Data Snooping in Equity Premium Prediction

Viktoria-Sophie Bartsch, Hubert Dichtl, Wolfgang Drobetz, and Andreas Neuhierl

First version: November 2015
This draft: May 2017

Abstract
We study the performance of a comprehensive set of equity premium forecasting strategies that have been shown to outperform the historical mean out-of-sample when tested in isolation. Using a multiple testing framework, we find that previous evidence on out-of-sample predictability is primarily due to data snooping. We are not able to identify any forecasting strategy that produces robust and statistically significant economic gains after controlling for data snooping biases and transaction costs. By focusing on the application of equity premium prediction, our findings support Harvey’s (2017) more general concern that many of the published results in financial economics will fail to hold up.

Keywords: Equity risk premium prediction; data snooping bias

JEL classification codes: G11, G12, G14

Acknowledgments: We thank Michael Halling, Alexander Hillert, Harald Lohre, Emanuel Mönch, and Tatjana Puhan for helpful comments.
1. Introduction

Does equity premium prediction pay off? While the in-sample predictability of the equity premium seems largely undisputed (Campbell, 2000), most investors are ultimately interested in whether forecasting strategies can deliver reliable out-of-sample gains. Recognizing the controversial debate regarding the out-of-sample performance of established stock return prediction models, Spiegel (2008) poses a challenging question: “Can our empirical models accurately forecast the equity premium any better than the historical mean?”

While Goyal and Welch (2008) suggest that a healthy skepticism is appropriate when it comes to predicting the equity premium out-of-sample, several subsequent studies contrarily conclude that some forecasting models provide ‘better’ results than the historical mean (Campbell and Thompson, 2008; Rapach, Strauss, and Zhou, 2010; Ferreira and Santa-Clara, 2011; Neely et al., 2014; Bätje and Menkhoff, 2016; Huang et al., 2016; among others). However, one challenge in answering the question of out-of-sample predictability is that almost all of these forecasting strategies are tested on a single data set. When many models are evaluated individually, some are bound to show superior performance by chance alone, even though they are not (Sullivan, Timmermann, and White, 1999). This bias in statistical inference is usually referred to as ‘data snooping’. Without properly adjusting for this bias in a multiple testing set-up, we might commit a type I error, i.e., falsely assessing a forecasting strategy as being superior when it is not. In fact, Harvey, Liu, and Zhu (2016) note that equity premium prediction offers an ideal setting to employ multiple testing methods. In addition, Campbell R. Harvey (2017) emphasizes in his recent 2017 AFA Presidential Address that “with the combination of unreported tests, lack of adjustment for multiple tests, and direct as well as indirect p-hacking, many of the result being published will fail to hold up.”

To the best of our knowledge, our study is the first to jointly examine the out-of-sample performance of a comprehensive set of equity premium forecasting strategies relative to the historical mean, while accounting for the data snooping bias. We construct a comprehensive set of 100 forecasting strategies...
strategies that are based on both univariate predictive regressions and advanced forecasting models, including strategies that adopt diffusion indices (Ludvigson and Ng, 2007; Neely et al., 2014) or combination forecast approaches (Timmermann, 2006; Rapach, Strauss, and Zhou, 2010), apply economic restrictions on the forecasts (Campbell and Thompson, 2008), predict disaggregated stock market returns (Ferreira and Santa-Clara, 2011), or model economic regime shifts (Henkel, Martin, and Nardari, 2011; Huang et al., 2016). We use these forecasting strategies to predict the monthly U.S. equity premium out-of-sample based on the most recent 180 months and track their out-of-sample performance for the subsequent month over the evaluation period from January 1966 to December 2015. We aim to answer Spiegel’s (2008) question, i.e., whether there are forecasting strategies that provide a significantly higher performance than the prevailing mean model. As performance measures, we use the mean squared forecast error and absolute as well as risk-adjusted excess returns.

Why is data snooping a concern in our analysis? Following the example in Hsu, Kuan, and Yen (2014), suppose these 100 models are mutually independent, and we apply a t-test to each model with the significance level of 5%. The probability of falsely rejecting at least one correct null hypothesis is $1 - (1 - 5\%)^{100} \approx 0.994$. Therefore, it is very likely that an individual test may incorrectly suggest an inferior model to be a significant one. This simple example emphasizes the importance of an appropriate method that can control such data-snooping bias and avoids spurious inference when many models are examined together.

To formally control for data snooping when testing for the possible superiority of a forecasting strategy, we apply Hansen (2005) test for superior predictive ability (SPA-test) that provides a multiple testing framework without data snooping bias. To identify as many forecasting strategies that can outperform the benchmark as possible, we further implement the stepwise extensions of the SPA-test recently proposed by Hsu, Hsu, and Kuan (2010) and Hsu, Kuan, and Yen (2014).

Our results show that many forecasting strategies outperform the historical mean when tested individually. However, once we control for data snooping, we find that no forecasting strategy can

---

3 See Rapach and Zhou (2013) for a survey of the literature on stock return forecasting.
outperform the historical mean in terms of mean squared forecast errors. With respect to return-based performance measures, we find marginal evidence for statistically significant economic gains at least on a risk-adjusted excess return basis when using the equity premium forecasts in a traditional mean-variance asset allocation, even after controlling for data snooping bias. In contrast, the benefits for a pure market timing investor are limited. Taken together, our findings strengthen the results of Goyal and Welch (2008) that the out-of-sample predictability of the equity premium is questionable.

The remainder of our paper is structured as follows: Section 2 reviews the literature on multiple testing. Section 3 provides an overview of equity premium forecasting strategies and a description of our methodological approach. Section 4 discusses our empirical results, and Section 5 presents various robustness checks. Section 6 concludes.

2. Literature review

In his 2017 AFA Presidential Address, Campbell R. Harvey (Harvey, 2017) provocingly states that the competition for top journal space spurs the publication of “an embarrassing number of false positives.” Although data snooping and the lack of adjustment for multiple tests has been identified as a major problem in financial economics, only few studies applied appropriate testing frameworks.

One of the first studies to address the problem of data snooping was Sullivan, Timmermann, and White (1999), who apply the multiple testing framework introduced by White (2000) in evaluating technical trading strategies. White’s (2000) “reality check” (RC) controls the family-wise error rate (FWER), i.e., the probability of at least one type I error given the pre-specified significance level $\alpha$. The RC-method was generalized to reduce the influence of poor models (Hansen, 2005), identify all significant models rather than only the best one (Romano and Wolf, 2005; Hsu, Hsu, and Kuan, 2010), and allow for more than one false rejection (Romano and Wolf, 2007; Hsu, Kuan, and Yen, 2014).

Utilizing these improved methods, Neuhierl and Schlusche (2011) study the performance of stock market timing rules and conclude that most market timing rules do not outperform a buy-and-

---

hold strategy after correcting for data snooping. More recently, Hsu, Lin, and Vincent (2017) examine the performance of popular cross-sectional return predictors and infer that most predictors are no longer significant after adjusting for data snooping bias. Applying the multiple testing framework to evaluate the out-of-sample performance of asset allocation strategies, Hsu et al. (2017) find that only a few portfolio strategies outperform the naïve 1/N diversification rule once controlling for data snooping.

Another strand of the data snooping literature focuses on controlling the false discovery rate (FDR), defined as the expected proportion of type I errors among all rejections. The FDR is less stringent than the FWER because it accounts for the number of tested strategies. Studies that implement the FDR testing framework include Barras, Scaillet, and Wermers (2010) on mutual fund performance, Bajgrowicz and Scaillet (2012) on technical trading rules, and Harvey, Liu, and Zhu (2016) on cross-sectional return predictability. All of these studies conclude that many earlier findings were likely biased by data snooping and are rendered insignificant once taking the necessary corrections.

3. **Empirical procedure**

3.1. **Set of forecasting strategies**

Data snooping tests are sensitive to the universe of forecasting strategies to which they are applied. To account for a complete set of forecasting strategies from which to draw, we consider both univariate predictive regression models and a comprehensive collection of advanced forecasting strategies. In putting together all these strategies, it is imperative to manage the trade-off between including too many ‘irrelevant’ strategies, thereby decreasing the power of the test, and including too few strategies, thereby overstating its statistical significance (Hansen, 2005). In line with Rapach and Zhou (2013), we survey forecasting strategies that have become popular in the literature. In total, we consider 28 univariate predictive regressions and 72 advanced forecasting techniques in our empirical analysis. In the following, we briefly summarize the relevant literature and the construction of the forecasting strategies. Table 1 provides a list of all 100 forecasting strategies.

[Insert Table 1 here]
**Univariate predictive regressions:** A univariate prediction regression is given as:

\[ r_{t+1} = \alpha + \beta x_t + \varepsilon_{t+1}, \]  

(1)

where \( r_{t+1} \) is the equity premium from period \( t \) to \( t+1 \), \( x_t \) a variable known at time \( t \) that is expected to predict the future equity premium, and \( \varepsilon_{t+1} \) a zero-mean disturbance term. The monthly (log) equity premium is defined as the continuously compounded stock return of the S&P 500 (including dividends) minus the log return on a one-month Treasury bill.

While it is impossible to construct the set of all conceivable predictors and combinations, we aim to build a set representative of the return predictability literature. Using the updated monthly data set provided by Goyal and Welch (2008), we compute 14 fundamental variables, including the dividend-price ratio (Campbell and Shiller, 1988), the book-to-market ratio (Kothari and Shanken, 1997), and interest rates (Fama and Schwert, 1977).

Neely et al. (2014) highlight the predictive power of technical indicators that stems from information frictions, for example if investors initially underreact to news due to behavioral biases and subsequently overreact as prices rise (Hong and Stein, 1999). Therefore, we augment our fundamental variables with various technical indicators based on popular trend-following strategies, i.e., moving averages (Zhou and Zhu, 2009), time-series momentum (Moskowitz, Ooi, and Pedersen, 2012), and volume data (Blume, Easley, and O'Hara, 1994).

In our empirical tests, we construct the technical indicators with different parametrizations and use the S&P 500 index (excluding dividends) as the price index and monthly volume data, where applicable. The detailed construction of the technical indicators is provided in appendix A. We follow Neely et al. (2014) and transform the technical indicators to point forecasts of the equity premium by using the respective technical indicator in the predictive regression model in equation (1).

---

5 Data are available from Amit Goyal’s webpage (http://www.hec.unil.ch/agoyal). A detailed description of the fundamental variables is provided in appendix A.

The out-of-sample predictions are generated by first estimating the regression model in equation (1) via OLS, and then using the fitted model to construct the equity risk premium forecast $\hat{r}_{t+1}$. We employ a rolling scheme to estimate the OLS parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ in order to capture uncertain model dynamics (Giacomini and White, 2006) and to comply with the stationarity requirement of Hansen (2005) SPA-test. The rolling estimation scheme ensures that the forecasting strategies do not nest the benchmark model, i.e., the recursive historical average equity premium, when applying the SPA-test (Elliott and Timmermann, 2016).

**Forecast restrictions:** Campbell and Thompson (2008) argue that the performance of univariate predictive regressions can be substantially improved by imposing weak restrictions on the signs of coefficients and return forecasts. Therefore, in one set of our strategies, we restrict the equity premium forecasts obtained from the univariate predictive regressions to be non-negative.

**Regime shifts:** As noted by Paye and Timmermann (2006) and Rapach and Wohar (2006), the data-generating process for stock returns is likely subject to substantial parameter instability due to structural breaks. Several forecasting strategies have been suggested to account for parameter instability. Building on work by Hamilton (1989), Guidolin and Timmermann (2007) estimate a multivariate Markov-switching model with four regimes – characterized as ‘crash’, ‘slow growth’, ‘bull’, and ‘recovery’ – and show that their model produces significant utility gains in asset allocation decisions. Exploiting the time-variation of fundamental variables, Henkel, Martin, and Nardari (2011) use a regime-switching vector auto-regression framework with two states that closely resemble the NBER-dated business cycles. They find that the historical average forecast is the best out-of-sample predictor in expansion periods, while fundamental variables provide useful information in recession periods.

Most recently, Huang et al. (2016) address the critique by Lettau and van Nieuwerburgh (2008) that regime-shifting models perform poorly out-of-sample due to unreliable estimates of both the timing and the size of regime shifts. Using a state-dependent predictive regression model that was introduced by Boyd, Hu, and Jagannathan (2005), they show that conventional predictive regressions are often misspecified and that their state-dependent approach is able to predict the equity premium in both bad and good times.
In our empirical analysis, we use the state-dependent predictive regression approach of Huang et al. (2016) with the full set of fundamental variables and technical indicators. As in Cooper, Gutierrez Jr., and Hameed (2004), the market states are identified based on past return information:

\[ r_{t+1} = \alpha + \beta_{\text{good}} x_t I_{\text{good},t} + \beta_{\text{bad}} x_t (1 - I_{\text{good},t}) + \epsilon_{t+1}. \]  

(2)

To proxy for the market state, the indicator variable \( I_{\text{good},t} \) takes the value of one when the past six-month (log) equity premium is non-negative, and zero otherwise (Huang et al., 2016).

**Combination forecasts:** Timmermann (2006) argues that combining individual forecasts is useful because it provides diversification gains compared to relying on forecasts from a single forecasting strategy. In addition, combining different forecasts also captures different degrees of adaptability of forecasting strategies to structural breaks, and alleviates the problem of model misspecification. Rapach, Strauss, and Zhou (2010) show that combinations of individual forecasts can deliver statistically and economically significant out-of-sample results due to reduced model uncertainty and parameter instability.

Combination forecasts are weighted averages of \( N \) individual forecasts that are estimated using the predictive regression in equation (1):

\[ \hat{r}_{\text{combination},t+1} = \frac{1}{N} \sum_{i=1}^{N} \omega_{i,t} \hat{r}_{i,t+1}. \]  

(3)

In our empirical analysis, we combine the individual forecasts based on either solely fundamental variables, solely technical indicators or all predictors. We use three simple averaging methods: the mean combination forecast sets \( \omega_{i,t} = \frac{1}{N} \); the median combination forecast is the median of \( \{\hat{r}_{i,t+1}\}_{i=1}^{N} \); and the trimmed mean combination forecast sets \( \omega_{i,t} = 0 \) for the individual forecasts with the smallest and largest value, and \( \omega_{i,t} = \frac{1}{N-2} \) for the remaining individual forecasts.

**Diffusion indices:** To avoid over-parametrization, several studies adopt a diffusion indices approach that assumes a factor structure for predictors and use estimates of the common factors as predictors in a predictive regression model. For example, Ludvigson and Ng (2007) extract three common
factors from a comprehensive set of macroeconomic and financial variables and find that the diffusion indices forecasts exhibit significant out-of-sample predictive power. In our empirical tests, we follow Stock and Watson (2006) and estimate the common factors using principal component analysis based on either the set of fundamental variables (PC-FUND), the set of technical indicators (PC-TECH), or all fundamental variables and technical indicators combined (PC-ALL). Rapach and Zhou (2013) emphasize that it is prudent to use a small number of common factors in prediction to avoid over-parametrization. Therefore, we use only the first principal component of either solely fundamental variables, solely technical indicators, or all predictors. The estimated principal components then serve as independent variables in the predictive regression model in equation (1).

**Sum-of-the-parts models:** The sum-of-the-parts (SOP) method proposed by Ferreira and Santa-Clara (2011) provides a stock market return forecast by separately forecasting the three components of the stock market return. It is one way of incorporating economic restrictions directly into the prediction. In particular, they decompose the return into the dividend-price ratio \( dp_{t+1} \), the growth rate of earnings \( ge_{t+1} \), and the growth rate of the price-earnings ratio \( gm_{t+1} \):

\[
    r_{t+1} = gm_{t+1} + ge_{t+1} + dp_{t+1} - r_{f,t+1}.
\]

(4)

Using this return decomposition, we assume no multiple growth, estimate the growth rate of earnings as a 20-year moving average of growth in earnings per share, and model the dividend-price ratio as a random walk. The return forecast of the SOP method can then be written as:

\[
    \hat{r}_{t+1}^{SOP} = \overline{ge}_t + dp_t - r_{f,t}.
\]

(5)

Expanding the work of Ferreira and Santa-Clara (2011), Bätje and Menkhoff (2016) develop an ‘extended’ sum-of-the-parts (ESOP) approach that combines the decomposition of the stock market return forecast with fundamental and technical indicators as well as combination forecasts. In a first step, the growth rate of the price-earnings ratio, \( \hat{gm}_{t,t+1} \), and the growth rate of earnings, \( \hat{ge}_{t,t+1} \), are estimated by univariate predictive regressions using solely fundamental variables or technical indicators, respectively. In a second step, the individual component forecasts are combined using simple averaging methods (mean, median, and trimmed mean). In a third step, the equity premium forecast is
obtained by summing the (combined) component forecasts, assuming that the dividend-price ratio follows a random walk:

\[
\hat{r}_{t+1}^{ESOP} = \hat{g}_{t+1}^{\text{Combination,FUND}} + \hat{g}_{t+1}^{\text{Combination,TECH}} + dp_t - r_{f,t}.
\] (6)

3.2. Implementation of the multiple testing framework

When considering a large number of possible forecasting strategies, data snooping is a natural concern (Lo and MacKinlay, 1990). To address this problem, Hansen (2005) proposes a test for superior predictive ability, or SPA-test, that allows for a comprehensive comparison of forecasting strategies, while ensuring that the results are robust to biases from data snooping.\(^7\) The SPA-test builds on White’s (2000) reality check but reduces the adverse influence of poor or irrelevant strategies on the power of the test. While the SPA-test can answer the question whether there is at least one superior forecasting strategy, if any, it is not able to identify all such strategies.

To address this shortcoming, Hsu, Hsu, and Kuan (2010) develop a stepwise extension of the SPA-test, or step-SPA-test, that is capable of identifying as many outperforming models as possible, while still removing poor models from consideration asymptotically. However, the ability to identify significant models is still limited due to the control of the familywise error rate (FWER), the probability of at least one false rejection given the pre-specified error rate \( \alpha \). In practice, when multiple testing involves a large number of hypotheses, incorrectly rejecting a few of them may not be a very serious problem. Controlling at least one false rejection then poses a very stringent criterion, and one may lower the rejection criterion and increase the test power by tolerating more false rejections. Hsu, Kuan, and Yen (2014) propose the step-SPA(\(k\))-test that asymptotically controls the FWER(\(k\)), the probabil-

\(^7\) A related multiple testing framework is based on the wild fixed-regressor bootstrap procedure developed by Clark and McCracken (2012). Their framework involves testing for equal predictive ability, i.e., whether the predictive ability of any forecasting strategy is the same as that of the benchmark model. As our main research interest is to assess whether any of the considered forecasting strategies in our set is indeed ‘better’ than the benchmark model, we have to test for predictive superiority, i.e., testing whether the predictive ability of any forecasting strategy is greater than the one of the benchmark model. As noted by Hansen (2005), this subtle distinction leads to a composite hypothesis, such that the null distribution is not unique but rather sample-dependent.
ity of at least \( k \) false rejections, where \( k \geq 2 \). They show that their step-SPA(\( k \))-test is consistent in that it can identify the violated null hypotheses with probability approaching one.

We employ these econometric methods to control for data snooping when comparing the out-of-sample performance of our set of forecasting strategies. The prevailing mean model of the equity premium, i.e., the recursive historical average equity premium since the beginning of the sample period, serves as the natural benchmark model, indicating a constant expected equity premium (Goyal and Welch, 2008). In the following, we briefly outline the implementation of these methods in our empirical analysis.

**SPA-test:** When testing for superior predictive ability in the presence of multiple alternative forecasts, we test the null hypothesis that the benchmark model, i.e., the historical mean, is not inferior to any alternative forecasting strategy:

\[
H_0: \max_{j=1,...,J} E(d_{j,t}) \equiv \mu_j \leq 0
\]  

(7)

where \( d_{j,t} \) is the difference of the performance measure of forecasting strategy \( j \) and the performance measure of the benchmark model at time \( t \). The performance of the forecasting strategies can be based on forecast errors as well as return-based measures. If the null hypothesis can be rejected, there is at least one forecasting strategy that outperforms the benchmark. Hansen (2005) proposes the studentized test statistic:

\[
\nu_t^{SPA} = \max \left( \max_{j=1,...,J} \frac{\sqrt{T} \bar{d}_j}{\hat{\omega}_j}, 0 \right)
\]  

(8)

where \( \bar{d}_j = T^{-1} \sum_{t=1}^{T} d_{j,t} \) denotes the average relative performance of forecasting strategy \( j \), and \( \hat{\omega}_j^2 \) is a consistent estimate of \( \omega_j^2 = \text{var}(\sqrt{T} \bar{d}_j) \). To reduce the influence of irrelevant strategies, at least asymptotically, Hansen (2005) advocates invoking a null distribution based on \( N(\hat{\mu}_j, \hat{\Omega}) \), where \( \hat{\mu}_j \) is an estimator for \( \mu_j \) given as

\[
\hat{\mu}_j = \bar{d}_j \mathbb{1}_{[\sqrt{T} \bar{d}_j/\hat{\omega}_j \leq \sqrt{2 \log T}]}.
\]

To approximate the distribution of the test statistic, we follow Hansen (2005) and implement the stationary bootstrap method of Politis and Romano (1994). In particular, for each strategy, we generate
$b = 1, ..., B$ resamples of $d_{j,t}$ by drawing geometrically distributed blocks with a mean block length of $q^{-1}$. We set the smoothing parameter $q = 0.5$ and generate $B = 10,000$ bootstrap resamples. The bootstrapped variables $d_{j,b,t}^*$ are re-centered about $\hat{\mu}_j$ as $Z_{j,b,t}^* = d_{j,b,t}^* - g(\bar{d}_j)$, where $g(\bar{d}_j)$ denotes the re-centering function, defined as $g(\bar{d}_j) = \bar{d}_j \mathbb{1}\{\sqrt{T\bar{d}_j/\hat{\omega}_j} - \sqrt{2\log\log T} \geq 0\}$. The studentized test statistic under the bootstrap is computed as $V_{b,t}^* = \max\left( \max_{j=1,...,J} \sqrt{T} Z_{j,b,t}^*, 0 \right)$, where $Z_{j,b}^* = T^{-1} \sum_{t=1}^T Z_{j,b,t}$.

A consistent estimate of the $p$-value is then given by:

$$\hat{p}_{SPA} = \sum_{b=1}^B \frac{\mathbb{1}\{V_{b,t}^* > V_{t}^{SPA}\} \mathbb{1}\{V_{b,t}^* > V_{b,t}^{SPA}\}}{B}.$$  \tag{9}

As shown by Hansen (2005), an upper and a lower bound for the $p$-value can be obtained by re-centering about $\hat{\mu}_j^U = 0$, which assumes that all competing forecasting strategies are as good as the benchmark model, and $\hat{\mu}_j^L = \min(\bar{d}_j, 0)$, which assumes that forecasting strategies that are outperformed by the benchmark model are ‘poor models in the limit’, respectively. A large difference between the upper and lower bound $p$-values is indicative of many poor forecasting strategies.

**Step-SPA-test:** If the null hypothesis of the SPA-test is rejected, we implement the stepwise extension of the SPA-test (step-SPA-test) developed by Hsu, Hsu, and Kuan (2010) to identify all additional significant forecasting strategies. First, we re-arrange the forecasting strategies in descending order of their test statistic and reject the top strategy if its test statistic is greater than the critical value, specified as the $1 - \alpha$ quantile of the empirical distribution bootstrapped from the entire set of forecasting strategies. Second, we remove $d_j$ of the rejected strategy and compute a new critical value bootstrapped from the subset of remaining forecasting strategies. We again reject the top strategy if its test statistic is greater than the new critical value and repeat this procedure until no further forecasting strategy can be rejected. All forecasting strategies that have been removed are then identified as superior strategies.

---

\(^8\) In our empirical analysis, we determine the critical values for the pre-specified error rate $\alpha = 5\%$. 

---

12
Step-SPA(k)-test: The step-SPA-test is able to successfully identify all superior strategies when the null hypothesis of the SPA-test is rejected, but is fairly conservative in doing so, as it controls the family wise error rate, i.e., the probability of at least one false rejection given the pre-specified error rate $\alpha$. However, when comparing a large set of forecasting strategies, as we do in our empirical analysis, one might be willing to tolerate a higher number of false rejections to increase test power and be able to better reject false null hypotheses. To accommodate this requirement, Hsu, Kuan, and Yen (2014) develop a refinement of the step-SPA-test, the step-SPA(k)-test, that asymptotically controls the probability of at least $k$ false rejections, with $k \geq 2$, less than or equal to a certain level $\alpha$.

The implementation of the step-SPA(k)-test is similar to the step-SPA-test: First, we re-arrange the forecasting strategies in descending order of their test statistic and reject all strategies with a test statistic greater than the critical value, specified as the $1 - \alpha$ quantile of the empirical distribution of the $k$-th largest test statistic bootstrapped from the entire set of forecasting strategies. Second, if the number of rejected strategies is less than $k$, the procedure stops and all strategies rejected in the first step are identified as superior to the benchmark model. Otherwise, we choose $k-1$ strategies from these rejected strategies, merge them with the remaining forecasting strategies that were not rejected in the first step and calculate a new critical value bootstrapped from this subset. We test all possible combinations of the $k-1$ strategies and determine the maximum critical value among all combinations. If the test statistic of any of the remaining forecasting strategies is greater than this maximum critical value, we add this strategy to the collection of rejected strategies and repeat this procedure until no further forecasting strategy can be rejected. In our empirical analysis, we set $k = 3$.

3.3. Measures of forecast performance

The most popular metric for evaluating the accuracy of point forecasts is the mean squared forecast error (MSFE) over the out-of-sample period. Therefore, in our first test we compare the performance of the forecasting strategies with the performance of the historical mean based on squared forecast errors $(r_t - \hat{r}_{jt})^2$, where $r_t$ is the realized (log) equity premium, and $\hat{r}_{jt}$ is the (log) equity premium forecast based on forecasting strategy $j$. 

13
However, as shown by Leitch and Tanner (1991), there is only a weak association between statistical measures of forecasting performance such as MSFE and economic forecast profitability, with forecasting strategies that outperform the benchmark model in terms of MSFE often failing to outperform when considering profit- or utility-based metrics. To assess whether the out-of-sample predictability is sufficiently large to be of economic value, we consider both absolute returns $r_{j,t}^{abs}$ based on the equity premium forecast of forecasting strategy $j$ and risk-adjusted excess returns $\frac{r_{j,t}^{abs} - r_{f,t}}{\sigma_j}$, where $\sigma_j$ is the volatility of the excess return of strategy $j$, as adequate performance measures.\textsuperscript{9}

We compute the absolute return $r_{j,t}^{abs}$ of an investor, who monthly allocates her portfolio between stocks and the risk-free asset $r_{f,t}$, using the (simple) equity premium forecast of strategy $j$:  

$$r_{j,t}^{abs} = w_{j,t}r_{j,t} + r_{f,t}$$

(10)

where $w_{j,t}$ is the proportion of total wealth allocated to the stock market. As the investor can then use the point forecasts as inputs for either a traditional mean-variance asset allocation or for a market timing decision, we choose $w_{j,t}$ for a mean-variance investor and a pure market timer, who is either fully invested in the market or holds the risk-free asset.

For a given coefficient of relative risk aversion $\gamma$ and a forecast of the equity premium variance $\hat{\sigma}_t^2$, a mean-variance investor chooses to hold $w_{j,t}^P = \frac{r_{j,t}}{\gamma \hat{\sigma}_t^2}$ of the risky asset. We set $\gamma = 5$ and estimate $\hat{\sigma}_t^2$ as a five-year rolling window of past monthly returns following Neely et al. (2014). Moreover, we impose portfolio constraints preventing investors from short-selling and levering more than 50%, so that $w_{j,t}^P$ is between 0 and 1.5.

A market timer is either fully invested in the stock market if the equity premium forecast is positive and reverts to holding the risk-free asset otherwise, i.e., she chooses:

\textsuperscript{9} The expected value of the loss function based on risk-adjusted excess returns equals the Sharpe (1994) ratio.
Given that most forecasting strategies involve frequent trading, a realistic evaluation of the performance of any forecasting strategy has to take transaction costs into account. We follow the choice in Balduzzi and Lynch (1999) and assume 50 basis points as roundtrip transaction costs.

4. Empirical results

Due to the availability of volume data, the sample period is from December 1950 to December 2015. We estimate all forecasting strategies using a rolling window of 180 months, and, after considering the initial estimation period, analyze the out-of-sample performance from January 1966 to December 2015. The prevailing mean model, i.e., the recursive historical average equity premium since December 1950, serves as our benchmark model.

4.1. Out-of-sample performance of forecasting strategies

Table 2 summarizes the out-of-sample performance of the historical mean and all forecasting strategies using the MSFE (panel A), the mean monthly absolute return (panel B) and the mean monthly risk-adjusted excess return (panel C) as performance measures. As explained in section 3.3, we assume 1) an investor with mean-variance preferences and relative risk aversion coefficient \( \gamma = 5 \), and 2) a market timer, who is fully invested in the stock market if the equity premium forecast is positive, and holds the risk-free asset otherwise. We impose roundtrip transaction costs of 50 basis points.

The results in panel A of Table 2 indicate that the historical mean generates lower forecast errors (MSFE of 19.35) than the average of all forecasting strategies (MSFE of 19.53). Our results reveal that none of the univariate predictive regressions is able to outperform the historical mean, largely confirming the results of Goyal and Welch (2008) and Neely et al. (2014). In line with the findings of Campbell and Thompson (2008), forecast restrictions improve upon the univariate predictive regressions, but, on average, do not outperform the historical mean. In contrast, state-dependent regressions

\[ w_{jt}^{MT} = \begin{cases} 1 & \text{if } \hat{r}_{jt} > 0 \\ 0 & \text{if } \hat{r}_{jt} \leq 0 \end{cases} \]
seem to worsen the performance of univariate predictive regressions. Moreover, contrasting the results of Neely et al. (2014), diffusion indices are not able to outperform the historical mean.\textsuperscript{10} Combination forecasts, by contrast, exhibit the lowest average MSFE of all forecasting strategies. The sum-of-the-parts models do not outperform the historical mean on average, but include the best of all forecasting strategies, the SOP-approach developed by Ferreira and Santa-Clara (2011) with a MSFE of 19.09. Overall, our results confirm that many forecasting strategies fail to outperform the historical mean when evaluated based on forecast errors.

The results in panel B.I of Table 2 indicate that an investor with mean-variance preferences can profoundly benefit from forecasting the equity premium. On average, most forecasting strategies – with exception of state-dependent regressions and diffusion indices – outperform the historical mean in terms of mean absolute returns. However, from panel B.II it becomes apparent that the historical mean serves as a more stringent benchmark model in a market timing context. Given the high positive average equity premium during the sample period, the prevailing mean model is identical to a buy-and-hold strategy. Consequently, the historical mean yields a mean absolute return of 0.8747\% per month over our out-of-sample period that is outperformed by only a few forecasting strategies. On average, most forecasting strategies – with exception of the sum-of-the-parts models – fail to outperform the historical mean based on mean absolute returns in a market timing context. Finally, the results in panel C of Table 2 indicate that, on average, most forecasting strategies generate higher mean risk-adjusted excess returns than the historical mean model, particularly for a mean-variance investor (panel C.I).

While the results in Table 2 provide a first indication as to which forecasting strategies might offer an improvement upon the historical mean, these analyses do not account for data snooping biases. To address these concerns, we apply Hansen (2005) SPA-test and its extensions. Since most advanced forecasting strategies claim to improve upon univariate predictive regressions, in a first step

\textsuperscript{10} Due to the rolling estimation scheme and the longer sample period, our results are not directly comparable to the results presented in Neely et al. (2014), who use an expanding windows scheme and data only up to 2011. When we apply a recursive estimation scheme, diffusion indices outperform the historical mean. However, as noted earlier, a recursive estimation scheme would violate the stationarity assumption of the SPA-test (see the discussion in Hansen, 2005).
we test each subset of advanced strategies separately, each time including the univariate predictive regressions in the test sample. However, as emphasized by Hansen (2005), testing different subsets of forecasting strategies is subject to data mining because the results do not incorporate the full set of strategies. Therefore, in a second step, we test the performance of all forecasting strategies jointly against the historical benchmark forecast. This latter framework imposes the most stringent test for superior predictive ability.

4.2. Test results based on squared forecast errors

In our first test, we assess whether any forecasting strategy can more accurately forecast the equity premium than the historical mean in terms of MSFE using Hansen (2005) SPA-test. Table 3 summarizes the results. Column (1) describes the set of forecasting strategies we draw from. Column (2) identifies the ‘most significant’ strategy, i.e., the strategy with the highest nominal $p$-value which results from a pairwise comparison of the strategy with the historical mean. In contrast to the $p$-values of the SPA-test, these $p$-values do not account for the entire set of strategies. Column (4) provides the consistent $p$-value and the lower and upper bound $p$-values of the SPA-test. If the consistent $p$-value is sufficiently small, we can reject the null hypothesis of the SPA-test, i.e., there is statistically significant evidence that at least one forecasting strategy is better than the historical mean in terms of MSFE. If the null hypothesis of the SPA-test is rejected, column (5) indicates the number of significant strategies identified by the step-SPA-test. We also implement the step-SPA(3)-test and record the number of respective significant strategies in column (6).

[Insert Table 3 here]

The first row in Table 3 summarizes the results of the SPA-test for the subset of univariate predictive regressions and forecast restrictions. The restricted forecast based on a volume-based indicator that signals the crossing of the one-month moving average of the ‘on balance’ volume and the respective twelve-months moving average, denoted as $VOL_{1-12}$ (rest.), is selected as the most significant strategy, improving upon univariate predictive regressions. However, the nominal $p$-value of 0.2642 indicates that this model is not able to outperform the historical mean when considered in isolation.
The consistent $p$-value of 0.9630 shows that the null hypothesis of the SPA-test cannot be rejected, i.e., there is no forecasting strategy than can outperform the benchmark.

Turning to the subset including the state-dependent regressions in the second row of Table 3, we note that all state-dependent regressions are dominated by a univariate predictive regression based on the term spread (TMS), which is selected as the most significant strategy in this subset. However, when we compare this strategy with the historical mean, its performance in terms of MSFE is not statistically different from the MSFE of the historical mean, as indicated by the nominal $p$-value of 0.5188. Accordingly, the null hypothesis of the SPA-test also cannot be rejected for this subset (with consistent $p$-value of 0.9968). Similar results apply for the subset of strategies that include the diffusion indices (fourth row).

In contrast, combinations forecasts (third row) can improve upon univariate predictive regressions, whereby the mean combination forecast based solely on fundamental variables, Mean (FUND), is selected as the most significant strategy (with nominal $p$-value of 0.2916). However, the consistent $p$-value of 0.9255 reveals that the null hypothesis of the SPA-test cannot be rejected, i.e., none of the combination forecasts is able to significantly outperform the historical mean when accounting for data snooping biases.

The results in the fifth row of Table 3 indicate that the SOP-approach also offers improvement upon univariate predictive regressions. It significantly outperforms the historical mean when considered in isolation (nominal $p$-value of 0.0232), but the null hypothesis of the SPA-test cannot be rejected at the 5% level of significance (consistent $p$-value of 0.2247).

Finally, when we turn to the results in the last row of Table 3, the SOP-approach is again selected as the most significant strategy when the full set of forecasting strategies is considered. However, there is no statistically significant evidence that any forecasting strategy is better than the historical mean in terms of MSFE. Even when tolerating up to three false rejections, none of the forecasting strategies is identified as superior.
Overall, the results in Table 3 indicate that many advanced forecasting strategies do not significantly improve upon univariate predictive regressions and do not outperform the historical mean once accounting for potential data snooping biases. Only Ferreira and Santa-Clara (2011) SOP-approach shows a marginal superiority compared to both univariate predictive regressions and the historical mean. Once correcting for data snooping biases, we are not able to identify any forecasting strategy that beats the historical mean out-of-sample in terms of MSFE.

4.3. Test results based on return-based performance measures

Table 4 shows the results of the data snooping tests using mean monthly absolute returns that an investor with mean-variance preferences and risk aversion coefficient $\gamma = 5$ (panel A), or a market timer who is either fully invested in the stock market or reverts to holding the risk-free asset otherwise (panel B), can generate over the out-of-sample period.

[Insert Table 4 here]

The results of panel A in Table 4 indicate that forecast restrictions (first row), forecast combinations (third row), and diffusion indices (fourth row) do not improve upon the univariate predictive regression based on TMS, which is selected as the most significant strategy in each subset (with nominal $p$-value of 0.0955). However, the null hypothesis of the SPA-test cannot be rejected for any of these subsets (all consistent $p$-values exceeding 5%). The state-dependent regression based on TMS (second row) marginally improves upon its univariate predictive regression (with nominal $p$-value of 0.0878). However, we cannot reject the null hypothesis of the SPA-test for this subset (with consistent $p$-value of 0.5785).

Turning to the subset including the sum-of-the-parts models (fifth row), the ESOP strategy using median combination forecasts, ESOP (Median), is selected as the most significant strategy in a pairwise comparison against the historical mean (with nominal $p$-value of 0.0012). However, we cannot reject the null hypothesis of the SPA-test at the 5% significance level (with consistent $p$-value of 0.1080). If we allow up to three false rejections, the step-SPA($J$)-test identifies all three ESOP models
as being superior to the historical mean in this subset. Finally, the null hypothesis of the SPA-test cannot be rejected for the full set of forecasting strategies with a consistent $p$-value of 0.1374.

If we assume market timing investor, panel B of Table 4 reveals that only the sum-of-the-parts models improve upon the univariate predictive regression based on TMS that is selected as the most significant model (nominal $p$-value of 0.4312) in each remaining subset. However, even when considered in isolation, the ESOP (Mean) strategy does not deliver a significant higher mean absolute return than the historical mean (nominal $p$-value of 0.1269). Consequently, we cannot reject the null hypothesis of the SPA-test neither for the subset including the sum-of-the-parts models (consistent $p$-value of 0.6333) nor for the entire set of forecasting strategies (consistent $p$-value of 0.6967).

As most investors not only consider returns but also risk in their performance evaluation, we additionally perform the data snooping tests using mean monthly risk-adjusted excess returns. Table 5 shows the results for an investor with mean-variance preferences and risk aversion coefficient $\gamma = 5$ (panel A), or a market timer who is either fully invested in the stock market or reverts to holding the risk-free asset otherwise (panel B).

[Insert Table 5 here]

As indicated by the results in panel A of Table 5, most advanced forecasting strategies fail to improve upon a univariate predictive regression based on the Treasury-bill rate (TBL), which is selected as the most significant strategy in the subsets including forecast restrictions (first row), state-dependent regressions (second row), and diffusion indices (fourth row). However, with a nominal $p$-value of 0.0905, this simple strategy does not significantly outperform the historical mean when considered in isolation. Therefore, as indicated by the large consistent $p$-values, all exceeding 5%, it is not able to outperform the historical mean when accounting for data snooping biases in any of these subsets. Turning to the subset including combination forecasts (third row), the most significant strategy in a pairwise comparison against the historical mean is the Mean (FUND) strategy (nominal $p$-value of 0.0410). However, we again cannot reject the null hypothesis of the SPA-test for this subset with a consistent $p$-value of 0.3087.
Nevertheless, the null hypothesis of the SPA-test is rejected for the subset including the sum-of-the-parts models (fifth row; with a consistent p-value of 0.0402). The ESOP (Median) strategy is selected as the most significant strategy with a nominal p-value of 0.0040, while the step-SPA(3)-test identifies all three ESOP strategies as superior to the historical mean. The null hypothesis of the SPA-test can marginally be rejected when accounting for all forecasting strategies under investigation (last row; with consistent p-value of 0.0528). The step-SPA(3)-test further confirms that all three ESOP strategies significantly outperform the historical mean after accounting for data snooping biases.

When evaluating the risk-adjusted excess performance of a market timer in panel B of Table 5, it becomes apparent that the univariate predictive regressions based on TMS is selected as the most significant strategy in all subsets exempt the one including the sum-of-the-parts models. However, even when considered in isolation, this univariate strategy does not significantly outperform the historical mean (with nominal p-value of 0.1754).

Again, the sum-of-the-parts models (fifth row) provide the only improvement upon the most significant univariate predictive regression. In particular, the ESOP (Mean) strategy exhibits the most significant performance in a pairwise comparison against the historical mean (with nominal p-value of 0.0211). Nevertheless, the outperformance of the sum-of-the-parts models is not robust when controlling for data snooping biases (with consistent p-value of 0.2086) in the respective subset. Consequently, the null hypothesis of the SPA-test can also not be rejected for the entire set of forecasting strategies (with consistent p-value of 0.2668). This result does not qualitatively change when tolerating up to three false rejections.

Taken together, our findings imply that investors who allocate their assets using a traditional mean-variance optimization procedure might benefit from forecasting the equity premium using the ESOP strategies rather than the historical mean at least on a risk-adjusted excess return basis. An investor aiming at timing the market might also benefit from the ESOP strategies’ forecasts. However, we are not able to rule out that the superior performance of the ESOP strategies in this latter application is merely due to luck.
5. Robustness checks

5.1. Variation of input parameters

Relaxation of short selling and leverage restrictions: As already discussed in section 3.3, we impose constraints on the weights used in the mean-variance optimization procedure. To determine whether an unconstrained investor might be able to gain more from the examined forecasting strategies, we repeat our empirical analyses without portfolio constraints. Somewhat unexpectedly, we observe an increase of the consistent $p$-values of the SPA-tests regardless of the performance measure used.

Coefficient of relative risk aversion: So far, we have assumed a conservative investor with a comparatively high coefficient of relative risk aversion ($\gamma = 5$). If we instead assume a more aggressive investor, indicated by a lower $\gamma$, we expect higher allocations to the stock market and probably higher portfolio returns (before transaction costs) for both the benchmark model and all forecasting strategies. If we repeat our empirical analyses with $\gamma = [1; 2]$ for the mean-variance investor, we still fail to reject the null hypothesis of the SPA-test under the mean absolute return criterion. However, using the step-SPA(3)-test, we are now able to identify up to three forecasting strategies (namely the ESOP strategies) that outperform the historical mean on a mean risk-adjusted excess return basis.

Transaction costs: On average, the forecasting strategies in our set have an average monthly turnover that is many times higher than the turnover of the historical mean model. To preclude that our results are not driven by the varying amount of the associated transactions costs, we repeat our analyses without considering transaction costs. As expected, the consistent $p$-values of the SPA-tests decrease. Nevertheless, our results remain largely unchanged, i.e., we fail to reject the null hypothesis under both the absolute and the risk-adjusted excess return criterion when assuming an investor purely timing the market. For an investor with mean-variance preferences, using the step-SPA(3)-test, we are now able to reject the null hypothesis at the 5% level of significance using mean risk-adjusted excess returns as the performance measure and identify all three ESOP strategies as significant.
5.2. Significance of forecasting strategies and publication date

So far, we have evaluated the out-of-sample performance of all forecasting strategies regardless of the order in which they were published. However, one might argue that the performance of any forecasting strategy should only be evaluated relative to the set of forecasting strategies that has been known at its publication date. In Figure 1, we plot the development of the consistent $p$-value of the SPA-test by sequentially updating the number of forecasting strategies using the mean absolute return that an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$ could achieve over the entire out-of-sample period as performance measure. The order of the forecasting strategies resembles the order in which they were published.\(^{11}\)

[Insert Figure 1 here]

As indicated by the scattered points, most forecasting strategies generate a higher mean absolute return over the out-of-sample period than the historical mean (dashed grey line) which confirms our earlier results. After including the univariate predictive regression based on the dividend-payout ratio (model #4), the maximum mean absolute return (dotted black line) increases to 0.72% per month and exceeds the return of the historical mean model. Including the univariate predictive regression based on TMS (model #11) increases the maximum mean absolute return further to 0.78% per month without rendering the consistent $p$-value (red line) significant. With an increasing number of forecasting strategies, but no ‘real’ improvement of the maximum performance, the consistent $p$-value increases gradually to a value of 0.59, displaying the effects of data snooping. The ESOP strategies (models #98-100) show remarkably high mean absolute returns, leading to a significant decrease in the consistent $p$-value. Overall, however, Figure 1 indicates that most forecasting strategies were not able to

\(^{11}\) We follow McLean and Pontiff (2016) and consider the publication date as the earlier of the actual publication date and the first time the paper appears on the SSRN. Based on their methodology, we develop the following ranking: 1. Univariate predictive regressions based on fundamental variables (see Goyal and Welch (2008) for a survey); 2. Forecast restrictions based on fundamental variables (Campbell and Thompson, 2005); 3. Combination forecasts based on fundamental variables (Rapach and Zhou, 2007); 4. SOP approach (Ferreira and Santa-Clara, 2008); 5. Univariate predictive regressions and forecast restrictions based on technical indicators, combination forecasts based on technical indicators and diffusion indices (Neely et al., 2009); 6. State-dependent predictive regressions based on fundamental variables and technical indicators (Huang et al., 2011); and 7. ESOP models (Bätje and Menkhoff, 2016). The order of the forecasting strategies within these groups is arbitrary.
significantly outperform the historical mean over the entire out-of-sample period, even without considering the number of forecasting strategies that were introduced afterwards.

5.3. Rolling sub-sample period analysis

One might argue that our findings support the view that the predictability of the equity premium is disappearing (Lettau and van Nieuwerburgh, 2008) or in fact illusory (Goyal and Welch, 2008). To address this question, we examine how the predictability of the equity premium has evolved over time. As of the end of each year, we use the out-of-sample performance of all forecasting strategies over the last ten years to conduct the SPA-tests, i.e., our first sub-sample period is from January 1966 to December 1975. Figure 2 illustrates our findings for an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$ using the mean absolute return over the respective ten-year sub-sample period as the performance measure.

[Insert Figure 2 here]

In panel A, we plot the mean absolute return over the respective ten-year sub-sample period of the benchmark model (dashed grey line) and each forecasting strategy $j$ (identified as scattered points). Confirming our previous results, there are many forecasting strategies that generate a higher mean absolute return than the historical mean model. We can identify some forecasting strategies that well outperform all other forecasting strategies, especially in the mid-80s to mid-90s and towards the end of the full sample period, as indicated by the outliers in the scatter plot. Moreover, panel A illustrates that forecasting the equity premium was especially beneficial during and after the global financial crisis when the historical mean model generated negative mean absolute returns. However, the performance of the forecasting strategies also became more dispersed during this period.

In panel B, we plot the development of the consistent $p$-value over the associated ten-year sub-sample period. If we presume that equity premium predictability is in fact disappearing over time, we would expect fewer or no forecasting strategies to outperform the benchmark model in the recent past, leading to a gradual increase of the consistent $p$-value. However, as shown in panel B, the consistent $p$-value fluctuates strongly between 0.7936 (ten-year sub-sample period ending in 2005) and 0.0301
(ten-year sub-sample period ending in 2014). We only observe five time windows with a statistically significant out-of-sample predictability, which are the ten-year periods ending in 1980, 1999, 2013, 2014, and 2015. During the time periods ending in 1980 and 1999, the observed predictability is highly sensitive to the selected window, i.e., predictability is no longer observable when we roll the ten-year window by only one year in either direction. Moreover, the best (or most significant) model varies over both time windows. While the univariate regression model based on the default yield spread (DFY) was the best model during the period ending in 1980, the forecasting model based on the T-bill rate (TBL) was the superior model during the period ending in 1999. Only the performance of the ESOP (Median) model seems to be more stable. In fact, this approach dominates statistically significant during the ten-year windows ending in 2013, 2014, and 2015.

6. Conclusions

In this study, we jointly examine the out-of-sample performance of (simple) univariate predictive regressions and a comprehensive set of advanced forecasting strategies relative to the historical mean. Almost all advanced forecasting strategies are documented in existing studies to be superior against the historical mean and also against some simple univariate predictive regressions. When we control in our empirical analysis for the data snooping bias within a multiple testing framework, however, we observe a statistically significant superiority only in very specific simulation setups, e.g., the ESOP (Median) prediction approach when we measure the economic profit of a mean-variance investor without short-sales and a 150% leverage restriction. Furthermore, our rolling window analysis reveals that even when we detect a data snooping resistant superior model, this superiority seems to be highly sub-sample dependent. Taken together, by focusing on the application of equity premium prediction, our findings support Harvey’s (2017) more general concern that many of the published results in financial economics will fail to hold up.
References


Bätje, F., and L. Menkhoff, 2016, Predicting the Equity Premium via its Components, Working paper.


Hsu, P.-H., Q. Han, W. Wu, and Z. Cao, 2017, Asset Allocation Strategies, the 1/N Rule, and Data Snooping, Working paper.


Tables

Table 1
List of forecasting strategies

<table>
<thead>
<tr>
<th>Forecasting techniques</th>
<th>Based on fundamental variables</th>
<th>Based on technical indicators</th>
<th>Based on all predictors</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate predictive regressions</td>
<td>14 fundamental variables</td>
<td>14 technical indicators</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>following Goyal and Welch (2008)</td>
<td>following Neely et al. (2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forecast restrictions</td>
<td>14 fundamental variables</td>
<td>14 technical indicators</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>following Goyal and Welch (2008)</td>
<td>following Neely et al. (2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>State-dependent predictive regressions</td>
<td>14 fundamental variables</td>
<td>14 technical indicators</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>following Goyal and Welch (2008)</td>
<td>following Neely et al. (2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Combination forecasts</td>
<td>Mean (FUND), Median (FUND),</td>
<td>Mean (TECH), Median (TECH),</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Trimmed Mean (FUND)</td>
<td>Trimmed Mean (TECH)</td>
<td>Mean (ALL), Median (ALL),</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Trimmed Mean (ALL)</td>
<td></td>
</tr>
<tr>
<td>Diffusion indices</td>
<td>PC-FUND</td>
<td>PC-TECH</td>
<td>PC-ALL</td>
<td>3</td>
</tr>
<tr>
<td>Sum-of-the-parts models</td>
<td></td>
<td></td>
<td>SOP, ESOP (Mean), ESOP</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(Median), ESOP (Trimmed</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean)</td>
<td></td>
</tr>
<tr>
<td>All forecasting strategies</td>
<td>46</td>
<td>46</td>
<td>8</td>
<td>100</td>
</tr>
</tbody>
</table>
This table reports the performance of the benchmark model and the descriptive statistics of the performance of the univariate predictive regressions, forecast restrictions, state-dependent regressions, combination forecasts, diffusion indices, and sum-of-the-parts models as well as all forecasting strategies combined over the out-of-sample period from January 1966 to December 2015. The benchmark model is the historical mean. The out-of-sample performance is evaluated based on the MSFE (panel A), the mean absolute return, and the mean risk-adjusted excess return. We assume I. an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$, and II. a market timer who is fully invested in the stock market if the equity premium forecast is positive and holds the risk-free asset otherwise. We impose roundtrip transaction costs of 50 basis points. The “best strategy” is the forecasting strategy with the highest performance measure, respectively.

<table>
<thead>
<tr>
<th>Panel A: MSFE</th>
<th>Benchmark</th>
<th>Univar. pred. regressions</th>
<th>Forecast restrictions</th>
<th>State-dep. pred. regressions</th>
<th>Combination forecasts</th>
<th>Diffusion indices</th>
<th>SOP models</th>
<th>All strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best strategy</td>
<td>TMS</td>
<td>TBL</td>
<td>MA2-12</td>
<td>Mean (FUND)</td>
<td>PC-TECH</td>
<td>SOP</td>
<td>SOP</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Mean absolute return</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Mean-variance investor</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>Best strategy</td>
</tr>
<tr>
<td>II. Market timer</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>Best strategy</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel C: Mean risk-adjusted excess return</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Mean-variance investor</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>Best strategy</td>
</tr>
<tr>
<td>II. Market timer</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>Best strategy</td>
</tr>
</tbody>
</table>
Table 3
Results of the data snooping tests based on MSFE

This table reports the results of the SPA-test, the step-SPA-test and the step-SPA(3)-test for the pre-specified error rate $\alpha = 5\%$ using the MSFE over the out-of-sample period from January 1966 to December 2015 as a performance measure. The benchmark model is the historical mean. Column (1) indicates the set of forecasting strategies the data snooping tests are applied to. The “most significant strategy” is the forecasting strategy with the lowest nominal $p$-value. The nominal $p$-values result from the pairwise comparisons of the most significant strategy with the historical mean. These $p$-values ignore the search over all strategies that preceded the selection of the strategy being compared to the historical mean, i.e., they do not account for the entire set of forecasting strategies. In addition, we report the consistent $p$-value of the SPA-test, as well as its lower and upper bounds. Finally, we state the number of outperforming strategies identified by the step-SPA- and step-SPA(3)-tests. “-” indicates that there is no outperforming strategy at the 5\% level of significance.

<table>
<thead>
<tr>
<th>Forecasting strategies</th>
<th>Most significant strategy</th>
<th>Nominal $p$-value</th>
<th>SPA $p$-value [lower bound; upper bound]</th>
<th># of strategies identified by step-SPA</th>
<th># of strategies identified by step-SPA(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>VOL/1-12 (rest.)</td>
<td>0.2642</td>
<td>0.9630 [0.8604; 0.9630]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions State-dependent pred. regressions</td>
<td>TMS</td>
<td>0.5188</td>
<td>0.9968 [0.9272; 0.9968]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Combination forecasts</td>
<td>Mean (FUND)</td>
<td>0.2916</td>
<td>0.9255 [0.6996; 0.9255]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Diffusion indices</td>
<td>TMS</td>
<td>0.5188</td>
<td>0.9934 [0.9236; 0.9934]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Sum-of-the-parts models</td>
<td>SOP</td>
<td>0.0232</td>
<td>0.2247 [0.1029; 0.2247]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>All strategies</td>
<td>SOP</td>
<td>0.0232</td>
<td>0.3304 [0.1983; 0.3325]</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 4
Results of the data snooping tests based on mean absolute returns

This table reports the results of the SPA-test, the step-SPA-test and the step-SPA(3)-test for the pre-specified error rate $\alpha = 5\%$ using the mean absolute returns over the out-of-sample period from January 1966 to December 2015 as a performance measure. The benchmark model is the historical mean. We assume an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$ (panel A), and a market timer who is fully invested in the stock market if the equity premium forecast is positive and holds the risk-free asset otherwise (panel B). We impose roundtrip transaction costs of 50 basis points. Column (1) indicates the set of forecasting strategies the data snooping tests are applied to. The “most significant strategy” is the forecasting strategy with the lowest nominal $p$-value. The nominal $p$-values result from the pairwise comparisons of the most significant strategy with the historical mean. These $p$-values ignore the search over all strategies that preceded the selection of the strategy being compared to the historical mean, i.e., they do not account for the entire set of forecasting strategies. In addition, we report the consistent $p$-value of the SPA-test, as well as its lower and upper bounds. Finally, we state the number of outperforming strategies identified by the step-SPA- and step-SPA(3)-tests. “-“ indicates that there is no outperforming strategy at the 5% level of significance.

<table>
<thead>
<tr>
<th>Forecasting strategies</th>
<th>Most significant strategy</th>
<th>Nominal $p$-value</th>
<th>SPA $p$-value [lower bound; upper bound]</th>
<th># of strategies identified by step-SPA</th>
<th># of strategies identified by step-SPA(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Mean-variance investor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>TMS</td>
<td>0.0955</td>
<td>0.5313 [0.3888; 0.5499]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions</td>
<td>TMS (st.dep.)</td>
<td>0.0878</td>
<td>0.5785 [0.4278; 0.5977]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Combination forecasts</td>
<td>TMS</td>
<td>0.0955</td>
<td>0.5324 [0.3937; 0.5508]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Diffusion indices</td>
<td>TMS</td>
<td>0.0955</td>
<td>0.5370 [0.3912; 0.5553]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Sum-of-the-parts models</td>
<td>ESOP (Median)</td>
<td>0.0012</td>
<td>0.1080 [0.0787; 0.1132]</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>All strategies</td>
<td>ESOP (Median)</td>
<td>0.0012</td>
<td>0.1374 [0.0975; 0.1449]</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Forecasting strategies</th>
<th>Most significant strategy</th>
<th>Nominal p-value</th>
<th>SPA p-value [lower bound; upper bound]</th>
<th># of strategies identified by step-SPA</th>
<th># of strategies identified by step-SPA(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel B: Market timer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions</td>
<td>TMS</td>
<td>0.4312</td>
<td>0.9546 [0.7912; 0.9659]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Forecast restrictions</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions</td>
<td>TMS</td>
<td>0.4312</td>
<td>0.9697 [0.8117; 0.9780]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>State-dependent pred. regressions</td>
<td>TMS</td>
<td>0.4312</td>
<td>0.9591 [0.8279; 0.9692]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Combination forecasts</td>
<td>TMS</td>
<td>0.4312</td>
<td>0.9576 [0.7955; 0.9677]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions</td>
<td>TMS</td>
<td>0.4312</td>
<td>0.9576 [0.7955; 0.9677]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Diffusion indices</td>
<td>TMS</td>
<td>0.4312</td>
<td>0.9576 [0.7955; 0.9677]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions</td>
<td>ESOP (Mean)</td>
<td>0.1269</td>
<td>0.6333 [0.3939; 0.6684]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sum-of-the-parts models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All strategies</td>
<td>ESOP (Mean)</td>
<td>0.1269</td>
<td>0.6967 [0.4435; 0.7345]</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 5

Results of the data snooping tests based on mean risk-adjusted excess returns

This table reports the results of the SPA-test, the step-SPA-test and the step-SPA(3)-test for the pre-specified error rate $\alpha = 5\%$ using the mean risk-adjusted excess returns over the out-of-sample period from January 1966 to December 2015 as a performance measure. The benchmark model is the historical mean. We assume an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$ (panel A), and a market timer who is fully invested in the stock market if the equity premium forecast is positive and holds the risk-free asset otherwise (panel B). We impose roundtrip transaction costs of 50 basis points. Column (1) indicates the set of forecasting strategies the data snooping tests are applied to. The “most significant strategy” is the forecasting strategy with the lowest nominal $p$-value. The nominal $p$-values result from the pairwise comparisons of the most significant strategy with the historical mean. These $p$-values ignore the search over all strategies that preceded the selection of the strategy being compared to the historical mean, i.e., they do not account for the entire set of forecasting strategies. In addition, we report the consistent $p$-value of the SPA-test, as well as its lower and upper bounds. Finally, we state the number of outperforming strategies identified by the step-SPA- and step-SPA(3)-tests. “-“indicates that there is no outperforming strategy at the 5% level of significance.

<table>
<thead>
<tr>
<th>Forecasting strategies</th>
<th>Most significant strategy</th>
<th>Nominal $p$-value</th>
<th>SPA $p$-value [lower bound; upper bound]</th>
<th># of strategies identified by step-SPA</th>
<th># of strategies identified by step-SPA(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Mean-variance investor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>TBL</td>
<td>0.0905</td>
<td>0.5365 [0.4056; 0.5577]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions State-dependent pred. regressions</td>
<td>TBL</td>
<td>0.0905</td>
<td>0.6053 [0.4576; 0.6280]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Combination forecasts</td>
<td>Mean (FUND)</td>
<td>0.0410</td>
<td>0.3087 [0.2309; 0.3239]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Diffusion indices</td>
<td>TBL</td>
<td>0.0905</td>
<td>0.5456 [0.4174; 0.5666]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Univariate pred. regressions Sum-of-the-parts models</td>
<td>ESOP (Median)</td>
<td>0.0040</td>
<td>0.0402 [0.0317; 0.0417]</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>All strategies</td>
<td>ESOP (Median)</td>
<td>0.0040</td>
<td>0.0528 [0.0399; 0.0550]</td>
<td>-</td>
<td>3</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Forecasting strategies</th>
<th>Most significant strategy</th>
<th>Nominal p-value</th>
<th>SPA p-value [lower bound; upper bound]</th>
<th># of strategies identified by step-SPA</th>
<th># of strategies identified by step-SPA(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel B: Market timer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>TMS</td>
<td>0.1754</td>
<td>0.8125 [0.6417; 0.8333]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>TMS</td>
<td>0.1754</td>
<td>0.8668 [0.6818; 0.8832]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>TMS</td>
<td>0.1754</td>
<td>0.8235 [0.6596; 0.8424]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>TMS</td>
<td>0.1754</td>
<td>0.8195 [0.6464; 0.8393]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Univariate pred. regressions Forecast restrictions</td>
<td>ESOP (Mean)</td>
<td>0.0211</td>
<td>0.2086 [0.1408; 0.2163]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All strategies</td>
<td>ESOP (Mean)</td>
<td>0.0211</td>
<td>0.2668 [0.1631; 0.2766]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figures

Figure 1. Development of the consistent \( p \)-value subject to the number of forecasting strategies. For each forecasting strategy \( j \), the mean absolute return over the out-of-sample period from January 1966 to December 2015 for an investor with mean-variance preferences and relative risk aversion coefficient \( \gamma = 5 \) is identified by the scattered points (left scale). For comparison, the current maximum mean absolute return, denoted as ‘best strategy’ (dotted black line), and the mean absolute return of the benchmark strategy (dashed grey line) are plotted on the left scale. Finally, the red line shows the development of the consistent \( p \)-value of the SPA-test including all forecasting strategies up to forecasting strategy \( j \) (right scale).
Figure 2. Rolling sub-sample analysis. In panel A, the mean absolute return of each forecasting strategy $j$ over the ten-year sub-sample period for an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$ is identified by the scattered points. For comparison, the mean absolute return of the benchmark model is plotted as a dashed red line. In panel B, the black line shows the development of the consistent $p$-value of the SPA-test over the respective ten-year sub-sample period. The dashed grey line identifies the 5% level of significance.
Appendix A: Definition of predictor variables

**DP**  Dividend-price ratio, calculated as the log of twelve-month moving sum of dividends paid on S&P 500 index minus log of stock prices

**DY**  Dividend yield, calculated as the log of twelve-month moving sum of dividends paid on S&P 500 index minus log of lagged stock prices

**EP**  Earnings-price ratio, calculated as the log of twelve-month moving sum of earnings paid on S&P 500 index minus log of stock prices

**DE**  Dividend-payout ratio, calculated as the log of twelve-month moving sum of dividends paid on S&P 500 index minus log of twelve-month moving sum of earnings

**RVOL**  Equity premium volatility based on twelve-month moving standard deviation estimator following Mele (2007)

**BM**  Book-to-market value ratio for the Dow Jones Industrial Average

**NTIS**  Net equity expansion, calculated as the ratio of a twelve-month moving sum of net equity issues by NYSE-listed stocks to the total end-of-year market capitalization of New York Stock Exchange (NYSE) stocks

**TBL**  Interest rate on a three-month Treasury bill

**LTY**  Long-term government bond yield

**LTR**  Return on long-term government bonds

**TMS**  Term spread, calculated as the long-term yield minus the Treasury bill rate

**DFY**  Default yield spread, calculated as the difference between Moody’s BAA- and AAA-rated corporate bond yields

**DFR**  Default return spread, calculated as the long-term corporate bond return minus the long-term government bond return

**INFL**  Inflation, calculated from CPI for all urban consumers, lagged by one month to account for the delay in CPI releases

**MAs-l**  Moving-average indicator, calculated as the difference between short-term (s) and long-term (l) moving averages of the stock price

\[
M_{s,l} = \begin{cases} 
1 & \text{if } M_{A_{s,t}} \geq M_{A_{l,t}} \\
0 & \text{if } M_{A_{s,t}} < M_{A_{l,t}} 
\end{cases} \quad \text{for } s = [1,2,3] \text{ and } l = [9,12]
\]

where \( M_{A_{j,t}} = \frac{1}{j} \sum_{i=0}^{j-1} P_{t-i} \) for \( j = s, l \)

**MOMm**  Momentum-indicator, calculated as the difference between the current stock price and the stock price \( m \) months ago

\[
M_{O,M_{t}} = \begin{cases} 
1 & \text{if } P_{t} \geq P_{t-m} \\
0 & \text{if } P_{t} < P_{t-m} 
\end{cases} \quad \text{for } m = [9,12]
\]

**VOLs-l**  Volume-based indicator, calculated as the difference between short-term (s) and long-term(l) moving averages of the ‘on balance’ volume (Granville, 1963)
\[ \text{VOLS}_{s,t} = \begin{cases} 1 & \text{if } \text{MA}^{OBV}_{s,t} \geq \text{MA}^{OBV}_{l,t} \\ 0 & \text{if } \text{MA}^{OBV}_{s,t} < \text{MA}^{OBV}_{l,t} \end{cases} \text{ for } s = [1, 2, 3] \text{ and } l = [9, 12] \]

where \( \text{MA}^{OBV}_{j,t} = \frac{1}{j} \sum_{i=0}^{j-1} \text{OBV}_{t-i} \) for \( j = s, l \)

where \( \text{OBV}_t = \sum_{k=1}^{t} \text{VOL}_k \text{D}_k \) where \( \text{D}_k = \begin{cases} 1 & \text{if } \text{P}_k \geq \text{P}_{k-1} \\ -1 & \text{if } \text{P}_k < \text{P}_{k-1} \end{cases} \)