國立臺灣大學管理學院財務金融學系

碩士論文

Department of Finance

College of Management

National Taiwan University

Master Thesis

網路搜尋量是否可以增進股票市場波動率的預測？國際實證

Can internet search volume improve volatility forecasting for the stock markets? International Evidence

陳蕙妤

Hui-Yu Chen

指導教授：王耀輝 博士

Advisor: Yaw-Huei Wang, Ph.D.

中華民國 101 年 6 月

June 2012

國立臺灣大學碩士學位論文
口試委員會審定書

網路搜尋量是否可以增進股票市場波動率的預測?

國際實證

Can internet search volume improve volatility forecasting for the stock markets? International Evidence

本論文係陳蕙妤君（R99723033）在國立臺灣大學財務金融學系、所完成之碩士學位論文，於民國 101 年 6 月 15 日承下列考試委員審查通過及口試及格，特此證明

口試委員：王耀輝（簽名）

（指導教授）

張森林

徐之強

系主任、所長 胡星陽（簽名）

誌謝

回顧過去兩年研究所生涯，要感謝許多人的支持、鼓勵與幫助。首先要感謝家人在我就學期間的付出，謝謝父母不論是在物質上，還是在精神上的資助與支持，也謝謝妹妹總是在我心情煩躁時聽我抱怨。

對於論文的完成，最需要感謝的人莫過於指導教授—王耀輝老師，謝謝老師細心地教導和耐心地指導，總是能及時發現錯誤，並且在我遇到問題時提供解決方法，讓我得以順利克服論文所遇到的所有困難。也謝謝兩位口委老師—張森林老師和徐之強老師對論文的指正與建議，使得我的論文更佳的完善。

感謝所有財金所的同學們，和你們一起分享經驗、討論問題總讓我獲益良多。尤其要謝謝恩偉、彥銘和宣叡給了我許多幫助，和你們在同一個指導老師門下感到很幸運。謝謝怡婷、佩汝和智閎總是耐心且不遺餘力地回答我的疑問，給我許多有用的建議。最後則要謝謝我的朋友們常常關心我、鼓勵我，偶而還要當我的垃圾桶。特別要謝謝相伊花了許多時間幫忙我檢查英文文法，讓我能及時訂正完論文。

此論文的完成，代表我的學生生涯告一個段落，我即將邁入職場，走向人生另一個階段。我將會持續地學習 、充實自己，踏實地往人生目標邁進！

摘要

在這篇論文中，我們使用 Google 搜尋量作為測量散戶投資者注意的媒介，來探討在不同的國家中，搜尋量和股票市場波動率之間的動態關係，以及檢驗搜尋量是否可以幫助預測波動率。我們發現搜尋量對預測未來實現波動率(realized volatility)一般是有用的。當有一個正的搜尋量衝擊，波動率並不會立即的反應而是在之後有正向的移動，但是波動率卻可以立即地影響搜尋量。當建立波動率預測的模型，搜尋量增加了有價值的信息，並且正面地影響未來的波動率。它還可以顯著地增進預測波動率的預測能力在樣本內，樣本外也可以但比較不顯著。在新興市場(emerging markets) 和新領域市場(frontier markets)，搜尋量可以增進預測波動率的現象變得較不明顯。而在我們的實證當中，有些國家沒有這個現象的可能原因除了市場的開發程度，還有較低頻率的資料、意義較不明確單一的搜尋關鍵字、較低的 Google 市佔率、國家的所在位置、較低的網路使用者普及率和較低的散戶投資者的比例。

關鍵字: 實現波動率, 預測, 散戶投資者, 網路搜尋量

Abstract

In this paper, we use Google search volume as proxy of retail investors' attention to study the dynamic relationship with stock market volatility and examine if it can help to forecast volatility in different countries. We find search volume is useful to predict future realized volatility generally. When there is a positive shock of search volume, realized volatility wouldn't react immediately but have positive movement later, while volatility can affect search volume immediately. Search volume adds valuable information for modeling volatility and influences future volatility positively. Search volume also can improve volatility forecasting in- and out-of-sample. But it becomes much more insignificantly in out-of-sample forecast evaluation. The phenomenon that search volume can improve volatility forecasting becomes more unobvious when turning to emerging markets and frontier markets. Besides the developed level of markets, there are some possible reasons, like lower frequency of data, less univocal search terms, lower market shares of Google, locations of countries, smaller penetration rate of internet users and lesser market shares of retail investors, can explain why this phenomenon becomes unobvious for some countries from our empirical results.

Key words: realized volatility, forecasting, retail investor, internet search volume

Contents

1. Introduction

In this paper, we use Google search volume (Da, Engelberg and Gao (2011)) as proxy of retail investors' attention to study the dynamic relationship with stock market volatility, test whether it can add more information for modeling volatility, examine if it can help to forecasting volatility in- and out-of-sample per country and compare these phenomenon in different markets.

In stock markets, huge movements catch investors' eyes. The model of Lux and Marchesi (1999) implies that volatility triggers search activity. And Merton (1987) establishes that investor attention may be relevant for stock pricing and stock liquidity. When the attention of investors increases, this may indicate trading activity increases. Many paper document that retail trades can make stock prices move (Kumar and Lee (2006), Dorn, Huberman and Sengmueller (2008), Kaniel, Sear and Titman (2008), Hvidkjaer (2008)). And Foucault, Sraer and Thesmar (2011) prove that trading activities made by retail investors are positively related to volatility, which can be regarded as behaviors of noisy traders. They estimate that volatility is driven by retail investors about 23% except fundamentals. Therefore, abnormal volatility attracts retail investors' attention and then causes retail investors invest in, which in turn makes volatility.

Nevertheless, measuring retail investors' attention is a hard work since it cannot be observed directly. In empirical studies, many proxies for attention have been employed, like published news announcements and headlines (Mitchell and Mulherin (1994), Berry and Howe (1994), Frieder and Subrahmanyam (2005), Barber and Odean (2008) and Yuan (2008), Fang and Peress (2009)), trading volume (Gervais, Kaniel, and Mingelgrin (2001), Barber and Odean (2008), Hou, Peng, and Xiong (2008)), advertisement expense (Grullon, Kanatas, and Weston (2004), Lou (2008), Chemmanur and Yan (2009)), price limits (Seasholes and Wu (2007)) and extreme returns (Barber and Odean (2008)). But using these proxies need critical assumption that if stock's name is mentioned or its return or turnover is extreme, then that indicate retail investors must pay attention to it, while this assumption cannot be guarantee in practice.

Internet search volume is proved to be a more direct and easier method to measure retail attention by Da, Engelberg and Gao (2011), who use search volume of stock tickers from *Google Trends*. They show that this method is timelier than other well-established attention proxies and mainly captures the retail investor attention. This method seems to be adequate for two reasons. First is that, nowadays, internet has become a popular way to search for information for individual investors. Since the usage of internet increased steadily worldwide during the recent decades, World Wide Web became accessible by nearly everyone and everywhere. And it is the largest pool which supply available information, freely or costly. Internet user usually choose search engine to seek information when needed. Second, an internet user will actively search a specific word only if he or she has interest in or demand for information about the object underlying the keyword.

Google search volume is the most popular proxy since Google search engine has the largest worldwide market shares, accounted for about 90.7% of all search engine on 2011.[1] Google search volume has significantly positive effects on trading activity, trading volume, stock liquidity and return volatility, both historical and implied (Vlastakis and Markellos (2010), Bank, Larch and Peter (2011)).

In addition, there are several evidences that internet search volume has power to forecast, such as unemployment rates, home sales, automotive sales (Choi and Varian

[1] Source: *StatCounter Global Stats* (http://gs.statcounter.com/)

(2009)) and influenza (Ginsberg et al. (2009)). In the financial field, Google search volume is documented to predict earnings (Da et al (2010a), Drake, Roulstone and Thornock (2011)), abnormal returns and trading volumes (Joseph, Wintoki, Zhang (2011)). Da et al. (2011) report that an increase in Google search volume predict higher stock prices in the short-run and reversals in the long run, which is consist with the attention theory of Barber and Odean (2008).

Nowadays is the era of globalization. Investors use international portfolio to diversify risks and increase profits. Besides to invest in developed markets and emerging markets, more and more investors like to invest in frontier markets. It is proved that when portfolio contains equities of frontier markets, both portfolio risk and returns can be improved (Jayasuriya and Shambora (2009)). In this point of view, we want to comprehensively explore not only the relationship between retail investors' attention measured by internet search volume and the stock market volatility but also the predict power of search volume for volatility forecasting in different markets with different development levels.

From the view of international portfolio, we'd like to use the same internet search engine, Google search volume index, measuring worldwide attentions of retail investors for three different MSCI indices, developed, emerging and frontier markets (DM, EM, FM) index, to test if search volume can increase different index volatility forecasting power. However, this method cannot be used since Datastream doesn't provide these three MSCI indices' intraday high and low prices, which are needed for realized volatility, and Google Trends also have not enough search volume data of MSCI frontier markets index.

Moreover, in stock markets, the main retail investors usually are local residents not foreigners. Instead, we focus on leading indices of each country, which belongs to markets of various development levels according to MSCI Market Classification.[2] And we should choose the internet search volume whose search engine has the highest market shares in its home country to measure local attentions of retail investors. In almost countries, Google is the most popular search engine. Baidu and Naver is leading search engine in China and South Korea respectively. But there are problems that they either have no English version or do not provide detail search volume data to be downloaded. Therefore, we use Google search volume to measure local attentions for each country's leading index and then test if search volume can improve volatility forecasting in different markets.

At first, we estimate a VAR model for every stock index to capture the dynamic relationship between Google search volume and stock index volatility. And then examine if past volatility can significantly influence present search volume (Granger (1969) and Sims (1972)) by Granger causality tests, see how volatility reacts over time to shock of search volume, and vice versa, by impulse response function and test how much of volatility can be explained by internet search volume by long-run variance decomposition under the VAR model. Next, we use three other regression models, AR(1), HAR and EGARCH, to rule out whether search volume has additional information for modeling volatility. Last, we compare the forecasting ability of the volatility models with and without lagged search volume in- and out-of-sample by using the mean squared error (MSE), the quasi-likelihood loss function (QL) and the R^2 of regression of the actual realized volatilities on their prediction.

We find past search volume is useful to predict future volatility generally and half of countries' Granger causality is bi-directional: high search activities follow high volatility, and high volatility follows high search activities. But, when there is a

[2] Source: http://www.msci.com/products/indices/market_classification.html

positive shock of search volume, volatility wouldn't react immediately but have positive movement later while volatility can affect search volume immediately. Throughout all countries, movement of volatility is contributed by search volume is ranging from 0.11% to 20.53%. As consistent with Foucault, Sraer and Thesmar (2011), search volume adds valuable information for modeling volatility and influences future volatility positively. Search volume also can improve volatility forecasting in- and out-of-sample. But it becomes much more insignificantly in out-of-sample forecast evaluation.

As the developed level of markets is lower, the phenomenon that search volume can help to forecast volatility becomes less obvious. Besides the developed level of markets, there are some possible reasons of why this phenomenon can't be seen from our tests and models in some countries. The proper reasons are lower frequency of data, less univocal search terms, lower market shares of Google, location of countries, smaller penetration rate of internet users and lesser market shares of retail investors.

The remainder of this paper is organized as follows. In Section 2, we describe the search volume data, data set of realized volatility and the statistics. Section 3 explains the method, models and tests that we use where section 3.1 studies the dynamic relationship between Google search volume and stock index volatility, section 3.2 examine whether the search volume can add valuable information to different volatility models and section 3.3 evaluates in- and out-of-sample volatility forecasts to examine if search volume can help to forecast future volatility. Section 4 is the results of tests and modeling. Finally, section 5 concludes.

2. Data

2.1 Stock index volatility

This study presents an analysis across market classifications, developed, emerging and frontier. We select 24 developed countries, 21 emerging countries and 24 frontier countries according to MSCI Market Classification.

From Datastream, we download the daily close prices, intraday high prices and intraday low prices of main stock index per country from June 2004 to February 2012. However, Datastream doesn’t provide all indices’ intraday high and low prices, especially for frontier markets, and some intraday high and low prices start from 2006/4/20 or later. Note that we use all three indices of USA since S&P 500, NASDAQ and DJIA are all important index in USA.

Table 1 contains the list of countries chosen from MSCI Market Classification, the name of leading index for each country and the start date of intraday high and low prices data provided by Datastream. For those countries without index name and start date mean that intraday high and low prices are not provided by Datastream. We will remove those countries without intraday high and low prices. For example, panel A shows that in developed markets, only New Zealand doesn’t have intraday high and low prices data. Panel B displays that 21 countries decrease to 16 countries in emerging markets. Panel C shows that in frontier markets, only 8 countries have intraday high and low prices data. Among these indices, the index price data of BDL, the leading index of Lebanon, has the shortest period starting from 2010/5/12.

Many researches use squared daily return as proxy for volatility. But any realized volatility measure calculated from only daily return will be noisy estimate. So we use the volatility proxy introduced by Parkinson (1980), which is much more accurate than the squared daily return. For stock index *i*, daily realized volatilities, $RV_{i,t}$, are

Table 1

List of countries with leading index and start date

This table displays the list of countries chosen from MSCI Market Classification, the name of leading index for each country and the start date of intraday high and low prices data provided by DataStream. For those countries without index name and start date mean that intraday high and low prices are not provided by DataStream. Panel A, B and C provide the list of developed, emerging and frontier markets respectively. Note that we use three leading indices, S&P 500, NASDAQ and DJIA, in USA.

Panel A: Developed Markets

Country	Index	Start Date
Australia	S&P/ASX 200	2004/1/2
Austria	ATX	2004/1/2
Belgium	BEL 20	2004/1/2
Canada	S&P/TSX COMPOSITE	2004/1/2
Denmark	OMXC20	2004/1/2
Finland	OMXH	2004/1/2
France	CAC 40	2004/1/2
Germany	DAX 30	2004/1/2
Greece	ATHEX COMPOSITE	2004/1/2
Hong Kong	HANG SENG	2004/1/2
Ireland	ISEQ	2004/1/2
Israel	TA 100	2006/4/20
Italy	FTSE MIB	2004/1/2
Japan	NIKKEI 225	2004/1/2
Netherlands	AEX	2004/1/2
New Zealand		
Norway	OBX price	2004/1/2
Portugal	PSI-20	2004/1/2
Singapore	STRAITS TIMES	2008/1/15
Spain	IGBM	2004/1/2
Sweden	OMXS30	2004/1/2
Switzerland	SMI	2004/1/2
United Kingdom	FTSE 100	2004/1/2
USA	S&P 500	2004/1/2
USA	NSADAQ	2004/1/2
USA	DJIA	2004/1/2

Table 1-Continued

Panel B: Emerging Markets

Country	Index	Start Date
Brazil		
Chile	IGPA	2006/4/20
China	SSE A SHARE	2004/1/2
Colombia	IGBC	2004/1/2
Czech Republic		
Egypt		
Hungary	BUX	2004/1/2
India	SENSEX	2004/1/2
Indonesia	IDX COMPOSITE	2004/1/2
South Korea	KOSPI	2004/1/2
Malaysia	KLCI	2004/1/2
Mexico	BOLSA	2004/1/2
Morocco		
Peru	IGBL	2006/4/20
Philippines	PSEi	2004/1/2
Poland		
Russia	RTS	2004/1/2
South Africa	FTSE/JSE ALL SHARE	2006/4/20
Taiwan	TAIEX	2004/1/2
Thailand	SET	2004/1/2
Turkey	ISE 100	2006/4/20

Table 1-Continued

Panel C: Frontier Markets		
Country	Index	Start Date
Argentina	MERVAL	2004/1/2
Bahrain		
Bangladesh		
Bulgaria	SOFIX	2004/1/2
Croatia	CROBEX	2007/11/23
Estonia		
Jordan		
Kazakhstan	KASE	2007/10/1
Kenya		
Kuwait		
Lebanon	BDL	2010/5/12
Lithuania		
Mauritius		
Nigeria	NSE 30	2009/12/16
Oman		
Pakistan	KSE 100	2007/6/4
Qatar		
Romania	BET	2004/1/2
Serbia		
Slovenia		
Sri Lanka		
Tunisia		
Ukraine		
United Arab Emirates		

defined by:

$$RV_{i,t} = \sqrt{\frac{\left(\log(h_{i,t}) - \log(l_{i,t})\right)^2}{4\log(2)}}, \qquad (1)$$

where $h_{i,t}$ ($l_{i,t}$) is the highest (lowest) price of index *i* on day *t*.

Because in Section 3.2 and Section 4, we will use daily return as input of EGARCH(1,1) model, we need to define daily return, $R_{i,t}$, first:

$$R_{i,t} = \log\left(\frac{P_{i,t}}{P_{i,t-1}}\right), \tag{2}$$

where $P_{i,t}$ is the close price of index *i* on day *t*.

Besides, we will use weekly search volume instead if daily search volume is not available for index. So we also have to define weekly realized volatility, $RV_{i,w}$, and weekly return, $R_{i,w}$:

$$RV_{i,w} = \sqrt{\sum_{j=t-4}^{t} RV_{i,j}{}^2}, \tag{3}$$

$$R_{i,w} = \log\left(\frac{P_{i,t}}{P_{i,t-5}}\right), \tag{4}$$

where *t* is Friday, *t-4* is Monday and *t-5* is last Friday.

2.2 Internet search volume

For internet search volume, we choose to use the search engine which has the highest market shares according to *StatCounter Global Stats*, which express global and each country's ranking of search engines' market shares. In global, Google has about 90.7% market shares of all search engines on 2011.

Table 2 presents top 2 search engines with market shares in each country on 2011. Panel A, B and C are developed, emerging and frontier markets respectively. From this table we find that in addition to China, where Baidu is the biggest search engine, and South Korea, where Naver is the most popular search engine, Google has the highest market shares in the other countries. And the market shares of Google are beyond 70% in almost all countries, except Hong Kong, Russia and Taiwan, where market shares are between 53.37% and 59.60%. Although Google is not the top one in China and South Korea, it still owns 30.73% and 34.16% market shares respectively, ranked at second.

Eventually, we use the same search engine, Google, to measure local attention of individual investors. There are two main reasons. First is the problem of language,

Table 2

Top 2 search engines on 2011

This table presents top 2 search engines with market shares in each country on 2011. Data source is *StatCounter Global Stats* (http://gs.statcounter.com/). Panel A, B and C are developed, emerging and frontier markets respectively. The country whose top 1 search engine has market shares below 70% or is not Google is indicated through bold numbers.

Panel A: Developed Markets

Country	Search Engine	Market Share	Search Engine	Market Share
Australia	Google	94.11%	bing	3.92%
Austria	Google	96.91%	bing	1.98%
Belgium	Google	98.08%	bing	0.83%
Canada	Google	91.83%	bing	4.79%
Denmark	Google	96.57%	bing	2.62%
Finland	Google	97.90%	bing	1.68%
France	Google	94.89%	bing	2.99%
Germany	Google	95.73%	bing	1.99%
Greece	Google	97.63%	bing	1.56%
Hong Kong	**Google**	**59.60%**	**Yahoo!**	**39.35%**
Ireland	Google	94.60%	bing	2.71%
Israel	Google	97.17%	bing	1.87%
Italy	Google	96.76%	Yahoo!	1.07%
Japan	Google	70.85%	Yahoo!	26.65%
Netherlands	Google	94.61%	StartPagina	2.54%
Norway	Google	93.77%	bing	3.00%
Portugal	Google	96.98%	bing	1.95%
Singapore	Google	85.91%	Yahoo!	11.12%
Spain	Google	96.48%	bing	2.28%
Sweden	Google	96.80%	bing	2.41%
Switzerland	Google	96.35%	bing	2.28%
United Kingdom	Google	91.78%	bing	4.40%
USA	Google	79.71%	Yahoo!	9.57%

Table 2-Continued

Panel B: Emerging Markets

country	Search Engine	Market Share	Search Engine	Market Share
Chile	Google	97.38%	bing	1.96%
China	**Baidu**	**65.51%**	**Google**	**30.73%**
Colombia	Google	96.63%	bing	2.61%
Hungary	Google	98.49%	bing	0.90%
India	Google	97.53%	Yahoo!	1.18%
Indonesia	Google	95.91%	Yahoo!	2.18%
South Korea	**Naver**	**55.72%**	**Google**	**34.16%**
Malaysia	Google	86.21%	Yahoo!	9.74%
Mexico	Google	92.75%	bing	5.20%
Peru	Google	97.80%	bing	1.54%
Philippines	Google	85.92%	Yahoo!	11.73%
Russia	**Google**	**54.99%**	**YANDEX RU**	**43.02%**
South Africa	Google	94.28%	bing	3.69%
Taiwan	**Google**	**53.37%**	**Yahoo!**	**45.43%**
Thailand	Google	99.21%	bing	0.58%
Turkey	Google	98.79%	bing	1.03%

Panel C: Frontier Markets

country	Search Engine	Market Share	Search Engine	Market Share
Argentina	Google	95.47%	bing	2.80%
Bulgaria	Google	98.56%	bing	0.79%
Croatia	Google	98.46%	bing	0.89%
Kazakhstan	Google	80.12%	YANDEX RU	17.77%
Lebanon	Google	94.58%	bing	2.81%
Nigeria	Google	88.96%	Yahoo!	4.94%
Pakistan	Google	94.67%	Yahoo!	2.46%
Romania	Google	97.62%	Yahoo!	1.13%

like Naver, the top one search engine in South Korea, is all in Korean without the version of English. Next, search engine does not provide detail search volume data to be downloaded, ex: Yahoo!.

Google provide Search Volume index, instead of effective total number, of search term publicly by *Google Trends*.[3] This index is a portion of Google web searches to compute how many searches have been done for the terms we enter, relative to the total number of searches done on Google over time. In this website, we can see graph of search volume index and download the search volume data globally, or in specific region, country or city, even in different periods. We also can compare search volume of several searching terms, up to five, when entering terms separated by comma ",".

After we signed into our Google Account, we could download two different modes of scaled data, relative and fixed. In relative mode, the data is scaled to the average search traffic for search term (represented as 1.0) during the time period we've selected while in fixed mode, the data is scaled to the average traffic during a fixed point in time (usually January 2004). Since the scale basis doesn't change with time in fixed mode, we can relate them in different time periods. Therefore, we all download the search volume data with fixed scaling.

Search volume could date back to January 2004. We'd like to download the highest frequency search volume data, the daily data. But the search volume data at daily frequency may has many missing data, or even not enough volume to show graph and to be downloaded. For those indices with above problems, we will make use of weekly search volume instead. We only consider trading days of the stock markets in order to match search volumes to the respective time series of volatility.

A search engine user may search for a specific index using its name, ticker symbols or moreover, the short name of its stock exchange. Since stock indices often have many names and ticker symbols, it is a problem to choose an appropriate search term for stock index. We need to find the most widely used search term for specific

[3] Source: http://www.google.com.tw/trends/.

stock index. In general, the short name of the index is preferred by individuals. Take the leading index, S&P/ASX 200, in Australia for example. Using its name as search term, there is not enough search volume to show graph. Search volume of "ASX 200" has a lot missing data before 2011. Finally, we set "ASX", which has correlation about 0.84 with "ASX 200" and far more often been searched, as keywords to download daily search volume data. [4]

For USA, we use all three leading index, S&P 500, NASDAQ, DJIA. The answer of the question which search term individuals use when looking for information about the stock index is easy, especially NASDAQ, which "NASDAQ" is used as keywords. For S&P 500, the number of search volume of "S&P" is about 2.1 times as often as the term "S&P 500". The correlation between the two search terms is 0.84. To DJIA, search volume of "DJIA" and "Dow Jones" amount to 15% and 46% respectively when compared to "Dow". And the pairwise correlations between these search volumes are remarkably high, all above 0.96. Therefore, we choose the search term that is most preferred by retail investors.

However, for some countries, we cannot find search volume of index while using its names or ticker symbols. At this time, we choose to use the name or short name of stock exchange where the index is traded. For example, "Bolsa de Madrid" is the Spanish name of the stock exchange of the leading index, IGBM, in Spain. We take away those countries which we cannot discover any search volume of the main index. And then we rearrange the sample period for each country. We also remove the countries whose number of observation is under 100, such as Portugal, whose search volume data has too many missing data before 2011.

Table 3 displays the list of countries in our sample with the name of leading

[4] Source: *Google Correlate* (http://www.google.com/trends/correlate/)

Table 3

List of countries in the sample with search term, start date and number of observation

This table contains the list of countries in our sample with the name of leading index, the search term used to measure local attention, the start date of sample period and the number of observation of realized volatility (search volume) in sample. For those name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. The search term which is not short name of relative index is indicated through bold type. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Country	Index	Search Term	Start date	Obs.
Australia	S&P/ASX 200	ASX	2005/7/7	1684
Austria	ATX	ATX	2008/9/19	180
Belgium	BEL 20	BEL	2004/1/9	425
Canada	S&P/TSX COMPOSITE	TSX	2005/7/19	1661
France	CAC 40	CAC	2007/1/2	1325
Germany	DAX 30	DAX	2006/1/2	1571
Hong Kong	HANG SENG	HANG	2009/2/9	760
Italy	FTSE MIB	MIB	2007/8/31	224
Japan	NIKKEI 225	NIKKEI	2005/11/1	1552
Netherlands	AEX	AEX	2007/1/2	1323
Singapore	STRAITS TIMES	STRAITS	2009/2/2	775
Spain	IGBM	**Bolsa de Madrid**	2006/10/2	1380
Sweden	OMXS30	**OMX**	2009/8/14	133
Switzerland	SMI	SMI	2007/11/9	225
United Kingdom	FTSE 100	FTSE	2006/1/3	1558
USA	S&P 500	S&P	2006/1/3	1551
USA	NASDAQ	NASDAQ	2005/1/3	1803
USA	DJIA	DOW	2005/1/3	1803

Table 3-Continued

Panel B: Emerging Markets				
Country	Index	Search Term	Start Date	Obs.
China	SSE A SHARE	A 股	2006/9/1	279
India	SENSEX	SENSEX	2007/10/1	1089
Malaysia	KLCI	**KLSE**	2010/8/2	385
Mexico	BOLSA	BOLSA	2005/1/3	1805
Peru	IGBL	**BVL**	2007/2/9	264
South Africa	FTSE/JSE ALL SHARE	JSE	2007/1/12	268
Thailand	SET	SET	2007/10/2	1077
Turkey	ISE 100	**IMKB**	2007/3/1	1239

Panel C: Frontier Markets				
Country	Index	Search Term	Start Date	Obs.
Croatia	CROBEX	**zagrebačka burza**	2007/11/30	222
Pakistan	KSE 100	KSE	2011/1/3	289
Romania	BET	**BVB**	2011/1/10	290

index, the search term used to search information about the index for measuring local attention, the start date of sample period and the number of observation of realized volatility (search volume) in sample. For those words in italic type mean that the data is at weekly frequency, ex: *Austria, China*. In general, individual investors like to use short name of index to search information. Only in Malaysia, people more prefer to use the ticker symbol. “KLSE” is one of ticker symbols of the leading index, KLCI, which we cannot find enough search volume of its short name. There are 6 countries, Spain, Sweden, Peru, Turkey, Croatia and Romania, where (short) names of stock exchange are used as search terms.

From Panel A of Table 3, there are 18 indices which belong to developed markets in our sample. Among these countries, there are 5 countries, Austria, Belgium, Italy, Sweden and Switzerland, which weekly data are used. The longest sample period is Belgium but the number of observation is only 425 since the data is at weekly frequency. Panel B of Table 3 shows that 8 emerging countries are included in sample and weekly data are used for China, Peru and South Africa. The last panel of Table 3 presents that in frontier markets, only Croatia with weekly data and 2 countries with daily data are contained in sample.

2.3 Summary statistics

Table 4 presents the descriptive statistics of realized volatility of each index, including mean, standard deviation, skewness, kurtosis and Jarque–Bera statistic (JB-Statistic), which is used to test the hypothesis that the data are from a normal distribution. A star, double star and triple star denote significance at 10%, 5% and 1% level, respectively. Panel A, B and C provide the summary statistics of developed, emerging and frontier markets respectively. We find that weekly data have higher standard deviation than daily data. And the volatility time series per country are all positively skewed and far from normally distributed.

Therefore, we logarithmically transform the realized volatility time series as suggested by Andersen, Bollerslev, Diebold and Ebens (2001) and Andersen et al. (2003). Table 5 shows that the log-RV time series are far better than RV data although log-RV time series mostly are not normally distributed.

Table 6 displays the descriptive statistics of search volume (*SV*) of each index. Just like the realized volatility data, the search volume time series are also heavily skewed and normality are all rejected with 99% confidence level. We therefore also take logarithms of search volume data (*log-SV*), whose descriptive statistics are showed by Table 7. Both skewness and excess kurtosis are significantly reduced, but normality is still rejected mostly.

Figure 1 displays the graphs of daily realized volatility (gray) and search volume (black) of the stock indices DJIA (USA), CAC 40 (France), SENSEX (India) and

Table 4

Summary statistics of realized volatility

This table presents the descriptive statistics of realized volatility of each index. Jarque–Bera statistic (JB-Statistic) is used to test null hypothesis that the data are from a normal distribution. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. A star, double star and triple star denote significance at 10%, 5% and 1% level, respectively. Panel A, B and C provide the summary statistics of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Australia	0.008	0.005	2.47	12.89	8564.19 ***
Austria	0.036	0.022	2.10	9.08	409.84 ***
Belgium	0.019	0.012	2.41	12.06	1866.21 ***
Canada	0.009	0.007	3.74	27.15	44197.68 ***
France	0.012	0.007	1.96	8.36	2433.92 ***
Germany	0.011	0.008	2.44	12.01	6861.55 ***
Hong Kong	0.009	0.005	1.57	7.09	841.62 ***
Italy	0.032	0.015	1.53	6.09	176.67 ***
Japan	0.009	0.007	4.22	32.25	59912.97 ***
Netherlands	0.011	0.007	2.22	9.98	3773.62 ***
Singapore	0.007	0.004	2.17	10.01	2190.80 ***
Spain	0.011	0.007	2.22	13.05	6932.78 ***
Sweden	0.024	0.011	1.80	8.18	220.28 ***
Switzerland	0.023	0.013	2.24	9.93	637.92 ***
United Kingdom	0.010	0.007	2.64	13.99	9645.66 ***
USA-S&P 500	0.010	0.008	2.97	15.74	12749.70 ***
USA-NASDAQ	0.009	0.007	2.99	16.65	16678.06 ***
USA-DJIA	0.009	0.007	3.40	20.53	26544.49 ***

Table 4-Continued

Panel B: Emerging Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
China	0.033	0.015	1.02	3.92	58.70 ***
India	0.013	0.008	2.45	13.20	5808.45 ***
Malaysia	0.004	0.003	3.84	28.18	11089.69 ***
Mexico	0.010	0.007	2.67	14.27	11686.64 ***
Peru	0.025	0.017	2.50	11.31	1034.28 ***
South Africa	0.025	0.012	1.70	6.52	267.10 ***
Thailand	0.009	0.006	2.94	17.12	10491.44 ***
Turkey	0.014	0.007	2.05	9.33	2931.09 ***

Panel C: Frontier Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Croatia	0.024	0.017	2.32	11.50	867.09 ***
Pakistan	0.008	0.004	1.07	4.50	82.09 ***
Romania	0.010	0.007	4.52	34.04	12586.15 ***

Table 5

Summary statistics of logarithms of realized volatility

This table presents the descriptive statistics of logarithms of realized volatility of each index. Jarque–Bera statistic (JB-Statistic) is used to test the hypothesis that the data are from a normal distribution. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. A star, double star and triple star denote significance at 10%, 5% and 1% level, respectively. Panel A, B and C provide the summary statistics of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Australia	-4.985	0.562	0.27	3.11	20.86 ***
Austria	-3.473	0.519	0.39	2.95	4.65
Belgium	-4.119	0.528	0.47	2.90	15.74 ***
Canada	-4.893	0.601	0.55	3.49	99.63 ***
France	-4.627	0.572	0.11	2.98	2.69
Germany	-4.701	0.600	0.22	2.99	12.36 ***
Hong Kong	-4.843	0.485	0.13	2.68	5.49 *
Italy	-3.544	0.443	0.22	3.07	1.77
Japan	-4.918	0.563	0.47	3.76	94.02 ***
Netherlands	-4.723	0.592	0.16	3.10	6.22 **
Singapore	-5.066	0.514	0.43	3.04	23.94 ***
Spain	-4.701	0.591	-0.01	2.98	0.04
Sweden	-3.823	0.417	0.24	3.20	1.55
Switzerland	-3.911	0.470	0.44	3.36	8.40 **
United Kingdom	-4.741	0.566	0.34	3.10	31.48 ***
USA-S&P 500	-4.867	0.650	0.36	3.07	33.18 ***
USA-NASDAQ	-4.842	0.570	0.44	3.31	65.27 ***
USA-DJIA	-4.950	0.621	0.50	3.31	83.05 ***

Table 5-Continued

Panel B: Emerging Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
China	-3.514	0.440	0.02	2.49	2.99
India	-4.520	0.555	0.24	2.96	10.18 **
Malaysia	-5.574	0.538	0.51	3.39	19.14 ***
Mexico	-4.737	0.557	0.26	3.41	33.53 ***
Peru	-3.892	0.589	0.26	3.68	8.01 **
South Africa	-3.807	0.442	0.30	3.19	4.51
Thailand	-4.825	0.523	0.46	3.32	43.11 ***
Turkey	-4.404	0.447	0.37	3.26	32.08 ***

Panel C: Frontier Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Croatia	-3.931	0.608	0.44	2.70	8.13 **
Pakistan	-5.078	0.773	-3.44	27.18	7581.95 ***
Romania	-4.773	0.521	0.72	4.47	51.04 ***

Table 6

Summary statistics of search volume

This table presents the descriptive statistics of search volume of each index. Jarque–Bera statistic (JB-Statistic) is used to test the hypothesis that the data are from a normal distribution. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. A star, double star and triple star denote significance at 10%, 5% and 1% level, respectively. Panel A, B and C provide the summary statistics of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Australia	2.329	0.594	1.94	13.44	8696.20 ***
Austria	0.263	0.163	4.63	30.98	6513.24 ***
Belgium	1.166	0.266	0.84	5.17	133.30 ***
Canada	1.767	0.689	4.42	33.75	70824.93 ***
France	2.720	1.765	5.33	44.73	102320.80 ***
Germany	1.812	1.107	5.85	51.40	162206.60 ***
Hong Kong	1.467	0.233	0.64	3.83	74.46 ***
Italy	0.507	0.156	1.89	9.35	509.50 ***
Japan	1.860	0.629	0.16	3.51	23.47 ***
Netherlands	1.883	1.173	5.22	38.64	75967.86 ***
Singapore	1.895	0.347	0.87	5.74	339.08 ***
Spain	1.228	0.294	1.55	9.24	2789.19 ***
Sweden	0.476	0.141	4.39	33.38	5541.17 ***
Switzerland	0.613	0.327	3.65	20.34	3318.72 ***
United Kingdom	1.124	0.575	5.95	54.89	183887.60 ***
USA-S&P 500	0.882	0.354	13.01	289.94	5361267.00 ***
USA-NASDAQ	0.657	0.255	3.98	31.66	66433.57 ***
USA-DJIA	1.460	1.073	5.53	54.04	204757.60 ***

Table 6-Continued

Panel B: Emerging Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
China	0.921	0.619	2.45	10.11	865.73 ***
India	2.049	1.111	3.85	33.44	44703.67 ***
Malaysia	0.647	0.131	0.88	4.14	69.97 ***
Mexico	0.742	0.166	0.28	3.15	24.94 ***
Peru	0.608	0.246	1.64	5.81	205.07 ***
South Africa	0.789	0.213	0.75	5.03	71.07 ***
Thailand	0.978	0.139	0.37	2.98	25.16 ***
Turkey	0.439	0.118	2.20	13.50	6680.54 ***

Panel C: Frontier Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Croatia	0.509	0.270	1.15	4.13	60.93 ***
Pakistan	0.460	0.113	0.80	3.05	30.55 ***
Romania	0.816	0.137	2.76	15.24	2170.11 ***

Table 7

Summary statistics of logarithms of search volume

This table presents the descriptive statistics of logarithms of search volume of each index. Jarque–Bera statistic (JB-Statistic) is used to test the hypothesis that the data are from a normal distribution. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. A star, double star and triple star denote significance at 10%, 5% and 1% level, respectively. Panel A, B and C provide the summary statistics of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Australia	0.817	0.237	0.21	4.49	167.50 ***
Austria	-1.435	0.400	1.39	6.88	170.89 ***
Belgium	0.128	0.223	0.11	2.68	2.56
Canada	0.522	0.279	1.72	8.22	2705.67 ***
France	0.896	0.408	1.41	6.90	1274.64 ***
Germany	0.508	0.364	1.84	9.02	3256.09 ***
Hong Kong	0.371	0.156	0.16	2.92	3.62
Italy	-0.719	0.273	0.65	3.82	21.90 ***
Japan	0.555	0.381	-0.73	2.86	138.06 ***
Netherlands	0.541	0.374	1.85	8.76	2580.74 ***
Singapore	0.623	0.179	0.09	3.46	7.84 **
Spain	0.179	0.225	0.27	4.18	96.43 ***
Sweden	-0.772	0.227	1.57	9.31	275.42 ***
Switzerland	-0.574	0.371	1.47	6.18	176.06 ***
United Kingdom	0.054	0.311	2.19	11.01	5413.15 ***
USA-S&P 500	-0.162	0.246	1.77	13.63	8110.50 ***
USA-NASDAQ	-0.472	0.306	0.84	5.64	734.06 ***
USA-DJIA	0.247	0.454	1.38	5.92	1216.91 ***

Table 7-Continued

Panel B: Emerging Markets					
Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
China	-0.236	0.521	0.75	3.55	29.72 ***
India	0.619	0.416	0.82	3.90	159.08 ***
Malaysia	-0.456	0.196	0.30	2.84	6.26 **
Mexico	-0.324	0.231	-0.48	3.66	100.90 ***
Peru	-0.564	0.350	0.76	3.18	25.87 ***
South Africa	-0.274	0.275	-0.34	3.56	8.54 **
Thailand	-0.032	0.142	-0.01	2.82	1.48
Turkey	-0.854	0.241	0.50	4.86	230.33 ***

Panel C: Frontier Markets					
Country	Mean	Std. Dev.	Skewness	Kurtosis	JB-Statistic
Croatia	-0.808	0.521	-0.03	2.46	2.70
Pakistan	-0.804	0.235	0.31	2.53	7.11 **
Romania	-0.214	0.146	1.62	8.04	431.59 ***

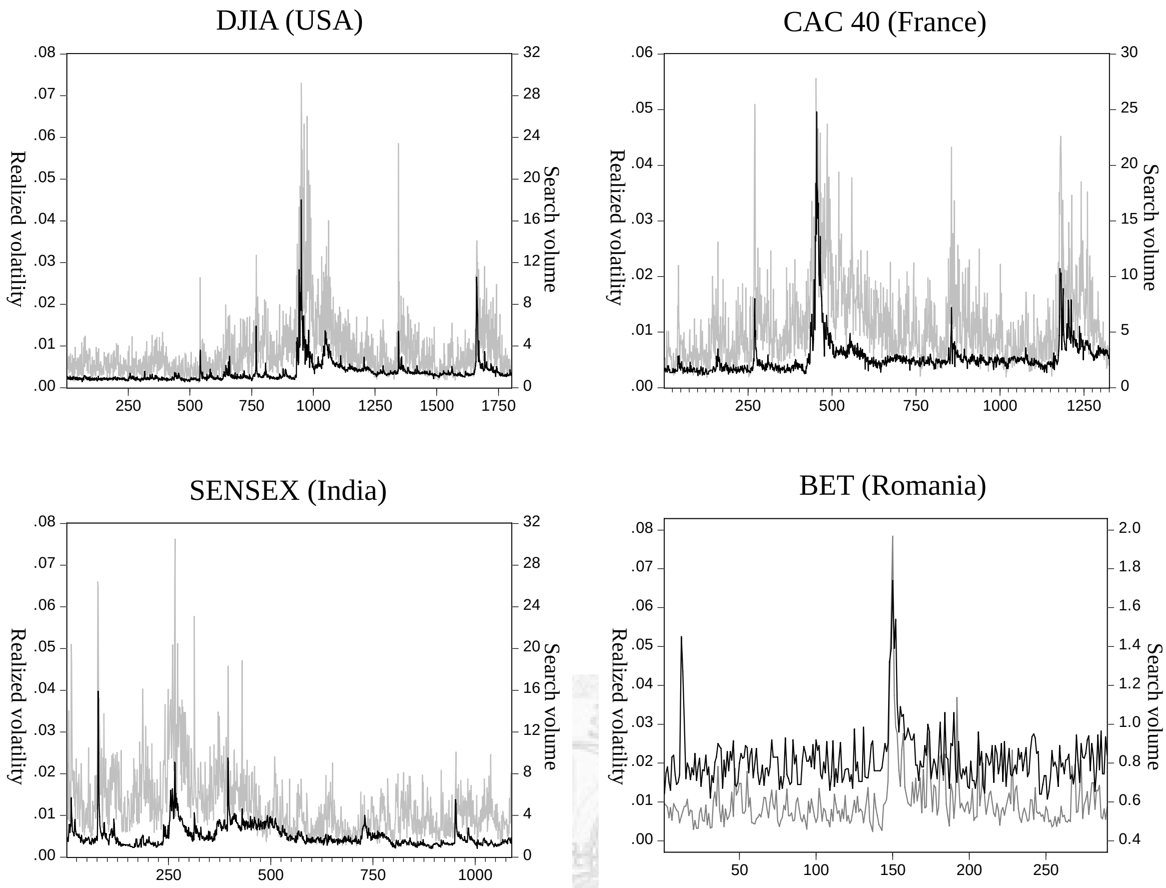

Figure 1:
Realized volatility and search activity
This Figure displays daily realized volatility (gray) and search volume (black) of the stock indices DJIA (USA), CAC 40 (France), SENSEX (India) and BET (Romania). The sample periods start from 2005/1/3/, 2007/1/2, 2007/10/1 and 2011/1/10 respectively and all end on 2012/2/28.

BET (Romania) where the sample periods start from 2005/1/3/, 2007/1/2, 2007/10/1 and 2011/1/10 respectively and all end on 2012/2/28. We choose 2 indices from developed markets and each from emerging and frontier markets. This can be seen from Figure 1 that realized volatility of stock index and search volume measuring local attention of index exhibit a strong co-movement over time. The correlation coefficients are 0.78, 0.69, 0.61 and 0.58 respectively.

3. Methods

3.1 Vector autoregressive model (VAR model)

In this section, we estimate a VAR model for every stock index to capture the dynamic relationship between Google search volume and stock index volatility. We study the dynamics of realized volatility and search volume by three ways: (1) Granger causality tests to examine if past volatility can significantly influence present search volume (Granger (1969) and Sims (1972)). (2) Impulse response function to see how volatility reacts over time to shock of search volume and vice versa. (3) Long-run variance decomposition to test how much of volatility can be explained by internet search volume.

First, we need to estimate a VAR(p) model:

$$log\text{-}RV_t = c_1 + \sum_{j=1}^{p} \beta_{1,j}\, log\text{-}RV_{t\text{-}j} + \sum_{j=1}^{p} \gamma_{1,j} log\text{-}SV_{t\text{-}j} + \varepsilon_{1,t}, \qquad (5)$$

$$log\text{-}SV_t = c_2 + \sum_{j=1}^{p} \beta_{2,j}\, log\text{-}RV_{t\text{-}j} + \sum_{j=1}^{p} \gamma_{2,j} log\text{-}SV_{t\text{-}j} + \varepsilon_{2,t}, \qquad (6)$$

where we decide the lag order (p) by Schwarz Criterion (SC), or named Bayes Information Criterion (BIC). The model with the lower value of SC is the one to be preferred to use.

The degree of freedom (df) used by Granger causality presented in Table 8 is just the optimal lag order (p) used for VAR model. The VAR model contains the first through sixth lags (*t-1~t-6*) of all the endogenous variables as p is 6. We could find that p is between 1 and 6 for daily data while weekly data have much smaller p, from 1 to 3. It makes sense that if t is this Friday and *t-6* is last Thursday, then it means this week and last week when we transform to weekly frequency. At this time, p is 1. So it is normal that lag order is smaller for weekly data, where Peru has largest p=3.

And then, we examine the following 3 tests under the optimal VAR model for every index.

3.1.1 Granger causality test

Granger causality test, approached by Granger (1969) is to see how much of the current *y* can be explained by past values of *y* and then to examine whether adding lagged values of *x* can improve the explanation. If the coefficients of lagged *x* are statistically significant, *y* is said to be Granger-caused by *x*. That is, if search volume has statistically significant information about future volatility by t-tests or F-tests then search volume is said to Granger cause stock market volatility. Note that the statement "*x* Granger cause *y* "does not imply that *y* is the effect or the result of *x*.

Here we use pairwise Granger causality tests to test whether an endogenous variable can be treated as exogenous in the VAR model. The null hypotheses are "*log-RV* doesn't Granger cause *log-SV*" and "*log-RV* doesn't Granger cause *log-SV*" to see whether realized volatility is useful in forecasting search volume and whether search volume is useful in forecasting volatility at the same time. If Chi-squared (Wald) statistic is larger than critical value, such that p-value is under 0.1 with 90% confidence level, then we can reject the null hypothesis.

3.1.2 Impulse response function (IRF)

Generally, impulse response refers to the reaction of any dynamic system in response to some external change. It traces the effect of a shock to one of the innovations on current and future values of the endogenous variables. Here we used to explore how volatility reacts over time to the shock of search volume, and vice versa.

To trace the response function, we set the number of period as 100 and use the Cholesky decomposition with the ordering, *log-RV log-SV*, due to the economically meaningful restriction of volatility being contemporaneously exogenous, i.e. volatility can affect search volume immediately, but search volume cannot contemporaneously affect volatility. This ordering intuitively indicates that abnormal volatility attracts retail investors' attention and then in turn makes volatility. On the other hand, search volume would not rise without a preceding event on the market.

3.1.3 Variance decomposition

While impulse response function trace the effects of a shock to one endogenous variable on to the other variables in the VAR model, variance decomposition separates the variation in an endogenous variable into the component shocks to the VAR model. Thus, the variance decomposition provides information about the relative importance of each random innovation in affecting the variables in the VAR model. We used this to examine the amount of information of search volume contributes to the volatility.

We set number of periods as 100 to capture long-term variance decomposition. Because of the economically meaningful restriction of volatility discussed before for tracing impulse response function, we use the same ordering, *log-RV log-SV*, such that volatility is contemporaneously exogenous.

3.2 Regression models

In this section, we use three other regression models to rule out whether search volume has additional information for modeling volatility. Here we only focus on the equation of interest, the volatility equation. We choose these regression models since they are commonly used to capture the time series properties of realized volatility and include lagged proxy of individual's attention to test whether retail investors' attention add information. Here, we only include search volume at one lag in these models, $log\text{-}SV_{t-1}$.

First, we estimate autoregressive models with first lag (AR(1)) and augment this with lagged search volume, $log\text{-}SV_{t-1}$, following Andersen, Bollerslev, Christoffersen and Diebold (2006) and Bollen and Inder (2002).

$$log\text{-}RV_t = c + \beta_1 log\text{-}RV_{t\text{-}1} + \gamma_1 log\text{-}SV_{t\text{-}1} + \varepsilon_t. \qquad (7)$$

Next, we estimate heterogeneous autoregressive (HAR) model of Corsi (2009), which could capture the long-memory properties of volatility very well. This HAR model has different lag length and augments with lagged search volume, $log\text{-}SV_{t\text{-}1}$,

$$log\text{-}RV_t = c + \beta_d log\text{-}RV_{t\text{-}1} + \beta_w log\text{-}RV^{w}_{t\text{-}1} + \beta_m log\text{-}RV^{m}_{t\text{-}1} + \gamma_1 log\text{-}SV_{t\text{-}1} + \varepsilon_t \qquad (8)$$

, where $log\text{-}RV^{w}_{t} = \frac{1}{5}\sum_{j=0}^{4} log\text{-}RV_{t\text{-}j}$ and $log\text{-}RV^{m}_{t} = \frac{1}{22}\sum_{j=0}^{21} log\text{-}RV_{t\text{-}j}$. That is, the model contains the realized volatility data of yesterday, previous week and previous month so it can explain the long-memory pattern of volatility well.

Since bad news usually cause higher volatility than good news, that is asymmetry, and in turn makes more attention of retail investors. The above two models don't consider asymmetry. Therefore, we estimate the EGARCH(1,1) model by augmented with lagged search volume, $log\text{-}SV_{t\text{-}1}$, (Nelson(1991)),

$$R_t = \lambda R_{t-1} + u_t, \qquad (9)$$

$$\log(\sigma_t^2) = \omega + \beta_1 \log(\sigma_{t-1}^2) + \alpha \left|\frac{u_{t-1}}{\sigma_{t-1}}\right| + \theta \frac{u_{t-1}}{\sigma_{t-1}} + \gamma_1 log\text{-}SV_{t\text{-}1}. \qquad (10)$$

The input of this model is not the realized volatility time series but the return data of index measured by equation (2). We augment the lagged search volume to the variance equation, which we interest in.

In all three models contain the previous day's search volume as an exogenous variable, AR(1)+SV, HAR+SV and EG+SV. We examine whether lagged search volume indeed add valuable information to the model by testing whether γ_1 is significantly different from zero.

3.3 Volatility forecasts

In this section we compare the forecasting ability of the volatility models with and without lagged search volume, $log\text{-}SV_{t\text{-}1}$, in- and out-of-sample. The models we use are the univariate AR(1), HAR and EGARCH models, which are simply equations (7), (8) and (10) with γ_1 equal to zero, and the respective augmented models including lagged search volume, AR(1)+SV (7), HAR+SV (8)and EGARCH+SV (10). We evaluate the forecasting ability by comparing realized volatility and its prediction following the literatures (e.g. Andersen et al. (2003), Ghysels et al. (2006), Ait-Sahalia and Mancini (2008)).

We use two robust loss functions to compare the volatility forecasting ability (Patton (2011)). They are the mean squared error (MSE) and the quasi-likelihood loss function (QL),

$$\text{MSE} = \frac{1}{n}\sum_{j=1}^{n}(RV_{t+j} - \widehat{RV}_{t+j|t+j-1})^2, \tag{11}$$

$$\text{QL} = \frac{1}{n}\sum_{j=1}^{n}\left[\frac{RV_{t+\mathrm{j}}}{\widehat{RV}_{t+\mathrm{j}|t+j-1}} - \log\left(\frac{RV_{t+\mathrm{j}}}{\widehat{RV}_{t+\mathrm{j}|t+j-1}}\right) - 1\right], \tag{12}$$

where $\widehat{RV}_{t+j|t+j-1}$ is the respective forecast of volatility based upon information available up to and including time t. If MSE and QL decrease after the models augment with lagged search volume, $log\text{-}SV_{t\text{-}1}$, then it indicates that search volume can improve forecasting ability. We also test that if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant.

In addition, we use the R^2 of regression of the actual realized volatilities on their prediction to compare the ability of volatility forecasts (Mincer and Zarnowitz (1969)),

$$RV_{t+\mathrm{j}} = c_0 + c_1\widehat{RV}_{t+\mathrm{j}|t+j-1} + e_t. \tag{13}$$

Search volume can help to improve volatility forecasting as the R^2 increase after the model augment with lagged search volume.

At the first, we make in-sample forecasts to evaluate one-step ahead forecasts of realized volatility. For in-sample analysis, we estimate the parameters in the sample period, where the total observations are used, and then using the same parameters to forecast one-step ahead volatilities. They are just the fitted values of the model. The total number of observations (Obs.) in the model for each index can be seen from the rightist column of Table 3.

While for out-of-sample analysis, we do not use the same parameters to predict volatilities. We set the window as 2/3 of total number of observations and then forecast volatility by rolling window. Take DJIA as an example. The number of total observations of DJIA is 1803 so the window is 1202. For the initial forecast, $\widehat{RV}_{1203}$, we estimate the models using the time series, t=1 to 1202. We then re-estimate the models using the time series, t=2 to 1203, for $\widehat{RV}_{1204}$. We repeat this action until the end of the period.

4. Empirical results

In this section, we show and discuss the results of tests and modeling by graphs, tables and statements. Note that in each table, we usually separate to three panels by developed, emerging and frontier markets, Panel A, B and C respectively. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. A star, double star and triple star denote significance at 10%, 5% and 1% level, respectively.

4.1 Dynamics of search volume and volatility (VAR model)

Under the VAR model for each stock index, we study the dynamic relationship between Google search volume and stock index volatility as following: (1) Whether search volume is useful in forecasting volatility by Granger causality tests, (2) How volatility reacts over time to shock of search volume and vice versa by impulse response function, (3) How much of volatility can be explained by search volume by long-run variance decomposition.

4.1.1 Whether search volume is useful in forecasting volatility?

The results of the Granger causality test are presented in Table 8. If Chi-squared (Wald) statistic is larger than critical value, such that p-value is under 0.01 with 90% confidence level, then we can reject the null hypothesis.

We focus on the indices of USA first, whose outcomes are presented at the bottom of Panel A. With 95% confidence level, only DJIA have bi-directional Granger causality: high search activities follow high volatility, and high volatility follows high search activities. Volatility of NASDAQ does not Granger cause its search volume and vise versa. For S&P 500 index, past volatility can significantly influence present search volume while past search volume can't influence present volatility. This may because the search terms "S&P" and "NASDAQ" are less univocal since "S&P" is often used as an abbreviation for the rating agency Standard & Poor's and "NASDAQ" have many meanings, such as the company, computer system and stock exchange market. Moreover, individuals pay most attentions on DJIA in USA since the amount of search volume of S&P 500, NASDAQ and DJIA are 1:1.4:4.2 according to *Google trend*. Therefore, we'd like to consider DJIA as the representative index of USA as we compare results among countries.

In stock markets, huge movements usually capture investors' attention. In this point of view, we expect the null hypothesis that *log-RV* doesn't Granger cause *log-SV* should be rejected. From the middle part of Table 8, we discover that past volatility time series do significantly affect present search volume data in general. With 90% confidence level, there are only 3 indices without the situation that stock shocks caused investors to pay attention in developed markets while in frontier market, there

Table 8

Granger causality test

This table provides the results of the Granger causality test. If Chi-squared (Wald) statistic is larger than critical value, such that p-value is under 0.01 with 90% confidence level, then we can reject the null hypothesis. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Null hypothesis:		log-RV doesn't Granger cause log-SV		log-SV doesn't Granger cause log-RV	
Country	df	Chi-sq	p-value	Chi-sq	p-value
Australia	6	6.94	(0.326)	20.93	(0.002)
Austria	1	1.66	(0.198)	13.08	(0.000)
Belgium	2	8.69	(0.013)	13.78	(0.001)
Canada	5	54.96	(0.000)	39.28	(0.000)
France	4	24.42	(0.000)	55.17	(0.000)
Germany	6	42.12	(0.000)	50.24	(0.000)
Hong Kong	4	14.01	(0.007)	1.04	(0.904)
Italy	1	2.01	(0.156)	2.44	(0.118)
Japan	5	20.36	(0.001)	31.50	(0.000)
Netherlands	5	40.87	(0.000)	28.92	(0.000)
Singapore	4	8.32	(0.080)	5.31	(0.257)
Spain	5	28.27	(0.000)	3.83	(0.574)
Sweden	1	4.77	(0.029)	0.71	(0.399)
Switzerland	1	3.94	(0.047)	26.15	(0.000)
United Kingdom	6	39.98	(0.000)	19.45	(0.004)
USA-S&P 500	4	33.17	(0.000)	9.06	(0.060)
USA-NASDAQ	5	10.94	(0.053)	2.72	(0.743)
USA-DJIA	5	27.26	(0.000)	31.38	(0.000)

Table 8-continued

Panel B: Emerging Markets

Null hypothesis:		log-RV doesn't Granger cause log-SV		log-SV doesn't Granger cause log-RV	
Country	df	Chi-sq	p-value	Chi-sq	p-value
China	2	0.62	(0.735)	12.62	(0.002)
India	5	23.49	(0.000)	15.45	(0.009)
Malaysia	2	3.08	(0.215)	2.40	(0.301)
Mexico	6	17.01	(0.009)	15.89	(0.014)
Peru	3	16.23	(0.001)	4.20	(0.241)
South Africa	2	2.12	(0.347)	4.90	(0.086)
Thailand	5	16.60	(0.005)	3.64	(0.603)
Turkey	3	22.45	(0.000)	17.62	(0.001)

Panel C: Frontier Markets

Null hypothesis:		log-RV doesn't Granger cause log-SV		log-SV doesn't Granger cause log-RV	
Country	df	Chi-sq	p-value	Chi-sq	p-value
Croatia	2	0.44	(0.803)	3.54	(0.170)
Pakistan	3	2.02	(0.569)	1.63	(0.653)
Romania	1	13.68	(0.000)	15.02	(0.000)

is only Romania has this phenomenon. Roughly, indices which volatility does not Granger cause search volume also have the same result inversely. That indicates that when abnormal volatility doesn't make individuals to search information, the abnormal search volume doesn't mean that individuals are going to invest in, which in turn does not affect volatility.

Then we turn to our interest that if search volume is useful in forecasting future volatilities. The results indicate that past search volume can influence present volatility generally since the null hypotheses have been rejected in most countries. And almost half of countries' Granger causality is bi-directional. The phenomenon, that search volume has statistically significant information about future volatility, is more unobvious as the developed level of markets is lower. In emerging markets, the p-values are larger than those in developed markets.

4.1.2 How volatility reacts over time to shock of search volume and vice versa?

Figure 2 displays the impulse response functions (IRF) of the VAR model for two selected index, DJIA (USA) and HANG SENG index (Hong Kong). Since the impulse response functions of most indices are alike, we just take IRF of DJIA as representative. There are some impulse response functions which are different from normal ones so we select IRF of HANG SENG, which is the most abnormal graph, to represent. Where realized volatility is gray line and search volume is black line. Upper graphs show how volatility reacts over time to a one standard-deviation shock of search volume (gray line) and volatility (black line), and lower graphs show reversely.

In general, such as IRF of DJIA from Figure 2, the impulse response functions implicate that volatility can affect search volume immediately, but search volume cannot contemporaneously affect volatility. The upper graph of IRF of DJIA displays that when there is a positive shock of search volume today, volatility wouldn't react immediately but have positive movement later. The intuition behind this is that abnormal volatility attracts retail investors' attention and then in turn makes volatility only after investors invest in. Both the response of volatility and search volume persist long, but the response of volatility declines faster.

However, some countries' impulse response functions are different from normal ones, especially Hong Kong and Peru, whose impulse response function is like the IRF of HANG SENG of Figure 2. Both the response of volatility and search volume persist shorter than DJIA and die out after about 40 days. Although the upper graph of HANG SENG's IRF shows that positive shock of search volume cannot contemporaneously affect volatility, however, the volatility decreases instead later.

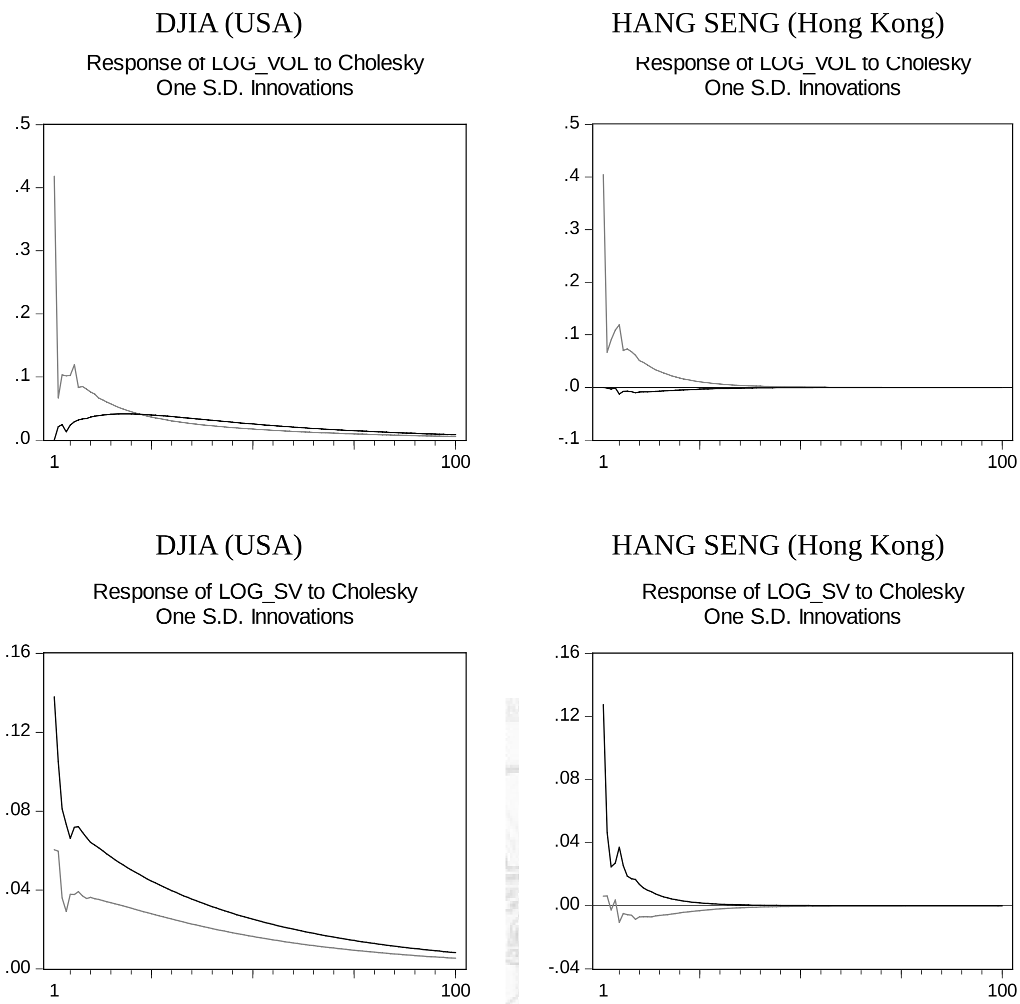

Figure 2:

Impulse response functions of DJIA (USA) and HANG SENG (Hong Kong)

This Figure displays the impulse response functions (IRF) of the VAR model for DJIA (USA) and HANG SENG index (Hong Kong). IRF of DJIA represents the normal situations while IRF of HANG SENG is the most abnormal graph. Where realized volatility is gray line and search volume is black line.

Spain, Malaysia and Thailand have similar situations and are consistent with the results of Granger causality that search volume cannot influence future volatility. The bottom figure of HANG SENG's IRF displays that search volume react immediately but declines later to positive shock of volatility. Mexico, Singapore and NASDAQ are alike but not consistent with the results of Granger causality.

4.1.3 How much of volatility can be explained by search volume?

Table 9 provides the variance decomposition's results with the ordering, *log-RV log-SV*, which are the amounts of variance of *log-RV* (left side) and *log-SV* (right side), where the unit is in percent. Take Australia for example, *log-SV* contributes 11.20% to the variance of *log-RV* while *log-RV* determines 6.71% amount the variance of *log-SV*.

Here we use the ordering such that volatility is contemporaneously exogenous. In this case, movement of search volume is contributed largely by volatility. From the right side of the table, we can see *log-RV* contributes from 0.55% for Singapore up to 37.47% for Switzerland to the variance of *log-SV*. In general, the contributions of volatility to the change of search activities in developed markets are larger than in emerging and frontier markets.

Turn to our interest that how much of volatility can be explained by attentions of retail investors, we see the left side of Table 9. Throughout all countries, movement of volatility is contributed by search volume is ranging from 0.11% for NASDAQ to 20.53% for France. If we ignore S&P 500 index and the countries which search volume does not Granger cause volatility then the lower bound of contributions of search volume to volatility increases to 3.36% (India). On the other words, for those countries which past search volume cannot influence present volatility, volatility can be explained little by search volume. In general, the contributions of search volume to the change of volatility in developed markets are larger than in emerging and frontier markets, too.

Table 9

Variance decomposition

This table provides the variance decomposition's results, which are the amounts of variance of log-RV (left) and log-SV (right). The biggest and smallest amounts of variance are indicated through bold numbers. The unit is in percent. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

	log-RV		log-SV	
Country	log-RV (%)	log-SV (%)	log-RV (%)	log-SV (%)
Australia	88.80	11.20	6.71	93.29
Austria	81.41	18.59	36.35	63.65
Belgium	81.33	18.67	26.23	73.77
Canada	83.73	16.27	19.43	80.57
France	79.47	**20.53**	17.04	82.96
Germany	87.92	12.08	17.33	82.67
Hong Kong	99.52	0.48	3.68	96.32
Italy	96.73	3.27	32.60	67.40
Japan	89.64	10.36	4.17	95.83
Netherlands	89.69	10.31	22.77	77.23
Singapore	97.45	2.55	**0.55**	99.45
Spain	99.29	0.71	6.11	93.89
Sweden	98.81	1.19	17.08	82.92
Switzerland	76.28	23.72	**37.47**	62.53
United Kingdom	96.27	3.73	33.20	66.80
USA-S&P 500	97.22	2.78	23.74	76.26
USA-NASDAQ	99.89	**0.11**	2.02	97.98
USA-DJIA	81.50	18.50	24.56	75.44

Table 9-continued

Panel B: Emerging Markets				
	log-RV		log-SV	
Country	log-RV (%)	log-SV (%)	log-RV (%)	log-SV (%)
China	80.58	**19.42**	4.52	95.48
India	96.64	3.36	**15.58**	84.42
Malaysia	99.20	0.80	**1.54**	98.46
Mexico	96.39	3.61	5.88	94.12
Peru	98.55	1.45	8.60	91.40
South Africa	86.16	13.84	4.40	95.60
Thailand	99.72	**0.28**	4.41	95.59
Turkey	96.10	3.90	5.96	94.04

Panel C: Frontier Markets				
	log-RV		log-SV	
Country	log-RV (%)	log-SV (%)	log-RV (%)	log-SV (%)
Croatia	92.33	**7.67**	**15.93**	84.07
Pakistan	99.44	**0.56**	**2.22**	97.78
Romania	93.02	6.98	12.78	87.22

4.2 Whether search volume has valuable information for modeling volatility?

Table 10 contains γ_1, the coefficient estimates of $log\text{-}SV_{t-1}$, and the corresponding p-value in each regression model, AR(1)+SV, HAR+SV and EGARCH+SV. Take France as an example. Search volume does add information for modeling volatilities in all models. The models, AR(1)+SV, HAR+SV and EGARCH+SV, predict that if search volume increase 100% today, volatility will increase 72.4%, 23.4% and 7% respectively in addition to the dynamic effects in volatility itself.

We focus on the indices of USA first, whose outcomes are presented at the bottom of Panel A. With 95% confidence level, for both S&P 500 and DJIA, search volume is helpful predictor of future volatility, except the AR(1) model augmented

Table 10

Coefficient estimates of lagged search volume in the regression models

This table contains γ_1, the coefficient estimates of $log\text{-}SV_{t\text{-}1}$, in each regression model and the corresponding p-value. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Model:	AR(1)+SV		HAR+SV		EGARCH+SV	
Country	Coef.	p-value	Coef.	p-value	Coef.	p-value
Australia	0.455	(0.000)	0.161	(0.004)	0.091	(0.000)
Austria	0.252	(0.022)	0.234	(0.083)	-0.040	(0.545)
Belgium	0.250	(0.063)	0.191	(0.037)	1.423	(0.000)
Canada	0.950	(0.000)	0.293	(0.000)	0.090	(0.000)
France	0.724	(0.000)	0.234	(0.000)	0.070	(0.000)
Germany	0.661	(0.000)	0.236	(0.000)	0.079	(0.000)
Hong Kong	-0.156	(0.251)	0.112	(0.304)	0.002	(0.960)
Italy	0.142	(0.239)	0.189	(0.065)	0.568	(0.081)
Japan	0.429	(0.000)	0.122	(0.001)	0.029	(0.021)
Netherlands	0.610	(0.000)	0.202	(0.000)	0.044	(0.010)
Singapore	0.017	(0.877)	0.116	(0.128)	0.058	(0.198)
Spain	-0.029	(0.757)	0.024	(0.675)	-0.005	(0.765)
Sweden	0.011	(0.945)	0.091	(0.420)	1.698	(0.007)
Switzerland	0.433	(0.001)	0.437	(0.000)	1.415	(0.000)
United Kingdom	0.412	(0.000)	0.205	(0.000)	0.040	(0.044)
USA-S&P 500	0.057	(0.542)	0.219	(0.000)	0.060	(0.000)
USA-NASDAQ	-0.060	(0.409)	0.063	(0.059)	-0.014	(0.086)
USA-DJIA	0.744	(0.000)	0.172	(0.000)	0.071	(0.000)

Table 10-continued

Panel B: Emerging Markets						
Model:	AR(1)+SV		HAR+SV		EGARCH+SV	
Country	Coef.	p-value	Coef.	p-value	Coef.	p-value
China	0.253	(0.000)	0.103	(0.017)	0.037	(0.437)
India	0.327	(0.000)	0.090	(0.007)	0.026	(0.046)
Malaysia	0.100	(0.613)	0.143	(0.330)	-0.030	(0.615)
Mexico	0.218	(0.007)	0.177	(0.000)	0.033	(0.010)
Peru	-0.200	(0.231)	0.009	(0.940)	-0.079	(0.616)
South Africa	0.038	(0.731)	0.184	(0.013)	0.162	(0.000)
Thailand	-0.044	(0.704)	0.032	(0.708)	0.043	(0.477)
Turkey	0.337	(0.000)	0.200	(0.000)	0.089	(0.004)

Panel C: Frontier Markets						
Model:	AR(1)+SV		HAR+SV		EGARCH+SV	
Country	Coef.	p-value	Coef.	p-value	Coef.	p-value
Croatia	0.290	(0.012)	0.052	(0.590)	0.175	(0.041)
Pakistan	0.300	(0.045)	0.171	(0.206)	-0.070	(0.045)
Romania	0.563	(0.032)	0.786	(0.002)	0.243	(0.323)

with lagged search volume for S&P 500. Search volume of NASDAQ does not add information to volatility model since the p-values are all above 0.05. The reason why search volume cannot help to predict volatility has discussed on section 4.1.1. These results are consistent with the results of Granger causality tests. It is optimal that consider DJIA as the representative index of USA when we compare results among countries.

The results indicate that search volume contains additional information about future volatility generally since $log\text{-}SV_{t-1}$ enters significantly in all models for most countries, ex: the leading indices worldwide (FTSE 100, CAC 40, DAX 30, DJIA). This consequence is consistent with the results of tests in VAR models, especially for countries whose Granger causality is bi-directional. And the coefficient estimates of

$log\text{-}SV_{t\text{-}1}$ are all positive indicate that search volume positively influences volatility, which is in line with Foucault, Sraer and Thesmar (2011).

On the other hand, for almost all of those countries which search volume do not Granger cause volatility, search volume neither does not add information for modeling volatility, like Hong Kong. To South Africa and Croatia, we could say search volume has additional information about future volatilities since lagged search volume enter insignificantly in only one model for each country, while Granger causality tests shows can't.

The phenomenon, that search volume has statistically significant information about future volatility, becomes more unobvious as the developed level of markets is worse. In frontier markets, the p-values are larger than those in developed markets and search volume contains significance about future volatility only in AR(1) model.

4.3 Does search volume help to improve volatility forecasts?

4.3.1 In-sample forecast evaluation

Table 11, Table 12 and Table 13 contain the in-sample forecasts evaluation of one-step ahead forecasts of realized volatility. The models are the univariate models (Uni.) and the respective augmented models (Aug.) including lagged search volume. Table 11, Table 12 and Table 13 display the comparison results of AR(1) vs. AR(1)+SV, HAR vs. HAR+SV and EGARCH vs. EGARCH+SV respectively. Forecasting ability are measured by the mean squared error (MSE, $\times 10^5$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (%) of the regression. P-value is result of test which testing if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant. The model is better when MSE decreases, QL decreases and R^2 increases.

We discuss the results of the USA indices first. From Panel A of each table, only

Table 11

In- sample forecast evaluation of AR(1) and AR(1)+SV models

This table compares the in-sample forecasts of AR(1) and AR(1)+SV model. Uni. means the univariate model, AR(1) here. Aug. means the augmented model with lagged search volume, AR(1)+SV here. Performance measures are the mean squared error (MSE, $\times 10^5$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (%) of the regression. P-value is result of test which testing if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant. The model is better as MSE decreases, QL decreases and R^2 increases. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Australia	2.05	2.00	(0.008)	11.80	11.59	(0.030)	32.32	34.93
Austria	16.81	15.07	(0.374)	5.33	5.17	(0.425)	64.05	68.11
Belgium	5.65	5.70	(0.700)	5.04	5.03	(0.474)	61.32	61.17
Canada	3.36	2.83	(0.004)	12.26	11.27	(0.000)	46.38	53.13
France	3.41	2.80	(0.000)	10.69	9.55	(0.000)	40.82	49.93
Germany	3.47	3.00	(0.000)	11.29	10.78	(0.003)	46.07	52.44
Hong Kong	1.85	1.85	(0.896)	10.27	10.27	(0.919)	16.97	16.91
Italy	11.81	11.62	(0.280)	4.62	4.60	(0.436)	51.52	52.48
Japan	3.17	3.10	(0.013)	12.25	11.72	(0.000)	33.82	36.40
Netherlands	3.16	2.76	(0.000)	10.76	10.04	(0.000)	42.94	49.64
Singapore	1.45	1.45	(0.291)	10.14	10.15	(0.417)	28.55	28.63
Spain	3.31	3.32	(0.003)	11.72	11.73	(0.921)	35.05	34.80
Sweden	5.74	5.77	(0.775)	4.20	4.22	(0.755)	53.96	53.80
Switzerland	6.99	6.96	(0.932)	4.60	4.41	(0.292)	56.77	56.94
United Kingdom	3.07	2.96	(0.015)	10.14	9.99	(0.088)	42.21	44.16
USA-S&P 500	3.64	3.63	(0.159)	13.07	13.05	(0.535)	46.96	47.15
USA-NASDAQ	2.81	2.83	(0.002)	10.92	10.92	(0.706)	44.17	43.87
USA-DJIA	3.22	2.67	(0.000)	12.90	11.67	(0.000)	44.62	51.64

Table 11- continued

Panel B: Emerging Markets								
	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
China	12.90	12.93	(0.999)	5.45	5.25	(0.129)	41.93	41.80
India	4.60	4.36	(0.024)	10.71	10.46	(0.081)	35.47	38.75
Malaysia	0.82	0.82	(0.498)	12.61	12.63	(0.891)	19.48	19.89
Mexico	3.52	3.52	(0.833)	12.02	11.98	(0.383)	30.11	30.15
Peru	23.96	23.81	(0.240)	14.06	14.10	(0.915)	27.09	28.01
South Africa	5.08	5.11	(0.430)	3.82	3.84	(0.933)	67.08	66.97
Thailand	2.56	2.57	(0.950)	10.20	10.21	(0.670)	33.85	33.82
Turkey	4.02	3.86	(0.001)	8.57	8.35	(0.011)	20.90	24.11

Panel C: Frontier Markets								
	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Croatia	18.09	18.33	(0.499)	11.06	11.03	(0.758)	42.91	42.59
Pakistan	1.50	1.48	(0.380)	20.66	20.39	(0.285)	18.91	19.20
Romania	3.92	3.75	(0.109)	10.97	10.77	(0.323)	36.89	40.65

Table 12

In- sample forecast evaluation of HAR and HAR+SV models

This table compares the in-sample forecasts of HAR and HAR+SV model. Uni. means the univariate model, HAR here. Aug. means the augmented model with lagged search volume, HAR+SV here. Performance measures are the mean squared error (MSE, $\times 10^5$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (%) of the regression. P-value is result of test which testing if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant. The model is better as MSE decreases, QL decreases and R^2 increases. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Australia	1.56	1.52	(0.002)	8.85	8.76	(0.054)	47.64	48.94
Austria	9.78	9.53	(0.325)	4.73	4.66	(0.666)	50.76	51.93
Belgium	5.76	5.64	(0.285)	4.74	4.70	(0.327)	62.04	62.71
Canada	2.45	2.25	(0.022)	9.35	9.15	(0.010)	58.71	61.23
France	2.89	2.68	(0.001)	8.91	8.65	(0.003)	48.88	52.11
Germany	2.82	2.53	(0.001)	8.90	8.64	(0.007)	54.45	58.76
Hong Kong	1.38	1.38	(0.127)	8.11	8.08	(0.419)	33.60	34.02
Italy	11.40	11.15	(0.375)	3.96	3.89	(0.360)	51.83	52.80
Japan	2.64	2.61	(0.026)	10.13	9.97	(0.018)	43.41	44.40
Netherlands	2.58	2.41	(0.002)	8.60	8.35	(0.007)	52.64	55.35
Singapore	1.05	1.05	(0.232)	7.76	7.73	(0.363)	44.30	44.53
Spain	2.91	2.90	(0.004)	9.71	9.70	(0.488)	42.35	42.54
Sweden	6.15	6.10	(0.633)	4.33	4.32	(0.743)	56.13	56.32
Switzerland	6.40	5.79	(0.373)	4.36	3.85	(0.015)	60.89	64.05
United Kingdom	2.46	2.31	(0.003)	7.76	7.62	(0.027)	52.80	55.13
USA-S&P 500	2.64	2.51	(0.001)	9.55	9.44	(0.049)	59.87	61.55
USA-NASDAQ	2.07	2.03	(0.002)	8.38	8.34	(0.091)	57.19	57.70
USA-DJIA	2.25	2.11	(0.001)	9.24	9.09	(0.013)	59.14	61.62

Table 12- continued

Panel B: Emerging Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
China	11.89	11.70	(0.462)	4.83	4.71	(0.176)	46.88	47.76
India	3.70	3.63	(0.066)	8.36	8.32	(0.328)	45.40	46.31
Malaysia	0.83	0.83	(0.583)	12.26	12.20	(0.554)	21.25	21.76
Mexico	2.93	2.91	(0.190)	10.12	10.06	(0.228)	41.85	42.19
Peru	22.21	22.22	(0.607)	12.65	12.65	(0.704)	31.96	31.92
South Africa	5.05	4.89	(0.197)	3.65	3.56	(0.231)	68.21	69.19
Thailand	2.37	2.37	(0.096)	9.09	9.09	(0.620)	38.51	38.57
Turkey	3.79	3.69	(0.004)	7.99	7.87	(0.044)	25.73	27.66

Panel C: Frontier Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Croatia	15.59	15.48	(0.267)	9.71	9.68	(0.690)	51.11	51.48
Pakistan	1.47	1.46	(0.533)	21.17	20.93	(0.265)	20.46	20.70
Romania	4.20	3.71	(0.086)	11.08	10.44	(0.096)	34.62	41.74

Table 13

In- sample forecast evaluation of EGARCH and EGARCH+SV models

This table compares the in-sample forecasts of EGARCH and EGARCH+SV model. Uni. means the univariate model, EGARCH here. Aug. means the augmented model with lagged search volume, EGARCH+SV here. Performance measures are the mean squared error (MSE, $\times 10^5$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (%) of the regression. P-value is result of test which testing if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant. The model is better as MSE decreases, QL decreases and R^2 increases. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Australia	2.86	2.82	(0.214)	14.63	14.31	(0.000)	49.95	51.41
Austria	43.96	41.10	(0.001)	8.65	8.57	(0.215)	64.07	63.84
Belgium	16.86	17.19	(0.898)	11.52	11.22	(0.532)	60.86	59.80
Canada	2.88	3.10	(0.064)	11.86	11.46	(0.000)	58.93	59.85
France	4.69	5.05	(0.008)	12.71	12.38	(0.001)	51.08	52.07
Germany	3.43	3.73	(0.005)	11.63	11.38	(0.006)	57.93	57.90
Hong Kong	5.76	5.76	(0.101)	19.46	19.46	(0.896)	29.81	29.76
Italy	36.47	57.91	(0.165)	8.94	9.00	(0.868)	49.48	47.48
Japan	6.91	6.87	(0.056)	22.60	22.36	(0.000)	46.46	47.61
Netherlands	4.69	5.38	(0.000)	12.55	12.49	(0.362)	56.52	55.46
Singapore	2.94	2.90	(0.016)	16.04	15.87	(0.006)	37.56	39.18
Spain	5.45	5.42	(0.000)	15.45	15.44	(0.095)	43.79	43.83
Sweden	9.13	22.77	(0.055)	6.51	8.46	(0.006)	46.05	37.39
Switzerland	85.73	61.02	(0.112)	10.15	6.76	(0.000)	42.78	52.59
United Kingdom	2.68	2.81	(0.010)	8.95	8.93	(0.597)	56.96	56.88
USA-S&P 500	3.29	3.43	(0.004)	12.40	12.26	(0.012)	63.69	64.18
USA-NASDAQ	3.44	3.42	(0.223)	13.12	13.12	(0.989)	58.97	58.45
USA-DJIA	2.43	2.52	(0.236)	10.95	10.29	(0.000)	62.33	64.48

Table 13- continued

Panel B: Emerging Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
China	20.35	19.33	(0.006)	7.74	7.51	(0.005)	38.58	41.85
India	7.06	7.19	(0.076)	12.83	12.67	(0.002)	43.54	44.08
Malaysia	1.05	1.05	(0.304)	16.94	16.96	(0.669)	24.35	24.51
Mexico	3.77	3.76	(0.630)	12.68	12.67	(0.761)	43.57	43.53
Peru	78.51	80.22	(0.207)	29.48	29.45	(0.827)	32.56	32.16
South Africa	7.34	7.40	(0.938)	6.05	4.68	(0.000)	59.12	63.14
Thailand	5.35	5.37	(0.140)	17.83	17.80	(0.097)	35.49	35.62
Turkey	6.50	6.53	(0.632)	11.93	11.79	(0.017)	26.39	27.44

Panel C: Frontier Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Croatia	91.72	77.53	(0.032)	17.17	17.10	(0.833)	53.06	53.48
Pakistan	2.22	2.33	(0.248)	20.57	21.36	(0.048)	13.42	8.06
Romania	4.14	4.08	(0.326)	13.24	13.02	(0.009)	38.98	40.09

DJIA's search volume can significantly improve the volatility forecasting in all three models with 95% confidence level. Search volume of S&P 500 can help to predict volatility both in HAR and EGARCH models. For NASDAQ, search volume is not helpful in forecasting future volatility except the HAR model. These results are consistent with the results of previous works, regression models and Granger causality tests. Just like what we do in previous section, we consider DJIA as the representative index of USA as we compare results among countries.

Throughout the results of comparison of in-sample volatility forecasts, we find in general, search volume can improve volatility forecasting since the loss functions, MSE and QL, reduce and R^2 increases in most countries. This conclusion is consistent with but not as statically significant as the results of regression models and Granger causality tests. Only half of countries in developed markets and two out of 8 countries in emerging countries show that search volume can significantly help to predict future volatilities. The valuation results of all three frontier countries are almost insignificant. That indicates that as the developed level of markets is worse, the phenomenon that search volume can help to forecast future volatility occur less.

4.3.2 Out-of-sample forecast evaluation

Table 14, Table 15 and Table 16 display the comparison results of the out-of-sample volatility forecasts of AR(1) vs. AR(1)+SV, HAR vs. HAR+SV and EGARCH vs. EGARCH+SV respectively. From these tables, we find that the results are consistent with the comparison results of in-sample volatility forecasts but are much more insignificant. Like Germany and Netherlands, both AR(1) and HAR models show search volume can help to forecast volatility in both in- and out-of -sample forecasts but they are less significant in out-of-sample forecasts. And some volatility models which can perform better when including search volume in-sample while can't out-sample, like AR(1) model in Australia, or inversely, such as EGARCH model in United Kingdom. Besides, in emerging markets, only EGARCH model show search volume can help to forecast future volatilities from Panel B of Table 16. That may due to only EGARCH model considers asymmetry.

We find that the models of countries which data is at weekly frequency usually have larger MSE than the models of countries with daily data both in- and out-of -sample, especially EGARCH model. And the models of countries with weekly data underperform those with daily data according to forecast valuations. Take Switzerland (weekly) and France (daily) for example. Both of them indicate that search volume is helpful in forecasting volatilities by the tests of VAR models and regression models. When we compare the forecasting ability of the volatility models, France shows that volatility can help to forecast volatility in- and out-of-sample, except EGARCH in

Table 14

Out-of- sample forecast evaluation of AR(1) and AR(1)+SV models

This table compares the out-of-sample forecasts of AR(1) and AR(1)+SV model. Uni. means the univariate model, AR(1) here. Aug. means the augmented model with lagged search volume, AR(1)+SV here. Performance measures are the mean squared error (MSE, $\times 10^5$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (%) of the regression. P-value is result of test which testing if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant. The model is better as MSE decreases, QL decreases and R^2 increases. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Australia	1.07	1.08	(0.454)	9.63	9.76	(0.087)	16.80	16.11
Austria	13.42	13.90	(0.557)	6.62	6.87	(0.335)	49.14	46.90
Belgium	5.40	5.45	(0.384)	4.63	4.59	(0.347)	42.76	42.10
Canada	1.30	1.32	(0.694)	10.94	11.18	(0.572)	24.95	24.29
France	2.85	2.33	(0.001)	9.47	8.83	(0.030)	38.43	49.76
Germany	3.08	2.41	(0.006)	11.08	10.18	(0.003)	40.37	51.69
Hong Kong	2.14	2.16	(0.306)	12.71	12.85	(0.310)	3.86	3.11
Italy	13.18	12.60	(0.659)	4.14	4.08	(0.752)	49.30	51.63
Japan	1.95	1.83	(0.015)	10.83	10.44	(0.180)	17.16	22.33
Netherlands	2.01	1.65	(0.015)	9.36	8.92	(0.081)	37.73	48.95
Singapore	1.06	1.07	(0.143)	9.46	9.48	(0.465)	9.49	8.90
Spain	3.51	3.45	(0.014)	10.89	10.77	(0.114)	20.30	21.88
Sweden	11.82	12.21	(0.012)	6.09	6.29	(0.003)	49.82	48.12
Switzerland	7.76	7.74	(0.995)	5.74	5.21	(0.304)	46.89	49.26
United Kingdom	2.11	1.92	(0.032)	9.60	9.24	(0.085)	28.31	34.74
USA-S&P 500	2.40	2.41	(0.811)	12.39	12.35	(0.776)	28.42	29.59
USA-NASDAQ	1.80	1.80	(0.957)	9.60	9.57	(0.387)	31.31	31.24
USA-DJIA	2.25	2.09	(0.121)	13.07	11.59	(0.000)	21.01	26.71

Table 14- continued

Panel B: Emerging Markets

	MSE (×10^5)			QL (×10^2)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
China	5.99	5.80	(0.106)	5.23	5.10	(0.128)	15.56	15.95
India	1.66	1.59	(0.028)	8.67	8.54	(0.353)	7.41	8.97
Malaysia	1.42	1.47	(0.014)	15.96	17.62	(0.002)	11.50	9.76
Mexico	2.17	2.17	(0.696)	11.18	11.17	(0.924)	15.95	15.65
Peru	17.77	17.97	(0.560)	11.32	11.56	(0.320)	21.23	20.22
South Africa	2.55	2.57	(0.217)	4.12	4.16	(0.130)	43.35	42.85
Thailand	1.36	1.36	(0.650)	7.93	7.96	(0.330)	30.65	30.56
Turkey	3.35	3.34	(0.917)	9.17	9.49	(0.104)	19.19	19.94

Panel C: Frontier Markets

	MSE (×10^5)			QL (×10^2)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Croatia	10.69	10.86	(0.450)	13.12	13.32	(0.149)	3.49	3.31
Pakistan	1.73	1.63	(0.021)	12.44	11.58	(0.034)	3.24	4.41
Romania	1.84	1.89	(0.447)	8.99	9.32	(0.417)	6.36	4.21

Table 15

Out-of- sample forecast evaluation of HAR and HAR+SV models

This table compares the out-of-sample forecasts of HAR and HAR+SV model. Uni. means the univariate model, HAR here. Aug. means the augmented model with lagged search volume, HAR+SV here. Performance measures are the mean squared error (MSE, $\times 10^5$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (%) of the regression. P-value is result of test which testing if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant. The model is better as MSE decreases, QL decreases and R^2 increases. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Australia	0.93	0.93	(0.669)	8.34	8.38	(0.666)	26.40	27.39
Austria	13.89	14.32	(0.492)	6.75	7.50	(0.138)	48.78	49.32
Belgium	5.66	5.67	(0.852)	4.60	4.61	(0.944)	38.61	38.12
Canada	1.21	1.22	(0.760)	9.69	9.88	(0.246)	29.88	31.55
France	2.67	2.44	(0.006)	8.66	8.32	(0.013)	41.68	45.95
Germany	2.68	2.36	(0.080)	9.43	9.06	(0.047)	47.60	52.84
Hong Kong	1.84	1.83	(0.103)	10.74	10.69	(0.423)	15.87	16.16
Italy	12.79	13.52	(0.706)	3.90	3.97	(0.765)	50.58	49.61
Japan	2.07	2.06	(0.719)	10.43	10.78	(0.076)	12.20	14.82
Netherlands	1.86	1.71	(0.052)	8.12	7.95	(0.217)	43.04	47.58
Singapore	0.89	0.90	(0.145)	7.46	7.55	(0.129)	22.50	22.98
Spain	3.05	3.02	(0.008)	9.20	9.17	(0.235)	29.10	29.93
Sweden	11.99	12.20	(0.736)	6.24	6.39	(0.428)	49.30	47.56
Switzerland	8.23	8.68	(0.854)	6.03	5.34	(0.241)	44.01	44.75
United Kingdom	1.87	1.77	(0.043)	8.16	8.03	(0.246)	37.05	40.93
USA-S&P 500	2.05	1.99	(0.496)	9.69	9.65	(0.790)	39.21	41.91
USA-NASDAQ	1.65	1.65	(0.997)	8.25	8.31	(0.247)	37.86	38.77
USA-DJIA	1.91	1.82	(0.046)	10.09	9.68	(0.000)	32.90	35.33

Table 15- continued

Panel B: Emerging Markets

Country	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
China	4.82	5.00	(0.339)	4.71	5.01	(0.150)	14.51	14.69
India	1.46	1.46	(0.905)	7.80	7.85	(0.403)	11.66	11.99
Malaysia	1.26	1.32	(0.025)	12.68	14.16	(0.003)	16.23	14.35
Mexico	1.94	1.98	(0.077)	10.04	10.33	(0.025)	24.20	24.43
Peru	17.47	17.58	(0.322)	10.89	11.00	(0.251)	22.69	22.33
South Africa	2.39	2.32	(0.527)	3.88	3.87	(0.954)	43.99	44.39
Thailand	1.39	1.39	(0.201)	7.86	7.88	(0.368)	30.25	30.10
Turkey	3.32	3.29	(0.718)	9.00	9.20	(0.200)	20.91	23.04

Panel C: Frontier Markets

Country	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Croatia	9.30	9.56	(0.053)	12.52	13.27	(0.089)	2.55	2.11
Pakistan	1.54	1.49	(0.154)	10.39	10.01	(0.168)	4.90	5.89
Romania	1.81	1.91	(0.340)	8.88	9.60	(0.229)	6.66	3.68

Table 16

Out-of- sample forecast evaluation of EGARCH and EGARCH+SV models

This table compares the out-of-sample forecasts of EGARCH and EGARCH+SV model. Uni. means the univariate model, EGARCH here. Aug. means the augmented model with lagged search volume, EGARCH+SV here. Performance measures are the mean squared error (MSE, $\times 10^5$), the quasi-likelihood loss function (QL, $\times 10^2$) and the R^2 (%) of the regression. P-value is result of test which testing if the differences between loss functions of the univariate models and ones of the respective augmented models are statically significant. The model is better as MSE decreases, QL decreases and R^2 increases. For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C provide the list of developed, emerging and frontier markets respectively.

Panel A: Developed Markets

	MSE ($\times 10^5$)			QL ($\times 10^2$)			R^2 (%)	
Country	Uni.	Aug.	p-value	Uni.	Aug.	p-value	Uni.	Aug.
Australia	3.24	2.75	(0.000)	19.09	17.71	(0.000)	38.35	36.94
Austria	19.52	47.13	(0.300)	11.20	10.49	(0.724)	53.95	33.23
Belgium	13.99	31.66	(0.000)	10.20	18.84	(0.000)	43.88	27.65
Canada	2.02	1.52	(0.000)	13.21	11.10	(0.000)	30.03	32.39
France	4.51	5.22	(0.000)	12.91	14.32	(0.000)	47.47	49.90
Germany	3.46	4.36	(0.000)	11.61	12.22	(0.001)	52.48	52.48
Hong Kong	4.84	5.37	(0.000)	18.72	20.51	(0.000)	16.49	17.69
Italy	23.56	51.46	(0.038)	6.60	7.35	(0.344)	60.24	59.64
Japan	8.04	6.51	(0.000)	30.40	26.07	(0.000)	17.84	18.82
Netherlands	3.47	3.41	(0.448)	13.77	13.58	(0.186)	47.82	49.47
Singapore	2.27	2.44	(0.000)	14.92	15.54	(0.000)	23.10	20.28
Spain	5.35	5.33	(0.322)	13.06	13.11	(0.136)	32.65	32.20
Sweden	25.04	30.97	(0.703)	10.43	13.37	(0.471)	32.74	35.11
Switzerland	13.85	6.65	(0.006)	12.37	5.83	(0.000)	48.51	62.07
United Kingdom	2.24	2.08	(0.002)	9.71	8.98	(0.000)	45.33	47.01
USA-S&P 500	3.34	3.43	(0.726)	13.81	11.91	(0.000)	42.05	41.36
USA-NASDAQ	3.88	3.91	(0.345)	15.12	15.31	(0.062)	33.29	33.70
USA-DJIA	2.49	2.79	(0.000)	12.32	13.03	(0.000)	35.52	37.72

Table 16- continued

Panel B: Emerging Markets

Country	MSE ($\times 10^5$) Uni.	MSE ($\times 10^5$) Aug.	MSE ($\times 10^5$) p-value	QL ($\times 10^2$) Uni.	QL ($\times 10^2$) Aug.	QL ($\times 10^2$) p-value	R^2 (%) Uni.	R^2 (%) Aug.
China	30.76	22.02	(0.001)	14.85	11.23	(0.003)	5.68	6.81
India	3.21	2.87	(0.000)	12.61	11.67	(0.000)	12.32	13.51
Malaysia	1.34	1.33	(0.801)	14.82	14.72	(0.823)	26.21	20.88
Mexico	3.23	2.97	(0.000)	15.06	14.22	(0.000)	22.79	21.80
Peru	70.01	67.86	(0.129)	31.10	29.99	(0.066)	17.50	17.37
South Africa	12.97	7.40	(0.000)	11.05	7.45	(0.000)	24.14	19.21
Thailand	4.64	5.26	(0.000)	21.06	22.34	(0.000)	25.33	23.06
Turkey	5.49	4.61	(0.000)	13.75	12.28	(0.000)	24.76	26.26

Panel C: Frontier Markets

Country	MSE ($\times 10^5$) Uni.	MSE ($\times 10^5$) Aug.	MSE ($\times 10^5$) p-value	QL ($\times 10^2$) Uni.	QL ($\times 10^2$) Aug.	QL ($\times 10^2$) p-value	R^2 (%) Uni.	R^2 (%) Aug.
Croatia	18.51	17.23	(0.144)	20.62	20.74	(0.947)	0.86	1.05
Pakistan	1.51	1.59	(0.632)	8.78	10.40	(0.119)	14.78	5.89
Romania	2.66	2.65	(0.937)	13.26	14.27	(0.245)	6.62	7.30

out-of-sample forecasts. But for Switzerland, only EGARCH model including search volume outperform the univariate model both in- and out-of sample and HAR with search volume perform better than the univariate model in-sample.

In frontier markets, the results of the out-of sample forecasts comparison are not consistent with the results of in-sample and neither with the previous results of VAR model and regression models, unlike developed and emerging markets. Take Croatia as an example. The Granger causality test shows search volume is not a useful predictor in forecasting volatility while regression models shows search volume can add additional information to both AR(1) and EGARCH models. But only EGARCH model in the in-sample forecasts can improve the forecasting ability of volatility. Inversely, Romania's search volume can Granger cause volatility but search volume cannot improve the forecasting power in addition to in-sample HAR model.

Overall, the HAR model augmented with search volume outperforms the other models both in- and out-of-sample since the MSE and QL are smallest and R^2 is largest for almost all countries. It indicates that HAR model augmented with search volume is the most optimal model to predict future volatility.

4.4 Why search volume can't help to forecasting volatilities in some countries?

From the above results, the leading countries in the world, like Canada, France, Germany, Japan, UK and USA all have the phenomenon that search volume can help to improve the forecasting ability of index volatility according to the tests and models. Besides the developed level of markets, we try to find why this phenomenon becomes unobvious in some countries from our tests and models, such as Hong Kong, Italy, Singapore, Spain, Sweden, Malaysia, Peru and Thailand.

Firstly, one possible reason is the lower frequency of data since lower frequency may ignore some information. It is documented that high frequency can forecast volatility more accurately because it can get more information about the reaction of price to movements. There are 9 countries with weekly data in our sample and only Switzerland and China have great results by all tests.

Next reason is the search terms are not very proper to use that we have discussed for S&P 500 and NASDAQ in section 4.3.1. The search terms that we use for each index can be seen from Table 3. Like Malaysia, we use "KLSE", one of ticker symbols, to search information about index since we can't find enough search data by short name of index. Some search terms are not univocal that have many meanings, like "SET" in Thailand, or have another important meaning, such as "STRAITS" which is short name of *The Straits Times* in Singapore. [5] And we find that in those

[5] *The Straits Times* (http://www.straitstimes.com/?a=1)

countries, Sweden, Peru, and Croatia, which we use both the (short) name of stock exchange as search terms and weekly data, search volume can't Granger cause volatility. Both Sweden and Peru even show that search volume hasn't valuable information about future volatility and can't help to predict future volatility

Third possible reason is that retail investors may not use Google to search information about the index. In our sample, there are only Hong Kong and China whose market shares of Google are below 70% from Table 2. But it's interesting that in Hong Kong, search volume is not useful in forecasting volatility while China shows that search volume is helpful to predict volatility. This may indicates that investors prefer use Google to search information about stocks in China.

Another probable reason is may be the locations of countries. From Table 17 we can see the continent for each country. We find that among those countries, where search volume can't help to improve the forecasting ability of volatility according to every test, half countries are in Asia. And among all Asian countries in our sample, only Japan and India have the good results that search volume is helpful to predict future volatilities by every test and model.

Last but not least reason is that individual investors usually search information about the stock index by newspaper or TV not by internet since the internet may not available for everyone, everywhere, and every time. We use penetration rate of internet users per country to check.[6] From Table 17, we could find penetration rates are consistent with the developed level of markets that higher in developed markets, lower in emerging frontier markets. Those countries which search volume can improve volatility forecasting usually have higher penetration rate. India has lowest penetration rate (10.2%) but search volume can significantly help to predict future

[6] Source: *Internet World Stats* (http://www.internetworldstats.com/stats.htm)

Table 17

Continent and penetration rate of internet users on 2011/12/31

This table presents the continent and penetration rate of internet users on 2011/12/31 for each country. Data source is *Internet World Stats* (http://www.internetworldstats.com/stats.htm). Penetration=internet users / population, (% of Population). For the name of countries in italic type mean that the data is at weekly frequency, ex: *Austria*. Panel A, B and C are developed, emerging and frontier markets respectively.

Panel A: Developed Markets

Country	Continent	Penetration
Australia	Australia	89.8%
Austria	Europe	74.8%
Belgium	Europe	81.4%
Canada	North America	81.6%
France	Europe	77.2%
Germany	Europe	82.7%
Hong Kong	Asia	68.7%
Italy	Europe	58.7%
Japan	Asia	80.0%
Netherlands	Europe	89.5%
Singapore	Asia	77.2%
Spain	Europe	65.6%
Sweden	Europe	92.9%
Switzerland	Europe	84.2%
United Kingdom	Europe	84.1%
USA	North America	78.3%

Panel B: Emerging Markets

Country	Continent	Penetration
China	Asia	38.4%
India	Asia	10.2%
Malaysia	Asia	61.7%
Mexico	Latin America	36.9%
Peru	Latin America	34.1%
South Africa	Africa	13.9%
Thailand	Asia	27.4%
Turkey	Europe	44.4%

Panel C: Frontier Markets

Country	Continent	Penetration
Croatia	Europe	59.2%
Pakistan	Asia	15.5%
Romania	Europe	39.2%

volatility in this country. That maybe because the internet users are the riches which are the main retail investors in stock market.

Besides above reasons, we think market shares of retail investors in stock markets is maybe one important reason. When huge movements of stock price catch retail investors' eyes and then make investors to invest in, the volatility may not change a lot if the market shares of individual are very low. However, it's so difficult to get the practical number of market shares of individual investors that we can't analyze if it is one of reasons.

5. Conclusion

In this paper, we use Google search volume (Da, Engelberg and Gao (2011)) as proxy of retail investors' attention to study the dynamic relationship with stock market volatility, test whether it can add more information for modeling volatility, examine if it can help to predict volatility in- and out-of-sample per country and compare these phenomenon in different markets.

We find search volume is useful to predict future volatility generally by Granger causality tests and half of countries' Granger causality is bi-directional: high search activities follow high volatility, and high volatility follows high search activities. However, when there is a positive shock of search volume, volatility wouldn't react immediately but have positive movement later, while volatility can affect search volume immediately by impulse response function. Throughout all countries, movement of volatility contributed by search volume is ranging from 0.11% to 20.53% by variance decomposition. From regression models, AR(1), HAR and EGARCH, we discover that search volume adds information to the volatility model and influences future volatility positively. Search volume also can improve volatility forecasting in- and out-of-sample by comparing the mean squared errors (MSE), the quasi-likelihood loss functions (QL) and the R^2s of the volatility models with and without lagged search volume. But it becomes much more insignificantly in out-of-sample forecast evaluation.

The phenomenon that search volume can help to predict future volatility becomes more unapparent when turning to emerging markets and frontier markets. Besides the developed level of markets, there are some possible reasons, like lower frequency of data, less univocal search terms, lower market shares of Google, locations of countries, smaller penetration rate of internet users and lesser market shares of retail investors, can explain why this phenomenon is unobvious for some countries from our empirical results.

However, to discuss the phenomenon in frontier markets is a little critical since there are only three frontier counties in our sample. For many frontier countries, we can’t find daily search volume but weekly data. And weekly data also have a lot missing data before 2011. In the future, we can do more study in frontier markets when there are enough search volume data.

Reference

Ait-Sahalia, Y. and Mancini, L., 2008, Out of sample forecasts of quadratic variation, *Journal of Econometrics* 147(1), 17-33.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F. and Diebold, F. X., 2006, Practical Volatility and Correlation Modeling for Financial Market Risk Management, in M. Carey and R. M. Stulz (eds), *The Risks of Financial Institutions*, University of Chicago Press, Chicago, Illinois, chapter 17, pp. 513-548.

Andersen, T. G., Bollerslev, T., Diebold, F. X. and Labys, P., 2003, Modeling and Forecasting Realized Volatility, *Econometrica* 71(2), 529-626.

Bank, M., Larch, M. and Peter, G., 2011, Google search volume and its influence on liquidity and returns of German stocks, *Financial Markets and Portfolio Management* 25, 239-264.

Barber, B. M. and Odean, T., 2008, All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors, *Review of Financial Studies* 21(2), 785–818.

Berry, T. D. and Howe, K. M., 1994, Public information arrival, *Journal of Finance* 49(4), 1331–1346.

Bollen, B. and Inder, B., 2002, Estimating daily volatility in financial markets utilizing intraday data, *Journal of Empirical Finance* 9, 551-562.

Chemmanur, T. and Yan, A., 2009, Advertising, attention, and stock returns, Working paper, Boston College and Fordham University.

Choi, H. and Varian, H., 2009, Predicting the present with Google Trends, Working paper, Google Inc.

Corsi, F., 2009, A Simple Approximate Long-Memory Model of Realized Volatility, *Journal of Financial Econometrics* 7(2), 174-196.

Da, Z., Engelberg, J. and Gao, P., 2010a, In search of earnings predictability, Working Paper, University of Notre Dame and University of North Carolina at Chapel Hill.

Da, Z., Engelberg, J. and Gao, P., 2010, In search of fundamentals, Working paper, University of Notre Dame and University of North Carolina at Chapel Hill.

Da, Z., Engelberg, J. and Gao, P., 2011, In Search of Attention, *The Journal of Finance* 66(5), 1461-1499.

Dimpfl, T. and Jank, S., 2011, Can internet search queries help to predict stock market volatility?, Working Paper, University of Tubingen.

Drake, M., Roulstone, D. and Thornock, J., 2011, Investor Information Demand: Evidence from Google Searches around Earnings Announcements, Working Paper, Brigham Young University, Ohio State University and University of Washington.

Dorn, D., Huberman, G. and Sengmueller, P., 2008, Correlated trading and returns, *Journal of Finance* 63, 885–919.

Fang, L., and Peress,J., 2009, Media coverage and the cross-section of stock returns, *Journal of Finance* 64, 2023–2052.

Foucault, T., Sraer, D. and Thesmar, D. J., 2011, Individual Investors and Volatility, *The Journal of Finance* 66(4), 1369-1406.

Frieder, L. and Subrahmanyam, A., 2005, Brand perceptions and the market for common stock, *Journal of Financial and Quantitative Analysis* 40(1), 57–85.

Gervais, S., Kaniel, P. and Mingelgrin, D.H., 2001, The high-volume return premium, Journal of Finance 56, 877–919.

Ghysels, E., Santa-Clara, P. and Valkanov, R., 2006, Predicting Volatility: Getting the Most out of Return Data Sampled at Different Frequencies, *Journal of Econometrics*1-2, 59-95.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L., 2009, Detecting influenza epidemics using search engine query data, *Nature* 457, 1012–1014.

Grullon, G., Kanatas, G. and Weston, J.P., 2004, Advertising, breath of ownership, and liquidity, *Review of Financial Studies* 17, 439–461.

Hou, K., Peng, L., and Xiong, W., 2008, A tale of two anomalies: The implications of investor attention for price and earnings momentum, Working paper, Ohio State University and Princeton University.

Hvidkjaer, S., 2008, Small trades and the cross section of stock returns, *Review of Financial Studies* 21, 1123–1151.

Jayasuriya, S., Shambora, W., 2009. Oops, we should have diversified! *Applied Financial Economics* 19, 1779-1785.

Joseph, K., Wintoki, M.B., Zhang, Z., 2011, Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search, *Journal of Forecasting* 27, 1116-1127.

Kaniel, R., Saar, G. and Titman, S., 2008, Individual investor trading and stock returns, *Journal of Finance* 63, 273–310.

Kumar, A., and Lee, C.M.C.2006, Retail investor sentiment and return co movements, *Journal of Finance* 61, 2451–2486.

Lou, D., 2008, Attracting investor attention through advertising, Working paper, London School of Economics and Political Science.

Lux, T. and Marchesi, M., 1999, Scaling and criticality in a stochastic multi-agent model of a financial market, *Nature* 397(6719), 498-500.

Merton, R. C., 1987, A simple model of capital market equilibrium with incomplete information, *Journal of Finance* 42(3), 483–510.

Mincer, J. A. and Zarnowitz, V., 1969, The Evaluation of Economic Forecasts, in J. A. Mincer (ed.), *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance, Studies in Business Cycles*, NBER.

Mitchell, M. L. and Mulherin, J. H., 1994, The impact of public information on the stock market, *Journal of Finance* 49(3), 923–950.

Patton, A. J., 2011, Volatility forecast comparison using imperfect volatility proxies, *Journal of Econometrics* 160(1), 246-256.

Seasholes, M.S., and Wu, G., 2007, Predictable behavior, profits, and attention, *Journal of Empirical Finance* 14, 590–610.

Vlastakis, N. and Markellos, R. N., 2010, Information demand and stock market volatility, *SSRN eLibrary*.

Yuan, Y., 2008, Attention and trading, Working paper, University of Iowa.