UNIVERSITY OF GOTHENBURG

SCHOOL OF BUSINESS, ECONOMICS AND LAW

Market efficiency and index fund flow

An empirical study of the relationship between passive investment and broad-market efficiency

Authors: Erik Larsson
Jacob Wergeland

Supervisor: Taylan Mavruk

University of Gothenburg
Department of Economics
Centre for Finance
Graduate School

Master Thesis in Finance

Gothenburg
June 2020

Abstract

An observable rise in the popularity of index funds have caused the index funds to, in 2017, capture 20% of total fund assets globally. A cornerstone of such passive investment is a belief in an efficiently priced security market. This paper aims to relate index fund flows with market efficiency during the period 2000-2019. Using S&P500 returns we estimate a market efficiency measurement called the Hurst exponent, using two accredited methods: the rescaled range analysis (RS) and the detrended fluctuation analysis (DFA). We find similar estimations as previous studies, wherein the S&P500 index have exhibited a slight mean-reverting return process, close to theoretical market efficiency. We further relate this time-varying market efficiency measurement of S&P500 to its index fund flows. Using a correlation filtering method to find index funds in the US targeting the S&P500 index, and aggregating these mutual funds individual flow, we obtain aggregate index fund flow. Conducting a Granger causality test on both fractional flow and dollar flow, we find a causality that market efficiency Granger cause index fund flow. We further estimate that a lesser degree of market efficiency have a negative impact on flow: the more long-term memory the index experience, the smaller level of flow. These results hold stronger for dollar flow rather than fractional flow.

Suggested keywords: *market efficiency*, *Hurst exponent*, *mutual fund flow*, *passive investments*

Acknowledgement

The classic pirate saying goes: *It is not the treasure chest at the X-marked spot that is most valuable; the real treasure is the friends acquired on the journey.* When we now find ourselves at the crossroads between academic school years and a professional career, we say this hold with absolute truth. Whether this friendship is characterized by knowledge, companionship, or everlasting-friends, does not matter; every student should cherish every moment.

In such dire times as the present with COVID-19 ravaging the world, where a smile might be rare, we would like to start this paper with a small and playful joke.

Question:
What is the difference between a government bond and a man?
Answer:
The government bond matures!

Before we let the reader dive into the text, we would like to direct our incomparable thanks towards our supervisor Taylan Mavruk for his valuable and helpful guidance. After all, the mason stands powerless without the forge. With that, we wish you a pleasant reading experience and hope this paper provides both interesting and educational aspects.

Best regards,
Erik & Jacob

Contents

Abstract **i**

Acknowledgement **ii**

1 Introduction **1**

2 Literature review **3**
2.1 Market efficiency 3
2.2 Fund flows and passive investments 4
2.2.1 Fees 5

3 Data **6**
3.1 Data management 6
3.1.1 Total net assets 7
3.1.2 Flow 8
3.2 Handling of fees 11

4 Methodology **12**
4.1 General method setup 13
4.2 Long-term dependency 14
4.3 The Hurst exponent 15
4.3.1 GARCH 15
4.3.2 Rescaled range analysis 16
4.3.3 Detrended fluctuation analysis 18
4.4 Endogeneity concerns 19
4.5 Estimation and control variables 23

5 Results **25**

6 Conclusion **33**

7 References **34**

A Appendix **37**
A.1 Data adjustments 37
A.2 Hurst histogram 37
A.3 Regressions - naked data 37

List of Figures

1 Histogram of returns 8
2 Flow time series 11
3 Evolution of Fees over time 12
4 Hurst exponent results 25
5 Hurst exponent and CI 27
6 Hurst histogram 37

List of Tables

1 Descriptive statistics for returns 7
2 Flow descriptive statistics 10
3 AR(1)-GARCH(1,1) 16
4 VAR specification - fractional flow 20
5 VAR specification - dollar flow 21
6 Granger Causality test 22
7 Annual statistics 24
8 Regressions - fractional flow 28
9 Regressions - dollar flow 29
10 Regressions (lagged Hurst) - dollar flow 30
11 Standardized Hurst coefficients 32
12 Data adjustments 37
13 Regressions, naked data- fractional flow 38
14 Regressions, naked data - dollar flow 39

1 Introduction

Over the last decade, in the wake of the great financial crisis, the financial markets have experienced an unprecedented rise in index funds as a popular investment vehicle; a rise supported by their cost-efficiency (Malkiel, 2003; Sirri & Tufano, 1998; Weissensteiner, 2019). Passive investments, typically referring to a broad market index investment scenario, [1] did in june 2017 grow up to 20% of total fund assets globally, compared to 8% in 2007 (Sushko & Turner, 2018). Although there may be considerable amounts of motivations for utilizing passive investing, one cornerstone, which should not be diminished, is that the market must be considered efficient for passive investing to perform (Wermers, 2000).[2] If the market would not be efficient, prowess investors would not choose passive investing (with its corresponding investment into something incorrectly priced) but instead change strategy and actively invest, improving the performance of their investing (Garleanu & Heje Pederson, 2019). As such, in the case of inefficient markets, some investors would deviate from this passive investing and would therefore change their allocation from the market distribution; assuming informed decisions and sufficient and adequate market-moving power, the efficiency of the market would thus improve.

Fama (1970, p.383) famously described market efficiency as a market where "security prices always 'fully reflect' all available information". Correctly priced securities thus bring a no-arbitrage trading or investing environment, wherein no parties suffer informational disadvantages. One important aspect of this no-arbitrage system of an efficient market is that no investor can be expected to persistently and systematically realize abnormal returns; and, as such, in an efficient market, only random (or rather unpredictable) fluctuations of the returns are possible (Kristoufek & Vosvrda, 2013). As such, it seems that the key motivational tool for a passive investor is a belief in an efficient market; where upon such a realization of an efficient market, inducing an inflow of passive money. Fund inflow and outflows have been a thoroughly researched topic within the finance lit-

[1] Explicitly, passive investments are investments with a purpose of replicating or obtaining the return of a market benchmark or index. Such an investment could be an ETF (exchange-traded fund) or mutual fund targeting an index. Most mutual funds replicate target index return by holding the index composite assets weighted by their index proportion (traditional passive investing), while other funds use derivative instruments (e.g. a synthetic ETF) to artificially replicate the return (synthetic passive investing). Important to note here is the implication for market efficiency and how the all-encompassing passive investing relates to it. As the synthetic passive investment vehicles or instruments have no market moving-impact on the underlying index, it should not be equalized with traditional passive investments, which posses such powers, with regard to market efficiency. It is self explanatory that a derivative transaction should have no price effect on the underlying asset, while a transaction of the underlying asset should affect the price of the underlying asset. Finally is the distinction between passive investment and passive management (an investment with no intent of actual monitoring), where the latter is often categorized as *dumb money*. Although passive investment infers a non-monitoring scenario as well, it is not identical to passive management and should not be confused as such.

[2] Such motivations could be time-constraints and similar hobby-related causes. For such investors, passive investing might prove easiest and most well-performing choice, regardless of the efficiency of the market.

erature. Understanding catalysts for fund flows have been essential for both practitioners and researchers. Previous studies has mainly focused on the impact of fees (Huang, Wei, & Yan, 2008; Sirri & Tufano, 1998), fund returns (Edelen & Warner, 2001) and volatility (Cao, Chang, & Wang, 2008).

So how do we start by examining this causality between market efficiency and index fund flows? And maybe more importantly, what is the direction of the causality? Several event studies have examined the phenomenon of abnormal returns realized by index composition changes (see, for example, Belasco, Finke, & Nanigian (2012) or Petajisto (2011)), clearly outlining an impact on prices, compared to non-constituents, due to index fund flows. Such impact suggests predictability of abnormal returns of single constituents, but does it indicate direction of the index or broad market efficiency? Intuitively, the overall index return cannot be predicted from abnormal returns realized by index constituents addition or deletion. As such, we have to utilize another type of measurement to determine broad market efficiency and how it relates to index fund flows. There exists various such ways to measure and test broad market efficiency: Approximate Entropy [ApEn] (Pincus, 1991; Pincus, 2008) to measure the irregularity and unpredictability of fluctuations in time-series data; Variance Ratio (Lo & MacKinlay, 1988); the Efficiency Index (Kristoufek & Vosvrda, 2013); or the Hurst exponent (Bariviera, 2011; Eom et al., 2008) for measuring long-term memory. Notably, as Belasco, Finke, & Nanigian (2012) suggest, the abnormal single constituents returns are in fact the liquidity premium corresponding to index inclusion. Maybe this hints that it is in fact liquidity that is related to market efficiency (it may be reasonable to assume that liquidity should be higher for an efficient market, i.e. unpredictable returns), but Bariviera (2011) find only a partial relationship between liquidity and market efficiency.

The majority of these previous studies have focused on testing, measuring, and ranking the efficiency of large indices of various securities; but, in light of the above discussion about the rise of passive investing, we will try to expand the existing body of literature by investigating the relationship between market efficiency and index fund flows, utilizing the Hurst exponent as a measurement for market efficiency. The expansion is two-fold: extending the econophysics literature by relating market efficiency to important market characteristics, building upon the liquidity-linkage by Bariviera (2011) and the predictability-linkage by Eom et al. (2008); and, further developing the fund flow literature by widening the studies from Huang, Wei, & Yan (2007), and Cao, Chang, & Wang (2008), and partly Sirri & Tufano (1998).

The vast majority of research within both market efficiency and flow have been conducted for the US-market, due to being among the most active and liquid markets. Further reasons to use the US-markets is the possibility to compare our findings to similar previous studies.

We aim to examine whether the S&P500 have exhibited market efficiency over last

two decades. We expect to find similar degree of efficiency for S&P500 as other developed markets.[3] We further aim to explain if, and subsequently how, index fund flow relates to market efficiency: the direction of the causality and the following impact. We hypothesize the causal direction that the degree of market efficiency affects the level of index fund flows, with the magnitude that a less efficient market would induce less flow.

We identify 633 index funds targeting S&P500 between 2000 and 2019. Using monthly index fund return and total net asset, we compute monthly aggregate flow values, both fractional flow and dollar flow. We estimate the Hurst exponent for S&P500, acting as a market efficiency measurement proxy. Overall, the S&P500 experienced a slight mean-reverting return process, close to market efficiency during our sample period. Identifying a causal relationship wherein a lesser degree of market efficiency negatively affects aggregate index fund flow. This relationship is characterized by a magnitude of one standard deviation change in the market efficiency measurement indicating a change in dollar flow by approximately 15% of the standard deviation in dollar flow.

2 Literature review

2.1 Market efficiency

In 1970 (Fama) famously postulated the efficient market hypothesis (EMH) which has been the dominant view of market functionality. Ever since creation, the EMH has been challenged. Grossman & Stiglitz (1980) examined the aspects of information accessibility and how this affects the EMH. Costless information is a requirement for efficient markets, and, for a functioning marketplace, this is per se impossible (Grossman & Stiglitz, 1980). With respect to the rise of passive investing in the recent decade, Fama (1991) posited that efficient markets gave rise to passive investing, since markets were theoretically already fully efficient. More recently, Weissensteiner (2019) instead theorized a reverse relationship wherein market-wide forecast errors were reduced by passive investing and efficiency improved. Kristoufek & Vosvrda (2013) found that, by estimating their Efficiency Index, the most efficient markets were Japan, Denmark, and Germany and the least efficient markets were Peru, Sri Lanka, and Slovakia. Suggesting that maybe the more globally integrated markets exhibit a larger degree of efficiency; nevertheless, the more local and less developed markets often exhibit more long term memory (i.e. a Hurst exponent greater than 0.5) while the US, UK, and other similar global markets were characterized by a reverse condition wherein they experienced anti-persistence (i.e. a Hurst slightly lower

[3]The expectation of a similar degree of efficiency for S&P500 as other developed market is based on previous studies findings, wherein such a similarity is found. Furthermore, Eom et al. (2008) conclude that the degree of efficiency is highly related to predictability (average hit-rate); thus, global markets should intuitively not exhibit easily available predictability.

than 0.5)(Kristoufek & Vosvrda, 2013)[4]. Bariviera (2011) found a similar result, wherein the Thai stock market nearly consequently experienced positive memory from 1975 until 2005. Cajueiro & Tabak (2004) elaborated further on emerging markets and computed Hurst coefficients for a wide range of market and compared them to the US market. They found over time mainly positive long term memory for all the examined markets, hence the US was closest to a Hurst of 0.5, concluding it to be the most efficient, whereas the emergent markets were less efficient.

Likewise, Eom et al. (2008) found a difference in market efficiency between emerging and more established markets. They utilized the Hurst coefficient and find that many of the less develop market places to be above market efficiency level of 0.5, hence a positive correlation. This implies that emerging markets exhibit more long-term memory than their developed counterparts. Their result is in general in line with previous studies, implying less efficiency in emerging markets than more developed markets; suggesting that a higher Hurst exponent corresponds to a lower degree of efficicency, rather than a Hurst exponent disconnected from 0.5. However, the Hurst levels estimated by Eom et al. (2008) for emerging markets were, in general, lower than other studies (see for example Cajueiro & Tabak (2004)). An explanation for this discrepancy could be due to differing underlying computational statistics between the papers,[5] which will be covered in this thesis method section.

2.2 Fund flows and passive investments

Numerous studies have examined factors affecting mutual funds flow. Generally, the mutual fund flows are affected by previous returns. While all mutual funds flow are sensitive to recent returns (Sirri & Tufano, 1998; Sapp & Tiwari, 2004), the sensitivity increases with lower participation costs (Huang, Wei, & Yan, 2007). Similarly, Warther (1995) found that market-wide aggregate inflows are strongly affected by simultaneous aggregate price movements. Cao, Chang, & Wang (2008) find that high frequency market volatility is negatively affected by aggregate flow, entailing evidence that a positive (negative) shock in flow decreases (increases) market volatility. Sirri & Tufano (1998) found that the flow into high-performing funds were disproportionately larger than the outflow from poor-performing funds. The large body of mutual fund flow studies mainly deal with panel data and intra-competition between the funds, and are often comprised entirely of

[4]The long term memory can be resembled by a momentum-factor (increases are likely to be followed increases), and anti-persistence by a more negative correlation than randomness prescribes (increases are more likely to be followed by decreases). The Hurst exponent is a method stemming from engineering, to find repeating patterns within a data. A Hurst of 0.5 indicates no memory whereas a higher or lower signals memory, either positive or negative memory.

[5]Eom et al (2008) uses detrended fluctuations analysis (DFA) to compute their Hurst coefficient whereas Cajeiro & Tabak (2004), Bariviera (2011) and Kristoufek & Vosvrda (2012) utilizes rescaled range statistics (RS). DFA is excluding short term memory in data at a greater level compared to the RS. Yielding a difference in results. This will be further explained in the method section

actively managed funds. Index funds flow studies widely concerns miss-pricing due to the flow, rather than trying to explain the flow. An inclusion into the substantial main indices may distort prices due to the large share of passive capital inflow caused by the relatively unconscious passive investing. Abnormal returns, both negative with the deletion from index and positive with the addition to index, affects securities bordering an index; between 1990 and 2005, the average excess (abnormal) return was from an index addition 8.8% and from index deletion -15.1% (Belasco, Finke, & Nanigian, 2012).

Nonetheless, Hortacsu & Syverson (2004) identify the importance of non-portfolio attributes in attracting investors for index funds. They further note that the low participation costs of index funds cause the fund with the lowest fee not to capture the entire homogeneous index fund market. On the same note, passive investing have shown to be the most cost efficient and optimal choice for investors irregardless of market condition due to the inability of fund managers to over time outperform their benchmark index (Malkiel, 2003). This introduces an important aspect of index fund flows. There is a fundamental difference between studying fund flows on a micro-level (fund specific flow) and on a macro-level (aggregate flow)(Cao, Chang, & Wang, 2008). Micro-level studies on mutual funds may examine differences between fund categories, often categorized the funds having different objectives and returns, where the outflow from one fund might be offset by inflow into another fund. Macro-level studies disregard specific fund flow, thus only examine market flows (Warther, 1995). Hortaçsu & Syverson (2004) found that index fund intra-competition were mainly driven by non-portfolio attributes, suggesting that micro-level studies on index funds relationship to some index measurement makes less sense since, generally, the reason of one index funds outflow should not be a another index funds' better return (after all index funds targeting the same index should have similar returns). The low sensitivity of fees for index funds combined with their cost efficiency suggest that index funds do not, in a performance setting, compete intravenous, like actively managed fund do. To conclude, when examining funds consisting of index holdings a macro-level approach is more suitable (Warther, 1995).

2.2.1 Fees

The subject of fees and its impact on flow has been revisited in the literature, Sirri & Tufano (1998) and Huang, Wei, & Yan (2007) are among the more noticed. The many properties of the fee component is not in the main scope for this thesis, although it is essential to keep in mind when discussing funds. The general effect found in previous literature indicate that fees has a negative impact on the fund flows, higher fees results in lower inflow Sirri &Tufano, (1998). This relationship comes with exceptions, the most prominent is the participation cost, including both search and marketing costs (Sirri &Tufano, 1998). Funds pay to gain investors and therefore raise the fees, however still attracting flow.

The fee structure is usually complex and includes several different fees, where the most common are the front and rear load fees, management fees and 12b fees.[6] For index funds the main aspect of differentiating is how to streamline operations in regards to fees, since the assets are bound to be index replicating. Cash management of the inflows and outflows are crucial when evaluating the overall performance (Elton, Grauber & Busse, 2004).

3 Data

From CRSP (WRDS), we obtain monthly data of total net assets (TNA) and fund return for 43 939 funds during the period 2000-2019, as well as both daily and monthly data for S&P500 for the period 1996-2019[7]. The daily data is used for the calculation of Hurst exponent and the monthly S&P500 data is utilized to generate returns for index funds, when specific fund returns are missing. To identify index funds tracking the S&P500 amongst all the acquired data we use a method suggested by WRDS (the provider of CRSP). By checking whether the funds have a return with a correlation of at least 99.5% ($\rho \geq 0.995$) to the S&P500, we isolate all index funds tracking the S&P500.[8] By utilizing this filtering method, we reduce the number of mutual funds down to 633; in other words, we identify 633 index funds which at some period between 2000 and 2019 target S&P500.[9]

We present introductory summary statistics for the identified index funds and S&P500 in Table 1. In figure 1 we show histograms over the S&P500 and identified index funds returns. We observe similar means and medians, while the index funds exhibit a wider distribution with more a larger outliers, resulting in a higher kurtosis and fatter tails. The lowest reported TNA from CRSP is 0.10, i.e. \$100 000, and is often reported the first period the index fund became active: why both the minimum and 1% value are the same.

3.1 Data management

Regarding missing values in both the return and total net assets (TNA) data of these identified index funds, numerous actions has been undertaken. As the remaining funds now target S&P500, we simply replace missing return values with the relevant return for S&P500. Although we noted previously, and as can be observed in Table 1, the index

[6] 12b fees are costs associated with distribution and marketing. In general this fee combined with the management fee is acknowledge as the expense ratio.

[7] The return data r_t and the total net assets data TNA_t indicate the return and total net assets for period t as of month-end.

[8] The index fund flag indicator provided by CRSP is deemed insufficient for isolating index funds targeting specific indices, why we instead use the suggested correlation filter approach.

[9] For some of these identified index funds we had "inadequate" number of monthly returns, down to as few as only 2 monthly returns. Although these sparse monthly returns produced a sufficient correlation against S&P500, for robustness, we did a name check and confirmed index funds targeting S&P500.

Summary statistics

	Panel A: returns			*Panel B: total net assets*
	S&P 500 *(D)*	S&P 500 *(M)*	Index funds *(M)*	Index funds *(M)*
N	240	5031	62111	70158
Max	11.55%	10.77%	34.06%	319624.1
99%	9.42%	3.43%	9.73%	70468.7
75%	2.97%	0.56%	3.22%	667.9
50%	0.055%	0.96%	1.11%	148.8
25%	-1.78%	0.47%	-1.70%	23.3
1%	-10.99%	-3.32%	-10.91%	0.1
Min	-16.94%	-9.03%	-35.33%	0.1
Mean	0.023%	0.42%	0.56%	2847.8
Std	1.129%	4.18%	4.46%	15304.9
Skew	-3.83	-59.9	-60.70	10.4
Kurt	11.8169	4.1068	5.1379	136.3

Table 1: Descriptive statistics for S&P500 and identified index funds from 2000-01-01 until 2019-12-31. Index funds were identified by filtering with a correlation of at least 99.5% towards S&P500. The TNA are in millions of USD. *(D)* and *(M)* represents daily and monthly data, respectively. The difference in the number of observations between the returns and TNA is due to that the missing values do not always align for the two variables.

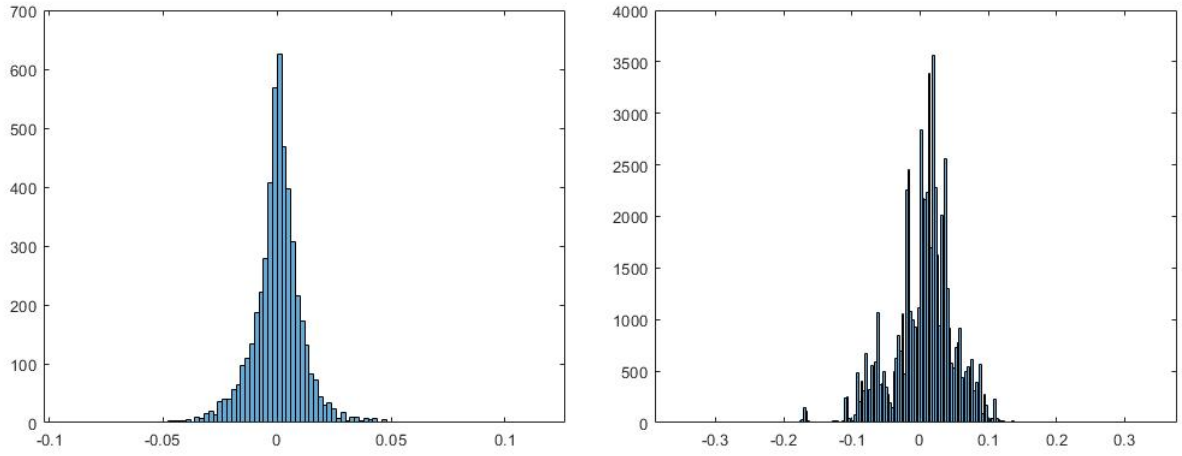
funds and S&P500 does not exhibit identical distributions, the high correlation threshold should be sufficient to not cause structural deficiencies. Below we outline procedures undertaken for the TNA and flow data.[10]

3.1.1 Total net assets

Several funds had reported a value of 0 for total net asset (TNA) for some months, which is unreasonable, as that would pertain to a non-existing fund; hence these values were treated as missing values. The missing TNA values were filled by three techniques: backward filling, forward filling and linear interpolation[11]. We utilize the Linear Interpolation method, and the other two methods for robustness check. Gaps of missing values ranging from 1 month up to 12 months were filled. Gaps of more than 12 months were treated as missing values and were not filled. To ensure this would not cause inconsistencies in the data, the number of gaps that would be filled from 13 months up to 24 months were one gap of 16 months and four gaps of 17 months. Limiting the fill gap method to 12 months does not greatly inhibit the data and suggest a removal of the possibility of an eventual

[10] In appendix A.1 we outline these procedures as well, but in a table format instead.

[11] For clarification, the linear interpolation fill method is simply a equal step-by-step increase to reach the end of the gap from the beginning of the gap. Explicitly, if Y_{start} is the value before the gap and Y_{end} is the value after the gap, and n is the number of missing values between these two data points, then the n_i:th missing value Y_{n_i} is given by $Y_{n_i} = Y_{start} + n_i \times \frac{Y_{end} - Y_{start}}{n+1}$. The backward and forward filling techniques are simply the next value after the gap carry backward and the previous value before the gap carry forward, respectively.



(a) S&P500 **(b)** Index funds

Figure 1: Histogram over the returns of S&P500 and identified index funds from 2000-01-01 until 2019-12-31. The return data depicted are in percentage and can be seen in Table 1.

structural bias in the data set. This procedure increases the number of total TNA data points by 2782, from 67 341 to 70 158.

3.1.2 Flow

In an attempt to not reinvent the wheel, and thus by convention of previous research (Sirri & Tufano, 1998; Huang, Wei, & Yan, 2007), we utilize a standard measurement variable for mutual funds flow. Explicitly, this can be defined as

$$FLOW_{i,t} = \frac{TNA_{i,t} - TNA_{i,t-1} \times (1 + r_{i,t})}{TNA_{i,t-1}} \tag{3.1}$$

where $TNA_{i,t}$ is the total net assets for the fund i for time period t, and $r_{i,t}$ is the return for fund i for the time period t. This standard procedure of estimating fund flows removes the intrinsic returns of the funds constituents, thus cleaning the total assets change from general price movements. This procedure for the funds flow computation is normally done for estimating cross-sectional relationships. We base our aggregate index fund flow computation formula on equation 3.1, but, as expected on an aggregate level, we modify the formula to estimate flow on a macro-scale. We adopt a similar method as Cao, Chang & Wang (2008) for calculating aggregate flow. The computation becomes fairly simple, where, for each time period, we estimate the dollar value flow for each fund and obtain a aggregate dollar value flow for each period. We subsequently adjust the aggregate dollar value flow by dividing by previous periods aggregate total net assets, creating a fractional

flow measurement. Mathematically, this procedure is equivalent to

$$FLOW_t^{AGG} = \sum_{i=1}^{N_t} FLOW_{i,t} \times TNA_{i,t-1} \times \sum_{j=1}^{N_{t-1}} \frac{1}{TNA_{j,t-1}} \tag{3.2}$$

where N_t is the number of index funds for period t. Explicitly, the dollar value flow is thus given by

$$FLOW_t^{AGG} = \sum_{i=1}^{N_t} FLOW_{i,t} \times TNA_{i,t-1} \tag{3.3}$$

As these flow equations rely on a difference operator, the resulting number of $FLOW_t^{AGG}$ values is one less than the number of $\sum_{i=1}^{N_t} TNA_{i,t}$ values. As such, we compute two different flow measurements: fractional flow using equation 3.2, and dollar flow using equation 3.3.

The reason why we utilize equation 3.2 and 3.3 for calculating $FLOW_t^{AGG}$ values is due to that some adjustments need to be done on individual fund flow level. Due to this nature of the flow data, we obtain a total of 70 255 final individual flow values, after all adjustments described below. Before any such adjustments to the flow data, we have 69 516 data points (i.e. naked individual flow data).[12] As the data exhibits irregularities regarding when different funds start and stop reporting data (we explain these as the inception and death of the fund); following Sirri & Tufano, 1998, we manually insert a 100% flow value for the month of the first reported TNA value (the Inception) and a -100% flow value for the month succeeding the last reported TNA value (the Death). This Inception/Death procedure adds 739 new flow values, raising total data points to 70 255. Likewise, some funds have long time periods of non-reported data in between data points; we treat funds with such data structure as dead if the gap is longer than 12 months. To reduce the impact of large outliers in the flow data, especially occurring in the second period of a funds reported data, wherein the first period TNA is typically a very low value and then a huge influx of money in the second time period's TNA value, causing an enormous flow value. We have winsorized the flow values in the first and the ninety ninth percentile. We winsorize the entire data set, and not per fund or per month, replacing 1405 number of Flow values.[13] Finally, we utilize equation 3.2 and 3.3 to compute our $FLOW^{AGG}$ values.

In Table 2 we present descriptive statistic for our calculated $FLOW^{AGG}$ values, providing us with several insights, and in figure 2 we show the evolution of $FLOW^{AGG}$ over time. The fractional flow data experience a clear positive mean of 0.35% (4.28% annualized) very similar to the median, possibly due to the much larger number of inflow

[12] In appendix A.3 we present regressions with the naked aggregate flow data. We advise the reader to read section 4 and 5 before engaging the naked data regressions.

[13] It is obvious that the ordering of the data adjustments matter. Naturally, we winsorize last as to not cause any 'damage' to the dataset.

Flow descriptive stats

Panel A: Aggregate mutual fractional fund flow

	N	Max	75%	50%	25%	Min	Mean	Std	Skew	Kurt
Flow	239	155	49	32	15	-70	35	33	0.78	4.76
Inflow	212	155	51	36	22	00	41	29	1.43	4.45
Outflow	27	-1	-6	-12	-19	-70	-15	14	-2.45	9.87

Panel B: Aggregate mutual dollar fund flow

	N	Max	75%	50%	25%	Min	Mean	Std	Skew	Kurt
Flow	239	8.48	3.81	2.21	0.89	-1.73	2.47	2.06	0.49	2.82
Inflow	212	8.48	3.99	2.54	1.46	0.00	2.86	1.85	0.72	2.96
Outflow	27	-0.02	-0.30	-0.44	-0.57	-1.73	-0.56	0.40	-1.51	4.78

Panel C: partial autocorrelations of aggregate mutual fund flow

Lag	1	2	3	4	5
Fractional flow	0.1105	0.0817	0.1525	-0.0501	0.1205
Dollar flow	0.5527	0.2733	0.2244	0.0609	0.1138

Table 2: Descriptive statistics for monthly aggregated flow values from Feb 2000 until Dec 2019. The values of fractional flow are in basis points and the values for dollar flow are in billions of dollar. Flow represent all flow data, while Inflow and Outflow represent the flow data corresponding to positive and negative flow, respectively.

observations. We observe that only (approximately) 11.3% of the flow data are outflows, and the last outflow occurs on Oct 2011. The extreme values are clearly observed for inflow, for both the fractional and dollar measurement. The fractional flow data seems to exhibit leptokurtic properties, meaning it contains fat tails and are more prone to outliers in comparison with a standard normal distribution, whereas the dollar fund flow show signs of the opposite, platykurtic, implying fewer and lower extreme values. We also observe a positive skewness for flow (more so for fractional flow than dollar flow), somewhat contradicting the daily S&P500 flow values computed by Cao, Chang & Wang (2008) over the period Feb 1998 to Dec 2003, who obtained a mean of -0.20 basis points with more outflows than inflows. They attribute their negative skewness to the large amounts of outflows that occurred during the dot-com bubble. Nevertheless, the evolution of our monthly aggregate fractional flows appear similar to Cao, Chang & Wang (2008)(the evolution of our fractional flow is showed in figure 2), suggesting that for an increasing fund sector the volatility in mutual fund flows occur more early than late for a given time span; why we also compute and test dollar flow. We contribute this *decreasing* fractional flow to the somewhat stable dollar flow. The large returns attributed to S&P500 during 2010-2019 is causing a size effect for fractional flow, wherein the total net assets is increasing more than dollar flow, resulting in a smaller relative flow.

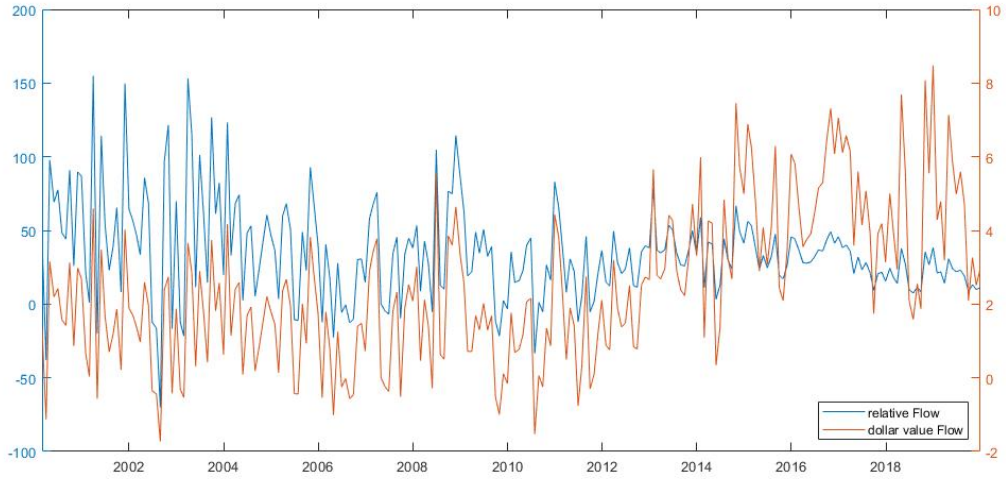


Figure 2: Estimated monthly aggregate Flow values from Feb 2000 until Dec 2019. Fractional flow (left hand axis) are presented in basis points, and dollar value Flow (right hand axis) are presented in billion of dollars.

3.2 Handling of fees

Data of different fees were obtained, including expense ratio, management fees and 12b fees. The return data obtained from CRSP handles the management and 12b fees, as they are included in the net asset value (NAV) of which the return is calculated (Center for research in security prices (CRSP), 2019). Equation 3.4 is used by CRSP to calculate the return for a specified fund.

$$r_t = \frac{NAV_t * Cumfact}{NAV_{t-1}} - 1 \tag{3.4}$$

where R_t is the return for time t, $Cumfact$ is a factor consisting of various distributions, such as dividends and splits that occur in the holdings, and NAV_t is the net asset value as of period-end t. This procedure is estimated to capture the effect of these events and yield a *fair* fund return. In conclusion, we are left with the expense ratio paid by investors, which will be utilized as the fee variable throughout this thesis.

Hortaçsu & Syverson (2004) find that index fund intra-competition were mainly related to other factors than fund fees. This suggests that the aggregating the index fund fee data does not cause harmful inconsistencies in the data, but rather provide us with a useful tool estimate the aggregate index fund flows to the general index fund fee level. As such, we collect yearly fee data for the identified index funds. Fee values are collected as percentage points and in the report fees are on average for all fund, i.e non weighted to capture the entire market of index fund and not suffer from large impact of leading index funds. The fee data experienced missing values in connection to inception and liquidation. These missing values were either filled by the previous years reported fee, or, if that value does no exist, the next years reported fee. Since the fees reported are yearly fees, they

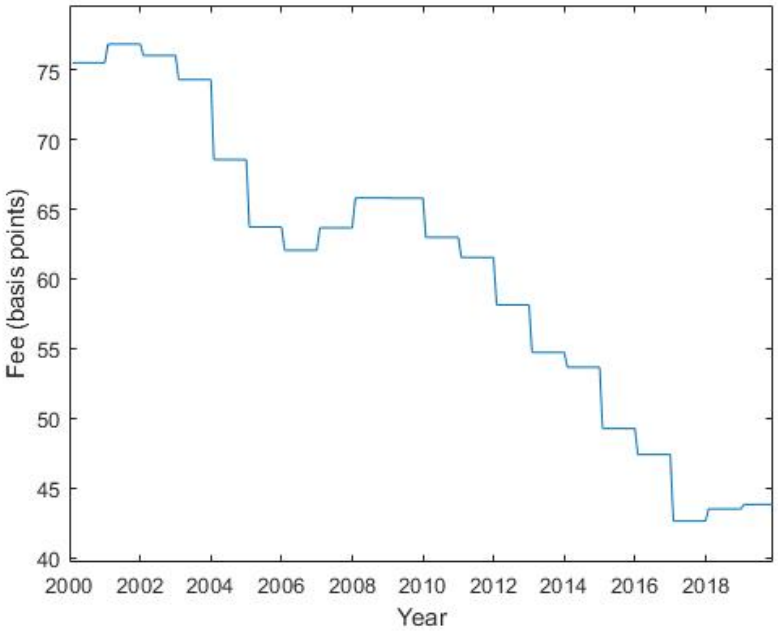


Figure 3: The evolution of average index fund yearly fee for the period 2000-01-01 to 2019-12-31. The Fee data are presented in basis points.

were divided by 12 to represent the cost for each month of holding the fund. Thus, we obtain piece wise monthly fee data, readjusting on a yearly basis. Finally we calculate the mean of each month to obtain an aggregated fee per month for all the index funds. Figure 3 displays a clear pattern in fee reduction over the past 20 years. The average cost of a S&P500 index fund has dropped roughly 30 basis points. Interestingly, a small upwards rebound is observed during periods of crisis (in 2002 and 2008-2010).

4 Methodology

Following the *recent* conventional papers on the evolution of market efficiency (i.e. time indexed/varying), we will utilize the Hurst exponent as a proxy for market efficiency (Bariviera, 2011; Eom et al, 2008; Kristoufek & Vosvrda, 2012). Explicitly, this thesis thus aims to examine the relationship between aggregate index funds flow (estimated on aggregated fund level) and the markets Hurst exponent (estimated on market level, i.e. the target index). This section is organized as follows: we will first setup up the general methodology as a martingale sequence for testing EMH, then introduce the concept of long-term dependency, and finally we will conduct a Granger causality test and specify our estimation model.

4.1 General method setup

As is custom (see for example Kristoufek & Vosvrda (2013) or Huang, Wei, & Yan, (2007)), we can define an efficient market in accordance to security prices conforming to a martingale sequence. Let C be the securities market expressed by the probability space $(\Omega, \mathcal{F}, P)$, where Ω is the sample space, $\mathcal{F}$ the event space (a subset of Ω). Further, let $\mathbb{P}$ be the real probability measure and the conditional probability as $\mathbb{P}[X_t|\mathcal{F}_t]$. As such,

A securities market $C = (\Omega, \mathcal{F}, P)$ is efficient if there exists P such that the time-series of prices $S = (S_t)_{t\geq 1}$ is a martingale process; i.e.,

$$\mathbb{P}[|S_t|] < \infty \qquad \mathbb{P}[S_{t+1}|\mathcal{F}_t] = S_t \qquad t \geq 0 \tag{4.1}$$

From here we can redefine the martingale price process to incorporate a random error term. Let $(\epsilon_t)_{t\geq 1}$ be random IID. variables with mean zero, that is $\mathbb{P}[\epsilon_t] = 0$. This incorporates innovations with no autocorrelation. Thus, we can reformulate the security price series into

$$\mathbb{P}[S_{t+1}|\mathcal{F}_t] = S_t + \mathbb{P}[\epsilon_{t+1}|\mathcal{F}_t] \tag{4.2}$$

Such a security market, where the security prices are described by the above, would be *efficient* as the market does not exhibit any pattern or memory to exploit (McCauley, Bassler, & Gunaratne, 2008): there exists no systematic way to beat the market. By utilizing this martingale feature of security prices, we obtain a robust model and avoid the random walk assumption of homoskedasticity (Kristoufek & Vosvrda, 2013). Converting price data to return data, and let the 1-period return be $r_t = \frac{S_t - S_{t-1}}{S_{t-1}}$ and μ be the drift of the return process, we obtain

$$\mathbb{P}[r_{t+1}|\mathcal{F}_t] = \mu + \phi r_{t-1} + \mathbb{P}[\epsilon_t|\mathcal{F}_t] \tag{4.3}$$

Modelling security returns in this manner (according to equation 4.3), the prices would adhere to a random walk, suggesting a weak form of EMH is present. Subsequently, such a return series residuals can be estimated to follow a white noise process by modelling the long-term dependency measurement the Hurst exponent, where a Hurst exponent equal to 0.5 corresponds to a random walk (Hsieh, 1993). This produces a variable effective as a test of the martingale sequence and long-term dependency. Important to note, is that a martingale sequence can exhibit memory in the sense we know the previous values and that the expected value is just this previous value (McCaule, Bassler, & Gunaratne, 2008). When we, in this paper, talk about a no-memory scenario, we rather mean that the innovation defining a step in the sequence cannot be predicted (i.e. it should be random).

4.2 Long-term dependency

The Hurst exponent is a popular estimate for assessing long-term memory of a time series. The method was originally invented by a hydrologist named Hurst who struggled to prevent the Nile River Dam of overflowing. The method was utilized within finance roughly a century after its first appearance and is today widely used to detect long term memory in data (Peters, 1991). The Hurst exponent generally has a support of $H \in (1, 0)$ with three general outcome estimations: $H = 0.50$ indicates a random and uncorrelated series with no memory, $H < 0.50$ indicates that the series exhibit anti-persistence or anti-memory, and $H > 0.50$ indicates long-term memory in the series Peters (1991).[14] Unfortunately, no known distribution exists for Hurst exponent, but these general outcomes have been asymptotically proven. The computation of Hurst exponent can be done by many different statistical techniques, the two most commonly used within finance and therefore utilized in this paper are; the rescheduled range (RS) and detrended fluctuations analysis (DFA). Both techniques are more thoroughly explained later in this chapter. One drawback of the RS Hurst exponent is the estimation of existing long-term memory due to short-term dependency in the data series (Grau-Carles (2000); Di Matteo, (2007)). The two main methods to avoid this issue is (i) to filter the return data through a GARCH-process, and (ii) to use the Detrended Fluctuation Analysis (DFA) method.

Bariviera (2011) and Cajueiro & Tabak (2004) [among others] bypass this issue by filtering the return series through an AR(1)-GARCH(1,1) process. This procedure removes the short-term dependencies (memory) found in the time-series (by construction of the autoregressive element), which, if left undisturbed, may instill long-term dependency (where none actually exist) in the Hurst exponent (Tabak & Cajueiro, 2004; Di Matteo, 2007). The lagged component of the main equation in the GARCH leaves the residuals as the "true and random" returns for which we aim to estimate long-term memory. Bariviera (2011) and Grau-Carles (2000) estimate the Hurst exponent using the DFA method, wherein the residuals of a locally detrended integrated time-series are used to compute the global fluctuations and repeated over multiple window lengths.[15] Di Matteo (2007) further criticize the conventional Hurst RS method for its sensitivity to outliers and presence of heteroskedasticity in the returns time series. In this paper, to avoid this spurious detection of long-term memory, we will estimate the Hurst exponent using both the RS

[14] The reason why the Hurst exponent, with its support of $H \in (0, 1)$, can act as a proxy for market efficiency is due to the measurement of the average fluctuation and how it relates to time periods. Unfortunately When the average fluctuation for a long window sizes is roughly equal to the average fluctuation for short window sizes, the return data is exhibiting frequent sign-changes ($+$ and $-$). Likewise, when the average fluctuation is larger for longer window sizes, the return data, by necessity, exhibit sign-trends wherein subsequent returns change sign less frequently. A white noise process have been empirically asymptotically discovered to equal a Hurst exponent of 0.5 (Anis & Lloyd, 1976).

[15] We here use the word integrated in the sense of the original paper from Peng et al. (1995), wherein a detailed description of the DFA method can be found, meaning a mean adjusted cumulative sum; i.e. a cumulative deviation from the mean.

approach (with filtered returns) and the DFA method. Furthermore, previous studies consistently estimate a higher Hurst exponent using the RS method than by using the DFA method (see for example Bariviera (2011)).

4.3 The Hurst exponent

The Hurst exponent is formally defined as a power law function (for reference, see for example Peters (1991)) of the type

$$\Phi_n = Cn^H \qquad as \ n \longrightarrow \infty \tag{4.4}$$

where H is the Hurst exponent, C is a constant, n is the number of observations in a partial time series, and $\Phi_n^{RS} = \mathbf{E}(RS)_n$ for RS (rescaled range analysis), or $\Phi_n^{DFA} = F_n$ for DFA (detrended fluctuation analysis). As such, distinguishing these methods is the computation of the left hand side (both of which will in detail be described in the following sections). To obtain the Hurst exponent H, a log-transformation is utilized[16], converting equation 4.4 into

$$log(\Phi_n) = log(C) + H \times log(n) \tag{4.5}$$

where, once Φ_n is computed, a simple linear regression can be run to estimate H. Let's define this inverse function for running equation 4.5 and estimating H, by inputting Φ_n on the left hand side, as $\Theta(\Phi_n)$. Explicitly, Φ_n is thus a vector with equal size as n.

When estimating a time varying Hurst exponent, reasonable assumptions regarding window size are of importance. Eom et al. (2008) used a window size of 60 months with 12 months rolling. In accordance with similar studies (see Cajuerio & Tabak, 2004; Bariviera, 2011), we are arguing for a Hurst window size corresponding to political cycles. As such, we utilize a window size of $N = 1008$ observations (252 trading days multiplied by four years)(let's call this the *global window length*[GWL]).[17] Calculating the Hurst exponent H_t thus requires $N = GWL - 1$ of previous observations including the observation on period t. As such, to compute the first Hurst exponent for Jan 2000, data is needed from Jan 1996; why we download a longer data set for S&P500.

4.3.1 GARCH

The Generalized Autoregressive Conditional Heteroskedasticity model, or GARCH, is a process to estimate the conditional variance of a time-series which exhibits heteroskedasticity. The GARCH-model is constituted by two equations: the mean equation, which

[16]By standard of convention, the log-operator is defined as the natural logarithm.

[17]The Hurst window size profoundly affects the estimated Hurst exponent. Thus, choosing the correct window length is a sensitive issue, why we follow previous studies. As with any econometric computation, more observations generally produce a more robust estimation; nonetheless, when estimating current day market efficiency, far historical returns should have a diminishing effect.

makes assumptions and models the underlying time-series, and the variance equation, which models the conditional variance of the underlying time-series. As previously discussed, we filter the return data through an AR(1)-GARCH(1,1) model for a better input variable into the RS analysis.

Thus, our model is specified as an AR(1)-GARCH(1,1) process

$$r_t = \mu + \Phi r_{t-1} + \epsilon_t \tag{4.6}$$

$$\epsilon_t = \sigma_t z_t, \quad z_t \sim N(0,1) \tag{4.7}$$

$$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{4.8}$$

where equation 4.6 is the mean equation with the specific AR(1)-term, and equation 4.8 is the variance equation, wherein the r_t is the return for time t and σ_t^2 is the conditional variance of the returns for time t, and μ, ω, α, and β are unknown model parameters that need to be estimated. For a reasonable behaving GARCH-process, some of the parameters require some restrictions: $\omega > 0$, $\alpha > 0$, and $\beta > 0$ to ensure positivity of the conditional variance, and $\alpha + \beta < 1$ to ensure stationarity. In Table 3, we present our estimation of the AR(1)-GARCH(1,1) process. We observe a clear significant autoregressive process in the S&P500 returns.

AR(1)-GARCH(1,1) estimation

		Value	SE	T-stat	P-value
AR(1)	Constant	0.0008	0.0001	7.4025	0.00
	AR(1)	-0.0494	0.0139	-3.5559	0.0004
GARCH(1,1)	Constant	0.00	0.00	6.3747	0.00
	GARCH(1)	0.8745	0.0061	142.34	0.00
	ARCH(1)	0.1050	0.0049	21.311	0.00

Table 3: Model estimation for AR(1)-GARCH(1,1), estimated using daily S&P500 returns from Jan 4th 1999 to Dec 31 2019; where the AR-value is the coefficient for the lagged return term in the mean equation, the ARCH-value is the coefficient for the lagged error term in the variance equation, and the GARCH-value is the coefficient for the lagged variance term in the variance equation.

4.3.2 Rescaled range analysis

The computation of the rescaled range analysis follows several steps, outlined below, and is performed for multiple partial series of the original full length series. That is, we first divide the full time series of length N into $j = 1, 2, ..., k$ non-overlapping partial time series with lengths n_i. There exists several ways to conduct this division: for example, $n_i = N, N/2, N/4...$, or n_i equal to the factors of N, where $i = 1, 2, ..., v$. For our

calculations we utilize the latter approach, due to its computational simplicity (following Weron (2002). We then execute the following steps for each n_i

1. Calculate the mean: $\mathbb{E}[X_j^{n_i}] = \frac{1}{n_i}\sum_{t=1}^{n_i} X_{j,t}^{n_i}$
2. Calculate mean-adjusted series: $Y_{j,t} = X_{j,t}^{n_i} - \mathbb{E}[X_j^{n_i}]$
3. Calculate cumulative deviations: $Z_{j,t}^{n_i} = \sum_{t=1}^{n_i} Y_{j,t}^{n_i}$
4. Calculate the range: $R_j^{n_i} = max(Z_j^{n_i}) - min(Z_j^{n_i})$
5. Calculate the variances: $(\sigma_j^{n_i})^2 = \mathbb{E}[(X_{j,t}^{n_i})^2] - \mathbb{E}[X_{j,t}^{n_i}]^2$
6. Compute the rescaled range and the average (expectation) for all partial time-series: $\frac{R_j^{n_i}}{\sigma_j^{n_i}}$ and subsequently $\mathbf{E}(RS)_n = \mathbb{E}\left[\frac{R_j^n}{\sigma_j^n}\right] = \frac{1}{k}\sum_{j=1}^{k}\frac{R_j^n}{\sigma_j^n}$

After obtaining $\mathbf{E}(RS)_n$, it is now possible to estimate the Hurst exponent in $\Theta(\Phi_n)$. However, such an estimation will have a significant deviation from its theoretical value for small window sizes n_i (Weron, 2002). To correct for this, we subtract the window sizes theoretical white noise approximation (this modification to the RS analysis was introduced by Anis & Lloyd (1976) with some slight modifications by Peters (1991)). As such, we obtain the *true* deviations from the white noise slope; the Hurst exponent can thus be calculated as 0.5 plus this deviation. The theoretical white noise approximation is given by (keeping the original left hand side notation from Anis & Lloyd (1976))

$$\mathbf{E}(R_n^{**}) = \begin{cases} \left(\frac{n-\frac{1}{2}}{n}\right)\frac{\Gamma\{\frac{1}{2}(n-1)\}}{\sqrt{\pi}\Gamma\{\frac{1}{2}n\}}\sum_{i=1}^{n-1}\sqrt{\frac{n-i}{i}} & \text{for n} \leq 340 \\ \left(\frac{n-\frac{1}{2}}{n}\right)\frac{1}{\sqrt{\frac{1}{2}\pi n}}\sum_{i=1}^{n-1}\sqrt{\frac{n-i}{i}} & \text{for n} > 340 \end{cases} \tag{4.9}$$

where $\Gamma\{\lambda\}$ is the gamma function evaluated at λ, and n are the window sizes n_i. Explicitly, the corrected version of the RS Hurst is thus calculated as[18]

$$H^{RS} = 0.5 + \Theta\left(\mathbf{E}[RS]_n\right) - \Theta(\mathbf{E}[R_n^{**}]) \tag{4.10}$$

Following Cajueiro & Tabak (2004) and Bariviera (2011), before calculating the Hurst exponent, we employ the AR(1)-GARCH(1,1) process to filter the returns for short-term dependency (see Table 3); the estimated residuals from the mean equation 4.6 are then divided by the conditional standard deviation from the variance equation 4.8. The resulting fraction is our filtered returns and is thus used to complete the Hurst calculation.

[18]The attentive and enlightened reader might here realize that $\mathbb{E}(R_n^{**})$ is not time-varying, but only a function of window sizes n. As such, in equation 4.10, we are subtracting a constant number $0.5 - \Theta(\mathbf{E}[R_n^{**}])$ from all estimated RS Hurst exponents. For a global window size of 1008 returns, and local window sizes $n = divisor(1008)$, this constant adds up to approximately -0.0688. This means that for our local window sizes n, $\Theta\left(\mathbf{E}[RS]_n\right)$ consistently overestimate the Hurst exponent by this value.

Formally we can define the filtered returns explicitly as:

$$\Omega_t = \frac{\epsilon_t}{\sqrt{\sigma_t^2}} \tag{4.11}$$

where Ω_t are the filtered returns, ϵ_t are the residuals from the AR(1)-GARCH(1,1) mean equation, and σ_t^2 are the conditional variances from the AR(1)-GARCH(1,1) variance equation. That is, we conduct the RS approach on the filtered returns Ω_t.

4.3.3 Detrended fluctuation analysis

Detrended fluctuation analysis (DFA) is a method to detect long term memory in data similar to the previously introduced RS approach, but differing in some aspects, which makes DFA a useful variable along the rescaled range analysis for robustness purposes. DFA was introduced by Peng et al. (1995) and foremost used within the medical field to find long term correlations in heart rate and DNA data. DFA was popularized within financial data by Kantelhardt et al. (2001) and is frequently used to examine long range dependencies. When detecting long range memory it is essential to filter out disturbances possibly causing spurious memory. In the RS Hurst approach short term memory was removed from the data by the filtering process. DFA instead identifies trends within each local window size, which can cause false dependencies, both long and short term (Eom et al, 2008) caused by externalities (Kantelhardt et al, 2001).

The computation of the detrended fluctuation analysis is in principal similar to the rescaled range analysis, but differs in some methodological aspects. Exactly alike, we first divide the full time series of length N into k non-overlapping partial time series (windows) with equal lengths n_i (utilizing the divisors approach here as well). For each n_i, we do the following. For each partial time series $j = 1, 2, ..., k$, we compute the mean $\bar{x_j} = \mathbb{E}[x_j]$ and subsequently the integrated series of $X_{j,t}$ (a cumulative mean-centered sum)

$$X_{j,t} = \sum_{i=1}^{t} [x_{j,i} - \bar{x_j}] \tag{4.12}$$

Withing each integrated series $X_{j,t}$, a fitted straight line is located to find the trend for each window. Let $\hat{Y}_{j,t}$ be this straight line fit, obtained from $X_{j,t} = \alpha + b\tau$, where $\tau = 1, 2, ..., n_i$. Then, the fluctuation F_n is computed as follows for each window length n_i

$$F_n = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (X_{j,t} - \hat{Y}_{j,t})^2} \tag{4.13}$$

Thus, we obtain an average fluctuation F_n for each integrated and detrended time series of length n_i, and we can now estimate the Hurst exponent H as $\Theta(F_n)$ (equation

4.5).[19] In contrast with the rescaled range analysis, the Hurst exponent estimated from the detrended fluctuation analysis have a support of $(0 < d \leq \infty)$. Similarly, a Hurst coefficient of $H < 0.5$ indicates negative memory or anti-persistence and $H > 0.5$ positive memory, and $d \approx 0.5$ is a sign of no memory. Although normal behaving return series still lie within $H \in (0, 1)$, a Geometric Brownian Motion process (a cumulative sum of the returns) would theoretically produce a Hurst exponent equal to 1.5.

4.4 Endogeneity concerns

The causal relationship between passive investing and market efficiency exhibits uncertainty regarding direction of causality. This duality of whether passive investing exists because of efficient markets or does passive investing affect market efficiency, raises concerns about the dependency in the structural regression.

To examine whether a variable has a *causal* relationship with another variable we choose to conduct a Granger causality test. The test procures a measurement of how well the dependent variable exhibits the same pattern as the key variable. By lagging the key variable, where p is the minimum and q the maximum lag of significance, we observe how well they can describe or predict the dependent variable. The model is thus given by a VAR(p,q)-model, specified explicitly as

$$\begin{bmatrix} F_t \\ H_t \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \psi_{11,p} & \psi_{12,p} \\ \psi_{21,p} & \psi_{22,p} \end{bmatrix} \begin{bmatrix} F_{t-p} \\ H_{t-p} \end{bmatrix} + ... + \begin{bmatrix} \psi_{11,q} & \psi_{12,q} \\ \psi_{21,q} & \psi_{22,q} \end{bmatrix} \begin{bmatrix} F_{t-q} \\ H_{t-q} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix} \tag{4.14}$$

where, F_t is the flow variable, H_t is the Hurst exponent, $\epsilon_{i,t}$ is a standard error term $\epsilon_{i,t} \sim N(0, \Sigma)$ where Σ is a 2×2 covariance matrix. The off-diagonal elements in equation 4.14 represent the cross-impact of flow and Hurst, while the diagonal elements are the autoregressive terms. We produce VAR-models for dependent variables fractional flow and dollar flow, with dependent variables RS Hurst as well as DFA Hurst. To determine

[19] We believe it is here appropriate to clarify the DFA-method with an example. Lets say the integrated series X_t have a length of 252 (let us also define this as the global window size). From this we choose a our window lengths n to be from 12 up to 1008 in increments of 1. That is we first divide X_t into non-overlapping windows of length 5, giving us $\lfloor \frac{252}{5} \rfloor = 50$ windows (let us call these windows local windows). For each local window we locate the straight line fit $\hat{Y}_t$. Next, for the calculation of F_5 we detrend X_t by its corresponding $\hat{Y}_t$ and complete the computation. We then repeat this process for F_6,F_7,...,F_{252} with local window sizes of $6, 7, ..., 252$. For a financial time series it is reasonable to set the minimum local window size to 5, corresponding to 1 trading week. This DFA technique is quite computational intensive as it estimates a vast number of regression. Explicitly, lets denote the total length of our required series to T, the global window length as GWL, and the maximum local window length as LWL_N, then the total number of estimates Ψ are equal to $\Psi = (T - GWL + 1) \times GWL \times \left(\sum_{n=LWL_1}^{LWL_N} \frac{n}{LWL_N} \right)$. To compute daily DFA in the preiod 2000-2019 with $GWL = 1008$ and $LWL_n = 5, 6, ..., 252$, we need T=6038 return data, from 8th Jan 1996 to 31th Dec 2019, resulting in approximately 641 million estimations. A more simple approach established by Weron (2002), which we utilize, is to choose LWL equal to the factors of the GWL, as well as limiting minimum LWL to 8. This is far less computational intensive and avoids the problem of unregular LWL causing last window to have fewer observations.

VAR specification - fractional flow

		DFA-VAR(3)		**RS**-VAR(1)	
	Dependent →	Flow	Hurst	Flow	Hurst
	Response ↓				
Constant		0.0055**	0.0127**	0.0089**	0.0161*
		(2.1027)	(2.1772)	(2.2137)	(1.8786)
	Flow	0.0982	0.0812	0.1056	0.0994
AR(1)		(1.5474)	(0.5761)	(1.6418)	(0.721)
	Hurst	0.0003	1.4905***	-0.0106	0.9699***
		(0.0104)	(23.35)	(-1.4457)	(61.646)
	Flow	0.0488	0.0954	-	-
AR(2)		(0.7768)	(0.6841)		
	Hurst	0.0453	-0.7232***	-	-
		(0.9483)	(-6.8149)		
	Flow	0.1437**	0.0326	-	-
AR(3)		(2.2949)	(-0.2343)		
	Hurst	-0.0528*	0.2021***	-	-
		(-1.8322)	(3.1607)		
NumParam, k		14		6	
LogLikelihood		1860.43		1868.38	
AIC		-3692.86		-3724.75	

Table 4: VAR model specification for fractional flow and Hurst exponent calculated using both rescaled range (RS) and detrended flucuation analysis (DFA). The best fit for DFA was a VAR(3) model and for RS a VAR(1) model. Flow was computed as specified in equation 3.2. Daily DFA and RS Hurst exponents was computed accordingly to section 4.3.2 and 4.3.3, respectively, and subsequently averaged on monthly basis. The values in parenthesis represent the above coefficients t-statistic.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.

the length of which to lag the key variable in the Granger Causality test (that is, to find optimal q), we conduct an Akaike Information Criterion-test (AIC). The AIC determines "information loss" given a specific model, and the best fit is the model which minimizes the AIC. The AIC is given by

$$AIC = -2(logL) + 2k \tag{4.15}$$

where $logL$ is the loglikelihood value of the estimated VAR(p,q)-model and k is the number of estimated parameters for the model. We run the AIC estimation for $p = 1$ and $q = 1, 2, ..., 12$. In Table 4 and 5 we present the VAR models, and their corresponding LogLikelihood and AIC values, for fractional flow and dollar flow, respectively. The best fit (lowest AIC value) obtained for fractional flow is for a DFA specification $q = 3$ and for a RS specification $q = 1$; and, the best fit (lowest AIC value) obtained for dollar flow is for a DFA specification $q = 3$ and for a RS specification $q = 2$. The off-diagonal elements in equation 4.14 is our main interest of study, where, for example, $\psi_{12,1}$

VAR specification - dollar flow

		DFA-VAR(3)		**RS**-VAR(2)	
	Dependent →	Flow	Hurst	Flow	Hurst
	Response ↓				
Constant		6.4417***	0.0213***	7.0607***	0.0172**
		(3.9386)	(2.8912)	(3.5449)	(1.9896)
	Flow	0.2875***	0.00	0.3591***	0.00
AR(1)		(4.5409)	(0.3181)	(5.734)	(0.0446)
	Hurst	-11.296	1.4818***	-27.307*	0.9613***
		(-0.7989)	(23.27)	(-1.8522)	(14.809)
	Flow	0.1414**	-0.0003	0.2309***	-0.0002
AR(2)		(2.1943)	(-0.8633)	(3.7402)	(-0.587)
	Hurst	27.11	-0.7083***	14.573	0.0034
		(1.1539)	(-6.693)	(0.9791)	(0.0515)
	Flow	0.1761***	-0.0004	-	-
AR(3)		(2.8273)	(-1.3749)		
	Hurst	-28.433**	0.1803***	-	-
		(-2.0007)	(2.817)		
NumParam, k		14		10	
LogLikelihood		399.08		391.15	
AIC		-770.16		-762.30	

Table 5: VAR model specification for dollar flow and Hurst exponent calculated using both rescaled range (RS) and detrended flucuation analysis (DFA). The best fit for DFA was a VAR(3) model and for RS a VAR(1) model. Flow was computed as specified in equation 3.3. Daily DFA and RS Hurst exponents was computed accordingly to section 4.3.2 and 4.3.3, respectively, and subsequently averaged on monthly basis. The values in parenthesis represent the above coefficients t-statistic. Dollar value flow are in billions of dollar.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.

correspond to dependent variable flow and independent response AR(1) Hurst in their respective tables.[20]

We note that H_{t-3}^{DFA} have a significant impact on $Flow_t$ (although only partially for fractional flow), possibly causing the best fit to be a VAR(3)-model for both fractional and dollar flow. Nonetheless, H_{t-1}^{RS} is not significant on fractional flow, it is partially significant at 10% on dollar flow. Moreover, all of the AR-terms on Hurst are highly significant. This is most likely due to that both DFA and RS are stationary processes. Economically this is intuitive, as one state of market efficiency highly depend on previous states of market efficiency.

Next, we run Granger Causality tests with the specified VAR models presented in Table 4 and 5. A Granger causality test estimates if either of the series Granger cause the

[20]To ensure the causality we also conduct a BIC-test, similar to the AIC. The BIC penalizes more for complex models than the AIC, and is given by $BIC = -2(logL) + k * ln(n)$, where n is the number of observations and otherwise the same notation as in equation 4.15 applies. The BIC-test yields similar results and confirm the direction of the causality.

Granger Causality test

Null hypothesis	Statistic	P-value	Causal direction
Panel A: fractional Flow			
DFA does not Granger cause FLOW	1.6538	0.1778	DFA $\nleftrightarrow$ FLOW
FLOW does not Granger cause DFA	0.2892	0.8332	
RS does not Granger cause FLOW	2.0638	0.1522	RS $\nleftrightarrow$ FLOW
FLOW does not Granger cause RS	0.5133	0.4744	
Panel B: dollar value Flow			
DFA does not Granger cause FLOW	4.6735	0.0034	DFA $\rightarrow$ FLOW
FLOW does not Granger cause DFA	1.3794	0.2498	
RS does not Granger cause FLOW	6.1413	0.0025	RS $\rightarrow$ FLOW
FLOW does not Granger cause RS	0.20499	0.8148	

Table 6: Granger causality test results for the VAR specified models. Panel A is conducted using VAR specification in Table 4 and Panel B is conducted using VAR specification in Table 5. The null hypothesis tests whether the VAR coefficients are jointly difference from zero, see equations 4.16 and 4.17 for explicit specifications. A significant p-value indicates rejection of the null hypothesis. The test statistics are computed using a F-test.

.

other one; i.e. if predictive power exists. The null hypothesis are stated as the off-diagonal elements (in equation 4.14) provide no joint significance.[21] Explicitly, this is

$$H_0^F : \psi_{12,p} = \psi_{12,p+1} = ... = \psi_{12,q} = 0 \tag{4.16}$$

and

$$H_0^H : \psi_{21,p} = \psi_{21,p+1} = ... = \psi_{21,q} = 0 \tag{4.17}$$

where H_0^F is interpreted as *HURST does not Granger cause FLOW*, and H_0^H as *FLOW does not Granger cause HURST*. In Table 6 we present test statistics for our Granger Causilty test. We find no predictive power for either fractional flow on RS or DFA, and neither for RS or DFA on fractional flow. Nevertheless, both RS and DFA seem to Granger cause dollar value flow, indicating a a causality running from market efficiency towards index fund flow. Furthermore, no observable simultaneity endogeneity is indicated, suggesting a one-direction causality relationship. This forecasting/predictive power dictate the direction of the relationship we want to study. We will as such continue this paper by examining the effect market efficiency have on aggregate index fund flow. A possible

[21] Another approach to test for Granger causality is the null hypothesis that the summed off-diagonal coefficients are equal to zero, $\sum \psi_{i,j} = 0$ for $j \neq i$. Such testing procedure examine if the predictor variable have any overall impact on the effect variable; in contrast with our current method which examines if at some lag p to q, the other series can be predicted. Although this particular sum-test approach could be of interest, we consider it out this papers scope, and to therein not conduct one. For the interested reader, this sum-procedure is performed by Kadapakkam, Krause, & Tse (2015) on ETFs.

explanation for why market efficiency Granger cause dollar flow, and not fractional flow, could be due to the size effect of fractional flow discussed in section 3.1.2.

4.5 Estimation and control variables

We base our estimation model on the causality found in the Granger causality test. Thus, to examine the relationship between passive investments and market efficiency, we regress $Hurst_t$ on $FLOW_t$, while controlling for variables that might affect Flow levels. Several variables beside flow and Hurst are included in the regression to increase explanatory power. The variables are generally consistent with the regression of Sirri & Tufano (1998), and Huang, Wei, & Yan (2007) as well as Edelen & Warner (2001), but adapted for index funds. Hence, variables related to specific fund returns (typically a ranking based on performance of the mutual funds) are omitted since they are very similar. We instead use market returns of the index, as it is reasonable to conclude that it is the index return which base an investment choice rather than the specific index fund return. When an investor analyzes whether to invest in an index fund, it is sensible to assume that the investor inquire the general index return and then buys the index fund available from an institution where they are existing customers. As such, we regress flow on the following variables:

- $Hurst$ - is included to examine the impact market efficiency have on index fund flow. Both contemporanous and lagged Hurst exponents are estimated.
- $Fees$ - lagged expense ratio is included in the regression on an aggregated level, to capture effects of fee levels differing for index funds over the time scope of the study. The fee value is equally weighted among all the index funds.
- $Log(TNA)$ - lagged logged aggregate total net assets is included to capture changes in the total market size of index funds.
- $Mkret$ - is included to capture effects of previous performance of the index fund market. We include contemporaneous as well as both 1-period and 2-period lagged market returns.
- $CumVol$ - lagged 12-month cumulative standard deviation is included to cover the effect of overall total riskiness of the market. This variable is created as a 12-month rolling window computation of the standard deviation.

Our specified final estimation thus arrive at

$$FLOW_t = \alpha + \beta_1 Hurst_t + \beta_2 Mkret_t + \beta_3 log(TNA_{t-1}) + \beta_4 Fee_{t-1} + \beta_5 CumVol_{t-1} + \epsilon_t \tag{4.18}$$

We conduct two series of regressions: the first with fractional flow as dependent variable, and the second with dollar value flow as dependent variable. Several regressions are

estimated per flow variable with varying regressors. These regressions are presented in section 5.

Annual statistics

Year	Flow (frac)	Flow (dollar)	TNA	Return	Fee	Hurst(RS)	Hurst(DFA)
2000	0.58%	1.8764	319.25	-0.7674%	6.29	0.4471	0.3965
2001	0.56%	1.7750	294.99	-0.9904%	6.41	0.4760	0.4348
2002	0.39%	1.7325	268.86	-2.0061%	6.34	0.4884	0.4452
2003	0.59%	0.8870	286.08	2.0593%	6.19	0.4889	0.4344
2004	0.48%	2.0230	359.74	0.7786%	5.71	0.5202	0.4691
2005	0.39%	1.4814	406.18	0.3123%	5.31	0.5014	0.4747
2006	0.08%	1.4125	458.34	1.1246%	5.17	0.4814	0.4393
2007	0.31%	0.6684	532.33	0.3641%	5.31	0.4519	0.4072
2008	0.51%	1.6435	480.24	-3.7423%	5.48	0.4362	0.4283
2009	0.23%	2.3668	415.74	2.0805%	5.48	0.4739	0.4797
2010	0.22%	0.8026	501.18	1.2312%	5.25	0.4998	0.4802
2011	0.22%	1.2997	596.47	0.1488%	5.13	0.5089	0.4817
2012	0.27%	1.0359	670.56	1.1846%	4.84	0.5075	0.4545
2013	0.41%	2.1999	870.87	2.3160%	4.56	0.4614	0.4107
2014	0.35%	3.4926	1094.3	0.9964%	4.47	0.4468	0.4062
2015	0.35%	3.9248	1264.3	0.0592%	4.10	0.4321	0.3703
2016	0.38%	4.2773	1423.1	0.8755%	3.95	0.4349	0.4086
2017	0.26%	5.3130	1833.1	1.5807%	3.55	0.4381	0.3935
2018	0.22%	4.4228	2138.6	-0.4070%	3.62	0.4752	0.4014
2019	0.18%	4.5541	2435.4	2.2948%	3.65	0.4916	0.4194

Table 7: Annual statistics of regression inputs. The values showcased are averaged per month on a yearly basis. Reported TNA and Flow (dollar)-values are in billions of US-dollars. The returns are averaged using S&P500 returns. The fees are in basis points and based on monthly holdings.

In Table 7, we present a summary of the main data used in the regressions averaged yearly. We observe a clear decreasing trend in the flow data. Looking at the TNA values, which instead portray an apparent increase, it seems reasonable to assume that the declining Flow values stem from the growing TNA, considering flow is a relative value. Cao, Chang & Wang (2008), find similar pattern in their flow data, for multiple fund categories between 1998 and 2003. The time period in this thesis suffers two majors set backs, return wise. Firstly, the dot-com crash in the early 2000s, and secondly, the great financial crisis in 2008-2009. The averaged monthly return of -3.7423% in 2008 would yield an annualized loss of almost 37%. Fees, as already mentioned, have declined in the last 20 years and are substantially lower in 2019. As argued by Huang, Wei, & Yan (2007) and Sirri & Tufano (1998) there is a connection between fund flow and the charged fee. The Hurst RS estimations vary around 0.45-0.5, while the Hurst DFA estimations vary around 0.4-0.45.[22] This discrepancy between RS and DFA values are in line with findings of Bariviera (2011).

[22]In appendix A.2 we preset histograms over the daily estimated Hurst exponents.

5 Results

Before we begin discussing the relationship between broad market efficiency and index fund flow, we will conduct a small examination of our estimated proxy for market efficiency (the Hurst exponent). In figure 4 we show graphically the evolution of the Hurst exponent during the period 2000-2019. As we can observe, H^{RS} is nearly consistently higher than H^{DFA}, where H^{RS} is hovering around 0.45-0.5, and H^{DFA} between 0.35 and 0.5. Explicitly, this means that S&P500 returns exhibits signs of unpredictable returns and market efficiency according to H^{RS}, while H^{DFA} rather indicates a slight mean-reverting mechanism of the returns. This procures a strange task to analyze such measurements impact. Notably, the discrepancy between the RS and DFA values mirror those of previous research (see for example Bariviera (2011)). Furthermore, these result are consistent with the theory behind both of these market efficiency measurements (see Grau-Carles, 2000).

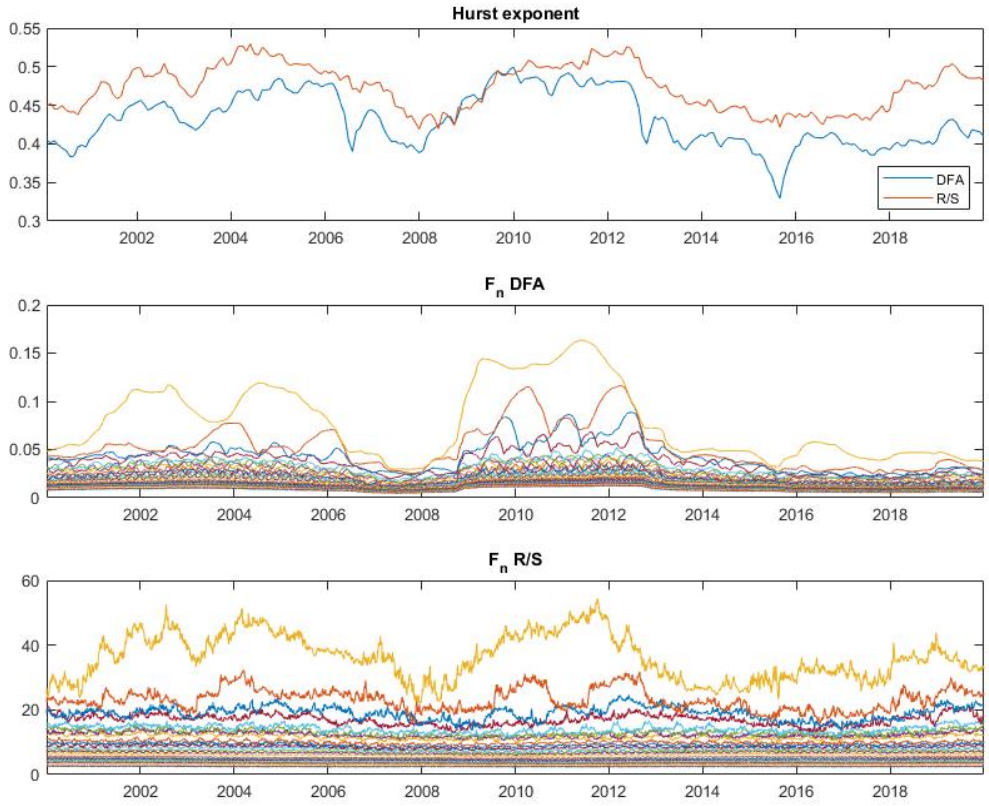


Figure 4: Hurst exponents and the underlying fluctuations (F_n) used in the computations. The Hurst exponents, RS and DFA, are computed accordingly to sections 4.3.2 and 4.3.3, respectively. Daily Hurst exponents are estimated, which then are averaged monthly. The second and third plots show the average fluctuations F_n and $\mathbf{E}(RS)_n$ used in the computations. The 24 lines represent the various window sizes, where lines higher up correspond to larger window sizes (see footnote 19).

Interestingly, in boom periods, the Hurst exponent seems to be trending downwards,

and vice versa, in periods of recession, the Hurst exponent is rising. Mathematically this makes sense, since a steady growing equity market exhibits small average fluctuations F_n (especially for larger window sizes), resulting in a lesser increase in average fluctuation for increasing window sizes. Likewise, for periods of high volatility and turbulence, the average fluctuation naturally increases causing an upward trend in the Hurst exponent. Economically, this response in the market efficiency is less intuitive. The asymptotically proven white noise series of a Hurst exponent equal to 0.5 seemingly appears arbitrary in an economic sense. Intuitively, it is reasonable that periods of high volatility should induce non-market predictability in returns, which we observe for the increasing of H_{DFA} closer to 0.5 during such periods. Conversely, such equity market volatility normally arises from intervals of *steady* market conditions followed by subsequent intervals of sharp downturns (e.g. the famous great financial crisis of 2008). Sufficiently long periods of downturns are thus periods of same-sign returns, exhibiting economic predictability and long-term dependence. Such reasoning further purports the notion that developing markets experience a lower degree of efficiency (Bariviera, 2011; Cajueiro & Tabak, 2004; Eom et al., 2008) given a historically higher volatility (Harvey, 1995; Umutlu, Akdeniz, & Altay-Salih, 2010).

In Figure 5 we show the estimated Hurst coefficients as well as the asymptotic empirical 95% confidence interval.[23] As such, Hurst exponents lying inside the confidence interval corresponds to an empirical estimation of market efficiency. We thus observe that the market efficiency measured by RS are for multiple periods non-distinguishable from efficiency, while market efficiency measured by DFA are consistently below market efficiency except for a brief period after the great financial crisis.

To examine the relationship between index fund flow and market efficiency we estimate three different set of regressions. In Table 8 we present our estimations on fractional flow and in Table 9 we present our estimations on dollar flow.[24] We also run a series of regressions with lagged Hurst exponents; these are presented in Table 10. The reasoning behind the lagged Hurst regressions is to provide robustness. We observe the same negative significance for the lagged Hurst as for the contemporaneous Hurst. This suggests that there exists no reverse causality relationship between flow and Hurst. For fractional flow we observe that both market efficiency measurements coefficient are most significant in regressions (B) and (F), whereas for dollar flow, all market efficiency measurement (both contemporaneous and lagged) are significant.

Noteworthy, we find a negative relationship between flow and contemporaneous market return. This is not in line with Edelen & Warner (2001), where daily fund flow are

[23]There exists no known distribution for the Hurst exponent, but Weron (2002) estimated an empirical confidence interval for both RS and DFA around a Hurst exponent eqaul to 0.5.

[24]In appendix A.3 we also present regressions run on naked aggregate flow data, wherein no adjustments on either retunr, TNA, or flow data has been performed. As we observe, we lose significance probably due to the large outliers in the naked flow data.

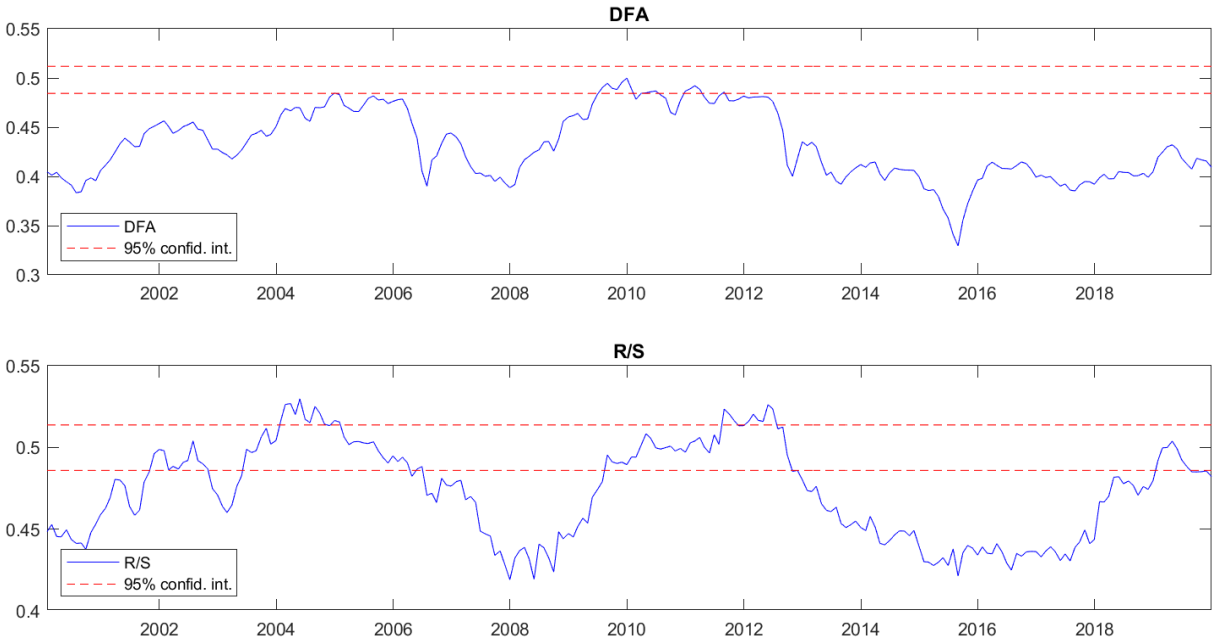


Figure 5: The evolution of the Hurst exponent over time, calculated using both DFA and RS. The red dotted lines represent a 95% empirical asymptotic confidence interval around $Hurst = 0.5$, estimated by Weron (2002).

positively correlated with concurrent market returns. They further conclude that it is not an attribute of simultaneous feedback trading, but rather a causality running from flow to returns within a day. Notably is the weak positive significance for $Fees_{t-1}$; a contra-intuitive relationship wherein flow increase with increasing fees. We suggest this is due to the fee variable acting as a time-proxy as we observe a clear negative trend in both fractional flow (see figure 2) and fees (see figure 3).[25] The same significant impact cannot be found in the dollar flow regressions. This is in contrast with studies done by Sirri & Tufano (1998) and Huang, Wei, & Yan (2007) who find a general negative relationship between fund flows and fees. A clear distinction from their work is that our thesis only analyze index funds flow whereas Sirri & Tufano (1998) and Huang, Wei, & Yan (2007) studied a broader spectrum of funds, which possibly are more sensitive to fees. Moreover, the logged and lagged TNA values are consistently positively significant in the regressions run on dollar flows, and partly negatively significant in the fractional flow regression. This observed difference could stem from the previous mentioned time factor as fractional flow is slightly decreasing over time whereas the dollar flow is steadily increasing. Overall we find no evidence that past year market volatility impacts current flow values.

[25]Including a trend variable in these regressions cause $Fees_{t-1}$ to become a negative non-significance.

Linear regression 1: fractional flow

	Panel A: DFA				*Panel B: RS*			
N = 239	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Intercept	0.0045**	0.02754***	-0.0251	-0.0211	0.0064**	0.0251***	-0.0354	-0.0327
	(2.17)	(3.43)	(-0.79)	(-0.67)	(2.17)	(3.86)	(-1.36)	(-1.26)
$Hurst_t$	-0.0023	-0.0147**	-0.0109	-0.0122	-0.0061	-0.0135**	-0.0113	-0.0126*
	(-0.47)	(-2.12)	(-1.31)	(-1.47)	(-0.97)	(-2.08)	(-1.64)	(-1.85)
$Mkret_t$	-	-0.0124**	-0.0133**	-0.0132**	-	-0.0126**	-0.0131**	-0.0130**
		(-1.99)	(-2.12)	(-2.11)		(-2.01)	(-2.10)	(-2.09)
$Mkret_{t-1}$	-	-	-	0.0037	-	-	-	0.0037
				(0.56)				(0.56)
$Mkret_{t-2}$	-	-	-	0.0024	-	-	-	0.0026
				(0.40)				(0.44)
$Log(TNA_{t-1})$	-	-0.0013***	0.0016	0.0014	-	-0.0011***	0.0023	0.0022
		(-3.15)	(0.95)	(0.83)		(-3.32)	(1.60)	(1.50)
$CumVol_{t-1}$	-	-	.0048	0.0055	-	-	-0.0009	-0.0009
			(0.28)	(0.32)			(-0.06)	(-0.06)
$Fees_{t-1}$	-	-	22.0462*	20.9689	-	-	26.6578**	26.1407**
			(1.66)	(1.58)			(2.23)	(2.19)
Adjusted R^2	-0.38%	8.24%	9.45%	8.99%	-0.13%	7.66%	9.54%	9.09%
F-statistic	0.22	6.24***	4.72***	3.91***	0.95	7.58***	5.38***	4.69***

Table 8: Linear regressions estimated on the dependent variable fractional $FLOW_t^{agg}$. $FLOW_t^{agg}$ is calculated as the sum of all index funds dollar value flow for month t adjusted by the previous month sum of all index funds total net assets (see equation 3.2). $Mkret_t$ is the S&P500 return from period t. $CumVol_t$ are the cumulative standard deviation from $t-11$ up to t using monthly S&P500 returns. The $Log(TNA_t)$, is the logged value of aggregated total net assets for all the index funds for month t. $Fees_t$ are the average fee level (expense ratio) of the index funds for month t. The Hurst exponents in panel A are calculated using the DFA-method and the Hurst exponents in Panel B the RS-method, both with a rolling window size of 4 years (1008 observations). The Hurst exponents are computed on a daily basis and then averaged over each month to obtain monthly estimates. The regression are estimated on a monthly basis and the numbers in parenthesis represent the t-statistic for the estimated coefficient above.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.

Linear regression 2: dollar flow

	Panel A: DFA				*Panel B: RS*			
N = 239	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Intercept	13.7845***	-16.0782***	-31.4213**	-31.2826**	14.9674***	-15.7781***	-37.4288***	-37.1942***
	(10.31)	(-4.83)	(-2.09)	(-2.08)	(7.84)	(-4.64)	(-3.02)	(-2.99)
$Hurst_t$	-26.1927***	-9.1362***	-8.3138**	-8.3271**	-26.4069***	-11.1122 **	-10.3066***	-10.3651***
	(-8.77)	(-3.04)	(-2.30)	(-2.26)	(-6.64)	(-3.11)	(-2.81)	(-2.75)
$Mkret_t$	-	-9.7575***	-10.0412***	-10.1331***	-	-9.5229***	-9.7557***	-9.8287***
		(-3.83)	(-3.94)	(-3.96)		(-3.68)	(-3.78)	(-3.78)
$Mkret_{t-1}$	-	-	-	0.8797	-	-	-	0.9735
				(0.37)				(0.41)
$Mkret_{t-2}$	-	-	-	-.8349	-	-	-	-.5472
				(-0.37)				(-0.24)
$Log(TNA_{t-1})$	-	1.6843***	2.5533***	2.5455***	-	1.7598***	2.9901***	2.9771***
		(9.02)	(3.05)	(3.04)		(9.98)	(4.31)	(4.28)
$CumVol_{t-1}$	-	-	3.4860	3.4207	-	-	-0.3548	-0.4249
			(0.48)	(0.48)			(-0.05)	(-0.06)
$Fees_{t-1}$	-	-	6392.223	6341.537	-	-	9550.218*	9487.402*
			(1.08)	(1.07)			(1.88)	(1.87)
Adjusted R^2	19.73%	43.42%	43.51%	43.08%	13.54%	43.77%	44.18%	43.75%
F-statistic	76.91***	52.91***	31.88***	23.26***	44.04***	53.82***	32.99***	23.66***

Table 9: Linear regressions estimated on the dependent variable dollar $FLOW_t^{agg}$. $FLOW_t^{agg}$ is calculated as the sum of all index funds dollar value flow for month t (see equation 3.3). $Mkret_t$ is the S&500 return for period t. $CumVol_t$ are the cumulative standard deviation from $t-11$ up to t using monthly S&P500 returns. The $Log(TNA_t)$, is the logged value of aggregated total net assets, expressed in $ billions, for all the index funds for month t. $Fees_t$ are the average fee level (expense ratio) of the index funds for month t. The Hurst exponents in panel A are calculated using the DFA-method and the Hurst exponents in Panel B the RS-method, both with a rolling window size of 4 years (1008 observations). The Hurst exponents are computed on a daily basis and then averaged over each month to obtain monthly estimates. The regression are estimated on a monthly basis and the numbers in parenthesis represent the t-statistic for the estimated coefficient above.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.

Linear regression 3: dollar flow, lagged Hurst

	Panel A: DFA				*Panel B: RS*			
N = 239	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Intercept	13.9142***	-15.5915***	-29.1311*	-28.7852*	15.7408***	-14.5046***	-35.3008***	-34.8036***
	(10.07)	(-4.72)	(-1.93)	(-1.91)	(8.70)	(-4.41)	(-2.92)	(-2.88)
$Hurst_{t-1}$	-26.4940***	-9.8331***	-9.0674**	-9.1521**	-28.0499***	-13.0412***	-12.1971***	-12.4051***
	(-8.60)	(-3.25)	(-2.48)	(-2.46)	(-7.44)	(-3.82)	(-3.48)	(-3.46)
$Mkret_t$	-	-9.6418***	-9.9188***	-10.0051***	-	-9.1592***	-9.3951***	-9.4662***
		(-3.80)	(-3.91)	(-3.93)		(-3.68)	(-3.66)	(-3.66)
$Mkret_{t-1}$	-	-	-	1.0370	-	-	-	1.3526
				(0.44)				(0.57)
$Mkret_{t-2}$	-	-	-	-0.7389	-	-	-	-0.523741
				(-0.33)				(-0.24)
$Log(TNA_{t-1})$	-	1.6703***	2.4362***	2.4173***	-	1.7326***	2.9127***	2.8873***
		(9.03)	(2.91)	(2.88)		(9.99)	(4.28)	(4.24)
$CumVol_{t-1}$	-	-	3.7741	3.4207	-	-	-0.5538	-0.6273
			(0.53)	(0.53)			(-0.08)	(-0.10)
$Fees_{t-1}$	-	-	5579.195	5469.133	-	-	9167.343*	9050.544*
			(0.94)	(0.92)			(1.83)	(1.82)
Adjusted R^2	20.55%	43.74%	43.74%	43.31%	15.73%	44.59%	44.92%	44.53%
F-statistic	73.89***	62.67***	38.00***	26.98***	55.40***	64.84***	39.83***	28.30***

Table 10: Linear regressions estimated on the dependent variable $FLOW_t^{agg}$. $FLOW_t^{agg}$ is calculated as the sum of all index funds dollar value flow for month t (see equation 3.3). $Mkret_t$ is the S&500 return for period t. $CumVol_t$ are the cumulative standard deviation from $t-11$ up to t using monthly S&P500 returns. The $Log(TNA_t)$, is the logged value of aggregated total net assets, expressed in $ billions, for all the index funds for month t. $Fees_t$ are the average fee level (expense ratio) of the index funds for month t. The Hurst exponents in panel A are calculated using the DFA-method and the Hurst exponents in Panel B the RS-method, both with a rolling window size of 4 years (1008 observations). The Hurst exponents are computed on a daily basis and then averaged over each month to obtain monthly estimates. The regression are estimated on a monthly basis and the numbers in parenthesis represent the t-statistic for the estimated coefficient above.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.

Although the RS coefficients are on average more significant than the DFA coefficients, we find very similar coefficients for both, suggesting that choice of method does not significantly alter the market efficiency's estimated effect on flow.[26] As such, we find that a lesser degree of market efficiency negatively impacts flow. But how do we further examine this phenomena? This market efficiency measurement is both stationary and continuous; flow and market efficiency are ultimately two completely different variables in terms of structure. Market efficiency should be seen as a current state of the market, a continuous condition of how inter-dependent the returns are. Flow, on the other hand, is an event, a one period happening. This touches upon an important issue; the levels of index fund flow should be affected by whether the market returns show signs of long-term dependency or short-term anti memory. Both states should indicate an inefficient market, but are otherwise completely different in nature.

In the regressions in Table 9 and 10, we observe that the higher the Hurst exponent is, the lower Flow becomes. This is reasonable, as the Hurst exponent generally increases when estimated over periods of high volatility with a justifiable simultaneous liquidation of index fund assets. This reasoning seems strengthened when concurrent market returns are included [(B) and (F)], which reduces the market efficiency coefficient. Nonetheless, the economic interpretation is more difficult. Although, that irregardless of the current state of the market efficiency, it exhibits a negative impact on flow.[27] Important here is the findings of Eom et al. (2008), that markets with higher Hurst exponent tends to be more predictable and exhibit a lower degree of efficiency. We thus argue that our findings, of a negative impact on Flow from a higher Hurst exponent, reinforce this phenomena. The more inefficient the market is (corresponding to a higher Hurst exponent), the lower the index fund Flow is. For clarification, our estimation of the degree of market efficiency for S&P500 is in line with other studies estimation of the degree of market efficiency in developed markets (Bariviera, 2011; Cajueiro & Tabak, 2004; Eom et al., 2008), wherein these developed markets generally obtain a Hurst estimation of around 0.40 and 0.50 (that is, a slight mean reversing process of returns).[28] Previous studies further find that

[26]Regarding the discrepancy between the significance of the coefficients between the DFA and RS approach, wherein RS coefficients overall are more significant, the conclusion should not be that the RS Hurst is a better measurement for market efficiency. Nonetheless, it seems that the RS market efficiency measurement more accurately describes concurrent index fund flows.

[27]Unfortunately, transforming the Hurst exponent in various ways in accordance to the 95% confidence interval, to try and isolate whether significant long-term memory or anti-persistence yield different effects on flow, produce non-significant results. Such non-significance in the current position of the market efficiency in respect to *true* efficiency (i.e. whether the Hurst exponent is below, inside, or above the confidence interval), strengthens our argument of a lesser degree of efficiency above than below the confidence interval.

[28]As of the writing of this paper, we have not found any other studies that show an equity market consistently exhibiting a Hurst exponent less than 0.4. This suggests that, while the theoretical support of the Hurst exponent is $H \in (0, 1)$, the empirical Hurst exponent support in finance applications is different. Maybe such a strong mean-reverting return process indicated by sub-0.30 Hurst exponent is not a plausible security market phenomena. Rather, the slight mean-reversing return process characterized by $H \approx 0.4$ appear standard for the current developed and considered *most efficient* markets.

Standardized coefficients

Regression	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Table 8	-0.0243	-0.1559**	-0.1152	-0.1289	-0.0537	-0.1183**	-0.0987	-0.1101*
Table 9	-0.45***	-0.16***	-0.14**	-0.14**	-0.37***	-0.16**	-0.15***	-0.15***
Table 10	-0.45***	-0.17***	-0.16**	-0.16**	-0.40***	-0.18***	-0.17***	-0.18***

Standard deviations

	RS	DFA	Flow (dollar)	Flow (fractional)
Std	0.0287	0.0348	2.06	33

Table 11: Standardized coefficients of the Hurst exponent in the estimated regressions for each table. The standardized coefficients are obtained by standardizing all variables in each regression to have a mean 0 and standard deviation of 1. These coefficients thus measure a one standard deviations change in the dependent variable with a one standard deviation change in the dependent variable multiplied by the coefficient. For reference, we also present the standard deviation of each measurement.The standard deviation of fractional flow is in basis points and the standard deviation for dollar flow is in billions of dollar.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.

developing markets display a higher Hurst exponent (lesser degree of efficiency). This touches upon *state* vs. *event* problematic, and further purports the notion that, irregardless of where a security market lies on the Hurst support scale, an increase in the Hurst exponent indicates a movement towards inefficiency; and, as of the findings of this paper, a reduction in index fund flow.

To examine the magnitude of the impact of market efficiency on Flow we utilize standardized coefficients (also called beta coefficients). In Table 11 we present the transformed $Hurst$ coefficients into standardized coefficients. We observe that the standardized coefficients for Table 9 and 10 are similar with only small differences between the regressions (except for regressions (A) and (E)). Examining the more significant coefficients against dollar flow indicates that an increase in the Hurst exponent of one standard deviation would result in an approximately decrease in dollar flow by 0.15 to 0.17% multiplied by one standard deviation in dollar flow. With a H^{RS} standard deviation of 0.0287 and dollar flow standard deviation of $2.06 billion, an increase in H^{RS} by one standard deviation roughly translates into an index fund outflow of $309 million. To put into perspective, the maximum one month difference in H^{RS} was an increase by 0.0247 that occurred in September 2009, corresponding to a simultaneous outflow of $997 million, the fifth largest outflow in our sample period.

The significant lagged Hurst exponents provide evidence that lead index fund flow can be predicted from current state of market efficiency. Ceteris paribus, a movement towards market efficiency, corresponding to a reduction in the Hurst exponent, indicates that aggregate index fund flow should be larger (in comparison with current month) in the impending month.

6 Conclusion

We find evidence that the S&P500 market index have exhibited signs a slight mean-reverting process (anti-persistence, or negative memory), close to market efficiency, between 2000 and 2019; characterized by $Hurst^{RS} \approx 0.5$ and a $Hurst^{DFA} \approx 0.4$. These findings are similar to previous studies of time-varying Hurst exponents. We further find a causality direction between market efficiency and index fund flows, where the degree of market efficiency negatively affect the level of flow. We find no indication of causality in the reverse direction. Our findings suggests that equity markets which are characterized by a higher Hurst exponent, a lower degree of efficiency, experience lower levels of aggregate index fund flow. For dollar flow, this relationship is characterized by a one standard deviation change in the market efficiency measurement corresponding to a change in dollar flow of (roughly) 15% of the standard deviation in dollar flow. The same relationship exists for fractional flow, albeit less significant, wherein the effect is (roughly) 12% of the standard deviation in fractional flow. Moreover, we find a similar significant relationship for previous month's Hurst exponent, indicating predictability for impending index fund flow from current month market efficiency. These results generally hold stronger for dollar flow than fractional flow.

Our findings suggests that dollar flow provides greater insight than fractional flow into the dynamics of macro-scale mutual fund flows. This discrepancy is not found in previous mutual fund flow studies, and highlights the size-effect present in a bull-markets, wherein the return dwarfs the dollar flow growth. Nonetheless, our estimated causality expands the possible significant control variables when evaluating mutual fund flows. The estimated market efficiency's effect on index fund flow bridge the theoretical gap between passive investing and degree of efficiency, suggesting that a random walk market (weak form EMH) induce larger flow. These implications are powerful for further studies regarding both fund flows and market efficiency, but also for practitioners trying to model mutual fund flows.

Generally, we encourage further studies examining the same relationship in other security markets in order to determine if our findings is a general index fund flow mechanism, or if it holds specifically for S&P500. As this paper studied broad-market efficiency, a natural continuation is to study the index fund flow and the efficiency of specific index constituents; i.e., the market impact of index fund flows on single company equity. Reasonably, large institutional actors providing index funds need to adjust their holdings to track the index (say once a month), and therein cause abnormal price movements. Such a study would prove useful for traders, highlighting arbitrage opportunities. Furthermore, as has been noted by Tiwari, Albulescu, & Yoon (2017), the stock market seems to exhibit signs of a multi-fractal nature. We therein propose further studies of multi-fractal Hurst exponents (MF-DFA) and its relation to index funds.

7 References

Anis, A.A. & Lloyd, E.H., (1976). The Expected Value of the Adjusted Rescaled Hurst Range of Independent Normal Summands. *Biometrika*, 63(1), pp.111–116.

Bariviera, A.F., (2011). The influence of liquidity on informational efficiency: The case of the Thai Stock Market. *Physica A: Statistical Mechanics and its Applications*, 390(23-24), pp.4426–4432.

Belasco, E., Finke, M. & Nanigian, D., (2012). The impact of passive investing on corporate valuations. *Managerial Finance*, 38(11), pp.1067–1084.

Cajueiro, D.O. & Tabak, B.M., (2004). The Hurst exponent over time: testing the assertion that emerging markets are becoming more efficient. *Physica A: Statistical Mechanics and its Applications*, 336(3-4), pp.521–537.

Cao, C., Chang, E.C. & Wang, Y., (2008). An empirical analysis of the dynamic relationship between mutual fund flow and market return volatility. *Journal of Banking and Finance*, 32(10), pp.2111–2123.

Center for research in security prices (CRSP), (2019). Survivor Bias Free US Mutual Fund Guide, https://wrds-www.wharton.upenn.edu/documents/1303/MFDB˙Guide.pdf?˙ga=2.213683028.1760127731.1585562190-838547392.1581322778

Di Matteo, T., (2007). Multi-scaling in finance. *Quantitative Finance*, 7(1), pp.21–36.

Edelen, R.M. & Warner, J., (2001). Aggregate price effects of institutional trading: a study of mutual fund flow and market returns. *Journal Of Financial Economics*, 59(2), pp.195–220.

Elton, E., Gruber, M. & Busse, J., (2004). Are investors rational? Choices among index funds. *Journal Of Finance*, 59(1), pp.261–288.

Eom, C. et al., (2008). Hurst exponent and prediction based on weak-form efficient market hypothesis of stock markets. *Physica A: Statistical Mechanics and its Applications*, 387(18), pp.4630–4636.

Fama, E.F., (1970). Efficient Capital Markets: A Review Of Theory And Empirical Work. *Journal of Finance*, 25(2), pp.383–417.

Fama, E.F., (1991). Efficient Capital Markets: II. *Journal of Finance*, 46(5), pp.1575–1617.

Grau-Carles, P., (2000). Empirical evidence of long-range correlations in stock returns. *Physica A: Statistical Mechanics and its Applications*, 287(3-4), pp.396–404.

Grossman, S. & Stiglitz, J., (1980). On the impossibility of informationally efficient markets. *American economic review*, 70(3), pp.393–408.

Harvey, C., (1995). Predictable risk and returns in emerging markets. *The Review of Financial Studies*, 8(3), pp.773–816.

Hortaçsu, A. & Syverson, C., (2004). Product Differentiation, Search Costs, and Competition in the Mutual Fund Industry: A Case Study of S&P 500 Index Funds. *The*

Quarterly Journal of Economics, 119(2), pp.403–456.

Hsieh, D.A., (1993). Chaos and Order in the Capital Markets: A New View of Cycles, Prices, and Market Volatility (Book Review). *The Journal of Finance*, 48(5), pp.2041–2044.

Huang, J., Wei, K.D. & Yan, H., (2007). Participation Costs and the Sensitivity of Fund Flows to Past Performance. *Journal of Finance*, 62(3), pp.1273–1311.

Kantelhardt, J.W. et al., (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3-4), pp.441–454.

Kadapakkam, P.-R., Krause, T. & Tse, Y., (2015) Exchange traded funds, size-based portfolios, and market efficiency. *Review of Quantitative Finance and Accounting*, 45(1), pp.89–110.

Kristoufek, L. & Vosvrda, M., (2013). Measuring capital market efficiency: Global and local correlations structure. *Physica A: Statistical Mechanics and its Applications*, 392(1), pp.184–193.

Malkiel, B.G., (2003). Passive Investment Strategies and Efficient Markets. *European Financial Management*, 9(1), pp.1–10.

McCauley, J.L., Bassler, K.E. & Gunaratne, G.H., (2008). Martingales, detrending data, and the efficient market hypothesis. *Physica A: Statistical Mechanics and its Applications*, 387(1), pp.202–216.

Peng, C. et al., (1995). Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 5(1), pp.82–87.

Petajisto, A., (2011). The index premium and its hidden cost for index funds. *Journal of Empirical Finance*, 18(2), pp.271–288.

Pincus, S.M., (1991). Approximate Entropy as a Measure of System Complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 88(6), pp.2297–2301.

Pincus, S.M., (2008). Approximate Entropy as an Irregularity Measure for Financial Data. *Econometric Reviews*, 27(4-6), pp.329–362.

Sapp, T. & Tiwari, A., (2004). Does Stock Return Momentum Explain the "Smart Money" Effect? *Journal of Finance*, 59(6), pp.2605–2622.

Sirri, E.R. & Tufano, P., (1998). Costly Search and Mutual Fund Flows. *Journal of Finance*, 53(5), pp.1589–1622.

Sushko, V. & Turner, G., (2018). The implications of passive investing for securities markets. *BIS quarterly review : international banking and financial market developments*, pp.113–131.

Tiwari, A.K., Albulescu, C.T. & Yoon, S.-M., (2017). A multifractal detrended fluctua-

tion analysis of financial market efficiency: Comparison using Dow Jones sector ETF indices. *Physica A: Statistical Mechanics and its Applications*, 483, pp.182–192.

Umutlu, M., Akdeniz, L. & Altay-Salih, A., (2010). The degree of financial liberalization and aggregated stock-return volatility in emerging markets. *Journal of Banking and Finance*, 34(3), pp.509–521.

Warther, V.A., (1995). Aggregate mutual fund flows and security returns. *Journal of Financial Economics*, 39(2), pp.209–235.

Wharton Research Data Services (WRDS), (2020). How to identify index funds. https://wrds-www.wharton.upenn.edu/pages/support/support-articles/crsp/mutual-fund/how-identify-index-funds/?˙ga=2.191050090.193790233.1585126468-1241545974.1580722954

Weissenteiner, A., (2019). Correlated noise: Why passive investment might improve market efficiency. *Journal of Economic Behavior and Organization*, 158, pp.158–172.

Weron, R., (2002). Estimating long-range dependence: finite sample properties and confidence intervals. *Physica A: Statistical Mechanics and its Applications*, 312(1-2), pp.285–299.

Printed references

Peters, E.E., (1991). *Chaos and order in the capital markets : a new view of cycles, prices, and market volatility*, New York: Wiley.

A Appendix

A.1 Data adjustments

Variable	Step	Procedure	Effect
Return			
	0)	Naked data	62 111
TNA			
	0)	Naked data	67 341
	1)	Fill	67 341 → 70 158
Flow			
	0)	Naked data	69 516
	1)	Inception & Death	69 516 → 70 225
	2)	Winsorize (1 and 99%)	70 255

Table 12: Data adjustments for downloaded mutual fund data. Data for a total of 43 938 funds were obtained from CRSP. These steps outline our data management procedure explained in section 3.1.

A.2 Hurst histogram

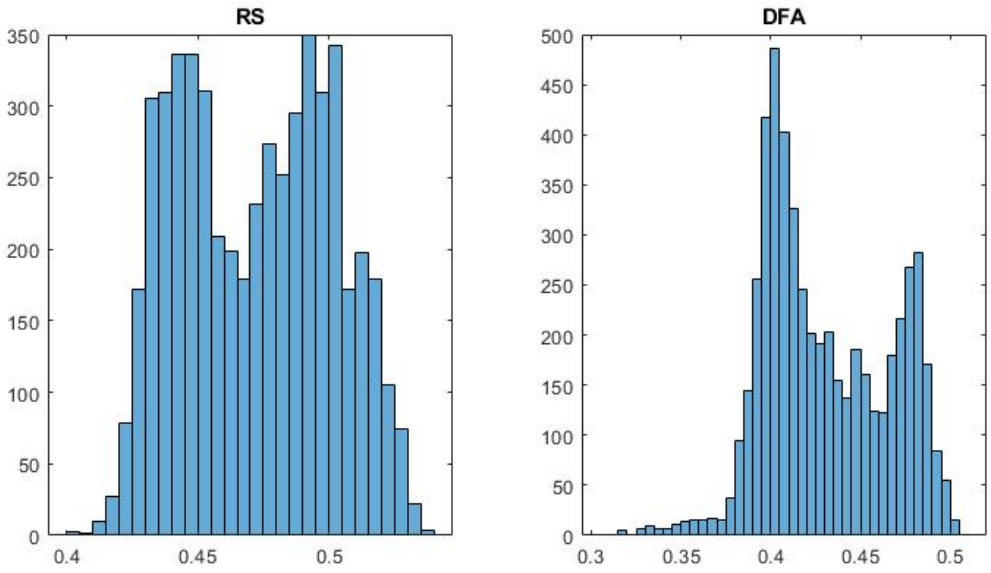


Figure 6: Histogram over daily Hurst estimates computed using a window size of 1008 during the period 2000-2019.

A.3 Regressions - naked data

Linear regression, naked data: fractional flow

	Panel A: DFA				*Panel B: RS*			
N = 239	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Intercept	0.0143*	0.0187	-0.2024**	-0.2333**	0.0216*	0.0252	-0.1720*	-0.1915*
	(1.75)	(0.83)	(-1.98)	(-2.31)	(1.79)	(0.92)	(-1.73)	(-1.97)
$Hurst_t$	-0.0230	-0.0260	-0.0017	0.0092	-0.0365	-0.0399	-0.0298	-0.0196
	(-1.23)	(-1.19)	(-0.07)	(0.39)	(-1.46)	(-1.36)	(-0.96)	(-0.69)
$Mkret_t$	-	0.0149	0.0125	0.0122	-	0.0164	0.0151	0.0147
		(0.66)	(0.56)	(0.53)		(0.73)	(0.69)	(0.66)
$Mkret_{t-1}$	-	-	-	-0.0378	-	-	-	-0.0358
				(-1.57)				(-1.50)
$Mkret_{t-2}$	-	-	-	-0.0181	-	-	-	-0.0156
				(-0.78)				(-0.69)
$Log(TNA_{t-1})$	-	-0.0002	0.0121**	0.0137**	-	-0.0002	0.0120**	0.0120**
		(-0.20)	(2.08)	(2.40)		(-0.13)	(1.97)	(2.18)
$CumVol_{t-1}$	-	-	-0.0496	-0.0554	-	-	-0.0410	-0.0407
			(-0.82)	(-0.93)			(-0.72)	(-0.71)
$Fees_{t-1}$	-	-	95.66**	103.72**	-	-	89.72**	93.10**
			(2.21)	(2.38)			(2.12)	(2.17)
Adjusted R^2	0.09%	-0.46%	0.85%	2.44%	0.46%	-0.03%	1.35%	2.61%
F-statistic	1.50	0.68	1.42	1.57	2.13	1.24	1.68	1.49

Table 13: Linear regressions estimated on the dependent variable fractional $FLOW_t^{agg}$. $FLOW_t^{agg}$ is calculated as the sum of all index funds dollar value flow for month t adjusted by the previous month sum of all index funds total net assets (see equation 3.2). No adjustments are done on the return, TNA, or flow data. $Mkret_t$ is the S&P500 return from period t. $CumVol_t$ are the cumulative standard deviation from $t-11$ up to t using monthly S&P500 returns. The $Log(TNA_t)$, is the logged value of aggregated total net assets for all the index funds for month t. $Fees_t$ are the average fee level (expense ratio) of the index funds for month t. The Hurst exponents in panel A are calculated using the DFA-method and the Hurst exponents in Panel B the RS-method, both with a rolling window size of 4 years (1008 observations). The Hurst exponents are computed on a daily basis and then averaged over each month to obtain monthly estimates. The regression are estimated on a monthly basis and the numbers in parenthesis represent the t-statistic for the estimated coefficient above.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.

Linear regression, naked data: dollar flow

	Panel A: DFA				*Panel B: RS*			
N = 239	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
Intercept	28.2045***	-42.2071**	-173.7096**	-196.0168**	29.2642***	-41.7072	-164.0102*	-178.5299*
	(3.33)	(-2.04)	(-2.19)	(-2.20)	(3.67)	(-1.39)	(-1.95)	(-1.95)
$Hurst_t$	-56.7264***	-17.4416	-3.4558	4.3331	-54.0161***	-20.9468	-14.7685	-7.2979
	(-3.09)	(-1.42)	(-0.26)	(0.29)	(-3.31)	(-0.95)	(-0.64)	(-0.28)
$Mkret_t$	-	2.3508	0.8832	1.1187	-	2.7874	1.9706	2.1526
		(0.16)	(0.06)	(0.07)		(0.19)	(0.13)	(0.14)
$Mkret_{t-1}$	-	-	-	-31.5603	-	-	-	-30.7215
				(-1.23)				(-1.17)
$Mkret_{t-2}$	-	-	-	-8.8723	-	-	-	-7.8688
				(-0.80)				(-0.71)
$Log(TNA_{t-1})$	-	3.9991***	11.3171**	12.5377**	-	4.1395**	11.0443**	11.7807**
		(2.69)	(2.39)	(2.37)		(2.46)	(2.30	(2.27)
$CumVol_{t-1}$	-	-	-26.3973	-30.0667	-	-	-24.4713	-23.7961
			(-0.81)	(-0.91)			(-0.81)	(-0.77)
$Fees_{t-1}$	-	-	56850.13**	62667.51**	-	-	55589.42**	58118.02**
			(2.00)	(2.02)			(1.99)	(1.97)
Adjusted R^2	2.98%	7.30%	7.32%	8.13%	1.68%	7.30%	7.44%	8.15%
F-statistic	9.58***	3.67**	2.27**	2.61**	10.98***	6.52***	4.03***	3.86***

Table 14: Linear regressions estimated on the dependent variable dollar $FLOW_t^{agg}$. $FLOW_t^{agg}$ is calculated as the sum of all index funds dollar value flow for month t (see equation 3.3). No adjustments are done on the return, TNA, or flow data. $Mkret_t$ is the S&500 return for period t. $CumVol_t$ are the cumulative standard deviation from $t-11$ up to t using monthly S&P500 returns. The $Log(TNA_t)$, is the logged value of aggregated total net assets, expressed in $ billions, for all the index funds for month t. $Fees_t$ are the average fee level (expense ratio) of the index funds for month t. The Hurst exponents in panel A are calculated using the DFA-method and the Hurst exponents in Panel B the RS-method, both with a rolling window size of 4 years (1008 observations). The Hurst exponents are computed on a daily basis and then averaged over each month to obtain monthly estimates. The regression are estimated on a monthly basis and the numbers in parenthesis represent the t-statistic for the estimated coefficient above.
***, **, and * indicates significance at 1%, 5%, and 10%, respectively.