(k\Delta t]} Cov(\xi_{1,i}^2, \zeta_{1,i}) = \Delta N_{1,k} Cov(\xi_{1,i}^2, \zeta_{1,i}) = \Delta N_{1,k} Cov(\eta_{1,i}^2, \nu_{1,i}) = \Delta N_{1,k}\kappa_1
$$

Similarly, (6) evaluates to $\Delta N_{2,k}^* g_{21} b_1 \kappa_2$, where κ_2 is the covariance in the microstructure components within ξ_2 and ζ_2.

Overall then, $Cov(r_{1,k}^2, v_{1,k}|\mathcal{F}_1) = \Delta N_{1,k}\kappa_1 + \Delta N_{2,k}^* \theta b_1 \kappa_2$

Since $\Delta N_{1,k}$ and $\Delta N_{2,k}^*$ are weakly positive, this means that fixed positive covariances in the microstructure components (and positive feedback coefficients g_{21} and b_1) generate a positive time-varying covariance between prices and volumes over fixed time periods. The conditional and unconditional correlations of $r_{1,k}^2$ with $v_{1,k}$ are harder to determine. However, we can say that the unconditional correlation must be weakly positive, since it is a positive function of the conditional covariance, which is always weakly positive. □

A.5 Lemma 1

$Cov(\Delta N^*_{2,k}, \Delta N^*_{2,k+\tau}) \sim Cov(\Delta N_{2,k}, \Delta N_{2,k+\tau})$

Let us restate the problem.

$N_2(t) :=$ number of asset 2 events in the interval $(0, t]$

$=$ e.g. 5 in Figure A.4 below

$N_2(k\Delta t) =$ number of asset 2 events in the interval $(0, k\Delta t]$

$=$ e.g. 4

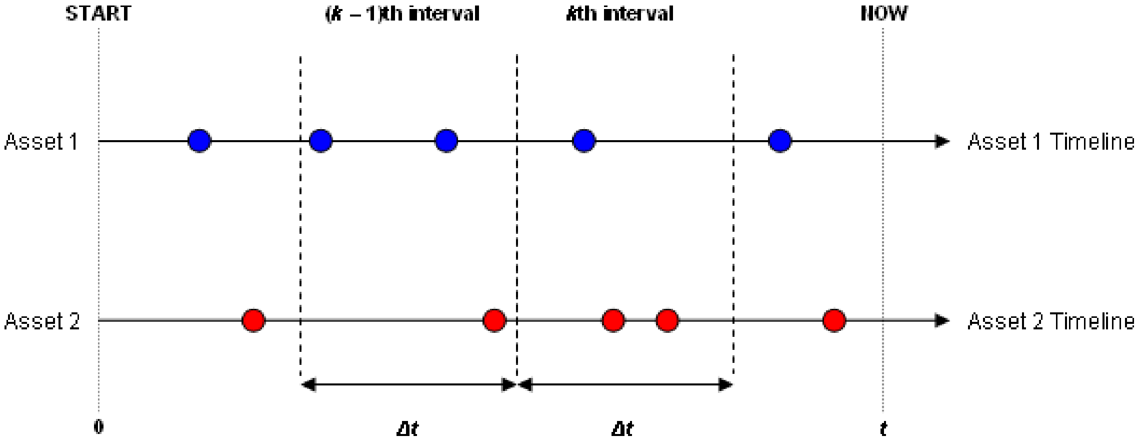
$\Delta N_{2,k} := N_2(k\Delta t) - N_2[(k-1)\Delta t]$

$=$ number of asset 2 events in the kth interval of length Δt

$=$ e.g. 2

$Cov(\Delta N_{2,k}, \Delta N_{2,k+\tau}) \sim C\tau^{2d_{x_2}-1}$ for some constant C as τ becomes large, by the definition of the LMSD process.

Figure A.4: Asset 2 Events in Fixed Time Intervals



$N_2(t_{1,N_1(t)}) :=$ number of asset 2 events which occur by the last asset 1 event in $(0, t]$

$=$ e.g. 4 in Figure A.4 above

$N_2(t_{1,N_1(k\Delta t)}) =$ number of asset 2 events which occur by the last asset 1 event in $(0, k\Delta t]$

$=$ e.g. 2

$\Delta N^*_{2,k} := N_2(t_{1,N_1(k\Delta t)}) - N_2(t_{1,N_1[(k-1)\Delta t)]})$

$=$ number of asset 2 events which occur

after the last asset 1 event in the (k - 1)th interval of length Δt and

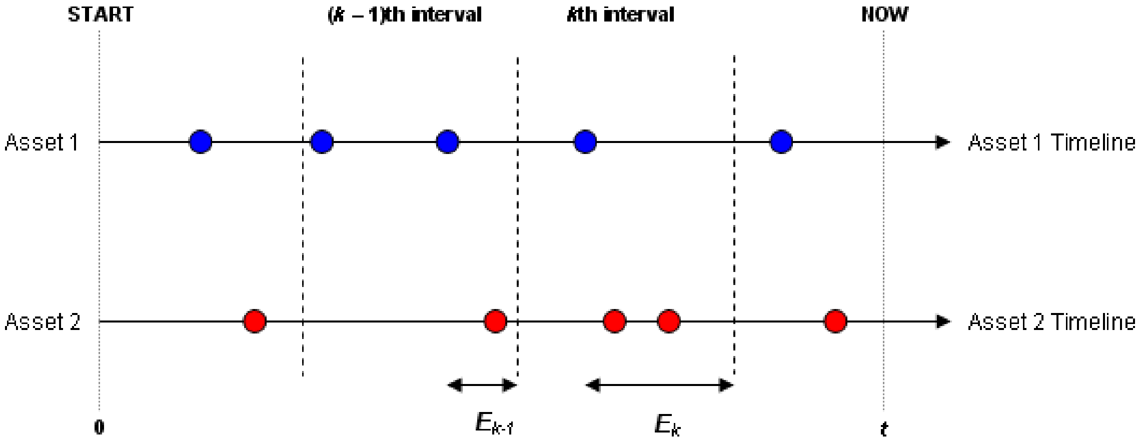
before the last asset 1 event in the kth interval of length Δt = e.g. 1

The problem is that, in the diagram, $1 = \Delta N^*_{2,k} \neq \Delta N_{2,k} = 2$. Intuitively, this occurs because the former quantity is not defined over a fixed time interval Δt but instead between last asset 1 events in consecutive intervals / the latter quantity has no dependence on asset 1.

A solution is to express $\Delta N^*_{2,k}$ in terms of $\Delta N_{2,k}$ and two "end effects"; an end effect is the number of asset 2 events which occur after the last asset 1 event in an interval. Denote the sequence of end effects as $\{E_k\}$. In Figure A.5 below, E_k is the end effect in the kth interval and E_{k-1} is the end effect in the $(k-1)$th interval.

Figure A.5: End Effects

End effects, shown by E_k and E_{k-1}, for Asset 2 with respect to Asset 1.



So: $\Delta N^*_{2,k} = \Delta N_{2,k} - E_k + E_{k-1}$

Formally, $\Delta N^*_{2,k} = \Delta N_{2,k} - [N_2(k\Delta t) - N_2(t_{1,N_1(k\Delta t)})] + [N_2((k-1)\Delta t) - N_2(t_{1,N_1((k-1)\Delta t)})]$

Then:

$$
\begin{aligned}
Cov(\Delta N^*_{2,k}, \Delta N^*_{2,k+\tau}) = Cov \Bigg\{ & \Delta N_{2,k} - \underbrace{[N_2(k\Delta t) - N_2(t_{1,N_1(k\Delta t)})]}_{E_k} \\
& + \underbrace{[N_2((k-1)\Delta t) - N_2(t_{1,N_1((k-1)\Delta t)})]}_{E_{k-1}}, \\
& \Delta N_{2,k+\tau} - \underbrace{[N_2((k+\tau)\Delta t) - N_2(t_{1,N_1((k+\tau)\Delta t)})]}_{E_{k+\tau}} \\
& + \underbrace{[N_2((k+\tau-1)\Delta t) - N_2(t_{1,N_1((k+\tau-1)\Delta t)})])}_{E_{k+\tau-1}} \Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
&= Cov(\Delta N_{2,k}, \Delta N_{2,k+\tau}) \\
&\quad - Cov(\Delta N_{2,k+\tau}, E_k) + Cov(\Delta N_{2,k+\tau}, E_{k-1}) \\
&\quad - Cov(\Delta N_{2,k}, E_{k+\tau}) + Cov(\Delta N_{2,k}, E_{k+\tau-1}) \\
&\quad + Cov(E_k, E_{k+\tau}) - Cov(E_k, E_{k+\tau-1}) \\
&\quad - Cov(E_{k-1}, E_{k+\tau}) + Cov(E_{k-1}, E_{k+\tau-1})
\end{aligned}
$$

We are only interested in the asymptotic covariances. We claim that, since each E_k is bounded from above by each $\Delta N_{2,k}$, then as $\tau \to \infty$, all of the covariance terms are of order $O(\tau^{2d_{x_2}-1})$.

Then we are left with $Cov(\Delta N^*_{2,k}, \Delta N^*_{2,k+\tau}) \sim Cov(\Delta N_{2,k}, \Delta N_{2,k+\tau})$. $\square$

Chapter 5: Conclusion and Further Research

5.1 Conclusion

An outline of the contribution of this thesis is provided in this section. A summary is presented first, with further detail in the final subsection.

5.1.1 Summary

This thesis has developed further understanding of the three main channels of information in high frequency data - times, prices and volumes, while also enabling fitting of the long memory property.

In terms of the time process, we provided an empirical comparison of the leading candidate models which allow for long memory. We found that the LMSD and MSMD models both have strengths which enable superior performance in forecasting ability, estimation speed and in-sample fit. For the volume process, we have conducted an empirical investigation showing when ARFIMA processes model the data well, and that deseasonalisation can artificially induce a pattern of decreasing memory with temporal aggregation. Finally, with respect to the price process, an extension to the theoretical system of proofs of Hurvich et al. (Deo, Hurvich, Soulier and Wang (2009), Deo, Hsieh and Hurvich (2010), Hurvich and Wang (2010)) has been made, to show that long memory exists in the squared returns of a cointegrated system of two assets. We have also included further proofs which explore volumes and connect prices to volumes. In addition, methods of processing the large high frequency datasets over multiple time horizons have been determined.

5.1.2 Detail

This thesis set out to gain insights into the channels of information at the highest frequencies. The following research gaps were identified:

1. A systematic empirical comparison of the estimation and forecasting properties of duration models, particularly those allowing for high order autocorrelation.
2. An investigation of the volume process, with a view to understanding how its properties vary across time horizons.
3. A baseline model connecting irregular times, prices and volumes, and enabling long memory and fractional cointegration.

In Chapter 2, we compared the ACD model of Engle and Russell (1998), the FIACD of Jasiak (1999), the LMSD model of Deo, Hsieh and Hurvich (2010) and an adapted form of Calvet and Fisher's MSM model (2004), which we called the MSMD model. We found that the FIACD model generally performed the worst, while overall, the ACD model was consistently strong so is arguably best. However, the MSMD was best in-sample, while the LMSD was arguably best out-of-sample, especially in the presence of high order serial correlation.

In Chapter 3, we investigated the volume process using univariate time series analysis rather than linking it to price volatility as is general practice. We found a pattern of stable long memory over all time horizons, which is theoretically consistent with the work of Chambers (1998). However, different ARFIMA orders were found over different time horizons, indicating that aggregation effects exist across horizons. Finally, we found that deseasonalisation can artificially induce a pattern of decreasing memory as the time horizon expands, which may invalidate further inference. This is problematic as researchers who concentrate on only one frequency may be unaware of this when deseasonalising.

In Chapter 4, we produced a set of propositions to extend a system enabling cointegration between high frequency prices (with irregular event times) created by Hurvich and Wang (2010). We extended to include volumes and link them with prices, allowing for irregular event times, long memory and fractional cointegration. We showed that the squared returns of the cointegrated system of prices features long memory, thereby extending the propagation mechanism of Deo, Hurvich, Soulier and Wang (2009) from a univariate to multivariate setting. We also demonstrated that volumes with a higher trend than prices can feature cointegration and that time-varying conditional correlation between variables over fixed intervals can be generated by constant conditional correlation at a tick level. This may enable better forecasting of time-varying correlation between variables, and also shows that such correlation may be purely an irregular time effect.

In the course of the research, we created a computing framework to process transaction-level datasets over multiple time horizons with a size of a billion data items, when research teams of 3 people might usually conduct similar implementation and analysis. We also produced a mathematical lemma which to the best of our knowledge is original; concerning two point processes, say A and B, the asymptotic autocovariance of a counting measure of the frequency of events of A with respect to events of B is actually the same as the asymptotic autocovariance of a counting measure which simply focusses on A.

5.2 Further Research

Further work is required to develop a greater understanding of high frequency data series, and may form part of a post doctoral research agenda. Principally, 1) the insights gained in this thesis can be extended, 2) work can also be done on applications of high frequency processes for times, prices and volumes, and 3) research can be conducted into what drives the information processes underlying high frequency data.

In terms of point 1), it would be interesting to investigate whether long memory in durations (or at least high order autocorrelation mimicking it) generated by the MSMD model can theoretically propagate to long memory in realized volatility, as shown for the LMSD model by Deo, Hurvich, Soulier and Wang (2009). The work could be further extended by considering different datasets and different distributions for the innovation terms driving durations processes, such as the Burr and Generalised Gamma distributions. It would also be useful to investigate alternatives to the duration modelling framework, such as intensity modelling as in the Autoregressive Conditional Intensity model of Russell (2001) or the Stochastic Conditional Intensity model of Bauwens and Hautsch (2006). Finally, it could be informative to apply the approach of Bialkowski et al. (2008) to split the volume process in our univariate analysis into systematic and idiosyncratic components (as is the case for the CAPM), but over many time horizons. Cross-sectional aggregation of volumes (across stocks) could also be further explored.

With respect to applications, it would be useful to compare the models of Chapter 2 in terms of their performance in forecasting realized volatility. It would also be interesting to analyse the relationship between cointegration from a transaction level and option pricing, hedging, pairs trading and index tracking as suggested by Hurvich and Wang (2010). Focussing purely on the irregular time process, the ideas of Prigent, Renault and Scaillet

(2000) could be explored further to examine option pricing via lattice models with irregular spacing between nodes.

Finally, it would be interesting to explore point process theory further to gain more insight into the structure of information processes which might underlie irregular times, prices and volumes. Operations on the information processes may be leading to properties in high frequency data; e.g. aggregation (”superposition” in Cox and Isham (1980)) of point processes leads to a new process - so following the logic in the other direction, we may be able to reverse engineer the information processes.

BIBLIOGRAPHY

Andersen, T. (1996). "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility". *Journal of Finance*, pages 169–204.

Andersen, T., Bollerslev, T., Diebold, F., & Labys, P. (2003). "Modeling and Forecasting Realized Volatility". *Econometrica*, volume 71(2):pages 579–625.

Andersen, T., Bollerslev, T., & Meddahi, N. (2004). "Analytical Evaluation of Volatility Forecasts". *International Economic Review*, volume 45(4):pages 1079–1110.

Andersen, T. G., & Bollerslev, T. (1996). "Heterogeneous Information Arrivals and Return Volatility Dynamics: Uncovering the Long-Run in High Frequency Returns". *SSRN eLibrary*.

Baillie, R., & Kapetanios, G. (2008). "Nonlinear Models for Strongly Dependent Processes with Financial Applications". *Journal of Econometrics*, volume 147(1):pages 60–71.

Baillie, R., Nijman, T., & Tschernig, R. (1994). "Temporal Aggregation of Fractionally Integrated ARMA Models". Technical report.

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). "Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity". *Journal of Econometrics*, volume 74(1):pages 3–30.

Barkoulas, J., & Baum, C. (1996). "Long-Term Dependence in Stock Returns". *Economics Letters*, volume 53(3):pages 253–259.

Bauwens, L., & Hautsch, N. (2006). "Stochastic Conditional Intensity Processes". *Journal of Financial Econometrics*, volume 4(3):page 450.

Bauwens, L., & Veredas, D. (1999). "The Stochastic Conditional Duration Model: a Latent Factor Model for the Analysis of Financial Durations". Technical report.

Bauwens, L., & Veredas, D. (2004). “The Stochastic Conditional Duration Model: a Latent Variable Model for the Analysis of Financial Durations”. *Journal of Econometrics*, volume 119(2):pages 381–412.

Beran, J. (1994). *Statistics for Long-Memory Processes*. Chapman & Hall.

Bialkowski, J., Darolles, S., & Le Fol, G. (2008). “Improving VWAP Strategies: A Dynamic Volume Approach”. *Journal of Banking & Finance*, volume 32(9):pages 1709–1722. ISSN 0378-4266.

Bollerslev, T., & Jubinsky, D. (1999). “Equity Trading Volume and Volatility: Latent Information Arrivals and Common Long-Run Dependencies”. *Journal of Business and Economic Statistics*.

Bollerslev, T., & Wright, J. (2001). “High-Frequency Data, Frequency Fomain Inference, and Volatility Forecasting”. *Review of Economics and Statistics*, volume 83(4):pages 596–602.

Breidt, F. J., Crato, N., & de Lima, P. (1998). “The Detection and Estimation of Long Memory in Stochastic Volatility”. *Journal of Econometrics*, volume 83(1-2):pages 325–348.

Brockwell, P., & Davis, R. (1991). *Time Series: Theory and Methods (Springer Series in Statistics)*. Springer-Verlag Berlin and Heidelberg GmbH & Co. K, 2nd rev edition.

Calvet, L., & Fisher, A. (2004). “How to Forecast Long-Run Volatility: Regime Switching and the Estimation of Multifractal Processes”. *Journal of Financial Econometrics*, volume 2(49-83).

Calvet, L., & Fisher, A. (2008). *Multifractal Volatility: Theory, Forecasting, and Pricing*. Academic Press. ISBN 0121500136.

Carr, P., & Wu, L. (2002). “Time-changed Lévy Processes and Option Pricing”. *Journal of Financial Economics*, volume 71(1):pages 113–141.

Carr, P., & Wu, L. (2003). "The Finite Moment Log Stable Process and Option Pricing". *The Journal of Finance*, volume 58(2):pages 753–778.

Chambers, M. (1998). "Long Memory and Aggregation in Macroeconomic Time Series". *International Economic Review*, volume 39(4):pages 1053–1072. ISSN 0020-6598.

Chan, E. (2009). *Quantitative Trading*. John Wiley and Sons, New Jersey.

Chan, N., & Palma, W. (1998). "State Space Modeling of Long-Memory Processes". *The Annals of Statistics*, volume 26(2):pages 719–740.

Clark, P. (1973). "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices". *Econometrica*, volume Vol. 41(1):pages p. 135–155.

Cox, D., & Isham, V. (1980). *Point Processes*. Chapman & Hall.

Daley, D., Rolski, T., & Vesilo, R. (2000). "Equivalence of Functional Limit Theorems for Stationary Point Processes and Their Palm Distributions". *Advances in Applied Probability*, volume Vol. 32.

Daley, D., & Vere-Jones, D. (2003). *Introduction to the Theory of Point Processes, Elementary Theory and Methods v. 1*. Springer, 2nd edition. ISBN 0387955410.

Darolles, S., & Le Fol, G. (2003). "Trading Volume and Arbitrage". *Working Papers*.

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford University Press, USA.

Deo, R., Hsieh, M., & Hurvich, C. M. (2006a). "The Persistence of Memory: From Durations to Realized Volatility". Technical report.

Deo, R., Hsieh, M., & Hurvich, C. M. (2010). "Long Memory in Intertrade Durations, Counts and Realized Volatility of NYSE stocks". *Journal of Statistical Planning and Inference*, volume 140(12):pages 3715 – 3733. doi:DOI:10.1016/j.jspi.2010.04.037. Special

Issue in Honor of Emanuel Parzen on the Occasion of his 80th Birthday and Retirement from the Department of Statistics, Texas A&M University - Emmanuel Parzen.

Deo, R., Hurvich, C., & Lu, Y. (2006b). "Forecasting Realized Volatility using a Long-Memory Stochastic Volatility Model: Estimation, Prediction and Seasonal Adjustment". *Journal of Econometrics*, volume 131(1-2):pages 29–58.

Deo, R., Hurvich, C. M., Soulier, P., & Wang, Y. (2009). "Conditions for the Propagation of Memory Parameter from Durations to Counts and Realized Volatility". *Econometric Theory*, volume 25(3):pages 764–792.

Diebold, F., & Inoue, A. (2001). "Long Memory and Regime Switching". *Journal of Econometrics*, volume 105(1):pages 131–159.

Douc, R., Roueff, F., & Soulier, P. (2008). "On the Existence of Some $\mathrm{ARCH}(\infty)$ Processes". *Stochastic Process. Appl.*, volume 118(5):pages 755–761. ISSN 0304-4149.

Easley, D., & O'Hara, M. (1992). "Time and the Process of Security Price Adjustment". *Journal of Finance*, volume 47(2):pages 576–605.

Engle, R. F. (2000). "The Econometrics of Ultra-High Frequency Data". *Econometrica*, volume Vol. 68(1):pages p. 1–22.

Engle, R. F., & Russell, J. R. (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Time Series Data". Technical report, Department of Economics, UC San Diego.

Engle, R. F., & Russell, J. R. (2005). "A Discrete-State Continuous-Time Model of Financial Transactions Prices and Times: The Autoregressive Conditional Multinomial-Autoregressive Conditional Duration Model". *Journal of Business & Economic Statistics*, volume 23(2):pages 166–180.

Epps, T., & Epps, M. (1976). "The Stochastic Dependence of Security Price Changes and Transaction Volumes: Implications for the Mixture-of-Distributions Hypothesis". *Econometrica*, volume 44(2):pages 305–321.

Fama, E. (1965). "The Behavior of Stock-Market Prices". *Journal of Business*, volume 38(1):page 34.

Fernandes, M., & Grammig, J. (2005). "Nonparametric Specification Tests for Conditional Duration Models". *Journal of Econometrics*, volume 127(1):pages 35–68.

Fernandes, M., & Grammig, J. (2006). "A Family of Autoregressive Conditional Duration Models". *Journal of Econometrics*, volume 130(1):pages 1–23.

Fleming, J., & Kirby, C. (2001). "Long Memory in Volatility and Trading Volume". Working paper series.

Friedman, M. (1953). *Essays in Positive Economics*. University of Chicago Press.

Gallant, A., Rossi, P., & Tauchen, G. (1992). "Stock Prices and Volume". *Review of Financial Studies*, pages 199–242.

Geweke, J., & Porter-Hudak, S. (1983). "The Estimation and Application of Long Memory Time Series Models". *Journal of Time Series Analysis*.

Granger, C. (1980). "Long Memory Relationships and The Aggregation of Dynamic Models". *Journal of Econometrics*, volume 14(2):pages 227–238.

Granger, C. (2001). "Long Memory Relationships and the Aggregation of Dynamic Models". *Essays in Econometrics: Collected Papers of Clive WJ Granger*, volume 14:pages 227–238.

Hamilton, J. (1994). *Time Series Analysis*. Princeton University Press.

Harris, L. (1987). “Transaction Data Tests of the Mixture of Distributions Hypothesis”. *Journal of Financial and Quantitative Analysis*, volume 22(2):pages 127–141.

Hurvich, C. M., & Wang, Y. (2010). “A Pure-Jump Transaction-Level Price Model Yielding Cointegration, Leverage, and Nonsynchronous Trading Effects”. *Journal of Business & Economic Statistics*, volume 28(4):pages 539–558.

Ishida, I., & Watanabe, T. (2009). “Modeling and Forecasting the Volatility of the Nikkei 225 Realized Volatility using the ARFIMA-GARCH model”. *Global COE Hi-Stat Discussion Paper Series*.

Jain, P., & Joh, G. (1988). “The Dependence between Hourly Prices and Trading Volume”. *Journal of Financial and Quantitative Analysis*, volume 23(3):pages 269–283.

Jasiak, J. (1999). “Persistence in Intertrade Durations”. *SSRN eLibrary*. doi:10.2139/ssrn.162008.

Kanzler, L. (1998). “GPH: MATLAB Module to Calculate Geweke-Porter-Hudak Long Memory Statistic”. Statistical Software Components, Boston College Department of Economics.

Karanasos, M., Psaradakis, Z., & Sola, M. (2004). “On the Autocorrelation Properties of Long-Memory GARCH Processes”. *Journal of Time Series Analysis*, volume 25(2):pages 265–282. ISSN 1467-9892.

Karpoff, J. (1987). “The Relation between Price Changes and Trading Volume: A Survey”. *Journal of Financial and Quantitative Analysis*, volume 22(1):pages 109–126.

Kon, S. (1984). “Models of Stock Returns - A Comparison”. *Journal of Finance*, volume 39(1):pages 147–165.

Künsch, H. (1986). “Discrimination between Monotonic Trends and Long-Range Dependence”. *Journal of Applied Probability*, pages 1025–1030.

Lamoureux, C., & Lastrapes, W. (1994). "Endogenous Trading Volume and Momentum in Stock-Return Volatility". *Journal of Business & Economic Statistics*.

Li, W., & Yu, P. (2003). "On the Residual Autocorrelation of the Autoregressive Conditional Duration Model". *Economics Letters*, volume 79(2):pages 169–175.

Liesenfeld, R. (2001). "A Generalized Bivariate Mixture Model for Stock Price Volatility and Trading Volume". *Journal of Econometrics*, volume 104(1):pages 141–178.

Lo, A., Mamaysky, H., & Wang, J. (2001). "Asset Prices and Trading Volume under Fixed Transactions Costs".

Lo, A., & Wang, J. (2001). "Trading Volume: Implications of an Intertemporal Capital Asset Pricing Model".

Lo, A., & Wang, J. W. (2000). "Trading Volume: Definitions, Data Analysis, and Implications of Portfolio Theory". Nber working papers, National Bureau of Economic Research, Inc.

Lobato, I., & Velasco, C. (2000). "Long Memory in Stock Market Trading Volume". *Journal of Business and Economic Statistics*, volume 18(4):pages 410–427.

Luenberger, D. (1998). "Investment Science".

Lux, T. (2008). "The Markov-Switching Multifractal Model of Asset Returns: GMM Estimation and Linear Forecasting of Volatility". *Journal of Business and Economic Statistics*, volume 26(2):pages 194–210.

Man, K., & Tiao, G. (2006). "Aggregation Effect and Forecasting Temporal Aggregates of Long Memory Processes". *International Journal of Forecasting*, volume 22(2):pages 267–281.

Manchaldore, J., Palit, I., & Soloviev, O. (2010). "Wavelet Decomposition for Intra-Day Volume Dynamics". *Quantitative Finance*, volume 10(8):pages 917–930.

Mandelbrot, B. (1963). "The Variation of Certain Speculative Prices". *Journal of Business*, volume 36(4):page 394.

Manganelli, S. (2002). "Duration, Volume and Volatility Impact of Trades". Ecb working paper no. 125, ECB.

Ng, W. L. (2008). "High Frequency Finance and Computational Market Micro-Structure Lecture Notes". Technical report.

Nieuwenhuis, G. (1989). "Equivalence of Functional Limit Theorems for Stationary Point Processes and Their Palm Distributions". *Probability Theory and Related Fields*, volume 81.

Ohanissian, A., Russell, J., & Tsay, R. (2008). "True or Spurious Long Memory? A New Test". *Journal of Business and Economic Statistics*, volume 26(2):pages 161–175.

Oomen, R. (2006). "Properties of Realized Variance under Alternative Sampling Schemes". *Journal of Business and Economic Statistics*, volume 24.

Pacurar, M. (2006). "Autoregressive Conditional Duration (ACD) Models in Finance: A Survey of the Theoretical and Empirical Literature". *SSRN eLibrary*.

Pierce, D., & Haugh, L. (1977). "Causality in Temporal Systems:: Characterization and a Survey". *Journal of Econometrics*, volume 5(3):pages 265–293.

Pindyck, R., & Rubinfeld, D. (1981). "Econometric Models and Economic Forecasts".

Press, S. (1967). "A Compound Events Model for Security Prices". *Journal of Business*, volume 40(3):page 317.

Prigent, J.-L., Renault, O., & Scaillet, O. (2000). "An Autoregressive Conditional Binomial Option Pricing Model". Fmg discussion papers, Financial Markets Group.

Renault, E., & Werker, B. (2011). "Causality Effects in Return Volatility Measures with Random Times". *Journal of Econometrics*, volume 160(1):pages 272–279.

Richardson, M., & Smith, T. (1994). "A Direct Test of the Mixture of Distributions Hypothesis: Measuring the Daily Flow of Information". *Journal of Financial and Quantitative Analysis*, volume 29(1).

Rodríguez-Poo, J. M., Veredas, D., & Espasa, A. (2007). "Semiparametric Estimation for Financial Durations". In W. P. Luc Bauwens, & D. Veredas, editors, "High Frequency Financial Econometrics. Recent Developments", Springer Verlag.

Russell, J. (2001). "Econometric Modeling of Multivariate Irregularly-Spaced High-Frequency Data". *Unpublished manuscript, University of Chicago, Graduate School of Business*.

Rydberg, T. (2000). "Realistic Statistical Modelling of Financial Data". *International Statistical Review*, volume 68(3):pages 233–258.

Tauchen, G., Zhang, H., & Liu, M. (1996). "Volume, Volatility, and Leverage: A Dynamic Analysis". *Journal of Econometrics*, volume 74(1):pages 177–208.

Tauchen, G. E., & Pitts, M. (1983). "The Price Variability-Volume Relationship on Speculative Markets". *Econometrica*, volume 51(2):pages 485–505.

Tsai, H., & Chan, K. (2005). "Temporal Aggregation of Stationary And Nonstationary Discrete-Time Processes". *Journal of Time Series Analysis*, volume 26(4):pages 613–624.

Tsay, R. S. (2005). *Analysis of Financial Time Series (Wiley Series in Probability and Statistics)*. John Wiley & Sons, 2nd edition. ISBN 0471690740.

Tschernig, R. (1994). “Long Memory in Foreign Exchange Rates Revisited”. *Institute of Statistics and Econometrics. Humboldt University of Berlin*.

Zaffaroni, P. (2004). “Contemporaneous Aggregation of Linear Dynamic Models in Large Economies”. *Journal of Econometrics*, volume 120(1):pages 75–102.