Also available at http://amc-journal.eu
ISSN 1855-3966 (printed edn.), ISSN 1855-3974 (electronic edn.)
ARS MATHEMATICA CONTEMPORANEA 9 (2015) 301–320

What can Wikipedia and Google tell us about stock prices under different market regimes?

Boris Cergol [*]
XLAB, Ljubljana, Slovenia

Matjaž Omladič [†]
Department for automation, biocybernetics and robotics,
Jozef Stefan Institute, Ljubljana

Received 21 October 2013, accepted 19 December 2013, published online 10 July 2015

Abstract

In less than five years a surprisingly high level of attention has built up in the possible connection between internet search data and stock prices. It is the main aim of this paper to point out how this connection may depend heavily on different regimes of the market, i.e. the bear market vs. the bull market. We consider three types of internet search data (relative Google search frequencies of company tickers, relative Google search frequencies of company names and page visits of Wikipedia articles about individual companies) and a substantial sample of companies which are members of the S&P 500 index. We discover two inverse patterns in stock prices: in the bear market what we propose to term a "merry frown" and in bull market a "sour smile", both clearly seen especially for the Wikipedia data. We propose market neutral strategies that exploit these new patterns and yield up to 17% in average annual return during our sample period from 2008 to 2013.

Keywords: Stock returns, internet search data, market regimes, trading strategies.

Math. Subj. Class.: 91G70, 97K80

1 Introduction

A byproduct of the increasingly widespread use of the internet is the data on internet activity of individual users. While most of this data is retained by the website owners and unavailable to the public either due to privacy or business reasons there are some exceptions. One such example is the Google Trends service which enables users to view the

[*] Operation part financed by the European Union, European Social Fund.

[†] Supported by a grant from the Slovenian Research Agency - ARRS.

E-mail addresses: boris.cergol@xlab.si (Boris Cergol), matjaz@omladic.net (Matjaž Omladič) relative frequencies of search queries entered into Google's search engine. Since becoming publicly available in 2006 Google Trends have attracted attention of researchers in various fields. In [12] the authors show that analysis of health-related search queries can lead to accurate estimates of influenza epidemics with a reporting lag of only one day which is almost two weeks sooner than traditional surveillance systems. Choi and Varian [6] apply a similar approach to estimating a number of economic indicators such as automobile sales or unemployment claims.

The relevance of internet search data for financial data analysis was first explored by Da, Engelberg and Gao [6] who considered the relative search frequencies of company tickers and names as proxies for investor attention in the US stock market. They show that search frequencies outperform existing measures of investor attention and that an increase in a company ticker's search frequency predicts a higher stock price in the following two weeks. In [2, 14] the authors obtain similar results in terms of future returns and additionally observe that an increase in a company name's search frequency is associated with a rise in trading activity and stock liquidity. The prevailing explanation for positive correlation between future stock returns and company-related relative search frequencies is based on the theory of Barber and Odean [3]. They suggest that attention-grabbing stocks experience short-term buying pressure from individual investors. This might simply be due to the fact that a single investor faces a difficult decision when deciding which of the thousands of available stocks to buy, while the decision of which stock to sell is much easier since it is usually limited to the few stocks that are part of his existing portfolio.

Google trends data has also been used in assessing investor sentiment. In [8] the authors construct the index which is a sum of relative search frequencies of economy-related terms associated with negative sentiment. This new index is able to predict values of existing investor sentiment indicators and has a perceptible impact on short-term future stock prices. In [18] a number of stock market index strategies are tested that profit from fluctuations of relative search frequencies of individual economy-related terms. Strategies of the same type are further explored in a related work [17] where Google trends data is replaced by the numbers of page visits to economy-related Wikipedia pages.

The main contribution of our paper is the addition of market regimes into the study of the connection between stock returns and internet search. If the reason for positive correlation between future returns and search frequencies is in fact in the cognitive bias of individual investors then we would expect that the effect would be even stronger in periods when investors face greater uncertainty and are even more prone to irrational decisions. We present a two-state hidden Markov model for the returns of the S&P 500 index. The model parameters are estimated by the Baum-Welch algorithm after which the most likely sequence of hidden states is found by the Viterbi algorithm. The first of the two states is characterized by low returns and high volatility and corresponds to what is commonly refered to as the "bear market" regime by investors. Conversely, the second state is characterized by high returns and low volatility and we label it the "bull market" regime.

We choose a sample of stocks that are members of the S&P 500 index and study the relation between their future short term returns and three different types of internet search data: the page visits of company-related articles on Wikipedia, the relative frequency of Google searches for company tickers and the relative frequency of Google searches for company names. To the best of our knowledge ours is the first study of this kind that takes Wikipedia data for individual companies into consideration. We also perform our analysis on daily data which is in contrast to most of the existing literature where financial applications of internet search is studied using weekly data.

We perform a number of cross-sectional Fama-MacBeth regressions where future stock returns are the explained variable and a single internet search variable is the regressor. This regressions are performed on a subsample of observations that belong to either the bear regime or the bull regime as well as on the entire sample. Our main result is that the market regime indeed has a strong influence on the relation between future stock returns and internet search data. In all three cases of internet search variables the future returns are higher in the bear regime compared to the bull regime given the same increase of the chosen internet search variable. This effect remains evident even after controlling for the factors of the Carhart four-factor model [5].

After controlling for the Carhart factors the Wikipedia page visits variable emerges as the one with the greatest influence on future stock returns. In fact, both of the Google search variables prove to be statistically insignificant. To our surprise, we also find very little evidence supporting the theory that an increase of investor attention to a given stock translates into a short term rise of the stock's price due to increased buying pressure. Instead we observe two different price patterns for which we propose the terms "merry frown" and "sour smile". A merry frown is a pattern of positive correlation between future stock returns and Wikipedia page visits that is observed only during the bear market. A sour smile is a pattern of negative correlation between future stock returns and Wikipedia page visits that is observed only during the bull market. Both patterns might be explained as a corrective investor counter-reaction to initial overpessimism in the bear market and to initial overoptimism in the bull market.

Economic significance of the merry frown and the sour smile is explored by constructing a market neutral strategy with long positions in stocks that are in the highest decile and short positions in stocks that are in the lowest decile with regard to Wikipedia page visits during the bear market. In the bull market the positions are reversed. We backtest the strategy for different trading frequencies (from 1 to 10 days) and observe that they generate positive returns which decrease with the length of trading frequency. The returns of the strategies are compared to random market neutral strategies generated by a Monte Carlo simulation and their statistical significance is established. We also find that returns of the strategies strongly increase if they are restricted to a subsample of stocks that are preferred by individual investors such as high volatility stocks, low market capitalization stocks or low price-to-book ratio stocks. In the best case, a trading strategy with daily trading frequency that is restricted to stocks with higher than median volatility yields an average annual return of 17% in our sample period.

The paper is organized as follows: Section 2 describes the data. The market regime model is presented in Section 3. In Section 4 we discuss the results of the Fama-MacBeth regressions. The trading strategies and backtest results are presented in Section 5. Section 6 concludes the paper.

2 Data description

The most important choice in the beginning of every statistical research is the choice of statistical population. We decided to limit our study to a sample of stocks that are included in the index S&P 500. More precisely, the stocks that were members of this index on June 7, 2013. Our choice is primarily motivated by the fact that the publicly accessible and freely available data on individual stocks is of highest quality for stocks listed on the

US stock market. Of course, the extension of our study to stocks listed on some less common stock markets remains a challenge for future investigation. We choose a sampling period from October 1, 2007, up to June 30, 2013. We restrict ourselves to this specific sampling period primarily because of the availability of website search data. Fortunately, this period includes very diverse market conditions including one of the greatest market crashes in history and the following rebounding growth. This gives us confidence that our findings would easily extend to future periods. For every stock in our sample we obtain the daily dividend and split adjusted closing prices from the Yahoo Finance website[1]. We additionally remove all stocks for which data is not available for the entire sample period. This mostly includes stocks that were members of the S&P 500 index on June 7, 2013, but were not yet publicly traded at the begining of our sample period.

We get the daily closing values of the index S&P 500 in the economic data section of the website of Federal Reserve Bank of St. Louis[2]. This data is obtained for a longer period from January 1, 2000, up to June 30, 2013, as required by the regime-switching model described in Section 3.

There are various choices for website search data that are worth testing for possible relations with fluctuations in stock prices. One of the possibilities is Google search data which is publicly available via the Google Trends Service[3]. In related studies, authors most commonly use relative search frequencies of stock tickers [7, 14] while some also consider relative search frequencies of company names [2]. The majority of studies rely on weekly data for these frequencies. This might be based on availability problems with Google data. When one requests a search frequency time series for a certain term the format of the returned series depends on the length of the period. For periods no longer than 3 months one gets the daily data, but for longer periods only weekly data is returned. An additional feature of the data so obtained is that it is normalized within the series to have a maximum value of 100. This has some advantages but also makes it difficult to compare values of series in different periods. This may have been the reason for most authors to restrict their studies to weekly data.

In order to overcome this difficulty, we acquire the three-month-period data every two months. From the data for the overlapping month we compute the quotient between the normalized factors of the two consecutive periods thus enabling us to concatenate the short period time series into one long period time series with the daily data. An additional problem arises with company names. Namely, one would need to know which name people are using for the company when searching for information about it. For instance, it is unlikely that most people would search for American Express Company by typing its full name into the search window but would instead just type American Express or simply AmEx. Accordingly, we replace company names in our sample with suitable abbreviations. By following the concatenating procedure described above we obtain a daily time series of relative search frequencies for both a company ticker and an abbreviated company name for each company in our sample. The sample period for this data is chosen to be from January 1, 2008, up to May 31, 2013.

Another source of internet search data that has recently been studied related to financial data is Wikipedia. For every article on Wikipedia, a time series of daily unique page visits

[1] http://www.finance.yahoo.com

[2] http://research.stlouisfed.org/fred2

[3] http://www.google.com.au/trends

can be obtained from the website Stats.Grok.Se[4]. This source of information has not gained as much attention as the Google Trends Data and the reason for this may lie in the fact that the available time series only span from January 2008. For every company in our sample, we find the Wikipedia article associated to the company and obtain a time series of unique daily page visits to this article in a sample period from January 1, 2008, up to May 31, 2013.

There are some comments we have to make that relate to data preprocessing of both Google search and Wikipedia page visits data. We first note that this data is not available throughout the chosen period for each company in our sampling period. Therefore, we have to exclude the companies with too much missing data. Our rule is to allow no more than 10% of missing data, while for the missing values we apply an imputation procedure that takes into account the weekly seasonality. Next we perform a detrendization of data using a sort of "longitudinal normalization", i.e. we divide the number for a given day by the average of the numbers for the last 56 days. The choice of length for this normalization period is similar to choices made by authors in related studies, for example in [7] where the length of the normalization period is 8 weeks. We also have to take care of the outliers. Their influence is reduced by taking a logarithm transformation of our data. Additionally, for each stock in our sample and each of the three variables, we perform a winsorization of the corresponding time series by limiting the range of data to its first percentile from below and to its 99th percentile from above.

There are strong seasonal effects on the weekly basis in both Google and Wikipedia data. We want to make data collected on different days of the week comparable by introducing a seasonal adjustment in the following way. We regress the data to the days-of-the-week (but one) as dummy variables to get average differences between the different days of the week which we add to the data of this particular day.

3 Market regimes

We intend to study the influence of market regimes on the relation between internet search data and stock returns. A market regime may be considered as a phase of persistent attributes observed in financial time series. This concept is most commonly used by investors when classifying the market into two phases: the bear market characterized by low returns and high volatility and the bull market characterized by high returns and low volatility. This dichotomous approach also lies at the core of our view on market dynamics. A systematic regime-switching time series model was first proposed by Hamilton [13] and the variants of this model are still being overwhelmingly used to study the regimes of economic and financial time series. However, recently, the problem of parameter estimation in financial regime-switching models has also been tackled by the Baum-Welch algorithm [19] which has previously been mostly used in engineering applications. We also decided to base our market regime estimation model on the Baum-Welch algorithm and refer the reader to [15] for a discussion of its advantages over Hamilton's algorithm.

We will determine the actual switching of the regimes only on the cumulative level and base it on the returns data of the S&P 500 index. The background of the model is a hidden Markov chain (denoted by $\mathcal{Q} = (q_t)_{t=1}^T$, where T is the length of the period) so that the market regimes are seen as states of this chain, the bear market becomes say state $i = 1$ and the bull market state $i = 2$. As a consequence the model has a 2×2

[4] http://stats.grok.se

transition matrix A defined in the usual way: $a_{ij} = p(q_{t+1} = j \,|\, q_t = i)$. Given an initial distribution of the states $\Pi = (\pi_1, \pi_2)$ we then have the Markov chain uniquely determined. Furthermore, we assume that the observations form a random sequence denoted by $\mathcal{O} = (o_t)_{t=1}^{T}$ where each o_t is the index return at time t and is determined randomly following a normal distribution $N(\mu_i, \sigma_i)$, for $i = 1, 2$, where the two parameters of this distribution depend on the state of the hidden Markov chain, i.e. on the market regime. We will call them the observation distributions and denote the corresponding sequence of distributions by $\mathcal{B}$. The entire model will be denoted by $\mathcal{M} := \{A, \Pi, \mathcal{B}\}$.

We first present the *forward algorithm* which helps computing the so called *forward variable*

$$\kappa_t(i) = p(o_1 o_2 \dots o_t, q_t = i \,|\, \mathcal{M}),$$

Here, κ is defined recursively:

$$\kappa_1(i) = \pi_i b_i(o_1),$$

for $i = 1, 2$, where $b_i(o_1)$ is the density of $N(\mu_i, \sigma_i)$ at the point o_1. Furthermore,

$$\kappa_{t+1}(j) = \left[\sum_{i=1}^{2} \kappa_t(i) a_{ij} \right] b_j(o_{t+1}), .$$

for $j = 1, 2$ and $t = 1, 2, \dots T - 1$ and $b_j(o_{t+1})$ is defined by analogy with the above. Likelihood $p(O|\mathcal{M})$ can be computed using the forward variable in the following manner:

$$p(O|\mathcal{M}) = \sum_{i=1}^{2} \kappa_T(i).$$

In the Baum-Welch algorithm we also need the *backward variable*

$$\varrho_t(i) = p(o_{t+1} o_{t+2} \dots o_T \,|\, q_t = i, \mathcal{M})$$

which we compute recursively using the *backward algorithm*. We first initialize $\varrho_T(i) = 1$ for $i = 1, 2$, and then let

$$\varrho_t(i) = \sum_{j=1}^{2} a_{ij} b_j(o_{t+1}) \varrho_{t+1}(j)$$

for $i = 1, 2$ and $t = T - 1, T - 2, \dots, 1$. Likelihood $p(O|\mathcal{M})$ can be computed using the backward variable in the following manner:

$$p(O|\mathcal{M}) = \sum_{i=1}^{2} \pi_i \varrho_1(i) b_i(o_1).$$

To initialize the Baum-Welch algorithm we choose a starting estimate of the model denoted by $\widehat{\mathcal{M}}_0$, and then we compute the likelihood $p(\mathcal{O} \,|\, \widehat{\mathcal{M}}_0)$ using the forward variable.

At this stage we start the iterative procedure consisting of four steps. The first step is to compute the forward variable $\widehat{\kappa}_t(i)$ and the backward variable $\widehat{\varrho}_t(i)$ based on the estimate of the model obtained on the previous iteration step $\widehat{\mathcal{M}}_k$. The final result of this step is the likelihood of transition from the state i to the state j given the model $\widehat{\mathcal{M}}_k$ and the observations $\mathcal{O}$ when time goes from t to $t+1$:

$$\begin{aligned}\widehat{\psi}_t(i,j) &= p(q_t = i, q_{t+1} = j|\mathcal{O}, \widehat{\mathcal{M}}_k)\\ &= \frac{p(q_t = i, q_{t+1} = j, \mathcal{O}\,|\,\widehat{\mathcal{M}}_k)}{p(\mathcal{O}\,|\,\widehat{\mathcal{M}}_k)}\\ &= \frac{\widehat{\kappa}_t(i)a_{ij}b_j(o_{t+1})\widehat{\varrho}_{t+1}(j)}{p(\mathcal{O}\,|\,\widehat{\mathcal{M}}_k)}.\end{aligned}$$

The first equation above is just the definition, in the second one we use the conditional formula and in the third one we express the likelihood with (the estimates of) the forward and backward variable. Next, we express the denominator of the last fraction above also using (the estimates of) the forward and backward variable:

$$p(\mathcal{O}\,|\,\widehat{\mathcal{M}}_k) = \sum_{i=1}^{2}\sum_{j=1}^{2}\widehat{\kappa}_t(i)a_{ij}b_j(o_{t+1})\widehat{\varrho}_{t+1}(j).$$

On the second step of the iteration procedure we need to estimate another likelihood, i.e.

$$\Gamma_i(i) = p(q_t = i\,|\,\mathcal{O}, \mathcal{M}) = \sum_{j=1}^{R}\psi_t(i,j).$$

Using the estimates of the first step $\widehat{\psi}_t(i,j)$ we compute the next estimate of the model $\widehat{\mathcal{M}}_{k+1}$. First, we compute the elements of the transition matrix A

$$\widehat{a}_{ij} = \frac{\sum_{t=1}^{T-1}\widehat{\psi}_t(i,j)}{\sum_{t=1}^{T-1}\Gamma_t(i)},$$

followed by the initial distribution Π

$$\widehat{\pi}_i = \Gamma_1(i)$$

and finally the observation distributions $\mathcal{B}$

$$\widehat{b}_j(s) = \frac{\sum_{t=1}^{T}\Gamma_t(j)'}{\sum_{t=1}^{T}\Gamma_t(i)}.$$

Here we understand $\Gamma_t(j)'$ given $o_t = s$.

On the third step of the iteration procedure we compute the likelihood using the new model $\widehat{\mathcal{M}}_{k+1}$. The fourth step is decisive: we compare the estimates of these likelihoods on the last two steps. If they are close enough, we stop the algorithm. If not, we proceed with another iteration starting with step one.

So, the final result of this algorithm is an estimate of the hidden Markov model $\mathcal{M} := \{A, \Pi, \mathcal{B}\}$. Based on this estimate, we want to give a prediction of the most probable state (bull or bear regime) for each point of time in the period. This will be done using the Viterbi algorithm. We first introduce, for $i = 1, 2$, the *Viterbi variable*

$$\delta_t(i) = \max_{q_1,q_2,\ldots,q_{t-1}} p(q_1, q_2, \ldots, q_{t-1}, q_t = i, o_1, o_2, \ldots, o_t\,|\,\mathcal{M})$$

which means the conditional likelihood of the most likely path of length t ending in state i given the model, where o_t are the the actual observed values of the index under consideration. We also need the value of the last but one state in this optimal path that ends in state i which we denote by $\tau_t(i)$.

We initialize by letting $\delta_1(i) = \pi_i b_i(o_1)$ and $\tau_1(i) = 0$ (an "empty state" on which nothing really depends) for $i = 1, 2$. The inductive steps of the algorithm go for $t = 2, 3, \ldots, T$. The dynamic programming approach yields

$$\delta_t(i) = b_i(o_t) \max_{i=1,2} (\delta_{t-1}(i) a_{ij})$$

together with

$$\tau_t(j) = \arg\max_{i=1,2} (\delta_{t-1}(i) a_{ij}).$$

At the end of the algorithm we terminate with the final optimal regime

$$q_T^* = \arg\max_{1=1,2} (\delta_{T-1}(i) a_{ij})$$

and then backtrack the whole optimal path

$$q_t^* = \tau_{t+1}(q_{t+1}^*)$$

for $t = T-1, T-2, \ldots, 1$.

Both the Baum-Welch algorithm and the Viterbi algorithm are not what we usually call online algorithms that would process the input data in the sequence they would be fed to the algorithm. This is a shortcoming since we are looking for a way to determining the market regime as a stopping time in the sense of martingale theory, i.e. the decision about a certain point in time can be made only based on the data of previous points in time. We are overcoming this obstacle by implementing it in an expanding window approach. The starting window in our approach will be the period from January 2, 2000 to January 2, 2008 (because January 2 is the first trading day in a year). On each step we expand the window by one trading day until we reach May 31, 2013 where the period that we are interested in ends. In each of these windows we run the Baum-Welch and the Viterbi algorithm and retain only the final optimal regime q_T^*. This way we determine the optimal market regime for each trading day of the period we are interested in a stopping time manner.

In Figure 1 we present the results of the algorithm described above. In this figure the daily values of the S&P 500 index are superimposed over two backgrounds – the red one corresponds to days of the bear market and the blue to days of the bull market according to our estimation. It is clear that our model is able to recognize quite well the bear market of 2008/2009 as well as the more pronounced market corrections in the following years. However it does perform less admirably when recognizing the beginning of the bull market in 2009. This is not unexpected since this period was characterized by extremely high returns as well as high, albeit decreasing, volatility. Such conditions are not well aligned with our model which assumes only two market regimes - the one with high returns and low volatility and the one with low returns and high volatility. An obvious solution to this problem would be extending our model to 4 regimes. However, fitting such a model would require a much larger data sample than we have available. For example, the authors of [16] fit a 4-regime model on 123 years of data. Due to this limitation, we decided for the more parsimonous 2-regime model.

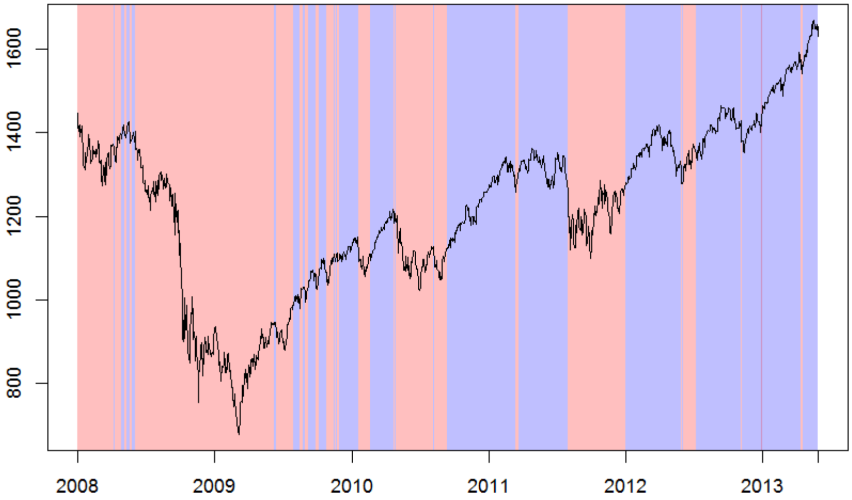

Figure 1: S&P 500 index during the bear regime (red background) and the bull regime (blue background)

4 Linear regression

In this section we study the connection between the stock returns and the data described in Section 2. Our main statistical tool will be the Fama-Macbeth cross-sectional linear regression [9]. This is a two-step procedure where a cross-sectional regression is performed for each time unit and then the time-series average of the estimated regression coefficients is calculated.

We first present the results of an analysis in which the explained variables are the cumulative future returns, while the search data is used as the regressor. For each k running from 1 to 15 we perform a Fama-Macbeth regression for the cumulative return from the time $T = t + 1$ to the time $T = t + k$ as the explained variable which we denote $r_{i,t+k}$ where i is the index spanning our entire selection of stocks. This way we allow for different periods of time that may be of interest ranging from 1 trading day to roughly 1 trading month. We perform an additional regression to test the contemporary return, i.e. the return observed on $T = t$. As described before we use three types of search data: Wikipedia page visits (denoted by $\mathrm{wiki}_{i,t}$), Google search queries for company tickers (denoted by $\mathrm{goog\_tickers}_{i,t}$) and Google search queries for company names (denoted by $\mathrm{goog\_names}_{i,t}$). All this data is taken at time $T = t$. The cross-sectional regressions performed for each time unit t and each $k = 0, 1, \ldots, 15$ are described by the following equations:

$$\begin{aligned} r_{i,t+k} &= \alpha_{i,t} + \beta_{\text{wiki},t}\, \text{wiki}_{i,t} + \varepsilon_{i,t+k}, \\ r_{i,t+k} &= \alpha_{i,t} + \beta_{\text{goog\_tickers},t}\, \text{goog\_tickers}_{i,t} + \varepsilon_{i,t+k}, \\ r_{i,t+k} &= \alpha_{i,t} + \beta_{\text{goog\_names},t}\, \text{goog\_names}_{i,t} + \varepsilon_{i,t+k}. \end{aligned}$$

There is an additional regressor we want to test for its influence which has a form of a dummy variable, i.e. the market regime. So, in practice we perform three actual regressions for each case of interest, one for the bear markets, one for the bull markets and one for the joint data independent of the regime. The results are presented in Table 1.

To make the presentation clearer we standardized the search data on every fixed date under consideration so that the regression coefficient has a simple interpretation. It gives an increase in the average return (positive or negative) given that the average internet search variable increases by one standard deviation. When we want to present the relation of these data to the length of the period we run into another difficulty, namely that the cumulative returns computed for different periods are not immediately comparable since their magnitude trivially depends on the period, so we decided to annualize them. There are three graphs in Figure 2. The first one presents dependence of annualized returns (based on regression coefficients) for the Wikipedia page visits, where the red line presents the data of the bear market, the blue one the data of the bull market and the purple one the joint data. A similar graph is presented for the Google search data for company tickers and the third one for the Google search data for company names.

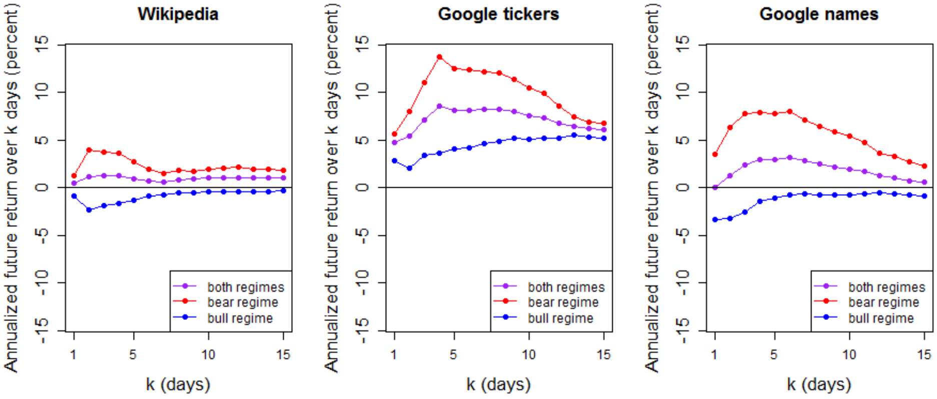

Figure 2: Changes in annualized future returns over k days after we observe a one standard deviation increase in individual search variables.

We first observe that the regression results in the case when market regimes are not taken under consideration differ substantially from the results when we do take them into account. Actually, in all the three cases of internet search variables we observe that a raise in the internet search variable is associated with higher future returns in the bear market compared to future returns in the bull market. This is confirming our starting hypothesis that market regimes have a strong influence on the connection between internet search data

	Wikipedia			Google tickers			Google names		
k	Both	Bear	Bull	Both	Bear	Bull	Both	Bear	Bull
0	**0.018***	0.010	**0.027*****	**0.017*****	**0.041*****	0.003	0.006	0.019	0.004
	(0.006)	(0.009)	(0.007)	(0.008)	(0.017)	(0.008)	(0.008)	(0.015)	(0.007)
1	0.002	0.005	-0.003	**0.019*****	0.022	0.011	0.000	0.014	**-0.014***
	(0.004)	(0.008)	(0.004)	(0.009)	(0.020)	(0.008)	(0.008)	(0.014)	(0.007)
2	0.009	**0.031*****	**-0.018******	**0.042******	**0.061***	0.016	0.010	**0.048*****	**-0.026*****
	(0.008)	(0.016)	(0.007)	(0.016)	(0.032)	(0.015)	(0.013)	(0.021)	(0.012)
3	0.015	**0.044*****	**-0.022*****	**0.082******	**0.125******	**0.040***	0.028	**0.089******	**-0.030***
	(0.012)	(0.022)	(0.010)	(0.024)	(0.044)	(0.022)	(0.018)	(0.030)	(0.017)
4	0.020	**0.057*****	**-0.026*****	**0.131******	**0.205******	**0.056*****	**0.046*****	**0.121******	-0.022
	(0.014)	(0.026)	(0.012)	(0.032)	(0.058)	(0.026)	(0.022)	(0.037)	(0.021)
5	0.018	**0.053***	**-0.027***	**0.156******	**0.235******	**0.079*****	**0.059*****	**0.148******	-0.021
	(0.017)	(0.031)	(0.014)	(0.039)	(0.069)	(0.031)	(0.027)	(0.047)	(0.025)
6	0.018	0.046	-0.021	**0.186******	**0.280******	**0.098******	**0.075*****	**0.184******	-0.019
	(0.020)	(0.035)	(0.016)	(0.047)	(0.083)	(0.037)	(0.033)	(0.056)	(0.029)
7	0.017	0.042	-0.021	**0.221******	**0.320******	**0.128******	**0.078*****	**0.191******	-0.018
	(0.024)	(0.040)	(0.019)	(0.055)	(0.095)	(0.045)	(0.037)	(0.061)	(0.033)
8	0.027	0.056	-0.018	**0.251******	**0.363******	**0.152******	**0.077***	**0.198******	-0.024
	(0.026)	(0.044)	(0.022)	(0.062)	(0.104)	(0.052)	(0.040)	(0.066)	(0.036)
9	0.032	0.063	-0.018	**0.277******	**0.387******	**0.181******	**0.077***	**0.205******	-0.028
	(0.029)	(0.048)	(0.023)	(0.066)	(0.112)	(0.057)	(0.044)	(0.070)	(0.040)
10	0.040	0.076	-0.015	**0.292******	**0.398******	**0.198******	0.075	**0.208******	-0.031
	(0.031)	(0.051)	(0.025)	(0.072)	(0.122)	(0.061)	(0.046)	(0.075)	(0.043)
11	0.047	**0.091***	-0.018	**0.312******	**0.414******	**0.223******	0.074	**0.206*****	-0.030
	(0.033)	(0.054)	(0.027)	(0.076)	(0.128)	(0.066)	(0.049)	(0.080)	(0.046)
12	0.052	**0.100***	-0.021	**0.314******	**0.392******	**0.243******	0.059	**0.169*****	-0.025
	(0.035)	(0.058)	(0.029)	(0.079)	(0.135)	(0.070)	(0.052)	(0.085)	(0.048)
13	0.053	0.099	-0.020	**0.325******	**0.371******	**0.278******	0.054	**0.168***	-0.035
	(0.037)	(0.061)	(0.031)	(0.084)	(0.143)	(0.076)	(0.054)	(0.089)	(0.051)
14	0.056	**0.107***	-0.022	**0.334******	**0.372*****	**0.290******	0.042	0.148	-0.042
	(0.039)	(0.063)	(0.033)	(0.088)	(0.150)	(0.081)	(0.056)	(0.092)	(0.053)
15	0.061	0.109	-0.016	**0.353******	**0.394*****	**0.306******	0.034	0.137	-0.049
	(0.042)	(0.068)	(0.035)	(0.094)	(0.159)	(0.087)	(0.059)	(0.097)	(0.056)

Table 1: Regression coefficients of internet search variables in Fama-MacBeth regressions where cumulative future stock returns are the explained variable. Table columns correspond to different regressor-regime combinations and table rows correspond to different horizons of future returns. Standard errors for regression coefficients are given in parentheses. Statistical significance at levels of 10%, 5% and 1% is denoted by *,**, and ***, respectively. Additionally, statistically significant results are printed in bold.

and stock returns. We also observe that there is a substantial difference between the importance of distinct internet search variables. Based on the analysis performed so far it seems that the possible influence of search data on stock returns is statistically the strongest for Google company tickers, followed by Google company names and finally Wikipedia page visits. It is also evident that the influence of search data on future returns is mostly short term with the largest absolute values of annualized returns (based on regression coefficients) attained for cases where $k \leqslant 10$.

We observe another phenomenon which is best seen in the case of Wikipedia page visits. In the bear market the values of returns first go up and then go back down so that they form a shape of a frown. During the bull market, on the other hand, we observe a mirror shape, i.e. a shape of a smile. Now, the interpretation of these shapes is in some sense the opposite of the usual meaning conveyed by these shapes. While the frown noticed means good news in bear times, the smile means bad news in bull times. So, we propose the two shapes to be called the “merry frown” and the “sour smile”. These shapes are not so easy to interpret. A possible explanation (taking into account also some other details of the Wiki shape) is that in bear markets investors are pessimistic and their overpessimistic reaction after increased attention perceived via the number of Wikipedia page visits on the first day, results in a counter-reaction in the days to follow and creates the merry frown. In the bull markets though investors are optimistic and their overoptimistic immediate reaction on the increased attention overturns into a sour smile.

In the next step we investigate whether the observed connection between internet search data and stock returns can be explained by including additional factors into our model. We replace our initial explained variable $r_{i,t}$ (future cumulative returns) by the so-called abnormal cumulative returns $ar_{i,t}$. These returns are obtained as residuals in a variant of the Carhart [5] four factor asset pricing model which is an extension of the well known Fama-French model [10]. The model is defined by the following equation:

$$r_{i,t+k} = r_t^{rf} + \beta_{1,i}(r_t^{mkt} - r_t^{rf}) + \beta_{2,i}\,\mathrm{HML}_t + \beta_{3,i}\,\mathrm{SMB}_t + \beta_{4,i}\,\mathrm{UMD}_t + ar_{i,t+k},$$

where r_t^{rf} is the risk-free rate of return (approximated by the daily rate of one month U.S. Treasury bills), $r_t^{mkt} - r_t^{rf}$ is the excess return of the entire stock market over the risk-free return, HML_t is the return difference between a portfolio of stocks with high and low book-to-market stocks, SMB_t is the return difference between a portfolio of small and big stocks in terms of their market capitalization and UMD_t is the return difference between a portfolio of stocks with high and low returns in the past year. The betas are estimated on a daily basis, using a rolling window of 120 days.

We repeat the cumulative return regressions described above for the case of abnormal cumulative returns and report the results in Table 2. The period dependencies of annualized abnormal returns are displayed in Figure 3. We see that results in the case of Wikipedia page visits variable are quite similar to those obtained before accounting for the Carhart factors. However the influence of Google search queries for company tickers and company names is greatly diminished. In fact, no statistically significant results at the 5% level are obtained for the company tickers variable regardless of the bear or bull market. This is in contrast to previous research which was performed on samples taken from earlier periods. Furthermore, we find little evidence that a rise in internet search variables corresponding to individual companies might directly translate into short-term buying pressure and consequently higher stock prices. The differences between the bear market and the bull market remain clearly visible, especially in the case of Wikipedia page visits. While our results

	Wikipedia			Google tickers			Google names		
k	Both	Bear	Bull	Both	Bear	Bull	Both	Bear	Bull
0	**0.014*****	0.006	**0.022*****	**0.011***	0.010	**0.015****	-0.000	-0.006	0.006
	(0.004)	(0.007)	(0.006)	(0.006)	(0.012)	(0.006)	(0.005)	(0.008)	(0.006)
1	0.002	0.006	-0.003	0.002	0.004	0.003	**-0.010****	-0.012	-0.009
	(0.003)	(0.006)	(0.003)	(0.006)	(0.011)	(0.006)	(0.005)	(0.008)	(0.005)
2	0.005	**0.018***	**-0.013****	0.005	0.006	0.003	**-0.016***	-0.009	**-0.022****
	(0.006)	(0.011)	(0.006)	(0.010)	(0.019)	(0.010)	(0.009)	(0.014)	(0.009)
3	0.012	**0.034****	**-0.015****	0.003	0.001	0.005	**-0.024****	-0.018	**-0.028****
	(0.008)	(0.015)	(0.008)	(0.015)	(0.027)	(0.013)	(0.012)	(0.020)	(0.013)
4	**0.018***	**0.048****	**-0.016***	0.017	0.030	0.006	-0.022	-0.009	**-0.031***
	(0.011)	(0.019)	(0.009)	(0.019)	(0.035)	(0.016)	(0.015)	(0.026)	(0.016)
5	0.019	**0.048****	-0.017	0.019	0.034	0.005	-0.022	0.000	**-0.038****
	(0.013)	(0.021)	(0.011)	(0.023)	(0.042)	(0.019)	(0.019)	(0.031)	(0.019)
6	0.020	**0.046***	-0.013	0.024	0.042	0.009	-0.024	0.004	**-0.044****
	(0.014)	(0.024)	(0.013)	(0.026)	(0.050)	(0.023)	(0.021)	(0.036)	(0.022)
7	0.024	**0.054****	-0.015	0.038	0.064	0.018	-0.021	0.011	**-0.043***
	(0.016)	(0.027)	(0.014)	(0.030)	(0.056)	(0.027)	(0.024)	(0.040)	(0.025)
8	0.027	**0.059****	-0.015	0.046	0.072	0.027	-0.025	0.012	**-0.052***
	(0.018)	(0.029)	(0.016)	(0.033)	(0.062)	(0.032)	(0.027)	(0.045)	(0.028)
9	0.029	**0.062****	-0.014	0.059	0.084	0.044	-0.032	0.007	**-0.062****
	(0.019)	(0.031)	(0.017)	(0.038)	(0.070)	(0.036)	(0.029)	(0.048)	(0.030)
10	0.031	**0.068****	-0.016	0.057	0.076	0.050	-0.043	-0.006	**-0.070****
	(0.020)	(0.033)	(0.019)	(0.041)	(0.077)	(0.039)	(0.032)	(0.052)	(0.032)
11	0.034	**0.074****	-0.016	0.066	0.089	0.058	-0.051	-0.017	**-0.075****
	(0.022)	(0.036)	(0.021)	(0.045)	(0.084)	(0.043)	(0.034)	(0.056)	(0.034)
12	0.036	**0.077****	-0.017	0.066	0.084	0.061	-0.058	-0.029	**-0.079****
	(0.023)	(0.039)	(0.022)	(0.048)	(0.087)	(0.046)	(0.037)	(0.061)	(0.036)
13	0.039	**0.081***	-0.013	0.070	0.081	0.072	**-0.067***	-0.041	**-0.085****
	(0.025)	(0.042)	(0.023)	(0.051)	(0.094)	(0.048)	(0.039)	(0.064)	(0.039)
14	0.040	**0.086***	-0.018	0.070	0.076	0.078	**-0.077***	-0.060	**-0.087****
	(0.027)	(0.044)	(0.024)	(0.054)	(0.101)	(0.050)	(0.041)	(0.067)	(0.040)
15	0.039	**0.084***	-0.018	0.072	0.074	0.085	**-0.085***	-0.075	**-0.088****
	(0.029)	(0.047)	(0.026)	(0.057)	(0.107)	(0.053)	(0.044)	(0.070)	(0.043)

Table 2: Regression coefficients of internet search variables in Fama-MacBeth regressions where cumulative abnormal future stock returns are the explained variable. Table columns correspond to different regressor-regime combinations and table rows correspond to different horizons of future returns. Standard errors for regression coefficients are given in parentheses. Statistical significance at levels of 10%, 5% and 1% is denoted by *,**, and ***, respectively. Additionally, statistically significant results are printed in bold.

show that a statistically significant dependence between future stock returns and internet search variables exists, we do note that the explanatory power of all the tested regressions as measured by the R^2 statistic is very low and only rises above 1% in a few cases. This is not unexpected if we take into account the fact that the regressions are predictive, that we are only using a single explanatory variable and that future stock returns are notoriously hard to predict. The questions whether our observations can nevertheless be used to obtain economic gains will be explored in Section 5.

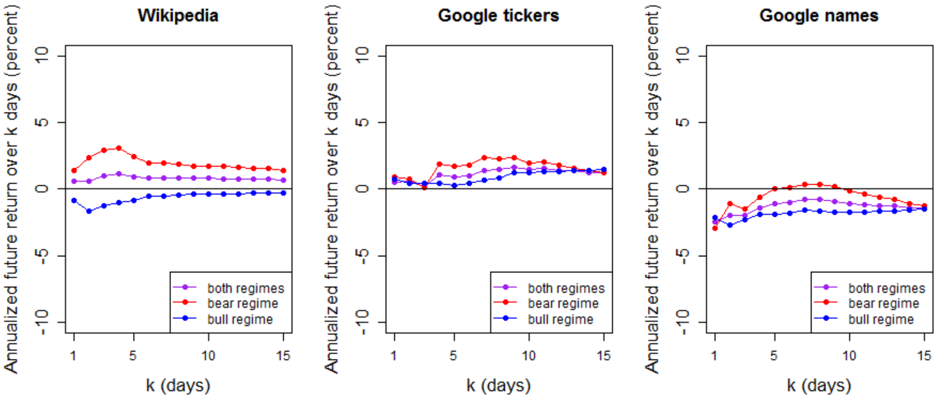

Figure 3: Changes in abnormal annualized future returns over k days after we observe a one standard deviation increase in individual search variables.

Based on these tests we conclude that the influence of the kind of attention noticed by Google search queries (either for company tickers or names) can be perceived also by other data that are more commonly applied by financial practitioners. On the other hand, it shows that Wikipedia page visits do indeed provide new information about the behavior of stock prices. Also, we perceive that the merry frown and sour smile effects persist for Wikipedia page visits even after controlling for the most commonly used asset pricing factors.

5 Trading strategies

In this section we want to verify how the results of Section 4 can be used, if at all, in forming trading strategies. In other words, we want to either statistically prove or disprove that internet data can increase our profits in financial markets. The evidence for the influence of internet search data on future stock returns is most compelling in the case of Wikipedia data, as shown in Section 4. Since we were also not able to find any examples in existing literature of this type of data being used in construction of trading strategies based on individual stocks, we decided to limit our analysis in this chapter only to Wikipedia page visits.

Our results show that in bear markets higher Wikipedia page visits are positively correlated with short term future return while in bull markets the corresponding correlation is negative. So we propose the strategy for bear markets to enter long position at the end of the trading day for stocks in the upper decile with respect to the most recent available data on Wikipedia page visits; and similarly to enter the short position at the end of the trading day for stocks in the lower decile with respect to these visits. In the bull market, the strategy is to do exactly the opposite. We propose that all the long positions and all the short positions are entered using the same weights with respect to the wealth that we are prepared to invest into this strategy. Since the data on Wikipedia page visits for any given day is only made available the following day we lag our Wikipedia variable for one day to ensure that the data would have been available at the time of our trading decision.

Of course there is a problem of determining the actual frequency of trading, this means for how long we should hold our positions. We know that we are talking about a short term effect, but what does short really mean in this particular context? To make this dilemma as clear as possible we are making a number of tests using some alternatives. Let f be the number of trading days between two consecutive trading decisions. For $f = 1, 2, \ldots, 8$ we are testing the fth strategy and give the result for three options. The first option is that we allow only trading in the bear market, the second one is that we allow trading only in the bull market and the third one is that we allow trading during both markets. In Table 3 we present the results obtained in percentage points of the annual return. It is clear that $f = 1$ is the best of the proposed strategies in all the three cases. It is also clear that the results are getting smaller with f increasing in the case of combined strategy and the bull-only strategy. However in the case of the bear strategy $f = 2$ and $f = 3$ are slightly better than $f = 1$. For f big enough the results of the strategies seem to become more or less random. The best of the three options tested is the combined application of both bear and bull strategy. It is also clear that the results of the bear strategy are better than the results of the bull strategy.

Trading frequency (days)	1	2	3	4	5	6	7	8
Bear	5.28	5.50	6.10	4.02	2.30	0.94	-0.55	2.21
Bull	2.76	1.33	-0.15	0.91	0.59	1.02	1.08	-2.53
Joint Bear & Bull	8.19	6.91	5.94	4.97	2.91	1.97	0.52	-0.37

Table 3: Average annual returns (in percentage points) of proposed trading strategies in relation to the trading frequency.

We also want to compare our strategies to suitable benchmark strategies. However, as we believe, the most usual benchmarks such as various indices are long-only strategies and the comparative testing with our strategies which include both long and short positions would not be fair. So we decided to compare it with random strategies using a Monte Carlo approach. Our control strategy is to choose in a uniformly random way 10 % stocks to be put in a long position and 10 % to be put in a short position. We created 1000 strategies of this type and computed the average yearly return for each of them. This produces a random sample of possible average yearly returns which we compare statistically to the average return of each of the strategies under consideration. As usual in this kind of situation, we perform a one-sample one-way Student t-test where we test the null hypothesis that the mean yearly return for the population of random strategies is equal to the return of our strategies against the alternative hypothesis which states that the mean yearly return for the population of random strategies is lower to the return of our strategies. As can be seen from results given in Table 4 we can reject this hypothesis for our joint bear and bull strategies for most of the trading frequencies considered.

In Figure 4 we want to present a slightly different view on the results of our strategies

Trading frequency (days)	1	2	3	4	5	6	7	8
μ_0	8.19	6.91	5.94	4.97	2.91	1.97	0.52	-0.37
μ	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25	-0.25
t value	-64.55	-54.78	-47.34	-39.91	-24.18	-16.96	-5.89	0.92
p value	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.82

Table 4: The results of a one-way Student t-test for testing the null hypothesis that the mean yearly return (μ) for the population of random strategies is equal to the return (μ_0) for our joint bear and bull strategy against the alternative hypothesis $\mu < \mu_0$. The p values are given in percentage points and rounded to two decimals.

compared to the random approach. Assume we invest a certain equity in the strategies above to be compared; and that we invest the same amount into each of the random strategies described in the previous paragraph. We compare the average of the randomly invested equity to the equity gained via the strategy under consideration for each day of our sample period. More interesting than the averages as such are the bands created around the averages using the daily standard deviation and its small multiples. We can see that the equity invested in our joint bear and bull strategy mostly stays in the area that is beyond the band which is three standard deviations above the average equity of random strategies.

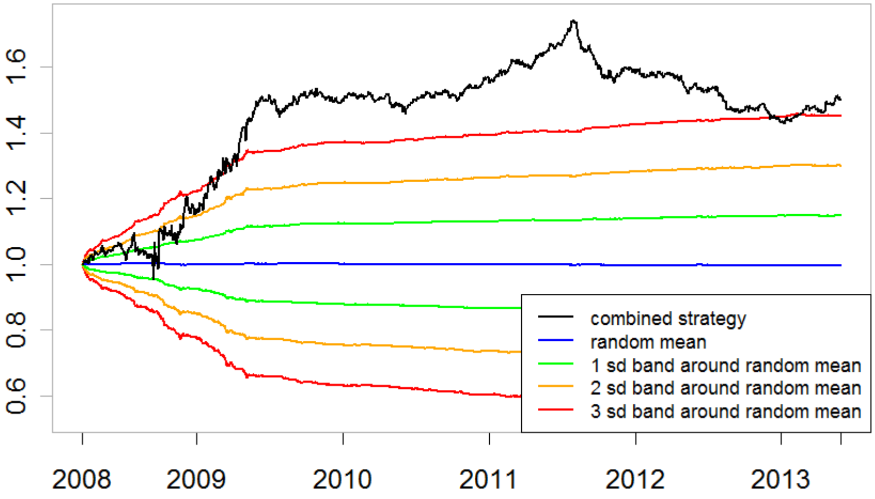

Figure 4: Equity curve of joint bear and bull strategy with trading frequency of one day compared to equities of random strategies represented by standard deviation bands around the mean equity.

In most cases the Wikipedia pages on individual S&P 500 companies contains only the most basic information. It is therefore safe to assume that this information source will mostly be utilized by individual investors since institutional investors have access to more sophisticated tools offering greater depth of information. Our hypothesis is that the influence of Wikipedia page visits on future stock returns will be higher for stocks that are likely to attract a higher proportion of individual investors. According to Barber and Odean [4] the individual investors generally have a tendency to tilt their stock investments towards high-beta, small and value stocks. In light of this result we construct three additional strategies based on our joint bear and bull strategy. In all these strategies we restrict our trading decision to a subsample of stocks that fall above or below the median of one of the following variables: volatility, market capitalization and price-to-book ratio. In the first strategy we choose a subsample of high volatility stocks, in the second one we choose a subsample of low market capitalization stocks and in the last one we chose a subsample of low price-to-book ratio. Volatility is calculated in a 20 trading day rolling window approach. The market capitalization and price-to-book ratio variables are obtained from the ADVFN service[5]. We present the results in Table 5. The results strongly support our hypothesis since all three subsample strategies outperform the full sample joint bear and bull strategy in cases of the most relevant trading frequencies ($1 \le f \le 5$).

Trading frequency (days)	1	2	3	4	5	6	7	8
Full sample	8.19	6.91	5.94	4.97	2.91	1.97	0.52	-0.37
High volatility subsample	15.35	16.40	10.13	8.53	4.12	7.15	-0.78	1.32
Low market cap subsample	17.43	13.41	9.07	6.86	4.53	5.49	0.31	1.12
Low price to book subsample	13.57	9.99	7.16	8.15	5.87	0.81	-0.64	3.30

Table 5: Average yearly returns (in percentage points) of subsample strategies compared to average yearly return of full sample joint bear and bull strategy.

All of the strategies presented up to this point have included both long and short positions. Since many investors face restrictions with respect to opening short positions in stocks, the question naturally arises whether our strategies can be adapted to be long-only. Let us consider the simplest possible adaptation which is the strategy where the investment rule during the bear market is to enter long positions at the end of each trading day for stocks in the upper decile with respect to the most recent available data on Wikipedia page visits. During the bull market, the strategy enters long positions for stocks which are in the lower decile with respect to the Wikipedia page visits. The annualized return of such a strategy in our sample period is 20.36%. Since the strategy is long-only, it is reasonable to compare its performance to that of the S&P 500 index whose annualized return during our sample period is merely 4.36%. The equity curves obtained by investing the same amount of wealth in both our adapted long-only strategy and the S&P 500 index are shown in Figure 5. The backtesting results favor the conclusion that even those investors who are restricted to only opening long positions might benefit from including the information about Wikipedia page visits in their investment decisions.

6 Conclusion

The key point of our paper is that it is essential to incorporate information about the market regime when studying the influence of internet search data on stock returns. This is clearly

[5] http://www.advfn.com

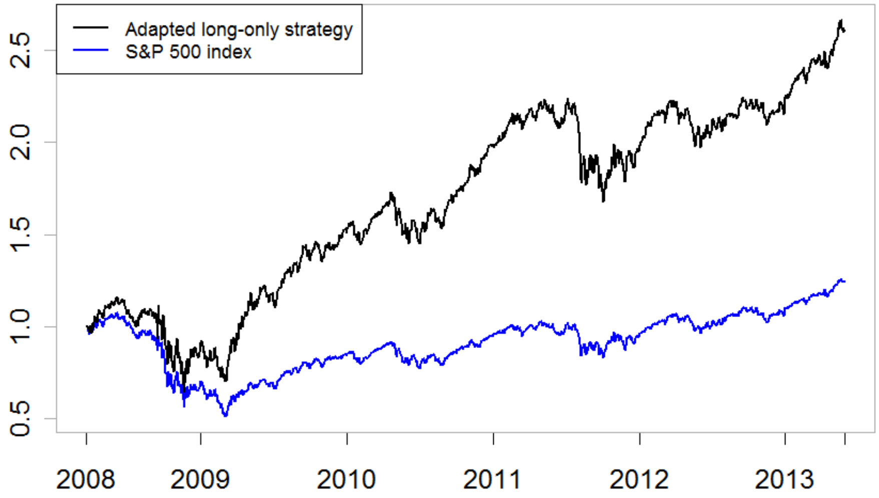

Figure 5: Equity curve of adapted long-only joint bear and bull strategy with trading frequency of one day compared to equity curve of S&P 500 index.

true for all the search variables considered since all show markedly higher correlations with future stock returns in the bear regime than in the bull regime. However, the distinction between the two regimes is especially singnificant in the case of the Wikipedia variable where we observe two inverse price patterns - a merry frown in the bear regime and a sour smile in the bull regime. Our regime estimation method is based on a hidden Markov model that only accounts for information revealed to us through the price fluctuations of the S&P 500 index. We suspect that even more interesting results might be obtained if search data were somehow included into the regime switching model itself, perhaps by building upon existing research into estimation of investor sentiment by internet search data such as [8].

After controlling for the Carhart factors the Wikipedia page visits variable emerges as the one with the most significant influence on future stock returns. Until recently this data set has been largely overlooked by researchers however we believe that it holds great potential for future applications. In a surprising turn, both of the Google search variables prove to be statistically insignificant for most periods of future return for stocks in our sample. This result is at odds with previous studies performed on earlier sample periods and warrants further research that would explain this discrepancy. We suggest that this might be caused by arbitrageurs already taking advantage of the effect of company-related Google search frequencies in line with the weak-form market-efficiency hypothesis.

We would like to make an additional point about Google Trends data with regard to future research. We noticed that previous studies have almost exclusively focused on relative search frequencies which is most likely due to the fact that individual time series obtained from the Google Trends service are normalized within series so that their values always span the interval from 0 to 100. In Section 2 we describe a straightforward approach that enables us to obtain full sample daily trends data regardless of normalization. A quite similar approach might be used to obtain data where non-relative search frequencies of two different terms can be compared. It would be interesting to know whether such data provides us with an even better proxy for investors' attention.

We also believe that the results presented in our paper may be of benefit to financial practitioners in at least two ways. Firstly, we show that Wikipedia can provide investors with insights into a stock's risk profile that are overlooked by existing asset pricing models such as the Carhart four-factor model. Secondly, the trading strategies presented in Section 5 may be of interest to speculative investors who are comfortable executing trading strategies with target investment holding periods of less than a week.

References

[1] A. Ang and A. Timmermann, Regime Changes and Financial Markets, *Annual Review of Financial Economics*, **4** (2012), 313-337.

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

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

[4] B. M. Barber and T. Odean, Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors, *Journal of Finance*, **55** (2000), 773-806.

[5] M. M. Carhart, On the persistence in mutual fund performance, *Journal of Finance*, **52** (1997), 57-82.

[6] H. Choi and H. Varian, Predicting the Present with Google Trends, *Economic Record*, **88** (2012), 2-9.

[7] Z. Da, J. Engelberg and P. Gao, In Search of Attention, *Journal of Finance*, **66** (2010), 1461-1499.

[8] Z. Da, J. Engelberg and P. Gao, The Sum of all FEARS: Investor Sentiment and Asset Prices, *Working Paper*, (2010).

[9] E. F. Fama and J. MacBeth, Risk, return and equilibrium: Empirical tests, *Journal of Political Economy*, **81** (1973), 607-636,

[10] E. F. Fama and K. R. French, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, **33**, 3-56.

[11] M. S. Drake, D. T. Roulstone and J. R. Thornock, Investor Information Demand: Evidence from Google Searches Around Earnings Announcements, *Journal of Accounting Research*, **50** (2012), 1001-1040.

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

[13] J. D. Hamilton, A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, *Econometrica*, **57** (1989), 357-384.

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

[15] S. Mitra, P. Date, Regime switching volatility calibration by the Baum-Welch method, *Journal of Computational and Applied Mathematics*, **234** (2010), 3243-3260.

[16] J. M. Maheu, T. H. McCurdy, Y. Song, Extracting Bull and Bear Markets from Stock Returns, *Working Paper*, (2009).

[17] H. S. Moat, C. Curme, A. Avakian, D. Y. Kenett, H. E. Stanley, and T. Preis, Quantifying Wikipedia usage patterns before stock market moves, *Scientific Reports*, **3** (2013)

[18] T. Preis, H. S. Moat and H. E. Stanley, Quantifying Trading Behavior in Financial Markets Using Google Trends, *Scientific Reports*, **3** (2013)

[19] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, *Proceedings of the IEEE*, **77** (1989), 257-285.

[20] N. Vlastakis and R. N. Markellos, Information demand and stock market volatility, *Journal of Banking & Finance*, **6** (2012), 1808-1821.