Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/19/20230159

Pattern Recognition of Stock Returns in the Very Short Run Leveraging High-Frequency Financial Data

Yan Gao[1,a,*]

[1]Ocean University of China, Shandong Province, China
a.rgao0829@gmail.com
**corresponding author*

Abstract: This paper uses high frequency stock trading data from 2007 to 2014 to study patterns of stock returns in the very short run. The paper verifies stylized facts of stock prices at low frequency with the exploration of high-frequency data and uses the concept of relative realized volatility (RRV) to measure volatility to understand the market uncertainty intraday. By providing a large number of empirical data facts, this paper advocates the use of ultra-high frequency data to study instantaneous real volatility, and demonstrates that long-term market volatility and the relationship between short and long term volatility can be implied by a simple RRV mean regression model.

Keywords: volatility, high-frequency, stock price, market crash

1. Introduction

High-frequency trading has become a dominant force in the U.S. capital market, accounting for more than 70% of U.S. daily dollar trading volume and half of all U.S. equity trading [1]. It is uncontroversial that high-frequency trading has changed the way of doing finance in the past several decades. High-frequency trading refers to fully automated trading strategies that trade very large volumes and hold them for very short periods, ranging from milliseconds to minutes or even hours [2]. The strategy of big data has been widely used in every aspects of research on HFT but the fact is that researchers still fail to use them very well and understand intraday volatility despite they have lots of trading and data. A great deal of effort has been spent on the modeling of volatility processes. Many volatility models have been constructed to verify and predict financial phenomena. For example, Bollerslev and Zhou used Deutsche marks. Sun.M reformulated the Heston stochastic volatility model as a high frequency evolution model with proportional increment of secondary variation. All of these studies, however, focused on quadratic changes in trading days and samples of prices every five minutes. In contrast, this article focuses on quadratic behavior within short intervals (100 seconds) during the trading day. By calculating realized volatility every 100 second interval, we are able to examine the behavior of random volatility over a trading day.

The study of volatility and its dynamic change is an important direction of finance. Volatility is commonly used to measure risk, drive the construction of an optimal portfolio, and determine the value of a company in the presence of various risk factors.This study focuses on the influence of high-frequency trading on stock price fluctuations and tries to find out the reasons for stock price changes in the short term. Traditional way of studying intraday stock price volatility has been bothered by market microstructure noise. People rely on the random walk price process to investigate the efficient stock price as well as calibrating the noise process. The price sampling frequency stops at 5 minute. However, this paper tries to sample prices as often as we can and treat the measurement error of prices as errors in variables to develop an equilibrium model of asset pricing. The paper assumes prices follow an affine-random walk model and we try to explain more data with specifications on the volatility process. Moreover, this paper will also study outliers of the intraday volatility equilibrium and testify the existence of information shock.

2. Literature Review

High-frequency trading (HFT) refers to computerized trading that seeks to profit from extremely brief market changes that people cannot take advantage of, such as the small change in the difference between the price at which a security is bought and the price at which it is sold, or the small difference between the price of a stock on different exchanges. HFT is fast and trades large volumes. Researchers build various kinds of trading models to explore the volatility of stocks. Volatility plays a key role in identifying investment opportunities. Researchers often expend a great deal of effort in modeling the volatility of stock prices. For example, the continuous time stochastic volatility (SV) model is widely used in the empirical financial literature. Heston is a popular SV model because it is simple and easy to analyze [3]. Fabien Guilbaud and Huyen Pham proposed a new idea to study optimal market-making strategies in limit order books (LOB) [4]. They assumed that a small agent would continuously submit limit sell orders at the best purchase price. Engle and Ghysels studied a series of historical data on long-term stock market volatility starting from the 19th century [5]. They found that both pure time series models and models driven by economic variables showed structural breaks in the entire sample of everyday data spanning about a century and a half. Bollerslev and Zhou estimated their model using the Deutschmark USD spot exchange rates [6]. Sun.M tried to explore the volatility of stock prices in a trading day, and improved Heston stochastic volatility model into a high frequency evolution model with quadratic variational proportional increment [7]. The strategy of big data has been widely used in many studies on HFT. Big data algorithms are often applied to trading decisions in HFT. Since algorithmic trading is conducted on a large scale, it will greatly affect market liquidity, resulting in increased volatility. Therefore, while big data algorithmic trading can improve efficiency and reduce transaction costs, it operates in a way that is potentially harmful to the liquidity and overall stability of forex markets.

Despite the fact that different models have different perspectives as well as strategies, all of these researches demonstrate that there is a strong link between market crash and high volatility, which is also the emphasis of this paper.

3. Data and Analysis

3.1. Initial Data Processing

The data comes from the TAQ NYSE trading data in the WRDS database and it contains 8 years of high-frequency (within second) price data. In this paper, the first and most basic thing to do is calculate log price returns per second, realized volatility and realized volatility over 100 second block and create a data frame to record the above numbers for each day. Realized volatility is theoretically a natural representation of the underlying composite volatility of the stock price process [8]. In practice, realized volatility has not worked as expected, and this failure is usually attributed to errors in stock prices. The market opens at 9:30 am and ends at 4 pm, so there are 23,401 seconds in between and hence 23,401 rows in the data frame.

For initial data process, this paper creates two big data matrices, one for log returns and the other for realized volatility. The log returns are for every second, the realized volatility is for every 100 second. The log returns is calculated by taking the log on the prices and taking the difference. The realized volatility is the squared log return, first the paper accumulates the squared log returns and then take the difference every 100 second. Besides, the RV is always positive for the reason that squared log returns of 100th second should always be larger than the squared log returns of the 1st second. RRV is also constructed to see the relative change during the day and it is defined by dividing RV by its daily mean.

Table 1: Basic statistics of log returns and RRV.

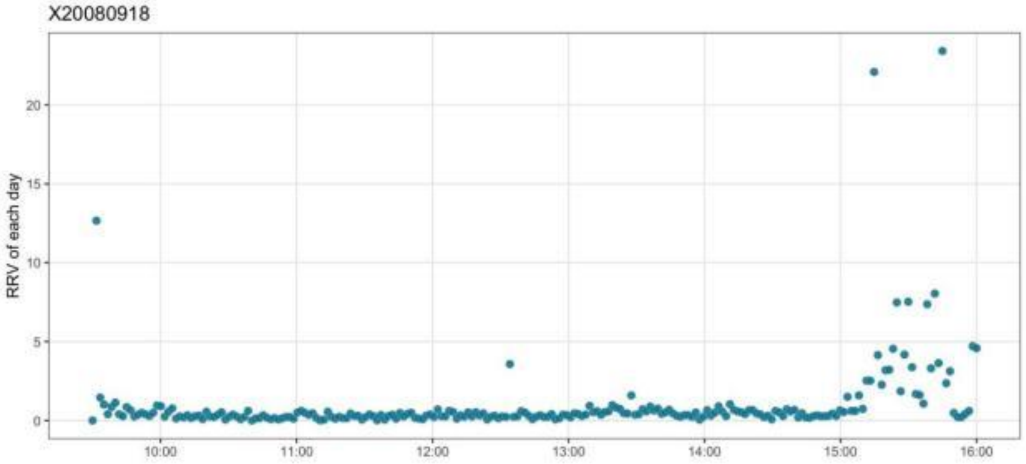
	mean	median	variance	standard variance	Q1	Q3
Log returns	0.073578	0	11.497856	3.390848	0	0
RRV	1	0.594919	6.686287	2.585785	0.287347	1.116850

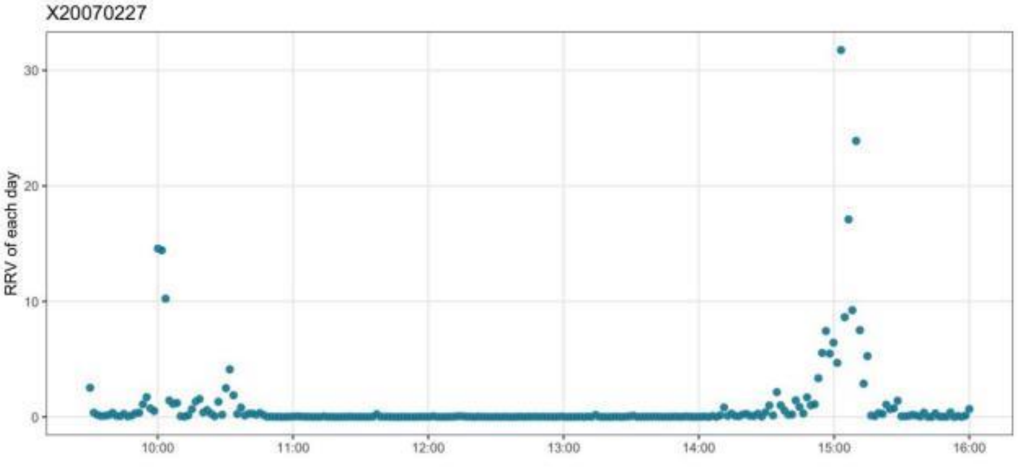
The plots of RRV for the following special days are done to see how RRV looks like and observe significant volatility clustering and jumps in the plot:

20080918 (Lehman Brothers bankruptcy)

20070227 (Chinese market crash)

20100506 (The Flash Crash)





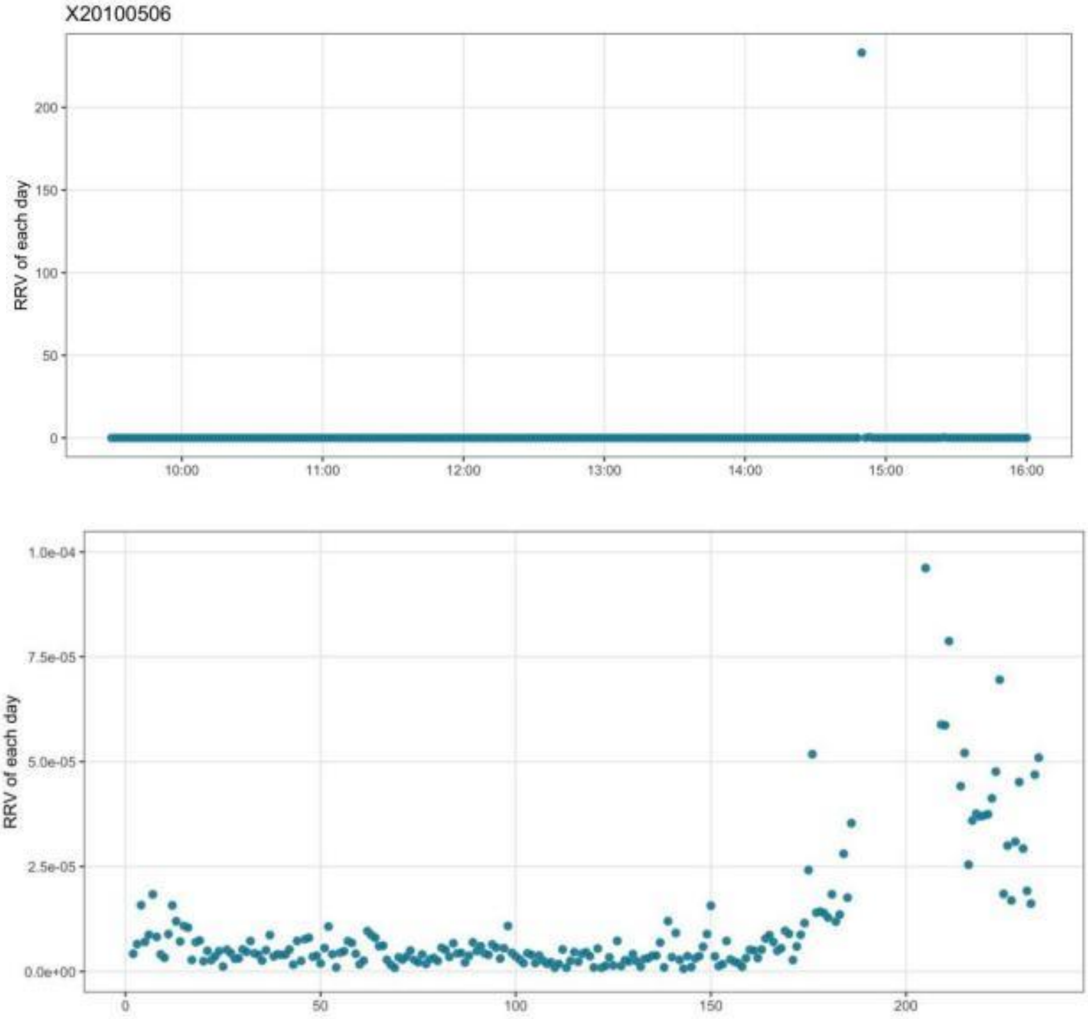


Figure 1: The plots of RRV in special days.

For the day 20100506, the y-axis is truncated to see the bottom data clearly.

It is observable that there are some significant volatility clustering and jumps in the plots. The RRV usually keeps at a low level after 9:30 a.m. But then it rises sharply at around 3 p.m. Moreover, it can be observed that one large fluctuation is often accompanied by another large fluctuation in the next period. The RRV begins to fall back around 4 p.m.

The FOMC announcement meeting is also very important because it usually talks about the interest rate and market conditions. It is held every Wednesday at the beginning of the quarter. RRVs these days are usually not very stable. This paper uses the FOMC announcement meeting schedule to sort out all the FOMC announcement days from 2007-2012. This paper plots the RRVs for these days and observe what happens around 2pm when the news releases on FFRs.

The paper has selected some FOMC announcement days which are:

20070131，20071031;

20080130，20080430，20081029;

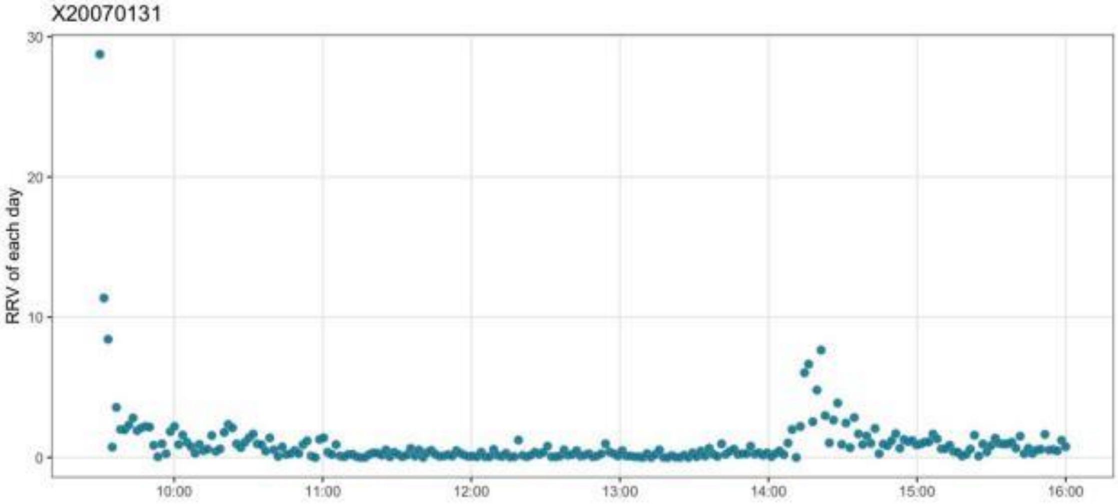
20090128，20090429;

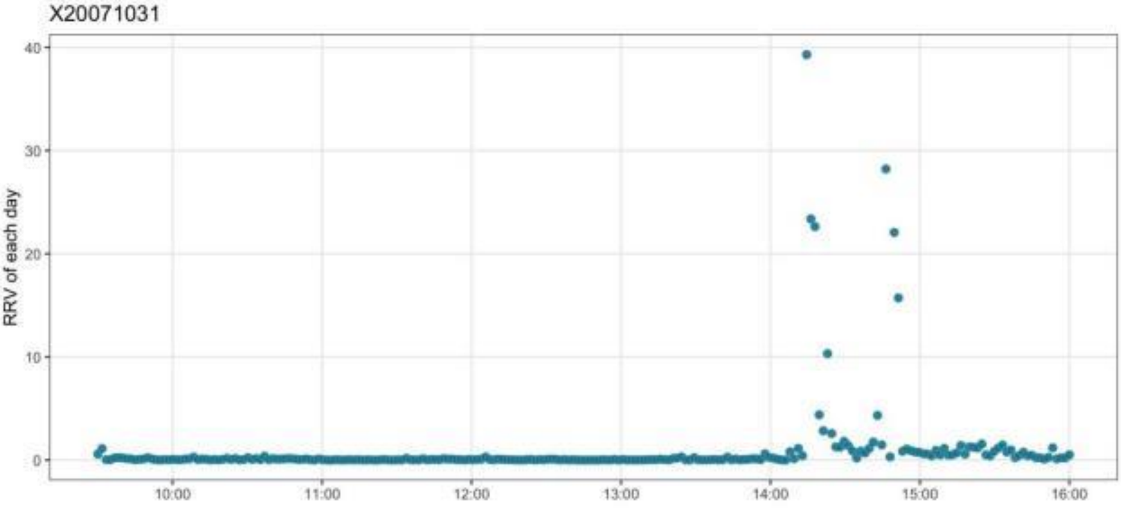
20100127，20100428;

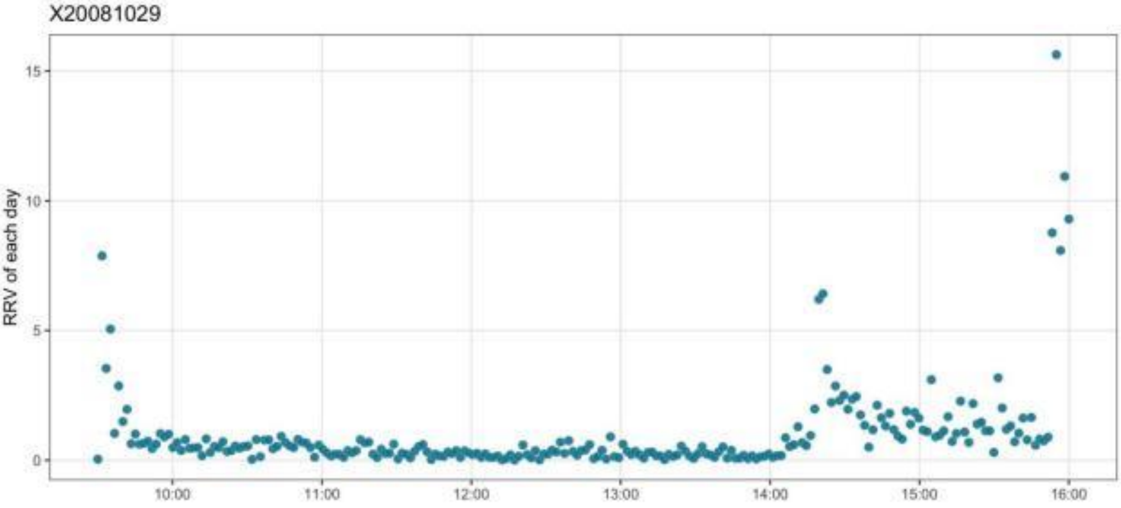
20110126，20110427;

20120125， 20120425， 20121024.

Some of the plots are presented as below:







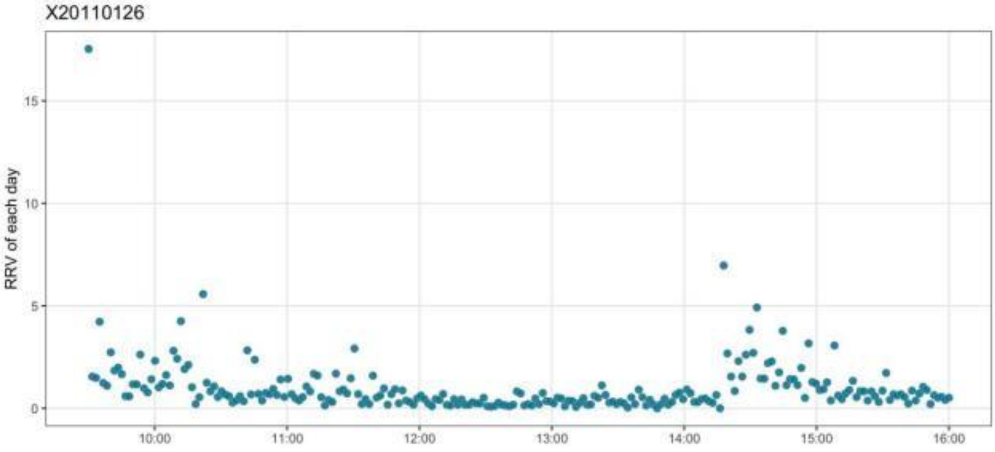


Figure 2: The images of RRV in FOMC announcement days.

3.2. Verifying Low-frequency Stylized Facts Using High-Frequency Stock Price Data

These are some important features studied in the research:

1. Irrelevance: asset prices have little correlation with returns at different times;

2. Heavy tail/super kurtosis: the empirical return distribution has fatter tails and more peak centers than the standard normal distribution;

3. Asymmetry/skewness: most observable returns are asymmetrical, favoring large deviations from the mean;

4. Jumps: large price movements demonstrate that empirical observations cannot be implied by a diffusion-type continuous process alone, but in combination with jumps;

5. Volatility aggregation: large price fluctuations in one period are often accompanied by equally large price fluctuations in the next period;

6. Leverage effect: When future volatility is supposed to be high, return tends to be negative (If the leverage increases, both the stocks and the bonds of the firm become more risky [9]).

These six features are the popular traits of low frequency financial data. This paper tend to verify them with high-frequency data.

For irrelevance, this paper calculates the correlation between log returns of consecutive days and see how the correlation changes over time.

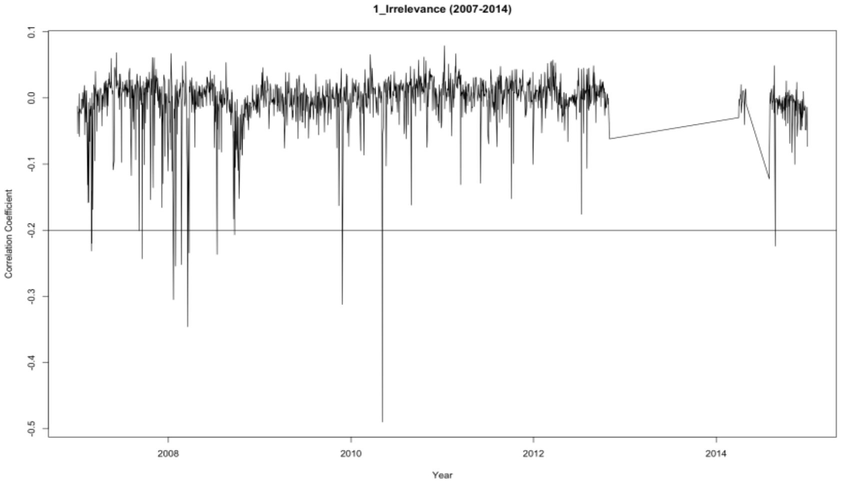


Figure 3: The plot of irrelevance between log returns of consecutive days.

The x-axis has been made to be the years and a threshold -0.2 has already been added. The data for 2013 is not available so there is a curve between 2012 and 2014. The curve starts from the last day of 2012 to the first day of 2014.

The average correlations over each year are shown below:

Table 2: The average correlations from 2007 to 2014.

Year	Average
2007	-0.008269280
2008	-0.019276533
2009	-0.005305432
2010	0.002987402
2011	0.007026777
2012	0.003859937
2014	-0.013502114

Most of the correlation coefficients fluctuate around 0 and the maximum is lower than 0.1. Only a few absolute values of specific correlations are close to 0.5. Therefore, the correlation coefficients are too small to show strong correlation. That means the irrelevancy really exists.

For fat-tail, this paper plots the distribution of log returns for the 7 years separately and examine any fat-tails.

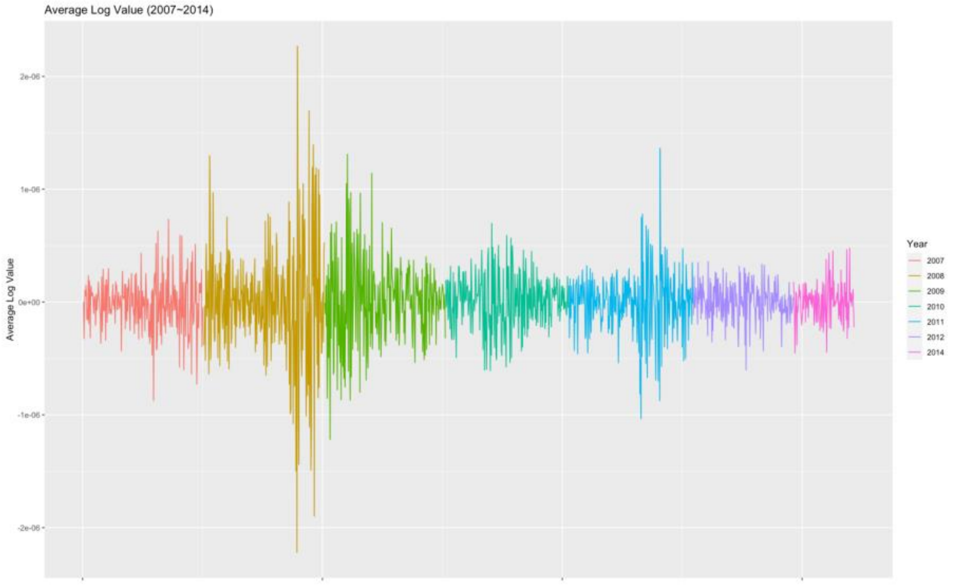


Figure 4: The plot of distribution of log returns over 7 years.

The daily average log returns of seven years have been showed in a whole histogram. The values of 2018 are larger than those of other years.

This paper also makes a histogram combining all the log returns in one year for the 6 years and the x-axis is the value of the average log returns and the y-axis is the number of days the value is in this range.

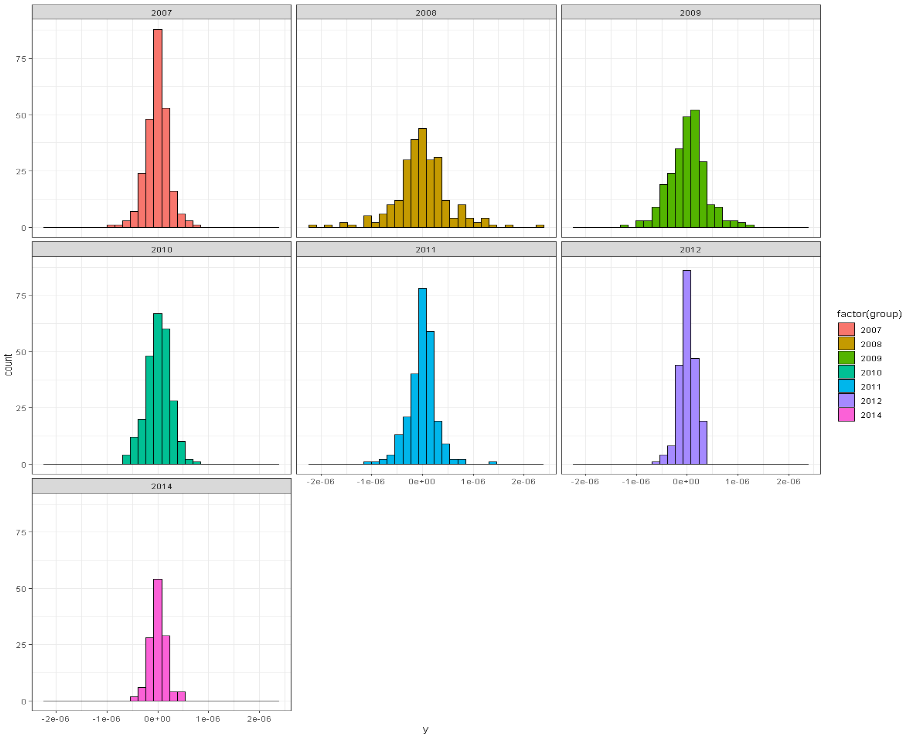


Figure 5: The histograms of log returns over 7 years.

A heavy tail is an empirical return distribution with fatter tails and more peak centers than the standard normal distribution. It can be observed that some plots have these features: high peaked centers and thicker tails. For example: 2007, 2010, 2011, 2012, 2014. In the center of the distribution, the values are extremely higher and less distributed than years of 2008 and 2009. The plots of log returns are more concentrated, not widely distributed, which means have fat-tails.

In the plot “1_irrelevance(2007-2014)”, asymmetry exists apparently. In the seven plots of the distributions of log returns, some plots show strong asymmetry, such as 2010, 2011 and 2012, and the other plots show weak symmetry.

For leverage, this paper simply plots the average daily log returns and the average daily RVs using the high-frequency data and see how they move in the same graph over the 7 years. This paper also calculates the correlation of two time series (RV number and daily log return) over months.

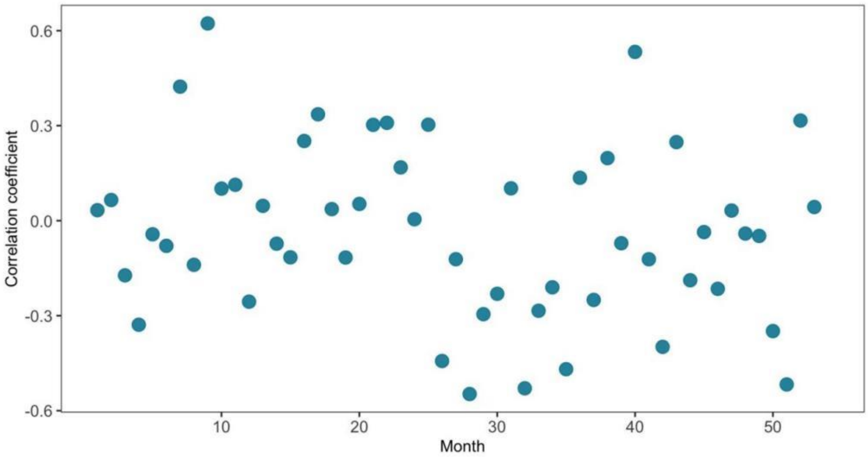


Figure 6: The plot of correlation of two time series.

The paper plots the correlation of two series (RV numbers and daily returns). Some daily log returns are negative but the RV doesn't seem to be significantly higher than the previous day. Log returns are negative in some days but RV numbers fail to be significantly higher than the previous days. Leverage effect can be used to explain the phenomenon that equity investors often overpay for risky stocks [10].

4. Distribution of RRV

Table 3: The distribution of RRV.

	[0,2.5)	[2.5,5)	[5,10)	[10,100)	[100,1000)	>1000
Distribution	3151	432	174	271	136	48
Total	312698	43092	12412	4850	364	48
Percentage	1%	1%	1.40%	5.60%	37%	100%
Days	1596	1595	1562	1399	176	18

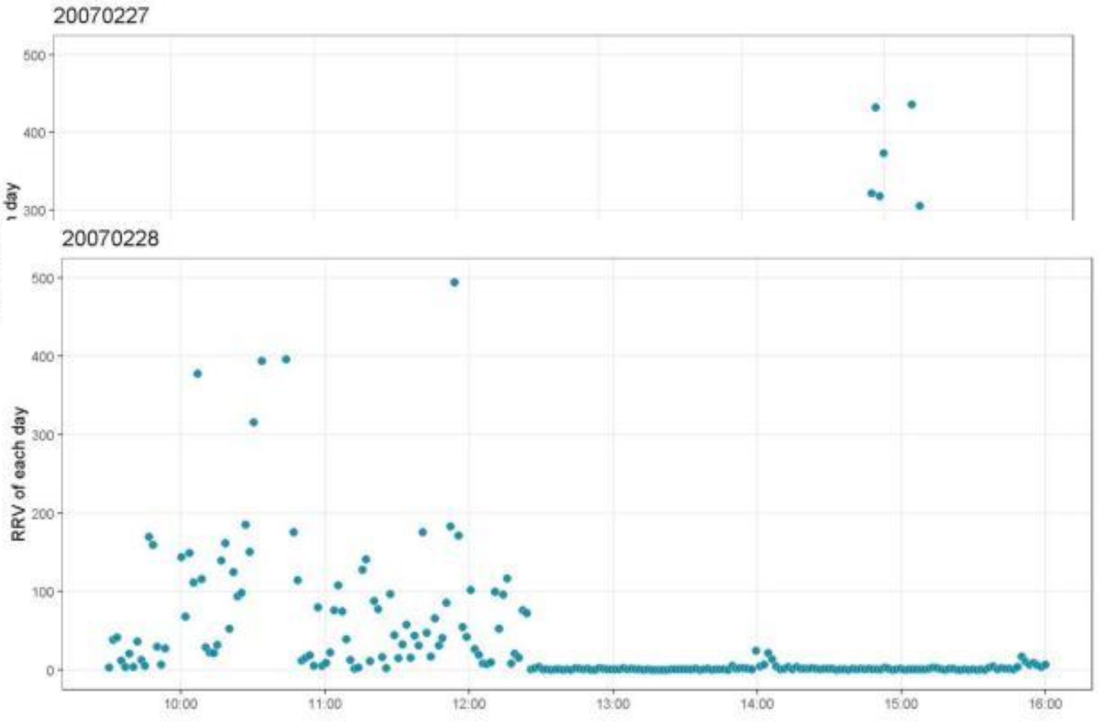
This paper divides the distribution of RRV into different intervals such as [0,2.5), [2.5,5), [5,10), [10,100), [100,1000), >1000. Two things need to be checked:

Extreme RRVs are concentrated in days with market crash events.

Extreme RRVs are followed by cascades of smaller RRVs, i.e., clusters of volatility also in the short run (The GARCH model demonstrates volatility clustering in the low frequency data but not mentioned the high-frequency data).

From the table above, there are 48 RRVs in the interval of >1000, 312698 RRVs in the interval of (0,2.5), 43092 RRVs in the interval of (2.5,5), 12412 RRVs in the interval of (5,10), 4850 RRVs in the interval of (10,100), 364 RRVs in the interval of (100,1000). The 48 RRVs larger than 1000 coming from 18 days are extreme RRV and they represent high volatility. Subsequently, this paper begins to focus on these 18 days having very large RRVs.

This paper plots the RRV of these 18 days and see if they have volatility cascades. The plots with apparent cascades are shown below:



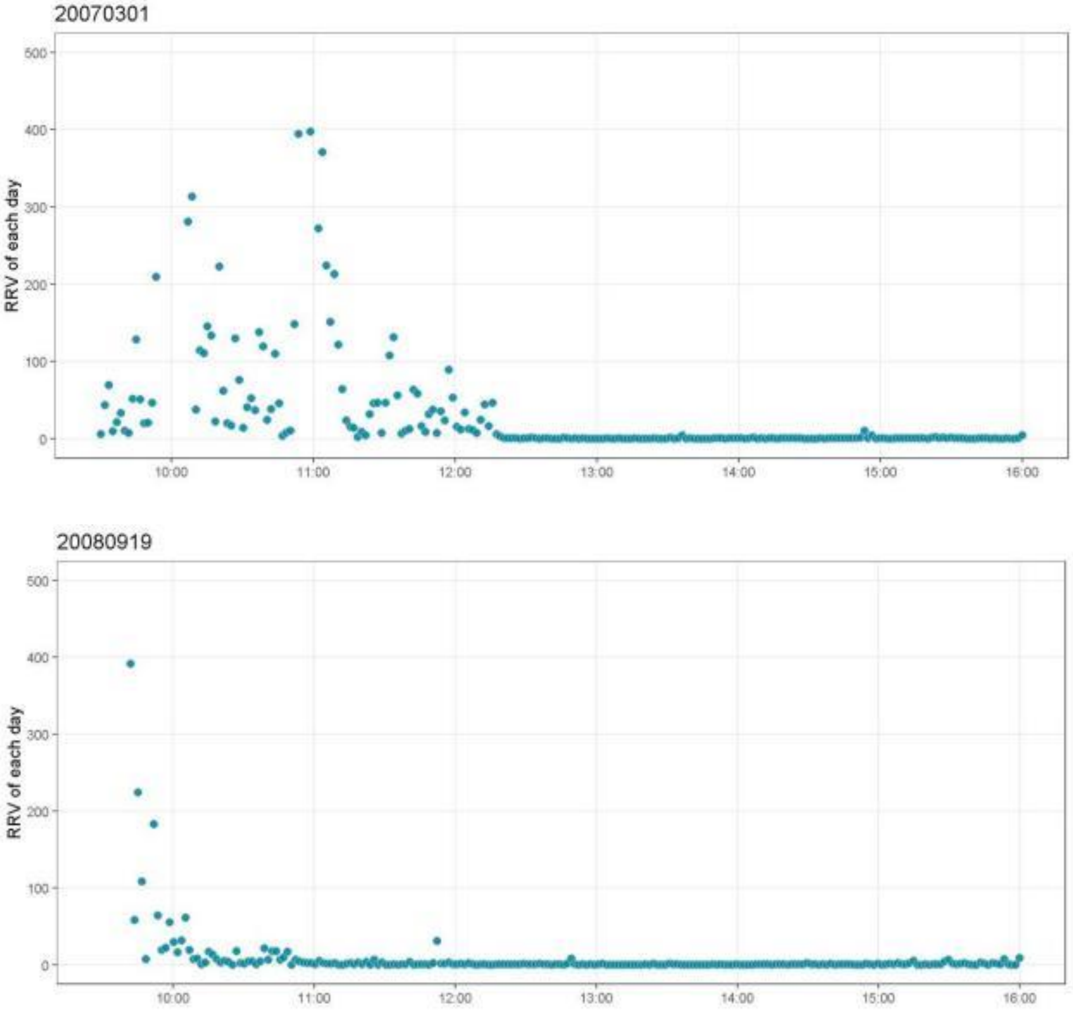


Figure 7: The plots of RRV with volatility cascades.

The plot of a regular day where there are no jumps but only the dynamic structure of RRVs is also included.

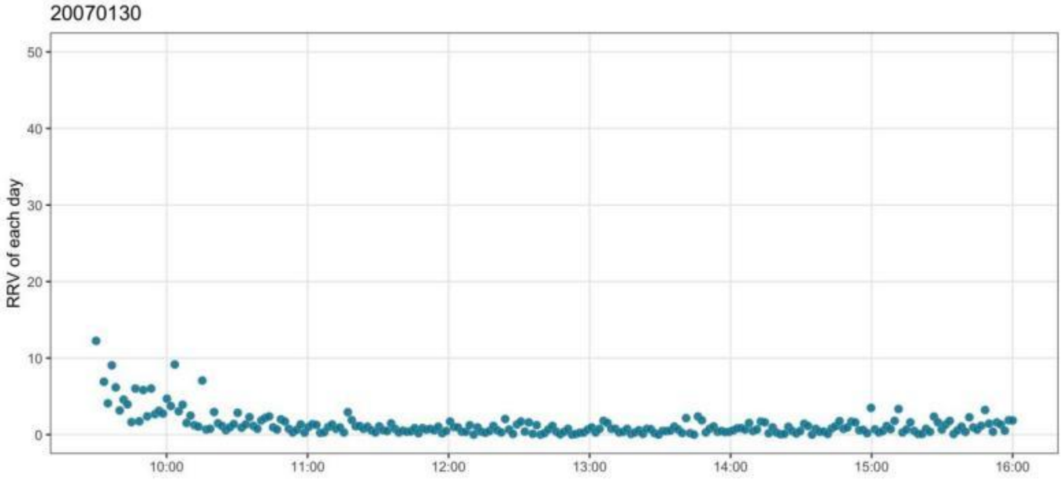


Figure 8: The plot of a regular day.

The 18 days with RRV larger than 1000 are special and deserve more research. The details are shown as below:

Table 4: The details about distribution of 18 days.

	[0,2.5)	[2.5,5)	[5,10)	[10,100)	[100,1000)	>1000
20070130	207	17	8	1	0	1
20070227	118	18	19	57	20	2
20070228	109	15	15	59	31	5
20070301	129	6	11	54	29	5
20070308	198	26	6	2	1	1
20070813	223	6	2	1	1	1
20070918	172	41	10	6	4	1
20071026	217	13	2	1	0	1
20080122	172	23	10	12	13	4
20080222	199	18	3	9	4	1
20080919	171	17	16	18	6	6
20081016	144	65	7	9	7	2
20090807	196	24	9	4	0	1
20090916	181	23	11	5	9	5
20091023	198	26	6	3	0	1
20091127	181	29	12	8	3	1
20100506	162	23	12	20	8	9
20111004	174	42	15	2	0	1
Sum	3151	432	174	271	136	48
Total	312698	43092	12412	4850	364	48
Percentage	1%	1%	1.40%	5.60%	37%	100%
Days	1596	1595	1562	1399	176	18
Time point	312698	43092	12412	4850	364	48

In order to calculate the percentage of RRVs in the 18 days, it is necessary to find out the number of RRV in each category. In these 18 days, there are 48 RRV in [1000,…), 136 RRV in [100,1000), 271 RRV in [10,100), 174 RRV in [5,10), 432 RRV in [2.5,5) and 3151 RRV in [0,2.5).

For example, there are 136 RRVs in [100,1000) for these days, that is about 136/364 = 37%, so the percentage of RRV in [100,1000) of 18 days is 37%. The outcome is shown in the plot above.

5. RRV Modeling

The last stage is RRV modeling. This paper starts with time series regressions to take advantage of the serial correlations. After that the paper tries to see if the error of the regression can be interpreted by some stochastic volatility processes.

First, this paper runs a regression with intercept using y= RRVt and x = RRVt-1. The paper uses the RRV of each 5-day period to run the regressions so RRVt refers to all RRV over a 5-day period and RRVt-1 refers to all RRV over the previous 5-day period. An equation like RRVt = intercept + slope * RRVt-1 + error is needed. The program will provide the intercept and slope. After that we can uncover beta using the slope of the regression and RRV_bar = intercept/(1-beta). Beta is the mean-reversion rate and RRV_bar is the long-term expected mean. These are important specifications of the RRV process.

This paper also plots the beta and RRV_bar changing over time.

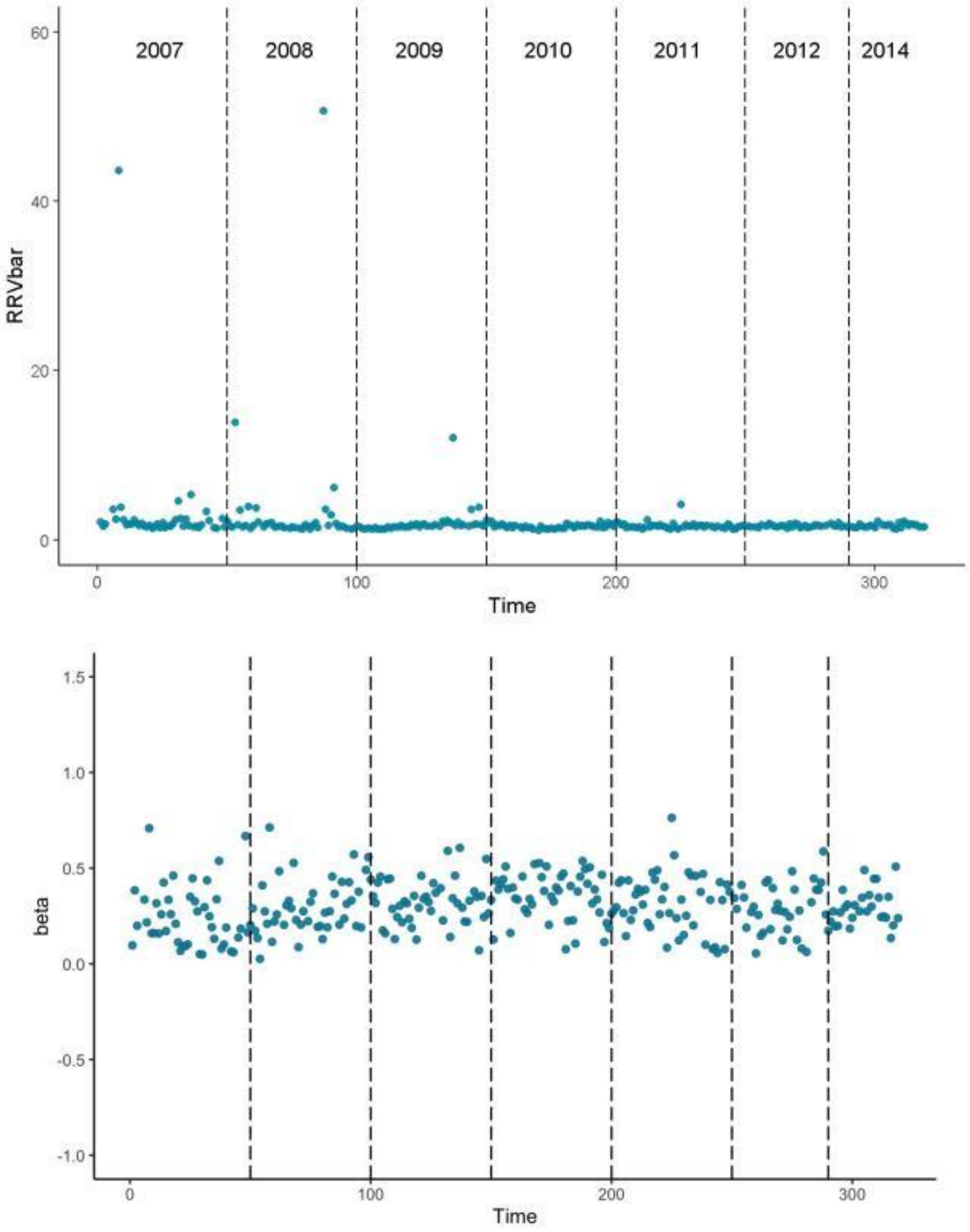


Figure 9: The plots of beta and RRV_bar.

Some RRV_bar and beta in the plot are deleted when they are not statistically significant and use p-value > 0.05 to decide.

From the beta graph, it can be observed that the reversion is between 0 and 1 and mostly similar no matter in crisis or not. For the RRV_bar graph, RRV_bar is supposed to be the long-term volatility. All the RRV_bar are positive and close to 0. However, in 2008 interval, it is observable that there are several extreme points larger than others. Apparently, 2008 financial crisis causes higher volatility so the RRV_bar becomes extreme. Besides, due to the flash crash of the stock market in China in 2007 and sovereign crisis in 2011, larger RRV_bars also happen in both intervals.

6. Conclusion

As noted above, the traditional approach to intraday stock price movements has been disrupted by the noise of market microstructure. It has therefore become customary to sample prices at most every five minutes when calculating daily RV1, since reducing the sampling frequency of stock prices to every five minutes helps to reduce the impact of such measurement errors. In contrast, in this work, realized volatility is calculated every 100 second interval, which is helpful for studying the behavior of the intraday random volatility path. Instead, this paper assumes that realized volatility measurements have quadratic differences of error, and introduces a hypothesis to justify substituting realized volatility for quadratic differences. The 100-second average RV is noisy, but is fairly close to the daily RV except for a few outliers. The results show that the intraday model balances higher volatility with faster mean reversion. In addition, it is pointed out that under certain conditions, few of the prevalence characteristics of low-frequency financial data apply to high-frequency data.

References

[1] *Zhang, F. (2010). The effect of high-frequency trading on stock volatility and price discovery. SSRN eLibrary.*

[2] *Zhang, L. (2006) Efficient estimation of stochastic volatility using noisy observations: a multi-scale approach", Bernoulli, 12, 1019-1043.*

[3] *Heston, S. (1993). A closed-form solution for options with stochastic volatility with applications to bonds and currency options. Review of Financial Studies, 6(2), 327–343.*

[4] *Guilbaud, F., & Pham, H. (2013). Optimal high-frequency trading with limit and market orders. Quantitative Finance, 13(1), 79-94.*

[5] *Engle, R., Ghysels, E., & Sohn, B. (2008). On the economic sources of stock market volatility. SSRN Electronic Journal.*

[6] *Bollerslev, T., & Zhou, H. (2002). Estimating stochastic volatility diffusion using conditional moments of integrated volatility. Journal of Econometrics, 109(1), 33–65.*

[7] *Sun, M. (2016). Modeling volatility using high-frequency data (Unpublished doctoral dissertation). UCLA, Los Angeles, CA.*

[8] *Ellickson, B., Sun, M., Whang, D., & Yan, S. (2018). Estimating a local Heston model. SSRN Archive.*

[9] *Schwert, G.W. (1989). Why does stock market volatility change over time? The Journal of Finance, 44(5), 1115–1153.*

[10] *Blitz, D., & van Vliet, P. (2007). The volatility effect: Lower risk without lower return. Journal of Portfolio Management, 102-113.*