Are Dutch mutual funds able to pick stocks? Evidence from a bootstrap approach

Master's thesis Economics (Financial Economics)

Master thesis 2020-2021

Student: Rens Siemen Eggink

Student number: S4821157

Supervisor: Dr J. Qiu (Jianying)

Hand-in date: July 06 2021

Abstract

This thesis applies a bootstrap residual resampling algorithm to assess if the performance of Dutch mutual equity funds is due to skill or luck. Unlike the parametric method, bootstrapping has the advantage of taking the non-normal alpha distribution into account that arises due to idiosyncratic risk-taking of individual mutual funds. This distinction is highly relevant since investors want to know if they can better take a passive or active investing approach. The study finds evidence of both positive and negative stock-picking skills among Dutch mutual funds, albeit luck explains the performance of most funds in the sample.

Keywords: Bootstrap; fund performance evaluation; Dutch mutual funds

Table of Contents

1. Introduction

During the Covid-19 pandemic, financial markets witnessed an enormous influx of inexperienced investors. Most of these investors participated in the stock market for the first time in their lives and might wonder whether they can better delegate their investments to an active stock-picking manager or whether a low-cost exchange-traded index fund (ETF) is the most sensible option. A natural question is then: Is an investment manager actively picking stocks able to compensate for his/her costs? Or is the performance merely a product of luck instead of skill? This question might arise to investors since, if luck instead of skill determines the manager's apparent outperformance, investors in actively managed funds might expect to trail a low-cost passive peer due to the manager's costs, crumbling the compounding effect of returns. In this search for skilled managers, Harvey and Liu (2020) argue that investors can make two types of mistakes. First, investors can delegate their investments to a manager that ex-post happens to be badly skilled. Second, investors can miss out on selecting a skilled manager, which implies incurring opportunity costs and a false negative.

While the academic literature has investigated this question in the United States, the situation in the Netherlands is unclear. Therefore, the thesis will first go through the findings of previous studies in the US and the UK, after which we will arrive at the research problem.

There has been much research into whether actively managed funds possess luck or skill in the U.S. (Kosowski, 2011). Even though there has been researching into active funds in the Netherlands, this has its limitations. Broeders, Van Oord and Rijsbergen (2019) evaluate active money management in the Netherlands. They find that the active money managers that get paid more for outperformance do not perform subsequently better than their peers. However, their research is limited to the spectrum of pension funds and does not cover the difference between (mis) fortune and skill. Scholtens (2005) investigated mutual fund performance. However, this paper mainly researches the differences in investment styles between socially responsible mutual funds and traditional mutual funds.

Otten and Schweitzer (2002) analyse the performance of Dutch and US funds, but they look at CAPM-alpha, which implies they do not cover whether performance is due to skill or luck and instead, only look at outperformance relative to a benchmark. Therefore, even though they find a CAPM alpha of 0.88% for the average fund, this does not necessarily mean that Otten and Schweitzer (2002) can attribute any skill to the outperforming fund managers since past positive alphas can also result from luck.

Therefore, this thesis elaborates on the former literature on mutual fund performance in the Netherlands by adding the skill or luck component. The following paragraph explains how to attribute mutual fund performance to either stock-picking skills or luck.

This study applies the bootstrapping method to a sample of Dutch mutual equity funds. Bootstrapping allows determining whether mutual funds truly possess (bad) skills or whether they have (bad) luck. This research aims to contribute to the mutual fund performance literature and has high practical relevance. Investors need to know whether mutual fund managers are genuinely skilled before making an informed investment decision. The thesis answers the following research question:

Can the performance of actively managed Dutch equity mutual funds be explained by skill or luck during the 1996-2021 period?

2. Literature review

Whether investors are best off taking a passive or active approach has highly practical relevance for investors. For example, Heuer, Merkle and Weber (2016) find that funds with positive past alphas receive capital inflows. However, those same funds tend to be more volatile and underperform less volatile funds over a more extended period. Therefore investors need to know whether strong past-performers are skilled – or just lucky.

Barras, Scaillet and Wemers (2010) find that investors tend to chase past hot funds and argue that investors do not recognise that positive alphas are often simply the product of luck and frequently misjudge luck for skill. Frazzini, Kabiller and Pedersen (2013) find that the even world's best-known stock picker, Warren Buffet, does not possess stock-picking skills since Buffet's alpha is insignificant after adjusting for exposure to systematic risk factors. Since some argue that even Mr Buffet is not a skilled stock picker, it would be interesting to see if high-paid mutual fund managers are.

2.1 Factor models

According to the Arbitrage Price Theory of Ross (1976), factors are a non-diversifiable source of returns, and therefore, those factors can be modelled in a linear function. The central idea is that in equilibrium, investors are only rewarded for taking the systematic risk, and no abnormal returns or arbitrage opportunities exist.

APT is silent on what these systematic factors are, and there have been continuing efforts of searching for the systematic factors. Stirred by the framework of Markowitz's (1959) Modern Portfolio Theory, Sharpe (1964) introduced CAPM (Capital Asset Pricing Model). Modern portfolio theory reports that variance of returns is a fair proxy for a portfolio's risk as long as investors only care about expected returns and return variance and are risk-averse. Tobin (1958) discovered the separation theorem, which states that all investors can have the same optimal portfolio positioned on the efficient frontier, which they can combine with a risk-free asset, as long as Markowitz's model's assumptions are valid. According to Fama (1976), non-diversifiable risk will disappear when such a portfolio is implemented. Moreover, a portfolios beta (β) will measure the degree to which the portfolios moves together with the market. The CAPM has also been used to evaluate mutual fund performance. For example, Jensen (1968) evaluated mutual fund performance in the US using CAPM alphas and finds that the active fund could not outperform the market from 1945 to 1964.

The ability of the CAPM to explain returns of individual stocks was discovered to have its flaws. However, Miller and Scholes (1972) predicted that the explanatory power would increase if one tested the CAPM on portfolios of stocks. Nonetheless, Reinganum (1981) found that market returns explain only a limited part of portfolio returns. Thus, instead of inferring market inefficiency, Reinganum suggested that alternative factors than only the market factor could play a role. Recently, Lai and Stohs (2015) note that the CAPM is incompetent in explaining stock portfolio returns.

Banz (1981) found that portfolios of stocks with a low market capitalisation delivered meaningful premia from 1936 to 1975 on top of the predicted CAPM return. Zaremba (2019) finds that the small cap premium is a worldwide phenomenon. However, the paper also notes that the small cap premium is not necessarily a free lunch since the strategy goes hand in hand with higher transaction costs and less liquidity. It is also noted that smallcap premium was less prominent after the 1980s due to decreased transaction costs. However, Ciliberti, Sérié, Simon, Lempérière and Bouchaud (2019) counterargue that the premium was very much alive during the start of the 21st century. Thus, the argument remains to be resolved.

A further well-known anomaly is the value premium. The value premium means that investors in companies that are ‘’financially distressed’’, as defined by a high BE/ME, i.e. high book value of equity compared to the market capitalisation, get rewarded with a premium over the return the investor would expect to achieve based on the CAPM. Market capitalisation means the market price of the share times all shares outstanding. However, Fama and French (1995) show that the value premium is not a free lunch either, as they demonstrate that higher long-term returns that result from a value tilt can be seen as compensation for risk since these companies are systematically more prone to fall into financial distress. For example, Heaton and Lucas (1997) find that investors in value, i.e. high BE/ME stocks especially suffer during liquidity crises.

Growth (low BE/ME) stocks are also called ‘’glamour stocks’’ since they attract much attention from the financial press. These stocks can be seen as expensive as investors pay a premium for the book value of equity relative to value stocks. On the other hand, value stocks can be seen as cheap, thanks to a relatively high book value of equity relative to the market capitalisation. However, critics argue that investors should only buy value stocks selectively to avoid the value trap, i.e. although the stock might appear cheap, it might be on sale for a good reason; just like a cigar butt, it might only have one puff left (Frazzini, Israel and Moskowitz, 2015).

However, Fama (2011) rejects the idea that mispricing drives the value premium since it assumes that investors do not learn from past mistakes. Nonetheless, Luo and Subrahmanyam (2019) contend that people derive utility from holding growth stocks, i.e. low BE/ME equities, initially pushing stock prices up but subsequent return down. They explain this by referring to growth stocks as glamour stocks, attracting investor attention, which affects the securities’ price. Thus, the argument remains to be settled.

Fama and French (1993) build the three-factor model that added the value (HML) and small cap (SMB) factors as additional non-diversifiable risk factors to CAPM’s market risk. The three factors are essential determinants of portfolio returns and can be seen as ‘’risk’’ factors since they are non-diversifiable by nature. The three-factor model has the advantage of having a higher explanatory power of stock portfolio returns than the CAPM.

The market factor (RM-RF) is computed by deducting the risk-free rate, typically a short-term loan issued by the government, such as a T-bill in the U.S., from the market portfolio's return. SMB returns reflect the returns of a hypothetical strategy in which an investor would simultaneously buy small stocks and short sell large stocks. Simply put, the strategy reflects the returns of small stocks minus large stocks. Finally, the value factor (HML) displays the returns of a strategy in which an investor short sells growth stocks and is long value stocks.

The four-factor model of Carhart (1997) adds the winners minus losers component. This means that stocks that have been past winners tend to continue to be winners. Returns from a WML (Winners minus Losers) strategy are computed by subtracting returns of past winners minus returns of past losers and is also called the momentum factor. Simply put, this strategy displays returns of a strategy for which an investor is short previous losers and long previous winners. Like the previous factors, there is again an academic debate whether the momentum factor compensates for risk or displays investor irrationality. According to Sinha (2016), slow response to news like earnings announcements can explain returns from this strategy.

Both factor models are used for the bootstrap procedure in the fourth chapter to find an appropriate benchmark. The essence of estimating the factor exposures of mutual funds is that factor returns are regressed on a mutual fund’s excess returns. The methodology that constructs the factor portfolios, which yield the factor returns, is elaborated in the third chapter.

Finally, the bootstrapping method allows distinguishing between skill and luck. For each fund, the residuals are collected based on the estimated coefficients. Then, the data are simulated under the null hypothesis that each fund has $\alpha_i = 0$, while bootstrapping on the residuals. Bootstrapping here means that residuals are taken with replacement out of the previously created pool of residuals and added to the estimated returns by the factor model under H_0 that $\alpha,$ i.e. skill, is absent and zero. In other words, this generates a return that is only due to luck, and therefore the resulting distribution of simulations is also called the luck distribution. Next, the factor model is estimated on the simulated returns for each fund. Then, the alphas from the OLS regressions on the simulated returns are sorted from high to low.

These steps are repeated 1000 times, resulting in a luck distribution, i.e. distribution of alphas purely explained by luck. Next, the empirical alpha is compared to the luck distribution. To give an example, if the p value is, for instance, 0.030, with a positive empirical alpha, this fund would have a 3% chance to obtain this performance under luck alone, which is very unlikely. This fund is then appointed skill since it performs significantly better than chance. Section 3.2.1. motivates the use of the bootstrap, and section 3.2.2 describes the technicalities concerning applying the bootstrap in more detail.

2.2 Bootstrapping

Kosowski, Timmermann, Wermers, and White (2006) find that most funds do not exhibit skill using bootstrapped p values. However, a subgroup of funds demonstrates skill, resulting in an alpha higher than the funds' costs. Cremers, Fulkerson and Riley (2019) note that the procedure of Kosowski et al. (2006) allows to ex-ante identify skilled fund managers and that the bootstrapping method is statistically robust relative to Carhart’s (1997) technique. Cremers et al. (2019) follows Kosowski et al. (2006) procedure and finds that mutual funds were skilled enough to cover costs in more recent years.

Cuthbertson, Nitzsche, and O'Sullivan (2008) separate the literature into two parts. They first identify funds that either under or outperform relative to a multifactor benchmark return. See chapter 3.1 for an elaboration on factor models. A factor is a stock’s characteristic explaining returns. In a factor model's equation, alpha (α) represents the excess return while accounting for the factor exposure of a portfolio. Investors who focus on maximising returns will look for high alphas because they offer higher returns than the predicted return of the factor model.

On the other hand, investors will bypass funds that exhibit negative alphas since they do worse than their benchmark. However, Riedl, Arno and Smeets (2017) dispute the assumption that investors in mutual funds aim to maximise returns. Instead, they find evidence that investors care, besides returns, about the fund’s ESG (Environmental, Social and Governance) profile.

The second topic touched upon by Cuthbertson et al. (2008) identifies if excess returns can be enduring over time and whether it is possible to identify those outperforming funds beforehand.

Identifying future star fund managers is crucial for investors who want to maximise returns while delegating their investments.

Studies into US mutual funds find little or no evidence of positive excess returns. However, Carhart (1997) finds evidence for negative excess returns. Studies in the UK have similar findings. See, for example, Blake et al. (1997) or Thomas and Tonks (2001). These papers account for factor exposure to compute the funds' alphas and find that most mutual funds produce negative alphas, and there are only a few or no funds with positive alphas.

Thus, the second point in the mutual fund literature is the ex-ante predictability of mutual fund alphas and whether those alphas are persistent in the future. It is essential to retail and institutional investors if past alphas predict future performance and indicate future excess fund returns. However, past winning mutual funds with stellar performance might not endure their performance since past performance can be a product of good fortune instead of a manager’s talent to pick the winning stocks.

To explore the performance of US mutual funds, Carhart (1997) ranked funds in deciles by observed alphas to see if the alphas stay positive or negative over time. He finds that, while positive past alphas fade away over time, mutual funds that have trailed the market continue to perform poorly. In addition, former papers focused on the US and UK money management industry find that past positive alphas do not persist. At the same time, Hoberg, Kumar and Prabhala (2018) provide evidence that the past performance of mutual funds can predict their future performance for the next twelve months. Carhart (1997) suggest that momentum explains why past one-year outperformance is not persistent. Funds for which we observed positive abnormal returns have been lucky to have many winners in their portfolio by accident.

Grossman and Stiglitz (1980) argue that active funds can only exist if markets do not sufficiently reflect relevant information for stock prices since it would not be sensible to pick stocks if the market were thoroughly efficient. They argue that money management makes equity markets more efficient, allocates capital more efficiently, and adds value to the economy.

To test whether a fund's alpha is due to skill or luck, Timmermann, Wermers and White (2006) use the bootstrapping technique. Their paper argues that bootstrapping has advantages over the method used by Carhart (1997) since Carhart's approach can lead to biased *p* values. More specifically, the alphas of mutual funds are often non-normally distributed, while Carhart's approach does not consider that the actual distribution of alphas is non-parametric. Kosowski et al. (2006) explain the nonnormality of alphas by the heterogenous risk-taking across funds. More specifically, concentrated portfolios of fund managers can explain nonnormality, which means that fund managers have only a small number of stock positions and make sector-specific bets.

As an additional motivation of the bootstrap, Golec (1992) explain the concentration of risk-taking of mutual funds as a principal-agent problem. They point out that the fund manager aims to maximise the capital inflows of new investors, who actively compare fund returns. This comparison stimulates the fund manager to adjust the portfolio and its risk depending on the performance compared to peers. This can lead to kurtosis and thick tails, also when mutual fund alphas are normally distributed.

Horowitz (2019) provides robust evidence that shows that bootstrap resampling of residuals has proper statistical outcomes, even if a regression model is not specified correctly. Moreover, Davison and Hinkley (1997) present residual resampling benefits quantile estimation when the estimated coefficients are not normally distributed. Therefore, this paper refers to resampled residuals as pseudo-residuals. Furthermore, this thesis will use a 5% significance threshold in line with Kosowski et al. (2006) to distinguish between skill or luck.

3. Method

3.1 Factor models

Sharpe (1964) introduced CAPM, in which a portfolio's exposure to market risk determines its expected returns. However, there exist multiple factor models in the literature. For instance, Fama and French’s (1992) three-factor model explains portfolio returns by including the tilt towards either small or large stocks and either value or growth stocks besides the portfolio’s loading on the market factor. Carhart (1997) extended this model with his momentum factor to his four-factor model. Chapter four will discuss which model has the most explanatory power for Dutch mutual funds based on multiple Newey-West (1987) regressions on an equal-weight portfolio of mutual funds and individual mutual funds. The following equation describes the three-factor model:

$$r_{i,t} = \alpha_i + \beta_i RMRF_t + SMB_t + HML_t + \epsilon_{i,t} \qquad 1)$$

SMB is the variable that shows the premium investors receive for investing in small companies' stock, HML shows the premium investors receive for investing in value stocks, and RMRF represents the market premium T_i shows the number of observations in months and α_i refers to the excess returns of the fund. Excess returns refer to the return that the model cannot explain and, therefore, might indicate skill or luck or indicate that the factor model is misspecified, as discussed in paragraph 5.2.3.

Carhart's (1997) four-factor model amplifies the three-factor model by the momentum factor, which implies that past winning stocks continue to win. In contrast, past loser stocks continue to lose. The following equation describes the four-factor model:

$$r_{i,t} = \alpha_i + \beta_i RMRF_t + SMB_t + HML_t + MOM_t + \epsilon_{i,t} \quad 2)$$

3.2 The Cross-Section Bootstrapping Method

This section describes the advantages of bootstrapping compared to traditional methods and provides an in-depth analysis of the bootstrapping procedure.

3.2.1 Motivation for the bootstrap

Kosowski et al. (2006) report that alphas of individual equity mutual funds are non-normally distributed. A non-normal distribution can be problematic if traditional statistical methods are used to analyse the data. For example, Ahad, Yin, Othman and Yaacob (2011) report that conventional statistical analysis of non-normally distributed data likely yields unreliable *p* values.

Bootstrapping tries to solve the issues caused by non-normality. Cuthbertson, Nitzsche and O'Sullivan (2008) adopt this approach because the empirical distributions of alphas of funds are non-normal. The bootstrapping method overcomes this problem by judging if the fund's tails' have good or bad luck or good or bad skill. Another advantage of bootstrapping is that it allows determining skill or luck for individual funds that possess highly non-normal idiosyncratic risk. For example, idiosyncratic risk investing in a mutual fund goes hand in hand with taking on company-specific risk, while investing in a passive index-tracker is usually only accompanied by market risk. In contrast to different approaches, bootstrapping does not assume idiosyncratic risk to have a known parametric, normal distribution. Unlike conventional methods, bootstrapping can also cope with non-normally distributed alphas in the distribution's end of the cross-section alpha distribution of funds (Kosowski et al., 2006).

This section will clarify the terms used in the former paragraph further. A fund's idiosyncratic risk is the specific risk of the fund that arises due to concentrated stock positions, just like a stock's returns entail firm-specific risk. A normal distribution means that values, or in this case, the alphas, would be distributed in a bell-shaped way, with both sides of the distribution symmetrical around the mean of the observed values. A normal distribution implies that values further located from the mean should be observed less often.

The data is considered non normally distributed if the empirical distribution of data significantly deviates from the bell shape. For example, Kosowski et al. (2006) find evidence that mutual fund alphas are non-normal distributed. Non-normally distributed data has multiple consequences regarding which statistical analysis is needed. For example, median and mean become dissimilar because the alpha distribution might not be symmetrical anymore. The distribution might also have fat tails, which means that the distribution has relatively many outliers. An outlier is an observation that lies relatively at the end of the distribution. It is an outlier because the probability of observing such a value is relatively low as expected under the assumption of normality.

Regular statistics, in this case, the parametric method, assume that the observations are approximately normally distributed. Thus, using such methods in a non-normal distribution might lead conclusions, or precisely, *p* values, to be biased. To illustrate, Ahad, Yin, Othman and Yaacob (2011) state that methods based on a normal distribution yield unreliable *p* values if the empiric distribution of alphas is non-normal. Unreliable *p* values are critical since these *p* values determine whether a fund statistically outperforms its benchmark. The bootstrapping method, on the other hand, assumes the distribution to be non-parametric. Thus, Bootstrapping allows the *p* values to be reliable when the empiric distribution of alphas is non-parametric distributed.

To provide a rationale for their bootstrapping approach, Kosowski et al. (2006) provide four arguments why the residuals of alphas of stock-picking mutual funds are subject to non-normality.

The first argument describes that funds usually hold highly concentrated positions in particular stocks or industries. Even though the central limit theorem argues that even stocks whose returns are non-normal distributed reach normality, this is only true if those stock positions are equally weighted. Since the portfolios are highly concentrated and the managers invest only in a limited number of equities, the fund's returns are non-normal due to the portfolio's exposure to firm-specific risks. Their second argument is that the market's returns can be non-parametric as well. Moreover, co-skewness of the market portfolio and individual equities might be possible. Third, individual equities have different magnitudes of time-series autocorrelation in returns. Fourth, managers can change their portfolio's risk when market conditions change. Finally, managers can adjust their portfolio's risk in anticipation of their performance relative to competing funds.

Since these four arguments point clearly towards non-normality, Kosowski et al. (2006) argue that bootstrapping is a more suitable solution than conventional methods in the cross-sectional distribution of mutual fund alphas. To sum up, bootstrapping is better able to cope with non-normality (Hesterberg, 2011).

3.2.2 Applying the bootstrap

Bootstrapping allows to compare the real-world alpha with an alpha distribution simulated from the sample, which is the luck distribution. Thus, bootstrapping allows determining whether past returns are a product of either (mis)fortune or (bad)skill. The bootstrapping approach consists of the following steps.

First, an appropriate factor model is estimated, and residuals are collected. Next, these residuals are resampled with replacement and added to the predicted returns by the estimated factor model, which yields a simulated time-series return based on luck. Next, the simulated returns are regressed to compute an "artificial" alpha. This process is done a thousand times for each fund. Finally, the simulations result in a pool of alphas depending on sampling variation. Lastly, the alphas are sorted from high to low, after which *p* values are generated by comparing the sorted alphas to the empiric alpha. The following paragraphs again explain the process in more detail to gain a deeper understanding.

First, each mutual fund's factor model’s coefficients are estimated with Newey-West (1987) standard errors, which resolve autocorrelation and heteroskedasticity related issues, and the model’s residuals are collected. This specific estimation method is used due to the residuals' statistical properties, elaborated in the results chapter. Residuals are the difference between the empirically observed value and ŷ, or the value predicted by the factor model under the H_0 of no outperformance (Studenmund and

Johnson, 2017). Next, the estimated coefficients are multiplied by the original factor returns. This step results in one time series return of fitted ŷ. The values of ŷ only depend on the factor coefficients of the individual mutual fund since under the H_0 that the constant, or the alpha, is zero. This allows obtaining the time series return predicted by the factor model under the absence of any skill. Then the residuals are resampled with replacement or simulated. This process creates a pseudo time series of resampled residuals, although the residuals do not have a time stamp anymore, unlike the return ŷ predicted by the factor model.

Next, the resampled residuals are added to the predicted factor return ŷ of the second step, while ŷ is maintained in its original chronological order. In other words, the simulated residuals are added to the returns predicted by the estimated coefficients of the first step. This process generates a pseudo-time-series return for each mutual fund. Then the study regresses the best fitting factor model on the pseudo-returns for each fund using Newey-West standard errors. This process results in an alpha distribution based on luck.

Further, the study performs the former steps 1000 times. This simulation process creates a distribution of 1000 simulated alphas solely dependent on sampling variation of the residuals, i.e. luck. The intuition behind the inference of skill versus luck is that the empiric alpha needs to be relatively extreme compared to the simulated alphas. In concrete, this means that the simulated alphas are ordered on value from high to low, and it is concluded that the fund is positively skilled if the empirical alpha lies above the 95th percentile of simulated alphas and that the manager has bad skill if the empirical alpha lies below the fifth percentile of simulated alphas. For example, if the empirical alpha has a higher value than all simulated alphas, it is a significant indication of skill because the empirical alpha is highly unlikely to arise from luck alone. Section 4.2.1. further illustrates the interpretation of bootstrapped *p* values.

3.3 Data

3.3.1 Retrieving the factor returns

The three-factor portfolio returns of the European stock market were derived from the Ken French Data Library. The following section describes the construction of the factor portfolios. The SMB factor is composed by sorting stocks based on size into two portfolios and the HML factor by sorting stocks based on book-to-market equity. Large stocks are the largest 10% in the market capitalisation.

On the other hand, the small stock portfolio includes stocks with a market cap located in the bottom ten per cent. This distinction results in 2x3 portfolios based on market cap and equity on the balance sheet relative to the market cap (HML). SMB returns are computed in the following way:

$$SMB = \frac{1}{3}(SV + SN + SG) - \frac{1}{3}(LV + LN + LG)$$

HML returns are computed in the following way:

$$HML = \frac{1}{2}(SV + LV) - \frac{1}{2}(SG + LG)$$

The portfolio returns are denominated in US dollars and include capital gains as well as dividends. Continuously compounding is not used. The excess market return is constructed by subtracting the market portfolio's returns by the one-month risk-free rate. The momentum factor is derived from the Ken French Data Library as well.

3.3.2 Mutual Fund Inclusion Criteria

The Thomson Reuters Datastream database is used to find the monthly returns of mutual funds. The returns are retrieved from Datastream for the period 1996/01/01 to 2021/31/03.

This thesis applies multiple filters to find funds in Datastream, in line with Cuthbertson et al. (2004). First, the fund must mainly invest in the Netherlands (>50% of exposure). Second, the funds have had to be alive for more than 36 months to account for the survivorship bias. Third, the funds had to be open-ended, which means that the investor can trade his stake in the fund for the price of which the underlying securities, i.e. mostly shares, in this case, are selling in the open market. The Datastream filters selected 19 Dutch funds. After testing for correlation between funds, one fund with the least number of months was dropped, which resulted in a final dataset of 18 domestic funds.

The dataset consists of both dead and alive funds to take the survivorship bias into account, as Rohleder, Scholz and Wilkens (2011) illustrated. Survivorship bias emerges when we assess fund performance that only incorporates surviving funds. Rohleder, Scholz and Wilkens (2011) note that only including surviving funds can overestimate fund performance since the leading causes of fund dying are lousy performance. Thus, only including surviving funds might be problematic. In line with Cuthbertson et al. (2008), this thesis incorporates a fund analysis if it is alive for at least 36. Table 1 shows the 18 funds and their period in the sample. The return includes annual management fees.

Moreover, the fund returns are derived in United States Dollars, in line with the imported factor portfolio returns from the Fama French Data Library. Finally, it is essential to note that no funds started after 2011. Lower demand for funds investing mainly in domestic equities due to a shift towards passive or global investing may explain the lack of fresh funds focusing on domestic equity. However, this is a premature conclusion, and it might be fruitful for future research to dive deeper into this phenomenon. Unfortunately, this study has to deal with a limited number of funds. This limitation might make it hard to find significant results and extrapolate findings into foreign markets. See the discussion chapter for an elaboration.

Table 1: Sample of Dutch mutual funds

Mutual Fund	***First datapoint***	***Months in Dataset***
Kempen Orange Fund NV	1-11-1995	305
Robeco Hollands Bezit	1-11-1995	300
Delta Lloyd Deelnemingen Fonds	1-11-1995	259
Zwitserleven Aandelenfonds	1-4-1996	264
FBTO Aandelenfonds Nederland	1-2-2000	180
BNP Paribas Netherlands	1-4-2000	218
Avero Achmea Nederlands Aandelenfonds	1-11-2000	93
Generali Aandelenfonds	1-3-2002	198
NN Dutch Fund	1-9-2002	223
NN Nederland Fonds	1-12-2002	220
Holland Fund	1-12-2003	126
De Goudse Nederlandse Aandelenfonds	1-1-2005	119
Nederlandse Aandelenfonds	1-6-2006	178
Add Value Fund NV	1-3-2007	169
Achmea Aandelenfonds Euro	1-9-2010	62
Teslin Darlin	1-1-2011	88
Teslin Midlin	1-1-2011	123
Teslin Todlin	1-1-2011	88

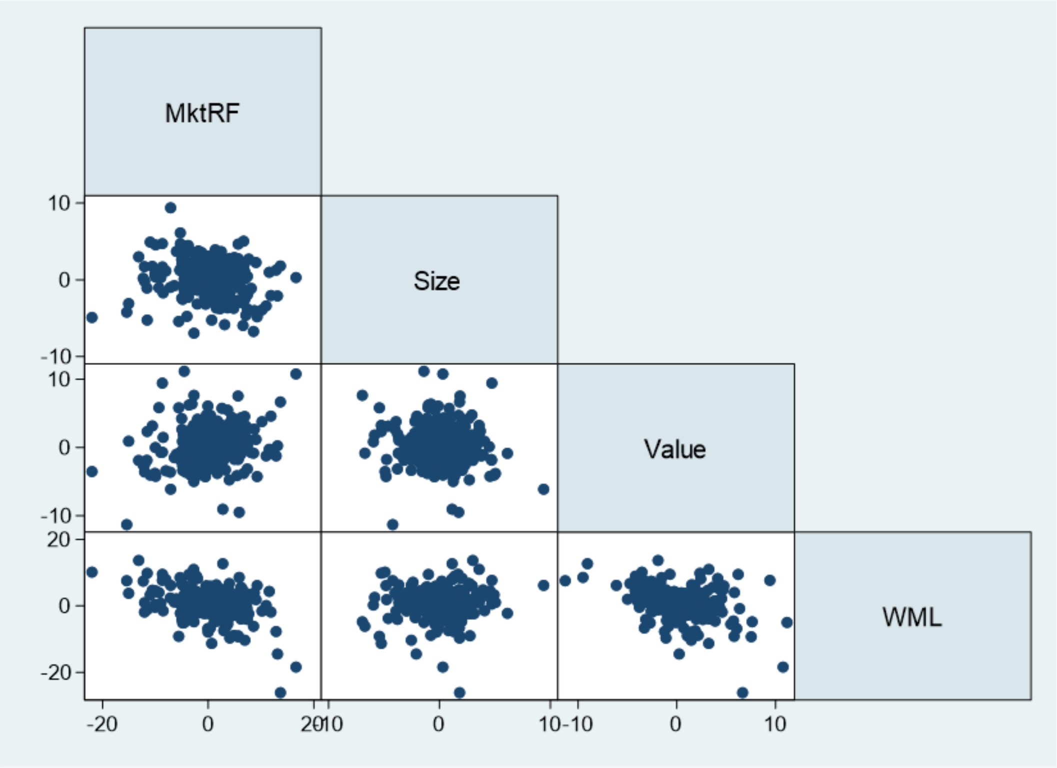
4. Results

Paragraph 4.1 describes the factor model’s monthly returns regressed on the mutual fund returns and subsequently selects which model fits the returns best. This regression allows assessing the best fitting model for each mutual fund. Furthermore, paragraph 4.2 infers based on the bootstrapped *p* values whether the mutual funds in the sample have (bad) stock-picking abilities or (mis) fortune. Finally, the discussion chapter debates and criticises the results and method.

4.1 Best fitting model

To perform the bootstrap, it is necessary to decide which factor model fits the fund returns best. This question is researched by regressing Carhart's (1997) and Fama and French's (1993) models on an equal-weighted portfolio of mutual funds' returns and the returns of each mutual fund. However, first, the four factors are quickly examined since this is relevant for evaluating fund performance. The scatterplot and the cross-correlation matrix show that Carhart's (1997) factors are reasonably evenly distributed.

Figure 1: Scatterplot of the 4-factor correlation matrix.



The scatterplot and cross-correlation matrix help to understand and interpreting tables 4 and 6. For example, the negative correlation in table 2 between HML and MOM might explain why a growth fund in tables 4 and 6 has a negative tilt on the MOM factor. Moreover, such a factor exposure might indicate that a fund is ‘’index-hugging’’ a passive factor tilted peer index or is not deviating from the standard factor exposure belonging to a specific factor tilt. For example, an actively stock-picking mutual fund with a value tilt following a momentum strategy, i.e. positive HML and MOM coefficients might exhibit a specific strategy that one could not have replicated by investing in a passive, one-factor tilted low-cost index fund. Therefore, it also shows how actively managed funds might deviate in their factor strategy from a passive ETF tilted towards only one factor. A passive, one factor tilted strategy automatically goes hand in hand with tilts on multiple other factors that might not necessarily synchronise with an investor’s preferences. For example, an investor that is simultaneously looking for adding value and momentum exposure to the portfolio might dismiss a passive value ETF because, as table 2 indicates, such a strategy tends to be contrarian. This also implies that an actively managed fund might add value here by implementing, for example, a momentum value strategy, which might not have been possible with a passive pure value fund.

Regarding the specifics of the factor correlations, the cross-correlation matrix in table 2 points out that market beta positively loads on the value factor and is contrarian, portfolios with high market risk load negatively on momentum. Moreover, it becomes clear that the value factor tends to be contrarian, loads positively on market risk and tilts towards large cap equities corresponding with Heaton and Lucas (1997). On the other hand, SMB slightly loads upbeat on momentum and has a growth tilt.

Table 2: Cross-correlation among the European factor returns

	Market	SMB	HML	WML
Market	1.000	-.1027	.2415	-.3775
SMB	-.1027	1.000	-.1038	.1029
HML	.2415	-.1038	1.000	-.3566
WML	-.3775	.1029	-.3566	1.000

Moreover, figure 2 gives extra confidence that the bootstrap is the prefered method since it shows that the returns on the factors are non normally distributed. Non normally distributed factor returns might be due to co-skewness of the factor portfolios returns. Kosowski et al. (2006) indicate that non-normality of factor returns can cause non-normality regarding the fund alphas. Therefore, these analyses give a first empiric motivation for performing the bootstrap. Especially the momentum factor has fat left tails with extreme outliers, while most of the returns of the value, market and Small minus Big factor concentrate in the centre of the distribution.

Figure 2: Histograms of Carhart's (1997) European factors.

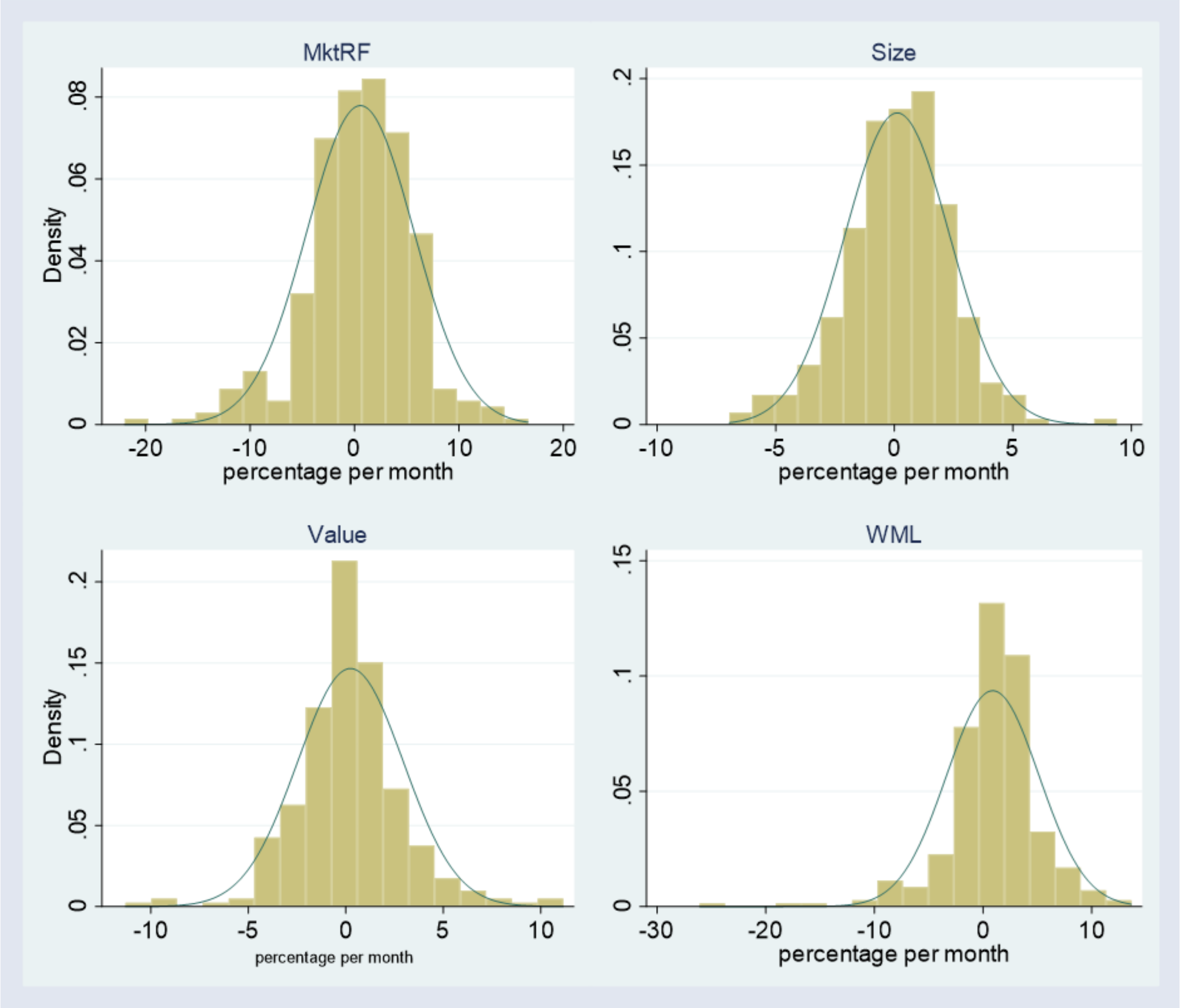


Table 3: Four-Factor loadings of the Equal-Weight portfolio of mutual funds.

	Coefficient	Standard error	P> |t|
MktRF	.7968	.0383	0.000
SMB	-.0714	.08209	0.385
HML	-.1140	.0715	0.112
MOM	-.1189	.0478	0.013
Constant	.2425	0.038	0.202
Adj. R^2	0.6469		

Table 2 displays the outcomes from Carhart's (1997) four-factor model run on the generated equal-weighted portfolio of the mutual funds in our sample. The Newey-West estimation (1987) conducted the regression.

Table 4 displays the regression of four factors on the average portfolio of mutual funds, and this regression delivers an adjusted R^2 of 64.49%. This R^2 value implies that the four-factor model explains most of the returns, but far from all. The constant term, or alpha, equals 0.2425, i.e. 0.2425% per month, but is not significant on a 5% threshold. The MktRF coefficient, i.e. the market factor, is highly significant. Besides the market factor, only the *MOM* coefficient has a significant t-statistic. The SMB factor is not significant, which is in line with the findings of Zaremba (2019) that the importance of

SMB in explaining returns has decreased over time due to decreased transaction costs and might therefore be a symptom of a diminishing liquidity risk premium. However, idiosyncratic differences between individual funds could also explain the absence of significant t-statistics regarding the SMB and HML factors of the equal-weight portfolio. In other words, the differences in factor tilt between the mutual funds might cancel each other out. Table 4 researches this ambiguity.

The lack of significant t-statistics for the SMB and HML coefficients for the equal-weighted portfolio displayed in table 3 motivates to regress Carhart's (1997) model and Fama and French's (1993) model on each mutual fund. These fund-specific regressions allow assessing whether the insignificant t-statistics are due to, for example, growth and value strategies cancelling out the average factor tilts.

Therefore, Table 4 displays the four-factor model regressed on each mutual fund. In this way, it is allowed to assess the factor tilts for each fund. First, it is essential to check whether the data incurred the problem of autocorrelation. Autocorrelation means that the error terms affect each other over time, i.e. the residuals depend on each other. Residual dependency means that the residuals have the same direction over time. For example, if residuals stay positive for a long time, the observed values are higher than the regression model predicted (Studenmund and Johnson, 2017). A positive test means that the residuals affect each other over time which is a symptom of autocorrelation. Since the test indicates a serial correlation, it is helpful to use standard errors of Newey-West (Smith and McAleer, 1994).

Table 4 displays four-factor exposures at the fund specific level. The positive coefficients on RM-RF, SMB, HML and MOM mean that an individual fund's portfolio is overweight on these factors relative to the market portfolio. A negative coefficient for momentum, for instance, implies that the fund follows a contrarian strategy and holds relatively much last years losers in its portfolio. Achmea Euro Aandelenfonds stands out with a relatively high adjusted R^2. The market factor is highly significant for each mutual fund and always displays a positive sign. The directions of the SMB HML and MOM coefficients are blended for the individual funds, which confirms the suspicion that individual fund factor tilts towards, for example, value or growth stocks, or small or large stocks, cancel each other out in the equal-weight portfolio in table 3. Therefore, table 4 explains the lack of HML and SMB significance for the average returns of fund's portfolios.

Regarding SMB, eight funds show a positive sign, of which five have significant t-statistics. Conversely, eleven funds show a negative sign, of which only four are significant. The result that funds have significant size tilts is somewhat surprising as Zaremba (2019) finds that the SMB factor lost influence after the 1980s. However, significant size tilts might also make sense in the light that funds specialise in equities of a specific size.

Regarding the MOM coefficient in table 4, seven funds yield a significant *t*-statistic. At the same time, only two hold a positive sign, hinting that the average fund tends to follow a contrarian strategy, which is corroborated with the significant negative coefficient for MOM in table 3. Conversely, HML exhibits positive coefficients regarding only six funds, which points to a value strategy. In comparison, HML delivers negative coefficients for 12 funds, indicating that most funds specialise in growth stocks. The finding that funds tend to specialise in small growth equities might also make sense in the light of Cremers et al. (2019), who find that especially small cap growth funds can skillfully pick stocks. However, this reasoning is weakened since HML only has two significant t-statistics, albeit those two are significant for funds with a growth tilt. Moreover, since only two funds hold a significant HML load, it is debatable whether the value factor adds extra explanatory power to the four-factor model concerning this dataset. Moreover, it also indicates that most fund managers pick stocks from both sides of the value spectrum.

Next, we check the distribution of alphas since the non-normal distribution is an essential justification for the bootstrapping procedure. Eleven funds have positive alphas; however, none yield significant t-statistics. Eight funds generate negative alphas, of which none are significant either. It is essential to note that Stata constructs these alphas based on the parametric assumption of normality. Even though it might seem impressive that the majority of funds produced positive alphas, it becomes less striking in the light of the findings of Rohleder, Scholz and Wilkens (2011) since they find that survivorship bias might explain the positive alphas because mostly the funds that yielded positive alphas in the past tend to survive, as they receive fund inflows.

Table 4: Four-factor loadings of the mutual funds in the sample___

	RM-RF		***SMB***		***HML***		***MOM***		***Constant***		
	β-Coefficient	t-stat	β-Coefficient	t-stat	β-Coefficient	t-stat	β-Coefficient	t-stat	Alpha %	t-stat	Adjusted R^2
Kempen Orange	.7832	18.29	.4915	5.36	.0159	0.20	-.0559	-1.05	.2114	1.00	0.5830
Delta Lloyd Deelnemingen Fonds	.6709	9.36	.3048	2.01	.0222	0.16	-.2582	-2.98	-.0303	-0.08	0.3362
Zwitserleven Aandelenfonds	.8207	17.13	-.4203	-4.13	-.1582	-1.72	-.0489	-0.84	.0111	0.05	0.5926
FBTO Aandelenfonds Nederland	.7589	14.39	-.1966	-1.63	-.2613	-2.62	-.1524	-2.37	-.0563	-0.20	0.6214
BNP Paribas	.7339	14.85	-.1617	-1.40	-.1466	-1.54	-.1294	-2.12	-.2802	-1.14	0.5892
Achmea Nederlands Aandelenfonds	.9016	8.62	-.1000	-0.51	.1525	0.81	-.2667	-2.37	-.7140	-1.42	0.5950
Generali	.7260	13.08	-.3944	-3.03	-.2809	-2.24	-.2109	-3.04	.1852	0.73	0.6024
NN Dutch Fund	.8165	12.04	-.2929	-1.85	-.2219	-1.48	.1542	-1.71	-.0531	-0.17	0.5000
NN Nederland Fonds	.7326	14.37	-.1190	-0.97	-.1490	-1.32	-.1012	-1.48	.1121	0.48	0.5768
Holland Fund	.6673	9.29	.3344	1.92	.2017	1.02	-.0131	-0.13	-.6184	-1.77	0.5527
De Goudse	.7264	9.30	.2039	1.07	-.0436	-0.21	-.0004	-0.00	-.0163	-0.04	0.5218
Nederlandse Aandelenfonds	.8118	15.23	-.1097	-0.84	-.1884	-1.52	-.0892	-1.12	.0952	0.38	0.6478
Add Value	.6981	11.03	.6014	3.85	-.0591	-0.40	-.1916	-2.02	.5329	1.74	0.5745
Achmea Aandelenfonds Euro	.6492	9.63	-.5533	-3.31	.1303	0.82	-.0247	-0.23	.5288	1.67	0.7430
Teslin Capital Darlin	.7239	7.59	.4879	2.14	-.1971	-0.98	.2478	-1.59	.6558	1.62	0.4676
Teslin Capital Midlin	.6696	9.97	.5100	3.19	.1005	0.74	-.1243	-1.14	.2880	0.98	0.5960
Teslin Capital Todlin	.6974	9.28	.2318	1.29	-.2478	-1.57	-.0800	-0.65	.3336	1.00	0.5288
Robeco Hollands Bezit	.8142	18.95	-.2972	-3.23	-.0958	-1.19	-.0957	-1.77	-.0399	-0.19	0.6083

* The coloms describe each fund's factor coefficients and adherent *t* statistics. A *t* statistic above 1.96 or lower than -1.96 indicates a significant factor tilt. The rows refer to the individual mutual funds. The adjusted R^2 refers to the explanatory power of the four-factor model for each fund. The regression is conducted via the standard errors of Newey-West (1987) to resolve heteroscedasticity and serial correlation of the residuals diagnosed in Stata. It is essential to note that alphas derived by the parametric method should not be interpreted as an indication of skill or luck since it assumes a normal distribution of alphas that Stata rejected. Alphas represent monthly percentages.

The thesis performs a Newey-West estimation on the three and four-factor models to assess which factor model fits the funds best. Table 5 shows the three-factor coefficients of the equal-weight portfolio, which is the hypothetical portfolio that tracks the average mutual fund's return. The three-factor model has slightly less explanatory power with an adjusted R^2 of 64.08% than Carhart's (1997) four-factor model's adjusted R^2 of 64.49%, albeit differences are not striking. Moreover, the SIC score has the lowest value for the four-factor model since the SIC score for the three-factor model is 1568.932 versus 1568.413 for the four-factor model. This distinction in the SIC score again hints that the four-factor model fits the data slightly better. To sum up, the thesis uses the four-factor model for each mutual fund as a benchmark in the bootstrap since the four-factor model's adjusted R^2 is highest for the majority of 13 funds, and the *t* statistics for momentum are significant for six funds.

Table 5: Three-factor loadings of the equal-weight portfolio of mutual funds

	β-Coefficient	Standard error	P> |t|
$MktRF$	.8271	.0366	0.000
SMB	-.0811	.0827	0.327
HML	-.0622	.0690	0.368
Alpha (% p.m.)	.1085	.1832	0.554
Adj. R^2	0.6408		

* Newey-West estimation on the cross-section, representing the average monthly returns of the mutual funds. $p < .05$ indicates significance. Note that the estimation method is based on the parametric assumption of normality and p values for the estimated alpha might therefore be unreliable as an indication for skill or luck. Alpha represent monthly percentages.

Table 6: Three-factor loadings of every mutual fund in the sample

	$RM - RF$		SMB		HML		*Constant*		
	β-Coefficient	t-stat	β-Coefficient	t-stat	β-Coefficient	t-stat	α (% p.m.)	t-stat	Adj. R^2
Kempen Orange Fund NV	.7974	19.64	.4869	5.32	.0403	0.53	.1483	0.73	0.5829
Delta Lloyd Deelnemingen Fonds	.7355	10.60	.2877	1.87	.1219	0.90	-.3399	-0.97	0.3154
Zwitserleven Aandelenfonds	.8328	18.22	-.4236	-4.17	-.1395	-1.57	-.0476	-0.21	0.5931
FBTO Aandelenfonds Nederland	.8091	16.54	-.2449	-2.03	-.2220	-2.23	.2109	-0.77	0.6115
BNP Paribas Netherlands	.7769	17.09	-.1992	-1.73	-.1147	-1.21	-.4045	-1.68	0.5825
Achmea Nederlands Aandelenfonds	1.017	10.73	-.2259	-1.16	.1176	0.61	-1.0457	-2.12	0.5739
Generali Aandelenfonds	.7874	14.92	-.4330	-3.27	-.2155	-1.71	-.0393	-0.16	0.5855
NN Dutch Fund	.8552	13.31	-.3198	-2.02	-.1371	-0.96	-.1864	-0.61	0.4956
NN Nederland Fonds	.7549	15.45	-.1304	-1.06	-.0903	-0.85	.0284	0.12	0.5745
Holland Fund	.6688	9.48	.3346	1.93	.2112	1.15	-.6322	-1.90	0.5563
De Goudse Nederlandse Aandelenfonds	.7265	9.49	.2039	1.08	-.0433	-0.22	-.0167	-0.05	0.5260
Nederlandse Aandelenfonds	.8276	16.08	-.1052	-0.80	-.1242	-1.13	.0384	0.15	0.6473
Add Value Fund	.7334	11.94	.6165	3.91	.07957	0.60	.4130	1.36	0.5666
Achmea Aandelenfonds Euro	.6504	9.76	-.5534	-3.33	.1461	1.03	.5057	1.71	0.7472
Teslin Capital Darlin	.7463	7.84	.4975	2.16	-.0630	-0.34	.4382	1.14	0.4580
Teslin Capital Midlin	.6876	10.52	.5051	3.16	.1816	1.56	.2021	0.71	0.5950
Teslin Capital Todlin	.7046	9.52	.2349	1.31	-.2045	-1.43	.2634	0.88	0.5320
Robeco Hollands Bezit	.8371	20.35	-.3060	-3.31	-.0595	-0.76	-.1500	-0.74	0.6055

* The coloms describe the three-factor coefficients and adherent t-statistics of each fund, while the rows refer to the individual mutual funds. The adjusted R^2 refers to the explanatory power of the model for each fund. The regression is conducted via the standard errors of Newey-West (1987) , accounting for heteroscedasticity and serial correlation among residuals. Alphas are based on the parametric assumption and do not indicate skill or luck.

Similar to the Carhart (1997) four-factor model, the regressions modify standard errors concerning autocorrelation and heteroscedasticity via a Newey-West (1987) estimation. Moreover, we check if the alphas are normally distributed. Same as the four-factor model, the alphas of eight of the funds in the sample are non-normal. On top of the non-normality of factor returns displayed in figure 2, this evidence gives extra confidence to proceed with the bootstrap method.

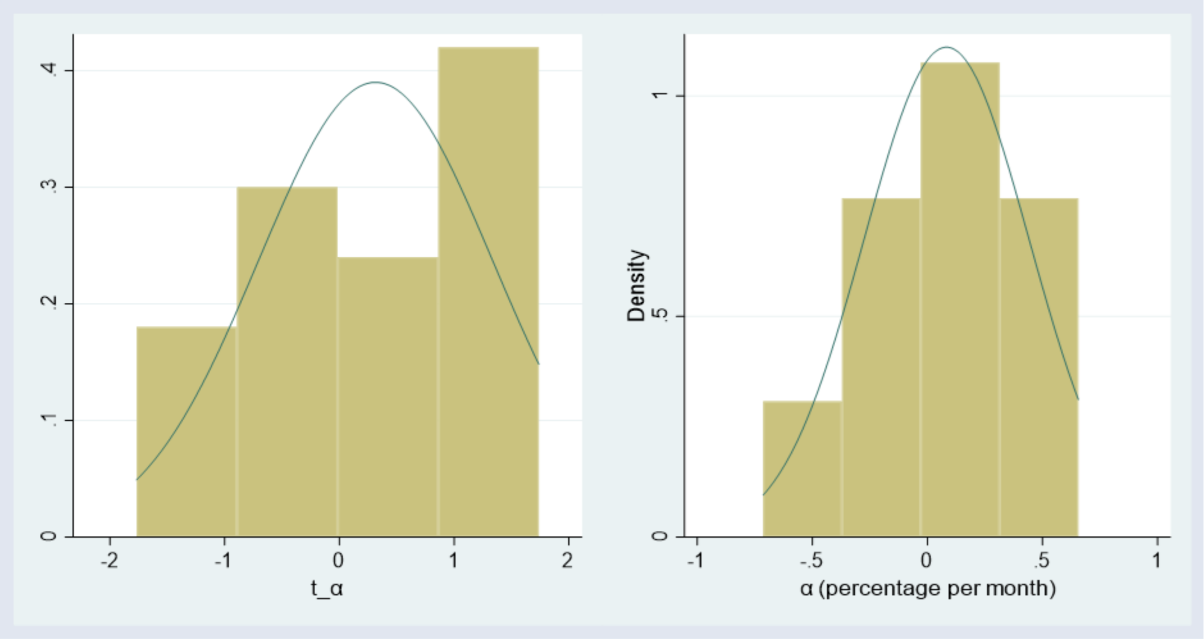
The equal-weighted mutual funds' portfolio yields a slightly positive alpha, but the t-statistic is not significant. However, the Mkt-RF coefficient, or market beta, is positive and highly significant and thus in line with the four-factor regression on the equal-weight portfolio. Similar to the Carhart four-factor model that table 3 displays, both the HML and SMB coefficients do not score significantly for the equal-weight portfolio. However, individual fund's factor might explain this by tilts cancelling each other out since the study finds significant t-statistics on the individual fund level, especially for the SMB factor. Most mutual funds exhibit a negative loading on the SMB factor, whereas most mutual funds load positively on the HML factor, which means their portfolio overweights value stocks.

Furthermore, only one mutual fund has a significant t-statistic on the HML factor, whereas ten funds yield significant t-statistics on the SMB factor. Thus, the lack of significance again questions the relevance of the HML factor for the sample. The blended signs of the SMB and HML coefficients corroborate the viewpoint that factor tilts of the individual funds might cancel each other out in the equal-weight portfolio. Table 5 displays the three-factor tilts of each mutual fund in detail.

The three-factor model's adjusted R^2 is quite strong for most mutual funds. However, Delta Lloyd Deelnemingen Fonds is an outlier with an adjusted R^2 of only 31.54%. The RM-RF coefficient always holds a positive sign.

From the analysis of both factor models, it is concluded that the four-factor model best fits both the individual funds and the equal-weight portfolio of mutual funds. The momentum factor slightly improves the adjusted R^2 of the equal-weight portfolio. Moreover, the momentum factor yields significant t-values, all with negative momentum tilts, for seven mutual funds and a significant negative tilt on the equal-weight portfolio. This result contradicts Sinha (2016) since the paper notes that momentum strategies tend to obtain higher risk-adjusted returns than contrarian strategies.

Figure 3: Histograms of the cross-section of parametric alphas and alpha t-statistics.



The left side histogram displays the distribution of t_α, the *t* statistics of monthly alphas.

The parametric method assumes the alphas to be normally distributed. However, figure 3 and multiple tests discussed in chapter 3 provide evidence that this is not the case. Thus, the study proceeds with the bootstrap resampling procedure.

4.2 Bootstrap of the entire period

The simulations resulted in 1000 alphas for every mutual fund in our sample, which comes down to a cross-section of 18,000 simulated alphas in total. Next, the thesis compared the empirical alpha to the ranked distribution from high to low of the cross-section of four-factor alphas. Additionally, the analysis results in a bootstrapped p-values based on sorted t_α's, allowing to make further inferences on whether a manager is skilled or lucky. The inference focuses on t_α because it has superior statistical properties (Brown, Goetzman, Ibbotson, Roger, Ross and Stephen, 1992). Section 4.2.1 thoroughly motivates the primary use of t_α to distinguish skill from luck. Performing both bootstrapping methods allows future research to gain a deeper understanding of both methods and gives this study a higher degree of confidence on whether a fund's performance is due to skill or fortune. Table 7 describes the bootstrap outcomes regarding the entire sample period in detail.

Table 7: Bootstrap simulation results for January 1996 to March 2021 period

Panel A: Results based on ordering α.

	Alphas > 0									**Alphas < 0**							
	Highest	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	Lowest
Panel A1*:* Parametric (regular) method, which implies an OLS estimation based on α.																		
Empiric α	.656	.533	.529	.334	.288	.211	.185	.112	.095	.011	-.016	-.030	-.040	-.053	-.056	-.280	-.619	-.714
p value	.109	.084	.100	.298	.329	.319	.469	.634	.707	.963	.966	.933	.851	.866	.841	.255	.079	.158
Panel A2: Bootstrap method based on α																		
p value	.042	.041	.073	.131	.155	.155	.263	.313	.350	.497	.506	.424	.428	.465	.401	.140	.039	.094

Panel B: Results based on ordering t_α

					Alpha > 0										**Alphas <0**			
Panel B1: Parametric (regular) method, which implies an OLS estimation based on α.																		
t_α (t-statistic of α)	1.621	1.738	1.674	1.048	.979	.999	.726	.477	.375	.047	-.043	-.084	-.188	-.169	-.202	-1.140	-1.769	-1.424
p value	.109	.084	.100	.298	.329	.319	.469	.634	.707	.963	.966	.933	.851	.866	.841	.255	.079	.158
Panel B2: Bootstrap method based on t_α																		
p value	.073	.024	.089	.148	.096	.181	.244	.293	.320	.491	.498	.430	.414	.456	.383	.115	.028	.043

Alphas represent monthly percentages. All panels are ordered from left to right from high parametric (standard method) four-factor alphas to parametric low four-factor alphas. Panel A1's first row displays the four-factor alphas. The second row displays the parametric *p* value, i.e. the *p* value derived from the Carhart (1997) regression. Panel A2 displays the bootstrapped *p* value, computed by comparing the empiric four-factor alpha to the simulated alpha distribution, or luck distribution, that was made by adding randomly resampled residuals with replacement to the return predicted by the four-factor model under H_0 that α, i.e. the constant, equals zero.

Panel B is also ordered from the highest empiric alpha to the lowest empiric alpha. This means that each column refers to the same fund, regardless of in which panel the data is situated. The first row in B1 refers to the t-statistic of the standard, parametric method obtained by OLS estimation with Newey-West adjusted standard errors. However, unlike panel A2, the bootstrapped *p* value in panel B2 is derived by comparing the empiric t_α to the ordered cross-section of simulated alphas *t* statistics. t_α based bootstrapping has superior statistical properties for the sample, which section 4.2.1 explains in more detail. Note that the second row of Panel B1 is 100% identical to the second row of panel A1. However, panel B1 allows conveniently comparing bootstrapped *p* values based on a simulated *t* statistic distribution to empiric, parametric *p* values.

Table 7 displays the bootstrap results based on the simulation process over the entire sample period, i.e. January 01, 1996, to March 01 2021. Every column across table 7 refers to the same fund. The first 11 funds produce a positive alpha, costs included, using four-factor returns as a personal benchmark.

Panel A1 shows the results from the parametric method, which were previously discussed in table 4. In addition, the table includes parametric results to compare the parametric method's p value to the bootstrap p value. Panel A2 displays the bootstrapped p values for each fund and concludes on skill versus luck by ordering simulated alphas derived via the residual resampling bootstrap method. The distribution of simulated alphas that result from luck is subsequently compared to each funds' empirical alpha.

Panel B2 displays the bootstrap outcomes derived by comparing a fund's empiric t_{α} to an ordered distribution, from high to low, of simulated *t* alphas. The simulated t_{α} distribution represents the results of a Newey-West estimation on a luck distribution simulated under the H_0 of no skill, or the absence of alpha. The bootstrapped p value represents a comparison of the resulting cross-sectional luck distribution to each mutual fund's alpha's empiric *t* statistic. Thus, panel B2 compares the empirical t_{α} or t-statistic of alpha, to the simulated t_{α} distribution of the cross-section.

4.2.1 Interpretation of bootstrapped p values

This paragraph dives further into the interpretation of bootstrapped p values. The paragraph first discusses the interpretation of panel A2 of Table 7, which displays bootstrapped *p* values based on α. Then, the paragraph discusses the interpretation of panel B2 of table 7, which displays bootstrapped *p* values based on t_{α}.

For example, a bootstrapped p value in panel A2 of Table 7, based on ordering α, of 0.041 for a (negative) positive empiric alpha fund provides evidence of (bad) skill because a *p* value of 0.041 means that the Carhart (1997) empiric alpha is (lower) higher than all but 4.1% of the simulations in the corresponding cross-section of simulated alphas purely based on luck, and is, therefore, evidence of (bad) skill since the fund's empiric alpha was (lower) higher than 95.9% of simulated alphas. This *p* value implies that the fund performed significantly (worse) better than luck, and therefore the performance is due to (bad) skill.

On the other hand, panel B2 of table 7 bases p values on ordering t_{α} or the t statistic of alpha. A bootstrapped p value, for example, .148, implies that 14.8% of the simulated alphas are higher (lower) in the case of a positive (negative) empiric t_{α} than the individual fund's parametric four-factor, empiric, t_{α} and thus indicates that the performance is due to (mis)fortune. Performance is due to luck because, for example, in the case of a positive empiric alpha, it is still somewhat likely, 14.8%, that a fund manager would have exhibited a higher alpha purely due to luck or sampling variation.

The parametric p value reflects the p value of the alpha from Carhart's (1997) four-factor regression of Table 4. Until recently, academics mainly used the parametric p value to judge a fund manager's performance. However, as paragraph 3.2.1. discusses, the parametric method assumes mutual fund alphas to have a normal distribution that does not correspond with the data.

There are noticeable variations in bootstrapped p values when applying the t_{α} compared to α. The advantage of ordering via t_{α} is that it accounts for heterogenous risk-taking and concentrated portfolio bets, giving it better statistical properties than α, giving more confidence that the performance is due to skill instead of luck. Technically, t_{α} scales alpha by the standard error and thus takes the consistency of alpha into account. Using t_{α} is especially important in this sample because non-normality arises due to idiosyncratic risk-taking among funds. Nonetheless, inferring skill or luck based on parametric t_{α} is not sufficient since the histogram of the OLS alphas in figure 3 indicates that the t_{α} is non-normally distributed. Even if fund performance is assed by t_{α}, the bootstrap continues to be essential to identify significant skills, mainly since mutual funds alphas are non-normal due to idiosyncratic risk-taking. Moreover, because the t_{α} takes the standard error of alpha into account, a simulation based on resampled t-statistics of alpha might have preferable and more precise statistical properties than simulating based on α. Also, t_{α} accounts for short-lived funds taking idiosyncratic risks associated with concentrated portfolios by scaling α by the standard error. To sum up, t_{α} is prefered to α due to concentrated portfolio bets and the presence of short-lived funds in the sample.

According to Horowitz (2003), ordering funds by t_{α} leads funds to be more often appointed (bad) skill than (bad) luck because the tails of the simulated distribution have thicker tails than under the parametric OLS assumptions because of the bootstrap residual resampling procedure. However, the result displayed in table 7 do not corroborate these findings since, ordering by α appoints two funds with positive skill and one fund with bad skill, while ordering by t_α leads to only one skilled fund and two poorly skilled funds.

For example, Teslin Midlin, i.e. the 5th fund, is assigned positive skills from the t_α bootstrap, but the α bootstrap indicates that the fund was lucky. This difference might be due to relatively high standard errors of alpha which leads to a relatively low *t* statistic of alpha, or more concretely, which might be due to a lack of consistency of performance.

Moreover, in the example of Add Value Fund N.V., the difference between the t_α bootstrapped p value of .024 is strikingly different from the insignificant p value of .084 indicated by the parametric method. This difference in p value means that a standard parametric OLS estimation assigns the performance of the fund to good fortune, while the more realistic t_α bootstrap indicates that Add Value Fund N.V. would only have a 2.4% chance of obtaining this alpha if pure luck or sampling variation determined performance. Therefore, the t_α bootstrap attributes the fund's performance to actual skill since it is improbable that such a t_α would arise from luck alone. This result corresponds to Cuthbertson et al. (2008) findings since they find evidence of positive skill.

It is essential to point out that, even though this thesis applies a 5% significance threshold in line with Kosowski et al. (2006), the p value threshold to distinguish skill from luck might depend on personal preferences for investors. For example, Riedl et al. (2017) find that investors care, besides returns, about synchronisation with ESG (Environmental, Social and Governance) goals and might use different thresholds for skill since it might not be their primary goal.

For the eleven funds that achieved a positive parametric alpha, the bootstrapped p values in panel A2 based on ordering simulated alphas and comparing this cross-section to the empiric alpha indicate the two funds' alphas with the highest two empiric parametric α was due to skill. This implies that their empiric alpha is higher than 95% of simulated alphas from the luck distribution. In comparison, luck explains the alpha of the following seven funds.

However, the t_α bootstrapped p values, which are statistically more robust since t_α scales alpha by standard error as explained in section 4.2.1., indicate that the performance of only Add Value Fund N.V. is due to stock-picking skills instead of luck.

The two funds with the lowest empiric alphas perform poorly due to actual bad skill since the p values derived from the t_α bootstrap simulations are significant. The finding of negative skill corresponds with Kosowski et al. (2006), who also use a 5% p value as a significance threshold. Moreover, as further elaborated in paragraph 5.2., their study has a higher number of mutual funds, making it more likely to be skilled.

The distinction in p values clearly shows the differences in inference between the parametric and the two bootstrapping methods since p values of the unrealistic parametric method indicate that no fund in the sample has positive nor negative skills.

4.3 Luck or skill during booms and busts

There is a wide variety of studies on mutual fund performance during crises. For example, Dong, Feng and Sadka (2018) provide evidence of opportunities to identify mispriced stocks after and during periods of crisis. Therefore, this study looks into the differences in fund performance during the recent volatility during the COVID-19 pandemic, the financial crisis and the internet bubble.

4.3.1 COVID-19 pandemic

Besides the health crisis that the COVID-19 pandemic triggered, it also led to volatility on financial markets, which had last been seen during the financial crisis, with one of the quickest bear markets and subsequent recoveries in financial history. Therefore, this thesis looks at whether Dutch mutual funds exhibited skill or luck during this unique period.

Regarding the equal-weight portfolio of funds, the SIC criterion is the lowest for the three-factor model. Moreover, the momentum factor is not significant for any fund. Also, the adjusted R^2 of each fund is the highest for the three-factor model. Therefore, the study proceeds the bootstrapping procedure for the COVID-19 sub-period with the three-factor model.

Table 8: Bootstrap simulation results for January 2020 to March 2021

Results based on t_α

Alphas > 0				**Alphas < 0**		
Panel A1: Parametric method						
α	0.874	0.834	0.406	-0.222	-1.42	-1.892
t_α	0.60	0.72	0.48	-0.22	-1.21	-1.72
p value	0.560	0.487	0.642	0.827	0.255	0.115
Panel A2: Bootstrap based on t_α						
p value	0.209	0.222	0.185	0.395	0.095	0.042

Alphas represent monthly percentages and the columns are ordered from high to low parametric (standard) three-factor alphas. Panel A2 displays the bootstrapped *p* value which compares the empiric t_α to the simulated t_α's. *P* values below 0.05 indicate bad skill for negative alphas, and *p* values below 0.05 indicate positive skill for positive alphas. Corresponding fund names to alphas are retrievable on the author's request.

The fund's performance with the most negative monthly alpha, i.e. -1.892% per month, is attributed to bad skill under the t_α bootstrap. The performance of the rest of the funds can be attributed to good and bad luck. Also, the performance of Add Value Fund N.V., the only t_α skilled fund of the entire sample period, as indicated by table 7, produces the highest alpha of 0.874% per month, although it is not significant for the COVID-19 sub-period.

To illustrate the difference in both inference procedures, the parametric method attributes the performance of the worst performing fund to bad luck, while the bootstrap *p* value attributes the negative alpha to actual bad skill. It is essential to note that the small number of funds in the simulations might affect the robustness of the bootstrapped *p* values since the inference compares the empiric alpha only to six simulations. Paragraph 5.2.1 further discusses this possible limitation.

4.3.2 Financial crisis and Euro crisis

Severe underlying problems in the financial sector became evident in 2007 when the US housing market bubble burst, which caused a liquidity crunch that spilt over to the global financial system and the real economy. As a result, equity markets in the United States alone lost $8 trillion in market capitalisation from Oktober 2007 to Oktober 2008 (Brunnermeier, 2009).

However, the crisis did not leave Europe untouched and brought several countries near bankruptcy. The European crisis only ended in July 2012 with the ECB president Mario Draghi's speech in the ECB pledged that it would "do whatever it takes to preserve the Euro" (De Haan, Oosterloo and

Schoenmaker, 2015). Ronald and Dol (2011) explain that the Dutch economy was inherently extra vulnerable due to the openness of the economy, a relatively large financial sector, and substantial private mortgage debts, which became problematic due to the housing market crash (Ronald and Dol, 2011). Therefore, this study looks if fund performances during Oktober 2007 to July 2012 were due to genuine (bad) stock-picking skills or (bad) luck.

The thesis again selects the best factor model for this sub-period. For each fund and an equal-weighted portfolio of mutual funds, the MOM factor is not significant. The adjusted R^2 is the highest for the three-factor model. Moreover, the SIC is the lowest for the three-factor model. These statistics all point out that the 3-factor model is better for this sub-period than the four-factor model.

Table 9: Bootstrap simulation results for Oktober 2007 to July 2012

Results based on t_α

	Alpha > 0	**Alphas < 0**											
Panel A1: Parametric method													
α	0.614	-0.120	-0.138	-0.151	-0.152	-0.191	-0.193	-0.413	-0.428	-0.440	-0.456	-0.519	-0.631
t_α	1.333	-0.283	-0.289	-0.345	-0.336	-0.422	-0.387	-1.041	-1.037	-1.156	-0.919	-1.155	-1.306
p value	0.261	0.824	0.804	0.787	0.818	0.714	0.696	0.404	0.438	0.389	0.419	0.374	0.242
Panel A2: Bootstrap based on t_α													
p value	0.099	0.418	0.405	0.379	0.395	0.355	0.371	0.172	0.165	0.128	0.206	0.150	0.122

Alphas represent monthly percentages, and the columns are ordered from high to low parametric three-factor alphas. Panel A2 displays the bootstrapped *p* value, which compares the empiric t_α to the simulated t_α's. *p* values below 0.05 indicate bad skill for negative alphas, and *p* values below 0.05 indicate positive skill for positive alphas. Corresponding fund names to alphas are retrievable on the author's request.

A substantial majority of 12 funds produced negative alphas during the sample period, while only the fund that exhibited skill over the entire sample period, Add Value Fund N.V., had a positive alpha of 0.614% per month. However, both the parametric and bootstrapped *p* values indicate that fund performance on both the positive and negative sides of the alpha spectrum is due to (bad) luck.

For example, the *p* value of 0.099 in panel A2 means that 9.9% of the alphas simulated under the H_0 that alpha equals zero, or no skill, are higher than the empiric alpha. Therefore it cannot be concluded under a 5% *p* value threshold that the positive alpha was genuinely due to skill during this period since there is still a 9.9% probability that the positive alpha could have been achieved purely by sampling variation or luck.

Again, substantial differences arise between the parametric and bootstrapped *p* values since the bootstrapped *p* values of all funds in the sample are lower than their parametric equivalents. For example, the best-performing fund's bootstrapped *p* value is 0.099, while the corresponding parametric *p* value equals 0.261.

4.3.3 Dotcom Mania

Stocks in the internet industry returned over 1000 per cent from the start of 1998 to the end of 2000. However, within two years, those seemingly stellar returns evaporated as the growth of the underlying business disappointed optimistic expectations.

The period in which internet stocks rose and subsequently fell is known as the internet bubble because, according to Ofek and Richardson (2003), stocks deviated from their fundamental value due to short-selling restrictions. Those restrictions were present in the form of high costs of borrowing shares to sell short, which subsequently led asset prices to mainly reflect the opinions of optimists since pessimists were often not able to participate in the price discovery process by selling high-flying internet stocks short.

However, for example, Bharath and Viswanathan (2006) argue that the rationality of investors can explain the up and downswings of stocks during this period. Therefore there is no definitive consensus if the internet sector stocks were in an asset price bubble, but the answer to this question lies beyond the scope of this study.

The examined period is demarcated from January 01 1998 until December 31 2001, consistent with Singh (2013). The period starts in 1998 because De Long and Magin (2006) argue that the bubble did not start until this period. The sample consists of four funds with at least 36 months of returns, and all four funds existed during the entire sample period.

The adjusted R^2 is the highest for the three-factor model regressed on the equal-weight portfolio of the four funds, while the SIC was the lowest for the three-factor model, namely 280.467 versus 283.744 for the four-factor model. Moreover, the MOM coefficient is not significant for the equal-weight portfolio, nor any fund. Therefore this sub-period uses the three-factor model since the statistical properties indicate the Fama French model to have the best fit.

Table 10: Bootstrap simulation results for January 1998 to December 2001

Results based on t_α

	Alphas > 0		**Alphas < 0**	
Panel A1: Parametric method				
α	.466	.155	-.1366	-.931
t_α	.781	.291	-.228	-.596
p value	0.439	0.772	0.821	0.554
Panel A2: Bootstrap based on t_α				
p value	0.170	0.371	0.444	0.274

Alphas represent monthly percentages, and the columns are ordered from high to low parametric three-factor alphas. Panel A2 displays the bootstrapped p value, which compares the empiric t_α to the simulated t_α's.

The empiric alpha is only compared to four simulations, of which one is it is own. This small number of simulations might induce limitations to the bootstrapped *p* values discussed in paragraph 5.2.1.

It is relevant to note the differences between the bootstrapped *p* values and the parametric *p* values. Both methods attribute the performance of both the funds with negative and positive alphas to luck. However, the *p* values differ substantially. For instance, the best performing fund has a parametric *p* value of 0.439 while it has a t_α bootstrapped *p* value of 0.170.

This section concludes that the substantial majority of funds in the sample during periods of booms and busts do not possess stock selection abilities and finds evidence that most fund performance can be attributed to good or bad fortune rather than (bad) stock-picking abilities. Only the worst performing fund during the COVID-19 pandemic exhibited bad skills. However, these conclusions should be interpreted cautiously since the samples consist of a limited number of funds that might affect the alphas' simulated distribution. The discussions section in paragraph 5.2.1. further debates these possible limitations.

4.4 CAPM bootstrap

The study selected the 4-factor model regarding the entire period since it had a higher adjusted R^2 and a lower SIC value relative to the three-factor model. Moreover, the momentum factor was significant for a substantial number of individual funds. However, the thesis additionally incorporates a CAPM-bootstrap to examine whether results are still robust under various models.

Table 11 examines the bootstrapped *p* values over the whole sample period of 1996-2021 using the CAPM of Sharpe (1964). Albeit one fund had a significant positive alpha according to the four-factor t_α bootstrap, the CAPM bootstrap yields a *p* value of ''only'' 0.066, which would attribute the performance to good fortune. Moreover, the four-factor bootstrap attributes the negative alpha performance of two funds to bad skill. However, the CAPM bootstrap attributes the performance of the worst three funds to bad skill.

Moreover, ten funds have a positive alpha concerning the four-factor model, but only eight funds have a positive CAPM alpha. These results again stress that finding an appropriate factor model benchmark is essential for evaluating mutual fund performance.

Table 11: CAPM Bootstrap simulation results for January 1996 to March 2021 period

Results based on ordering t_{α}

	Alphas > 0								**Alphas < 0**									
	Highest	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16	#17	Lowest
Panel A1: Parametric method																		
Empiric α	.580	.466	.379	.336	.227	.220	.068	.012	-.003	-.155	-.165	-.195	-.236	-.261	-.374	-.497	-.588	-.943
t_{α}	1.52	1.50	1.29	1.10	1.08	0.76	0.28	0.05	-0.01	-0.66	-0.65	-0.95	-0.77	-0.75	-1.37	-2.11	-1.75	-2.25
p value	.132	.136	.201	.276	.281	.446	.782	.957	.993	.508	.518	.344	.440	.453	.171	.036	.083	.027
Panel A2: Bootstrap based on t_{α}																		
p value	.072	.066	.080	.167	.160	.167	.369	.459	.511	.221	.251	.161	.203	.220	.061	.021	.025	.007

Alphas represent monthly percentages. All panels are ordered from left to right from high parametric four-factor alphas to parametric low four-factor alphas. Panel A1's first row displays the CAPM alphas. The third row displays the parametric *p* value. Panel A2 displays the bootstrapped *p* value based on t_{α}. Each column refers to the same fund, regardless of the panel in which the data is situated.

5. Conclusion and discussion

5.1 Conclusion

This research examines if the performance of actively managed Dutch equity mutual funds is explained by skill or luck during the 1996-2021 period. The mutual funds in the dataset invest mainly in equity listed in the Netherlands and have at least 36 months of returns between 1996/01/01 to 2021/31/03. The Carhart (1997) four-factor model is selected to extract the alpha that cannot be explained by the four risk factors but is created by stock-picking skills or luck. These European four-factor returns were derived from the Ken French Data Library. Table 4 provides evidence that the average fund follows a contrarian strategy. The direction of the average fund’s four-factor alpha is positive but not significant.

Next, we bootstrap *p* values via the residual resampling method for each of the individual mutual funds. Adding the resampled residuals of the four-factor model to the predicted returns under the null of no outperformance allows simulating a distribution of alphas and alpha *t* statistics for each fund. This process results in simulated distributions that are a result of pure luck.

The thesis presents evidence of positive skill for one mutual fund and negative skill for two mutual funds from January 1996 to March 2021. Previous studies have used the standard parametric method, which looks at the fund’s generated alpha relative to a factor model. However, the parametric method has the disadvantage of not separating the effect of luck from genuine stock-picking skills since the assumption of normality is violated due to concentrated stock portfolios.

The residual resampling bootstrapping method is used to solve the statistical challenges related to non-normality. This technique assumes the alphas to be non-parametrically distributed. Thus, this method allows comparing the empirically observed alpha of each fund to a cross-section of simulated alphas, which is based on luck or sampling variation alone. This study provides evidence that the best (worst) performing funds have higher (lower) alphas than expected by the luck distribution and therefore finds evidence of positive and negative skills among Dutch equity mutual funds, albeit the performance of the majority of mutual funds can be attributed to (bad) luck.

Moreover, the only fund that is appointed skill according to the t_α bootstrap, namely Add Value Fund N.V, significantly loads positively on SMB while having a significant negative load on the MOM factor and a slight tilt towards growth stocks, albeit the *t* statistic is not significant. These loads imply that the fund specialises in contrarian smallcap stocks while picking from both the growth and value universe with a slight preference for growth stocks. These factor tilts correspond with Cremers et al. (2019) findings that most skilled funds pick from the small cap growth universe.

Contact by e-mail with the fund's management corroborated the factor tilts. The management indicated that the fund holds a relatively concentrated portfolio of only 12 stocks and specialises in small cap equities with a strong competitive moat, with the latter perhaps explaining the slight growth tilt of the fund, since one might expect competitive, growing companies to sell at a premium relative to the value of equity as indicated by the balance sheet. Moreover, the fund's focus on competitive businesses hints that it has a positive tilt on the Quality Minus Junk (QMJ) factor of Asness, Frazzini and Pedersen (2019), which has implications for future research on Dutch mutual funds, as paragraph 5.2.3 further explains. Moreover, the management's indication that the fund has a concentrated portfolio again stresses the importance of applying the bootstrap, although non-normality was already formally indicated by testing for it in Stata.

Moreover, the *p* values that resulted from the α bootstrap, the t_α bootstrap and the parametric method all conclude differently on whether funds are skilled. Thus, the used method appears to be highly relevant for assessing the performance of actively managed Dutch Equity mutual funds and stresses the importance of using the bootstrap technique to distinguish between skill and luck.

5.2 Discussion

This study aims to assess if alphas of Dutch equity mutual funds are the product of either (mis)fortune or stock picking abilities. This paper's main results imply that, using the t_α bootstrap and the four-factor model as a benchmark, the fund performance of the second top fund is the result of skill, while the performance of the other nine positive alpha funds is due to luck. Moreover, the worst two performing funds are negatively skilled, while the other six underperform the four-factor model due to bad luck.

However, the study and methodology have their limitations, which gives room for future research to improve the academic literature. The first suggestion for future research would be comparing funds with global exposure to funds that invest solely in the Netherlands since Fortin and Michelson (2005) find that global funds more often add value to investors versus domestic-only funds since global funds can pick stocks from a more comprehensive array of equities.

5.2.1 Sample size

This study's sample size is limited since the data consists of only 18 actively managed mutual funds. Furthermore, next to the sample size, the number of funds is not evenly distributed over time since not all funds survived the entire time in the sample or entered later than 1996. The sample size might limit the study's external validity since we compare a fund's empirical alpha to a pool of alphas simulated using the residuals of the sample's mutual funds. Since a limited number of funds determine the pool of simulated alphas, the sample size likely affects the bootstrapped *p* values. In this line of reasoning, the results of this research are not comparable to the results of studies on mutual fund performance in non-Dutch markets with larger sample sizes.

Kosowski et al. (2006) discuss the effect of small samples on the bootstrapped *p* values and conclude that the residual resampling bootstrap method is especially essential in small samples since samples with a limited number of mutual funds tend to have a relatively non-normal distribution of alphas. The highly non-normal alpha distribution might lead the parametric *p* values to differ substantially from the adherent non-parametric bootstrapped *p* values. Thus, a small sample makes the inference procedure more likely to conclude that a fund is (un)fortunate rather than having positive or bad stock-picking abilities.

The additional challenge for funds to score significant *p* values in the non-parametric method increases the strength of the evidence of both positive and negative skills. Nonetheless, this line of thought can be disputed since the sample size of this study is relatively small and differs over time. However, Chernick (2008) finds that samples with limited funds can still deliver valid bootstrapped outcomes. Still, small-sized samples have more variability than the bootstrap results because observations can recurrently appear due to the resampled nature of the simulation. After all, the resampling procedure replaces residuals in the original pool from which it randomly picks the next residual.

5.2.2 Time-varying beta

In line with Kosowski et al. (2006), the bootstrap algorithm assumes that factor loadings are fixed over the entire sample period. In other words, the bootstrap uses a stationary beta and unconditional approach. However, mutual funds might alter their factor loadings throughout market cycles, implying that fund managers participate in factor timing. For example, Andreu, Sáez and Sarto (2018) find evidence that funds take on more market risk during stable markets while decreasing market risk during highly volatile market periods.

Therefore, it would be relevant for future research to see if mutual funds are still skilled if changing factor loadings over time are controlled for. This could provide evidence on whether funds are truly skilled stock-pickers or rather skilled factor timers. Therefore, bootstrapping using a conditional model that allows the betas to vary over time might be a fruitful path for future research on the performance of Dutch mutual funds.

5.2.3 Joint hypothesis problem

This study finds evidence for stock-picking abilities using the four-factor model as a benchmark and the bootstrapping technique. However, this does not necessarily have implications for market efficiency. For example, market efficiency cannot be validated or rejected since it is unsure if Carhart’s (1997) model fits the funds' returns best. More specifically, the factor model might be misspecified, and an alternative model might be more appropriate. For example, Martin and Puthenpurackal (2008) find that the four-factor model cannot explain the positive alpha of Mr Warren Buffet’s publicly listed investment vehicle Berkshire Hathaway. However, Frazzini et al. (2013) find that controlled for QMJ (Quality minus Junk), Berkshire Hathaway’s alpha is not statistically significant. Contacting the best performing fund corroborated the idea that QMJ might explain the skill, as the fund implicitly indicated to focus on these kinds of equities. However, as Frazzini et al. (2013) indicate, consistently harvesting factor premia outside the four-factor model might be a skill itself and does not make a fund’s performance less impressive.

In other words, a stock-picker’s apparent skill relative to a factor model is not necessarily evidence of stock-picking abilities but rather explained by consistently harvesting alternative factor premia that the four-factor model might not capture. Specifically, this results in the joint hypothesis problem of Fama (1970) since it is unclear if the positive and negative alphas are due to markets inefficiencies or an incorrectly specified factor model used to explain returns.

Therefore, it would be fruitful for future research to see if the alphas are still robust after controlling for alternative risk models. However, adding additional factors should be done selectively since the extension of traditional factor models has also been criticised by Feng, Giglio and Xiu (2017). They find that many new factors are repackaged traditional risk factors and do not significantly increase the model's explanatory power.

5.2.4 Survival bias

The minimum requirement for inclusion in the dataset is that a fund has to exist for thirty-six months. The lack of short-lived funds limits the probability that the simulation resamples the same residual a high number of times. However, this also gives room for critics to debate whether the sample is a fair reflection of all funds in the Dutch mutual fund market. As a consequence of surpassing the short-lived funds, survival bias could notably play a role in the results of this study since Hanke, Keswani, Quigley and Zagonov (2018) note that short-lived funds produce lower alphas than their surviving peers.

A way to increase the internal validity of a minor sample-sized bootstrap procedure is to expand the number of bootstrap simulations. However, for sample periods with a limited number of funds, the parametric method might be best even when it is considered that the alpha distribution is non-normal. Cuthbertson et al. (2008) also consider the relationship between the validity of the bootstrapping procedure and the survival bias. Again, the answer depends on the sample size whether survival bias significantly affects the bootstrap outcomes. They argue that, since dead funds usually perform lousily, the surviving funds’ performance looks relatively impressive and increase the probability of being appointed unlucky instead of having bad skills. Technically, this difference is explained by the added poor performing, short-lived funds situated in the far left of the alpha distribution.

Most short-lived funds die due to investors withdrawing money from poor performers (Ha and Ko, 2017). Therefore, the fund that the bootstrap method now marks as negatively skilled might have been marked as unlucky if the short-lived mutual funds had been included. Therefore, a sample that includes short-lived funds likely causes inference procedures to more often conclude a fund to have bad stock-picking abilities.

Moreover, Galagedera, Fukuyama, Watson and Tan (2020) report that it is possible that a tiny number of short-lived outperforming funds had genuine stock-picking abilities but fused with other mutual funds as a result of their positive alphas achieved in the past. As the actively managed mutual fund market in the Netherlands has a relatively limited number of active funds caused by a high degree of consolidation, it seems possible that there are stock-picking abilities among short-lived mutual funds, which have been omitted from this study. Furthermore, it is possible that the funds in the right outer end of the empirical alpha distribution would become skilled had we included short-lived funds since the bootstrapped *p* value for outperformers is close to the 5% threshold. In light of the earlier thoughts on the consequences of the absence of short-lived funds in our sample and the limited number of funds in our dataset, it would be a fruitful path for future papers to look into differences regarding bootstrapped *p* values between samples that include and exclude short-lived funds.

6. References

Ahad, N. A., Yin, T. S., Othman, A. R., & Yaacob, C. R. (2011). Sensitivity of normality tests to non-normal data. *Sains Malaysiana*, *40*(6), 637-641.

Andreu, L., Matallín-Sáez, J. C., & Sarto, J. L. (2018). Mutual fund performance attribution and market timing using portfolio holdings. *International Review of Economics & Finance*, *57*, 353-370.

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2019). Quality minus junk. *Review of Accounting Studies*, *24*(1), 34-112.

Bharath, S. T., & Viswanathan, S. (2006). Is the internet bubble consistent with rationality? *Available at SSRN 943609*.

Broeders, D. W., van Oord, A., & Rijsbergen, D. R. (2019). Does it pay to pay performance fees? Empirical evidence from Dutch pension funds. *Journal of International Money and Finance*, *93*, 299-312.

Brunnermeier, M. K. (2009). Deciphering the liquidity and credit crunch 2007-2008. *Journal of Economic perspectives*, *23*(1), 77-100.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, *52*(1), 57-82.

Ciliberti, S., Sérié, E., Simon, G., Lempérière, Y., & Bouchaud, J. P. (2019). The Size Premium in Equity Markets: Where Is the Risk?. *The Journal of Portfolio Management*, *45*(5), 58-68.

Cremers, K. M., Fulkerson, J. A., & Riley, T. B. (2019). Challenging the conventional wisdom on active management: A review of the past 20 years of academic literature on actively managed mutual funds. *Financial Analysts Journal*, *75*(4), 8-35.

Cuthbertson, Nitzsche, & O'Sullivan (2008). UK mutual fund performance: Skill or luck? *Journal of Empirical Finance*, *15*(4), 613-634.

De Haan, J., Oosterloo, S., & Schoenmaker, D. (2015). *European financial markets and institutions*. Cambridge University Press.

DeLong, J. B., & Magin, K. (2006). *A short note on the size of the dot-com bubble* (No. w12011). National Bureau of Economic Research.

Dong, X., Feng, S., & Sadka, R. (2019). Liquidity risk and mutual fund performance. *Management Science*, *65*(3), 1020-1041.

Fama, E. F., & French, K. R. (1992). The cross-section of expected stock returns. *The Journal of Finance*, *47*(2), 427-465.

Fortin, R., & Michelson, S. (2005). Active international mutual fund management; can managers neat the index? *Managerial Finance*.

Frazzini, A., Kabiller, D., & Pedersen, L. H. (2018). Buffett's alpha. *Financial Analysts Journal*, *74*(4), 35-55.

Feng, G., Giglio, S., & Xiu, D. (2017). Taming the factor zoo. *Chicago Booth research paper*, (17-04).

Galagedera, D. U., Fukuyama, H., Watson, J., & Tan, E. K. (2020). Do mutual fund managers earn their fees? New measures for performance appraisal. *European Journal of Operational Research, 287*(2), 653-667.

Golec, J. H. (1992). Empirical tests of a principal-agent model of the investor-investment advisor relationship. *Journal of Financial and Quantitative Analysis*, 81-95.

Ha, Y., & Ko, K. (2017). Why do fund managers increase risk? *Journal of Banking & Finance*, *78*, 108-116.

Hanke, B., Keswani, A., Quigley, G., & Zagonov, M. (2018). Survivorship bias and comparability of UK open-ended fund databases. *Economics Letters*, *172*, 110-114.

Harvey, C. R., & Liu, Y. (2020). False (and missed) discoveries in financial economics. *The Journal of Finance*, *75*(5), 2503-2553.

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(6), 497-526.

Hoberg, G., Kumar, N., & Prabhala, N. (2018). Mutual fund competition, managerial skill, and alpha persistence. *The Review of Financial Studies*, *31*(5), 1896-1929.

Horowitz, J. L. (2019). Bootstrap methods in econometrics. *Annual Review of Economics*, *11*, 193-224.

Karolyi, G. A., Lee, K. H., & Van Dijk, M. A. (2012). Understanding commonality in liquidity around the world. *Journal of financial economics*, *105*(1), 82-112.

Kenneth R. French – European Factor returns Data Library. Retrieved from: https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Kosowski, R. (2011). Do mutual funds perform when it matters most to investors? US mutual fund performance and risk in recessions and expansions. *The Quarterly Journal of Finance*, *1*(03), 607-664.

Lai, T. Y., & Stohs, M. H. (2015). Yes, CAPM is dead. *International Journal of Business*, *20*(2), 144.

Luo, J., & Subrahmanyam, A. (2019). The affect heuristic and stock ownership: A theoretical perspective. *Review of Financial Economics*, *37*(1), 6-37.

Ofek, E., & Richardson, M. (2003). Dotcom mania: The rise and fall of internet stock prices. *The Journal of Finance*, *58*(3), 1113-1137.

Otten, R., & Schweitzer, M. (2002). A comparison between the European and the US mutual fund industry. *Managerial Finance*.

Riedl, Arno, and Paul Smeets. 2017. Why do investors hold socially responsible mutual funds?. *Journal of Finance 72*, 2505-2550.

Rohleder, M., Scholz, H., & Wilkens, M. (2011). Survivorship bias and mutual fund performance: Relevance, significance, and methodical differences. *Review of Finance*, *15*(2), 441-474.

Sharpe, O. W., & Miller, M. (1964). CAPM. *Equilibrium*, *7*.

Scholtens, B. (2005). Style and performance of Dutch socially responsible investment funds. *The Journal of Investing*, *14*(1), 63-72.

Singh, V. (2013). Did institutions herd during the internet bubble? *Review of Quantitative Finance and Accounting*, *41*(3), 513-534.

Sinha, N. R. (2016). Underreaction to news in the US stock market. *Quarterly Journal of Finance*, *6*(02), 1650005.

Smith, J., & McAleer, M. (1994). Newey–West covariance matrix estimates for models with generated regressors. *Applied Economics*, *26*(6), 734-7d39.

Studenmund, A. H., & Johnson, B. K. (2017). *A practical guide to using econometrics*. Pearson Education Limited.

Zaremba, A. (2019). The cross section of country equity returns: a review of empirical literature. *Journal of Risk and Financial Management*, *12*(4), 165