Predicting the equity premium out-of-sample: Are there superior forecasting strategies?

Viktoria-Sophie Bartsch[a], Hubert Dichtl[b], Wolfgang Drobetz[c], and Andreas Neuhierl[d,‡]

First version: January 2017

Abstract

This study analyzes the performance of a comprehensive set of equity premium forecasting strategies that have been shown to outperform the historical average forecast out-of-sample and to improve upon conventional predictive regressions when tested in isolation. Controlling for potential data snooping biases using Hansen's (2005) SPA-test and its stepwise extension, we find that only the sum-of-the-parts model proposed by Ferreira and Santa-Clara (2011) outperforms the historical average forecast in terms of mean squared forecast errors. However, several advanced forecasting strategies are able to produce statistically significant economic gains when used in a traditional mean-variance asset allocation, even after controlling for data snooping biases. In contrast, the benefits for an investor aiming at timing the market are limited.

Keywords: Equity risk premium prediction; economic variables; technical indicators; forecast combination; diffusion index; regime shifts; asset allocation; data snooping bias

JEL classification codes: G11, G12, G14

[a] Faculty of Business, Hamburg University, Moorweidenstr. 18, 20148 Hamburg, Germany.

[b] Faculty of Business, Hamburg University, Moorweidenstr. 18, 20148 Hamburg, Germany.

[c] Faculty of Business, Hamburg University, Moorweidenstr. 18, 20148 Hamburg, Germany.

[d] Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556, USA.

‡ *Acknowledgments*: We would like to thank Alexander Hillert for helpful comments.

1. Introduction

While in-sample predictability of the equity premium is largely undisputed (Campbell, 2000), out-of-sample predictability is still a controversy. Although the merits of out-of-sample validation is disputable (Inoue and Kilian, 2004), most investors are ultimately interested whether the prediction of the equity premium pays off. This is also the point of view we take in this study, which jointly examines the out-of-sample performance of simple predictive regressions and advanced forecasting strategies in order to assess whether any of these strategies is truly superior to the historical average forecast out-of-sample.

Traditionally, academic studies focused on the predictive ability of fundamental variables such as the dividend-price ratio (Campbell and Shiller, 1989), book-to-market ratio (Kothari and Shanken, 1997), or interest rates (Fama and Schwert, 1977) in a simple predictive regression framework. While the existing evidence is predominantly in-sample[1], recently an increasing number of studies have tested the out-of-sample predictability of the equity premium. One of the most influential studies is the one by Goyal and Welch (2008). They examine both the in-sample and out-of-sample performance of many popular fundamental variables and conclude that, based on the out-of-sample validations, most conventional predictive regressions seem unstable and do not benefit investors in timing the market.

Taking up the challenge posed by Goyal and Welch (2008) to "explore more variables and/or more sophisticated models," several recent studies have proposed more advanced forecasting strategies. For example, one way to improve upon conventional predictive regressions is to exploit the predictive ability of alternative predictors such as lagged industry returns (Hong, Torous and Valkanov, 2007) or technical indicators (Neely et al., 2014). However, when considering alternative predictor variables, out-of-sample predictability might be especially challenging to uncover due to model uncertainty (i.e., the uncertainty which forecasting strategy is effectively "the best") and parameter instabil-

[1] See Rapach and Zhou (2013) for a survey of the literature on stock return forecasting.

ity (i.e., changing parameters due to business cycle fluctuations) (Pesaran and Timmermann, 1995). Therefore, a new strand of literature on out-of-sample equity premium predictability aims to improve forecasting performance by directly addressing these two challenges. These more advanced forecasting techniques include strategies based on diffusion indices (Ludvigson and Ng, 2007; Neely et al., 2014), combination forecasts (Timmermann, 2006; Rapach, Strauss, and Zhou, 2010), forecast restrictions (Campbell and Thompson, 2008; Ferreira and Santa-Clara, 2011), or regime-shifts (Henkel, Martin, and Nardari, 2011; Dangl und Halling; 2012; Huang et al., 2016). All of these strategies have been shown to improve upon conventional predictive regressions and outperform the historical average forecast when tested in isolation.

So far, a systematic comparison of these advanced forecasting strategies is yet pending. Our paper is the first to jointly reexamine the out-of-sample performance of conventional predictive regressions and advanced forecasting strategies in order to assess whether any of the forecasting strategies is truly superior to the historical average forecast out-of-sample. When comparing the performance of several forecasting strategies, data snooping concerns naturally arise. As noted by Sullivan, Timmermann, and White (1999), most forecasting strategies are tested on a single data set, and their perceived outperformance may be purely from chance rather than due to any genuine merit. To control for data snooping biases when testing for a possible superiority of certain forecasting strategies, Hansen (2005) proposes a test for superior predictive ability (SPA) that allows for the necessary correction. In our empirical analysis, we use Hansen's (2005) SPA-test as well as its stepwise extension by Hsu, Hsu, and Kuan (2010) to ensure that our results are robust against data snooping biases.

We analyze the performance of a comprehensive set of equity premium forecasting strategies that have been shown to outperform the historical average forecast out-of-sample and to improve upon conventional predictive regressions when tested in isolation. When controlling for data snooping biases, we find that only the sum-of-the-parts model proposed by Ferreira and Santa-Clara (2011) outper-

forms the historical average forecast in terms of mean squared forecast errors. In line with the results of Leitch and Tanner (1991), several forecasting strategies are able to produce statistically significant economic gains when used in a traditional mean-variance asset allocation, even after controlling for data snooping biases. However, the benefits for an investor aiming at timing the market are limited.

The remainder of our paper is structured as follows: Section 2 gives an overview of equity premium forecasting strategies and how of we implement the selected strategies in our empirical analysis. Section 3 briefly describes the data snooping testing approaches. Section 4 discusses our main empirical results, and section 5 presents various robustness checks. Section 6 concludes.

2. Forecasting strategies

As noted by Sullivan, Timmermann, and White (1999), data snooping tests are especially 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 conventional predictive regression models and a comprehensive collection of advanced forecasting strategies. In doing so, it is imperative to trade-off between including too many 'irrelevant' strategies, thereby decreasing the power of the test (Hansen, 2005), and including too few strategies, thereby overstating its statistical significance. Following Rapach and Zhou (2013), we survey forecasting strategies that have become popular in the literature.

Conventional predictive regressions: A simple predictive regression model is given as follows:

$$r_{t+1} = \alpha + \beta x_t + \varepsilon_{t+1} \tag{1}$$

where r_{t+1} is the equity premium from period t to t+1, x_t a variable available at time t that is expected to predict the future equity premium, and ε_{t+1} a zero-mean disturbance term. We define the monthly (log) equity premium as the continuously compounded stock return of the S&P 500 (includ-

ing dividends) minus the log return on a risk-free bill. Using the updated monthly data set provided by Goyal and Welch (2008),[2] we construct a set of 14 fundamental variables representative of the literature. A detailed description of the variables is provided in Appendix [A].

The out-of-sample predictions of 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 $\hat{r}_{t+1}$. In our empirical analysis, we employ a rolling (instead of a recursive) 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's (2005) SPA-test.

Alternative predictors: Hong, Torous, and Valkanov (2007) propagated the predictive ability of lagged industry returns to forecast the broader stock market due to gradual information diffusion across markets. We use monthly returns on 38 value-weighted industry portfolios taken from Kenneth French's website, albeit we have to drop five industries due to missing observations.[3] The lagged industry returns are then used as alternative predictors in equation (1).

Neely et al. (2014) affirmed the predictive power of technical indicators. One of the most popular trend-following strategies that has been successfully applied to various markets is based on moving averages. A simple moving average indicator *MAs-l* generates a signal based on the cross-over of two moving averages with short length s and long length l, respectively, and is defined as

$$S_{i,t} = \begin{cases} 1 \; if \; MA_{s,t} \geq MA_{l,t} \\ 0 \; if \; MA_{s,t} < MA_{l,t} \end{cases} \text{ where } \; MA_{j,t} = \frac{1}{j}\sum_{i=0}^{j-1} P_{t-i} \text{ for } \; j = s, l \tag{2}$$

[2] Data are available at Amit Goyal's webpage, at http://www.hec.unil.ch/agoyal/.

[3] As in Hong, Torous and Valkanov (2007), the five industries excluded from the analysis are GARBG (sanitary services), STEAM (steam supply), WATER (irrigation systems), GOVT (public administration), and OTHER.

Moskowitz, Ooi and Pedersen (2012) study 58 liquid securities and document a strong relationship between a security's next month excess return and its past return. We construct an indicator *MOMm* that captures this time-series momentum:

$$S_{i,t} = \begin{cases} 1 \; if \; P_t \geq P_{t-m} \\ 0 \; if \; P_t < P_{t-m} \end{cases} \tag{3}$$

In addition to historical prices, volume data play a significant role in technical analysis. For example, Blume, Easley, and O'Hara (1994) demonstrate that volume data provide further information that cannot be deduced from the price statistic alone. In order to use the information contained in the volume data, we follow Granville (1963) and construct a technical indicator *VOLs-l* based on the 'on-balance' volume:

$$S_{i,t} = \begin{cases} 1 \;\; if \;\; MA_{s,t}^{OBV} \geq MA_{l,t}^{OBV} \\ 0 \;\; if \;\; MA_{s,t}^{OBV} < MA_{l,t}^{OBV} \end{cases} \text{ where } MA_{j,t}^{OBV} = \frac{1}{j}\sum_{i=0}^{j-1} OBV_{t-i} \text{ for } j = s, l \tag{4}$$

The on-balance volume OBV_t combines both volume and price information:

$$OBV_t = \sum_{k=1}^{t} VOL_k \, D_k \text{ where } D_k = \begin{cases} 1 \;\; if \;\; P_k \geq P_{k-1} \\ -1 \;\; if \;\; P_k < P_{k-1} \end{cases} \tag{5}$$

In our empirical tests, we construct the moving-average and volume-based indicators based on short lengths $s = 1, 2, 3$ and $l = 9, 12$ months, respectively. For the momentum indicator, we consider price information lagged by $m = 9, 12$ months. For all technical indicators, we use the S&P 500 (excluding dividends) as the price index P_t and monthly volume data VOL_k from Yahoo Finance.[4] To allow for a direct comparison, we follow Neely et al. (2014) and transform the technical indicators S_t to point forecasts of the equity premium by replacing x_t in the predictive regression model in equation (1) by the respective indicator $S_{i,t}$ in equations (2), (3), and (4).

[4] Volume data are available at http://de.finance.yahoo.com.

Diffusion indices: To avoid over-parametrization, several studies adopt a diffusion indices approach that assumes a factor model structure for the variables $x_{i,t}$ and use estimates of the common factors as predictors in a simple predictive regression model. For example, Ludvigson and Ng (2007) extract three common factors, denoted as 'volatility', 'risk premium' and 'real' factors, from a comprehensive set of macroeconomic and financial variables and find that the subsequent diffusion indices forecasts exhibit significant out-of-sample predictive power. We follow Stock and Watson (2006) and estimate the common factors using principal component analysis based on either the full set of fundamental variables (PC-FUND), based on technical indicators (PC-TECH) or fundamental variables and technical indicators combined (PC-ALL). As noted by Rapach and Zhou (2013), for forecasting purposes it is prudent to keep the number of common factors small to avoid over-parametrization. Therefore, we consider only the first principal component. The extracted principal components then serve as independent variables in the predictive regression model in equation (1).

Forecast restrictions: Campbell and Thompson (2008) argue that the performance of conventional predictive regressions can be substantially improved by imposing weak restrictions on the signs of coefficients and return forecasts. In our empirical analysis, we restrict the equity premium forecasts obtained from the conventional predictive regressions to be non-negative.

Combination forecasts: Timmermann (2006) argues that combining individual forecasts proves fruitful as it provides diversification gains compared to relying on forecasts from a single forecasting strategy, captures different degrees of adaptability of forecasting strategies to structural breaks and guards against model misspecification. Rapach, Strauss, and Zhou (2010) show that combinations of individual forecasts deliver statistically and economically significant out-of-sample results due to reduced model uncertainty and parameter instability. The combination forecasts are weighted averages of *N* individual forecasts that are estimated using the predictive regression model in equation (1):

$$\hat{r}_{combination,t+1} = \sum_{i=1}^{N} \omega_{i,t}\, \hat{r}_{i,t+1} \tag{6}$$

In our empirical analysis, we combine the individual forecasts based on the full set of fundamental variables ($N = 14$) and 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.

Regime shifts: As noted by Paye and Timmermann (2006) and Rapach and Wohar (2006b), the data-generating process for stock returns is subject to substantial parameter instability due to structural breaks. To account for parameter instability, several forecasting strategies have been suggested. 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 document that their model produces significant utility gains in asset allocation decisions. Exploiting the time-variation of several 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 expansions, while fundamental variables provide useful information in recessionary periods.

Most recently, Huang et al. (2016) use a state-dependent predictive regression model that was introduced by Boyd, Hu, and Jagannathan (2005) and addresses the critique by Lettau and Van Nieuwerburgh (2008) that regime-shifting models perform poorly out-of-sample due to unreliable estimates of the timing and the size of regime shifts. Their results indicate 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. Switching between a finite number of states, Dangl and Halling

(2012) test a time-varying model that allows for parameters to evolve as random walks from period to period. Their model is able to significantly outperform the historical average forecast out-of-sample.

In our analysis, we apply the state-dependent predictive regression approach of Huang et al. (2016) to the full set of fundamental variables. Following Cooper, Gutierrez, and Hameed (2004), the market states are identified based on past return information:

$$r_{t+1} = \alpha + \beta_{good} x_t I_{good,t} + \beta_{bad} x_t (1 - I_{good,t}) + \varepsilon_{t+1} \tag{7}$$

To proxy for the market state, we follow Huang et al. (2016) and construct the indicator I_t to take the value of one when the past six-month (log) equity premium is non-negative, and zero otherwise. As noted by these authors, this indicator-based identification of states coincides with the results obtained by Henkel, Martin, and Nardari (2001) using Bayesian estimation.

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, namely 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}):

$$r_{t+1} = gm_{t+1} + ge_{t+1} + dp_{t+1} - r_{f,t+1} \tag{8}$$

Using this return decomposition, we follow the simplest version of the SOP method by assuming no multiple growth, estimating the growth rate of earnings as a 20-year moving average of growth in earnings per share, and modelling the dividend-price ratio as a random walk:

$$\hat{r}_{t+1}^{SOP} = \overline{ge}_t + dp_t - r_{f,t} \tag{9}$$

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, $\widehat{gm}_{i,t+1}$, and the growth rate of earnings, $\widehat{ge}_{i,t+1}$, are estimated by simple 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). Finally, the equity premium forecast is obtained by summing the (combined) component forecasts, assuming that the dividend-price ratio and the risk-free rate follow a random walk:

$$\hat{r}_{t+1}^{ESOP} = \widehat{gm}_{t+1}^{combination,FUND} + \widehat{ge}_{t+1}^{combination,TECH} + dp_t - r_{f,t} \tag{10}$$

3. Testing methods

When considering a large number of possible forecasting strategies, data snooping is a natural concern (Lo and MacKinlay, 1990). In the context of the out-of-sample equity premium predictability, various testing procedures have been developed to avoid spurious statistical inference. For example, Rapach and Wohar (2006a) apply the McCracken (2007) MSE-F statistic that tests the null hypothesis that the mean squared forecast error (MSFE) of the historical average forecast is less than or equal to the minimum MSFE of all considered forecasting strategies against the one-sided (upper-tail) hypothesis. Computing critical values for the maximum statistics using a bootstrap procedure, they conclude that out-of-sample equity premium predictability is reasonably robust to data snooping concerns. More recently, Neely et al. (2014) implement a modified version of White's (2000) reality check based on a wild fixed-regressor bootstrap procedure developed by Clark and McCracken (2012) to show that their forecasting strategy based on diffusion indices has significant out-of-sample predictive power.

However, most of the applied test statistics involve tests for equal predictive ability, i.e., testing whether the predictive ability of some forecasting strategy is the same as the one 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 in fact sample-dependent.

Building on earlier work by White (2000), Hansen (2005) propose a test for superior predictive ability (SPA-test) that allows for a comprehensive comparison of forecasting strategies, while ensuring that the results are robust to data snooping effects and that including poor or irrelevant strategies do not influence the power of the test. The predictive ability of each forecasting strategy is defined in terms of its expected loss $E(L_k)$. 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 a first test we compare the performance of the forecasting strategies with the performance of a benchmark model based on the MSFE loss function $L_{k,t} = \left(r_t - \hat{r}_{k,t}\right)^2$, where r_t is the realized (log) equity premium, and $\hat{r}_{k,t}$ is the (log) equity premium forecast based on forecasting strategy *k*.

However, as shown by Leitch and Tanner (1991), there is only a weak relationship between MSFE and 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. Therefore, we consider both the (negative) absolute return, $L_{k,t} = -r^*_{k,t}$ based on the equity premium forecast of strategy *k*, and the risk-adjusted excess return $L_{k,t} = -\frac{r^*_{k,t} - r_{f,t}}{\sigma_k}$, where σ_k is the volatility of the excess return of strategy *k*, as adequate loss functions.

The loss values are then transformed into relative performance variables, defined as $d_{k,t} = L_{0,t} - L_{k,t}$, where $L_{0,t}$ denotes the loss function of the benchmark model. The historical average of the equity premium serves as a natural benchmark model that indicates a constant expected equity premi-

um (Goyal and Welch, 2008). In our empirical application, we calculate the historical average forecast as the average of the equity premium over the same rolling window used to estimate the conventional predictive regressions and advanced forecasting strategies.

Next, when testing for superior predictive ability, we test the null hypothesis that the benchmark model is not inferior to any alternative forecasting strategy:

$$H_0: \max_{k=1,\dots,m} E(d_{k,t}) \equiv \mu_k \leq 0 \tag{11}$$

If the null can be rejected, there is at least one forecasting strategy that outperforms the benchmark. As a test statistic, Hansen (2005) proposes the studentized test statistic:

$$T_n^{SPA} = max\left(\max_{k=1,\dots,m} \frac{n^{1/2}\bar{d}_k}{\widehat{\omega}_k}, 0\right) \tag{12}$$

where $\bar{d}_k = n^{-1}\sum_{t=1}^{n} d_{k,t}$ denotes the average relative performance of forecasting strategy *k*, and $\widehat{\omega}_k^2$ is a consistent estimate of $\omega_k^2 = var(n^{1/2}\bar{d}_k)$. To ensure that poor forecasting strategies do not asymptotically influence the test statistic, Hansen (2005) advocates invoking a null distribution based on $N(\hat{\mu}, \widehat{\Omega})$, where $\hat{\mu}_k$ is an estimator for μ_k given as $\hat{\mu}_k = \bar{d}_k \mathbb{1}_{\{n^{1/2}\bar{d}_k/\widehat{\omega}_k \leq -\sqrt{log\ log\ n}\}}$.

To approximate the distribution of the test statistic, we follow Hansen (2005) and use the stationary bootstrap of Politis and Romano (1994). For each strategy, we generate $b = 1, \dots, \mathrm{B}$ resamples of $d_{k,t}$ by drawing geometrically distributed blocks with a mean block length of q^{-1}. We set the smoothing parameter $q = 0.5$ and generate $\mathrm{B} = 10{,}000$ bootstrap resamples. The bootstrapped variables $d_{k,b,t}^*$ are re-centered about $\hat{\mu}_k$ as $Z_{k,b,t}^* = d_{k,b,t}^* - g(\bar{d}_k)$, and the studentized test statistic under the bootstrap is calculated as $T_{b,n}^{SPA^*} = max\left(max_{k=1,\dots,m} \frac{n^{1/2}\bar{Z}_{k,b}^*}{\widehat{\omega}_k}, 0\right)$, where $\bar{Z}_{k,b}^* = n^{-1}\sum_{t=1}^{n} Z_{k,b,t}^*$. A consistent estimate of the *p*-value is then given by:

$$\hat{p}_{SPA} = \sum_{b=1}^{B} \frac{\mathbb{1}_{\left\{T_{b,n}^{SPA^*} > T_n^{SPA}\right\}}}{B} \tag{13}$$

where the null hypothesis is rejected for small *p*-values. As shown by Hansen (2005), an upper and a lower bound for the *p*-value can be obtained by re-centering about $\hat{\mu}_k^u = 0$, which assumes that all competing forecasting strategies are as good as the benchmark model, and $\hat{\mu}_k^l = min(\bar{d}_k, 0)$, 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.

If the null hypothesis is rejected, we employ the stepwise extension of the SPA-test (step-SPA-test) developed by Hsu, Hsu and Kuan (2010) in order to identify additional significant forecasting strategies. First, we re-arrange the forecasting strategies in descending order of their average relative performance $\bar{d}_k$ and reject the best model if the studentized test statistic is greater than the critical value bootstrapped from the entire set of forecasting strategies. Second, we remove $\bar{d}_k$ of the rejected model and compute a critical value bootstrapped from the subset of remaining forecasting strategies. We again reject the top model if the studentized test statistic is greater than the new critical value and repeat this procedure until no further forecasting strategy can be rejected.

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 (including the historical average forecast) 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. Most recent studies analyze the out-of-sample forecasts in terms of the Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OOS}) and the Clark and West (2007) MSFE-adjusted statistic (Ferreira and Santa-Clara, 2011; Neely et al., 2014). The R^2_{OOS}

measures the proportional reduction in MSFE relative to the historical average forecast. A positive value indicates that the respective forecast strategy outperforms the historical average in terms of MSFE, and vice versa. The MSFE-adjusted statistics is comparable to the McCracken (2007) MSE-F statistics in so far as it tests the null hypothesis that the MSFE of the historical average forecast is less than or equal to the MSFE of the respective forecasting strategy against the (one-sided) upper-tail hypothesis. The MSFE-adjusted statistics can be interpreted as usual t-statistics where the significance level is assessed according to standard normal critical values.

Table I summarizes the out-of-sample results of the conventional predictive regressions and advanced forecasting strategies. For the sake of brevity, we only present results for those forecasting strategies that exhibit either a positive R^2_{OOS} or a significant MSFE-adjusted statistic, thus indicating an outperformance of the historical average forecast in terms of MSFE.

[Insert Table I here]

As reported in panel A of Table I, only the conventional predictive regression models based on the long-term return (LTR) and the term spread (TMS) exhibit a positive R^2_{OOS}. Interestingly, the MSFE-adjusted statistics indicate that, despite the negative R^2_{OOS} statistics, the MSFEs of the conventional predictive regressions based on the net equity expansion (NTIS), the Treasury-bill rate (TBL) and the default yield (DFY) are also significantly less than that of the historical average forecast.[5] Overall, the fundamental variables show only limited out-of-sample predictive ability, confirming the earlier results of Goyal and Welch (2008).

[5] Although these results seem surprising at first, they are plausible when comparing nested models, as in our empirical analysis. As shown by Clark and West (2007), the conventional predictive regression framework is expected to produce noisier estimates of the equity premium than the historical average forecast. Therefore, the difference between the MSFE of a conventional predictive regression forecast and the MSFE of the historical average forecast is upward biased. The MSFE-adjusted statistic accounts for this bias under the null hypothesis, such that the MSFE-adjusted statistic is able to reject the null even if the R^2_{OOS} is negative.

Turning next to the advanced forecasting strategies in panel B of Table I, we find that, except for the return on the radio and television broadcasting industry portfolio (TV), none of the lagged industry returns exhibit a positive R^2_{OOS} or a significant MSFE-adjusted statistic. Moreover, contrasting the results of Neely et al. (2014), neither technical indicators nor diffusion indices are able to outperform the historical average forecast (and thus are not included).[6] Combination forecasts, by contrast, exhibit significantly lower MSFEs than the historical average forecast, as indicated by their positive R^2_{OOS} and significant MSFE-adjusted statistics. With each R^2_{OOS} exceeding 0.60%, these forecasting strategies seem to outperform the conventional predictive regressions in terms of MSFE. In line with the findings of Campbell and Thompson (2008), forecast restrictions improve upon the conventional predictive regressions by either strengthening the predictive ability of those fundamental variables that already significantly outperform the historical average forecast in panel A, or uncovering the previously unrecognized predictive ability of conventional predictive regressions based on the equity premium volatility (RVOL) or the long-term government bond yield (LTY).

In contrast, state-dependent regressions seem to worsen the performance of conventional predictive regressions. For example, the R^2_{OOS} of the predictive regression based on TMS decreases from 0.24% to -0.56% when considering market states. Finally, all sum-of-the-parts models outperform the historical average forecast at the 1% level of statistical significance. Taken together, our results suggest that combination forecasts, forecast restrictions, and the sum-of-the-parts models outperform both the historical average forecast and conventional predictive regressions in terms of MSFE, as indicated by their positive R^2_{OOS} and/or significant MSFE-adjusted statistics.

[6] We emphasize that, due to the rolling estimation scheme and a 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 their recursive estimation scheme, both technical indicators and diffusion indices significantly outperform the historical average forecast. However, as noted earlier, a recursive estimation scheme would violate the stationarity assumption of the SPA test (see the discussion in Hansen, 2005).

While the results in Table I provide a first indication as to which advanced forecasting strategies might offer an improvement upon conventional predictive regression models, these analyses neither account for data snooping biases (Lo and MacKinlay, 1990), nor are they able to establish whether there are any forecasting strategies that in fact exhibit predictive superiority (as opposed to equal predictive ability). Therefore, we apply Hansen's (2005) SPA-test, comparing both the conventional predictive regressions and the advanced forecasting strategies with the performance of the historical average forecast. Since most advanced forecasting strategies claim to improve upon conventional predictive regressions, we first test each subset of advanced strategies separately, each time including the conventional predictive regressions in the test sample. 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, we further test the performance of conventional predictive regressions and all advanced forecasting strategies jointly against the historical benchmark forecast.

4.1. Forecast evaluation based on MSFE

In our first test, we assess whether any forecasting strategy can more accurately forecast the equity premium than the historical average forecast in terms of MSFE using Hansen's (2005) SPA-test. Table II shows the results. Column (1) describes the set of forecasting strategies we draw from. Column (3) gives the loss values of the benchmark model, the most significant model (the model with the highest *t*-statistic), and the best model (the model with the lowest loss value). The 'nominal' *p*-values in column (4) result from a pairwise comparison of the most significant and the best model with the benchmark. In contrast to the *p*-values of the SPA-test, these *p*-values do not account for the entire set of strategies. Column (5) gives 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 average forecast in terms of MSFE.

[Insert Table II here]

The first row in Table II summarizes the results of the SPA-test for the subset of conventional predictive regressions and lagged industry returns. Using the lagged return of the radio and television (TV) portfolio as an independent variable in a simple predictive regression model leads to the most significant and best results in terms of MSFE, thus outperforming the conventional predictive regressions. However, the 'nominal' *p*-value of 0.3594 indicates no outperformance with respect to the historical average forecast, even when considered in isolation. Consequently, the null hypothesis of the SPA-test cannot be rejected for this subset (as indicated by a consistent *p*-value of 0.9988).

Turning to the subset including the technical indicators in the second row of Table II, we note that all technical indicators are dominated by a conventional predictive regression based on the term spread (TMS), which is selected as the most significant and best model in this subset. However, when we compare the most significant and best model with the historical average forecast, its performance in terms of MSFE is not statistically different from the MSFE of the historical average forecast, as indicated by 'nominal' *p*-value of 0.4268. Accordingly, the null hypothesis of the SPA-test also cannot be rejected for any of these subsets (with consistent *p*-value of 0.9963). Similar results apply for the subsets including the diffusion indices (third row) or the state-dependent predictive regressions (sixth row). In contrast, combinations forecasts (fourth row) can improve upon conventional predictive regressions. Interestingly, the most significant and best models are not identical. While the median combination forecast is the most significant model (with 'nominal' *p*-value of 0.0761), the mean combination forecast generates the lowest MSFE over the out-of-sample period (with minimal loss value of 19.2403). However, the consistent *p*-value of 0.5468 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 average forecast when accounting for data snooping biases.

The results in the fifth row of Table II suggest that similar conclusions can be drawn for the subset including forecast restrictions. The restricted forecast based on the Treasury-bill rate (TBL) is selected as the most significant and best model, improving upon conventional predictive regressions. The 'nominal' *p*-value of 0.0969 indicates that this model is able to outperform the historical average forecast when considered in isolation. However, the consistent *p*-value of 0.7473 shows that the null hypothesis of the SPA-test can once again not be rejected.

The results in the seventh row of Table II indicate that the SOP-approach offers improvement upon conventional predictive regressions. It significantly outperforms the historical average forecast when considered in isolation ('nominal' *p*-value of 0.0058) as well as when accounting for the entire subset of forecasting strategies (consistent *p*-value of 0.0556). The stepwise-SPA-test reveals that the SOP-approach is the only strategy that significantly outperforms the historical average forecast in this subset.

Finally, when we turn to the results in the eighth row of Table II, the SOP-approach is also selected as the most significant and best model when the full set of forecasting strategies is considered. However, we can no longer reject the null hypothesis of the SPA-test (consistent *p*-value of 0.1660), i.e., there is no statistically significant evidence that any forecasting strategy is better than the historical average forecast in terms of MSFE. However, we note that the spread between the upper and lower bound *p*-values is relatively large and points towards the inclusion of many poor forecasting strategies (Hansen, 2005). If we assume that the forecasting strategies with worse performance than the historical average forecast are indeed poor models in the limit, the null hypothesis of the SPA-test can again be rejected, as indicated by the lower bound *p*-value of 0.0085.

Overall, the results in Table II indicate that many advanced forecasting strategies do not significantly improve upon conventional predictive regressions and do not outperform the historical average

forecast once accounting for potential data snooping biases. Only Ferreira and Santa-Clara's (2011) SOP-approach shows a clear superiority compared to both conventional predictive regressions and the historical average in terms of MSFE.

4.2. Forecast evaluation based on profit-based measures

As noted by Leitch and Tanner (1991), there is only a weak association between MSFE and forecast profitability. In order to assess whether out-of-sample predictability is sufficiently large to be of economic value, we compare the performance of both the conventional predictive regressions and advanced forecasting strategies with the historical average forecast (i) in terms of the absolute mean return and (ii) the risk-adjusted excess return using Hansen's (2005) SPA-test.

We compute the absolute return $r^*_{k,t}$ of an investor who monthly allocates her portfolio between the stock market and cash $r_{f,t}$, based on the (simple) equity premium forecast of strategy *k*:

$$r^*_{k,t} = w_{k,t} r_t + r_{f,t} \tag{14}$$

where $w_{k,t}$ is the proportion of total wealth allocated to the stock market. As investor can then use the point forecasts either as inputs for a traditional mean-variance asset allocation decision or for a market timing decision, we choose $w_{k,t}$ accordingly:

- *Mean-variance asset allocation*: A mean-variance investor sets the optimal weight on the stock market as:

$$w^P_{k,t} = \frac{\hat{r}_{k,t}}{\gamma \hat{\sigma}^2_t} \tag{15}$$

where γ is the coefficient of relative risk aversion, and $\hat{\sigma}^2_t$ a forecast of the (simple) equity premium variance. We set $\gamma = 5$ and estimate $\hat{\sigma}^2_t$ as a five-year rolling window of past monthly returns following Neely et al. (2014). Moreover, we impose realistic portfolio constraints prevent-

ing investors from short selling and taking more than 50% leverage by limiting $w_{k,t}^{P}$ to lie between 0 and 1.5.

- *Market timing*: A market timer is fully invested in the stock market if the equity premium forecast is positive and reverts to holding cash otherwise:

$$w_{k,t}^{MT} = \begin{cases} 1 \; if \; \hat{r}_{k,t} > 0 \\ 0 \; if \; \hat{r}_{k,t} \leq 0 \end{cases} \tag{16}$$

Since most forecasting strategies involve frequent trading, a realistic assessment of the performance of any forecasting strategy relative to the benchmark model has to take transaction costs into account. In particular, we follow Balduzzi and Lynch (1999) and assume 50 basis points as turnover-dependent costs.

Table III displays the results for a loss function based on mean absolute returns (panel A) and risk-adjusted excess returns (panel B) when assuming a mean-variance investor. All in all, the results confirm that economic forecasting evaluation is important, as many forecasting strategies significantly outperform the historical average forecast in terms of both absolute and risk-adjusted measures when considered in isolation, as indicated by the sufficiently small 'nominal' *p*-values.

[Insert Table III here]

More specifically, the results of panel A in Table III indicate that lagged industry returns (first row), technical indicators (second row), diffusion indices (third row), and forecast restrictions (fifth row) do not improve upon the conventional predictive regression based on TMS, which in each subset is selected as the most significant and best model. The conventional predictive regression forecasts based on TMS generate a mean absolute return of -0.7824% per month that is significantly higher than the mean absolute return of the historical average forecast (with 'nominal' *p*-value of 0.0283). How-

ever, the null hypothesis of the SPA-test cannot be rejected for each of these subsets (all consistent *p*-values exceeding 10%).

In contrast to the conventional predictive regression based on TMS, the mean combination forecast (fourth row) significantly outperforms the historical average forecast when considered in isolation (with 'nominal' *p*-value of 0.0109). In addition, we can (albeit marginally) reject the null hypothesis of the SPA-test for the subset that includes combination forecasts (with consistent *p*-value of 0.1005), i.e., at least the mean combination forecast can significantly outperform the historical average forecast after accounting for data snooping biases. The results in the sixth row indicate that a state-dependent regression based on TMS is selected as the best model, delivering a higher mean absolute return of 0.8108% per month, and significantly outperforming the historical average forecast when considered in isolation (with 'nominal' *p*-value of 0.0302). However, none of the state-dependent regressions is able to significantly outperform the historical average forecast when the entire set of strategies is considered (with consistent *p*-value of 0.3136).

Turning to the subset including the sum-of-the-parts models (seventh row), we note that a sum-of-the-parts model is selected as both the most significant and best model, thus outperforming conventional predictive regressions. While the ESOP model based on median combination forecasts, ESOP (Median), is selected as the most significant model in a pairwise comparison against the historical average forecast (with 'nominal' *p*-value of 0.0011), the ESOP model using mean combination forecasts, ESOP (Mean), yields the highest mean absolute return of 0.9305% per month over the out-of-sample period. In addition, we can reject the null hypothesis of the SPA-test at the 5% level of significance (consistent *p*-value of 0.0124). Finally, the null hypothesis of the SPA-test can also be rejected for the full set of forecasting strategies (with a consistent *p*-value of 0.0346). The step-SPA-test further confirms that all three ESOP models significantly outperform the historical average forecast.

Panel B of Table III summarizes the results of the SPA-test for a loss function based on the risk-adjusted return measure. For the subsets including lagged industry returns (first row), technical indicators (second row), diffusion indices (third row), forecast restrictions (fifth row), and state-dependent regressions (sixth row), the conventional predictive regression based on TBL turns out to be the most significant and best model, yielding a risk-adjusted excess return of 0.1071% per month (with a 'nominal' *p*-value of 0.0234). Therefore, this simple model significantly (albeit at the margin) outperforms the historical average forecast when considered in isolation. As indicated by large consistent *p*-values, all exceeding 10%, however, it is not able to outperform the historical average forecast when accounting for data snooping biases in any of these subsets. For the subset including combination forecasts (fourth row), the most significant model in a pairwise comparison against the historical average forecast is a mean combination forecast (with a 'nominal' *p*-value of 0.0019). We can even reject the null hypothesis of the SPA-test for this subset (with a consistent *p*-value of 0.0195).

The null hypothesis of the SPA-test can also be rejected for the subset including the sum-of-the-parts models (seventh row; with a consistent *p*-value of 0.0099). The ESOP (Median) model is selected as the most significant and best model, yielding a highly significant risk-adjusted excess return of 0.1617% per month ('nominal' *p*-value of 0.0008). Finally, the null hypothesis of the SPA-test can even be rejected when accounting for all forecasting strategies under investigation (eighth row; with consistent *p*-value of 0.0265). The step-SPA-test further confirms that all three ESOP models and the mean as well as the median combination forecasts significantly outperform the historical average forecast after accounting for data snooping biases.

Next, we assume a market timing investor. Table IV again shows the results for a loss function based on mean absolute returns (panel A) or risk-adjusted excess return (panel B). It becomes apparent that the historical average forecast serves as a more stringent benchmark model both in terms of absolute and risk-adjusted excess returns in a market timing context as opposed to a traditional mean-

variance asset allocation setting. Given the highly positive average equity premium during the sample period, the historical average forecast is almost identical to a buy-and-hold strategy and invested in cash for only seven months during the out-of-sample period following the early 1980s recession (April to October 1982). Accordingly, it yields a mean absolute return of 0.8459% and a risk-adjusted excess return of 0.1001% per month over our out-of-sample period.

[Insert Table IV here]

Panel A of Table IV reveals that, except for the sum-of-the-parts models, none of the advanced forecasting strategies is able to improve upon the conventional predictive regression based on TMS if the mean absolute return measure is used. This conventional prediction model delivers a higher mean absolute return than the historical average forecast (0.8884% vs. 0.8459% per month); however, the difference is not statistically significant even when considered in isolation (with 'nominal' *p*-value of 0.3024). As a result, we cannot reject the null hypothesis of the SPA-tests for each subset in the first to sixth row.

Only the sum-of-the-parts models are able to improve upon conventional predictive regressions and outperform the historical average forecast when considered in isolation with a mean absolute return of 0.9909% per month (with 'nominal' *p*-value of 0.0873) for the ESOP (Mean) model. Nevertheless, with a consistent *p*-value of 0.4607, we cannot reject the null hypothesis of the SPA-test for the subset including the sum-of-the-parts models. Accordingly, the null hypothesis of the SPA-test can also not be rejected for the entire set of forecasting strategies (with consistent *p*-value of 0.6539).

When evaluating the risk-adjusted performance measure in panel B of Table IV, in most subsets the conventional predictive regressions based on TMS or the dividend-payout ratio (DE) are selected as the most significant and best model, respectively. But when considered in isolation, these conven-

tional models do not significantly outperform the historical average forecast (with 'nominal' *p*-values of 0.1266 and 0.1797, respectively).

Again, the sum-of-the-parts models (seventh row) provide the only improvement upon the most significant and best conventional predictive regressions. In particular, the ESOP (Mean) model exhibits the most significant performance in a pairwise comparison against the historical average forecast (with 'nominal' *p*-value of 0.0165), while the ESOP (Median) model yields the highest risk-adjusted excess return of 0.1565% per month over the out-of-sample period. Nevertheless, the outperformance of the sum-of-the-parts models is not robust to controlling for data snooping biases (with consistent *p*-value of 0.1408) 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 0.2841).

Taken together, our findings imply that investors who allocate their assets using a traditional mean-variance optimization procedure can benefit greatly from forecasting the equity premium using the ESOP models rather than the historical average forecast. The ESOP models produce significantly higher mean absolute and risk-adjusted excess returns that are robust to data snooping biases. A market timing investor might also benefit from the ESOP model forecasts; however, we cannot rule out that the superior performance of the ESOP model is merely due to luck.

5. Robustness

To assess whether our results are robust to the choice of the out-of-sample period, we split our out-of-sample period in two sub-samples of equal lengths (January 1966 to December 1990 and January 1991 to December 2015) and repeat our analyses for the full set of forecasting strategies. The results in panel A of Table V show that the SOP model is selected as the most significant model in terms of MSFE in both sub-samples. However, as indicated by the difference in 'nominal' *p*-values between the sub-samples, the superior performance of the SOP model in terms of MSFE is predominantly driv-

en by its performance in the first half of the out-of-sample period. This finding coincides with the results of Ferreira and Santa-Clara (2011), who also report better out-of-sample performance in their earlier sub-sample. Accordingly, the null hypothesis of the SPA test is only rejected in the first half of the out-of-sample period (with consistent *p*-value of 0.0219 vs. 0.9950 in the second half).

[Insert Table V here]

Panel B of Table V summarizes the results for a mean-variance investor. We note that, regardless of the performance measure, one of the sum-of-the-parts models outperforms the historical average forecast in either sub-sample, supporting our previous results. While in the first half of the out-of-sample period, an ESOP (Median) model is selected as the most significant and best model in terms of mean absolute return (with 'nominal' *p*-value of 0.0139), in the second half it is an ESOP (Mean) model that is selected (with 'nominal *p*-value of 0.0022). Considering risk-adjusted excess returns, the ESOP (Mean) model delivers the most significant and best performance in the first half of the out-of-sample period (with 'nominal' *p*-value of 0.0009), and the ESOP (Median) model delivers the highest risk-adjusted excess return of 0.1940% per month in the second half of the out-of-sample period (with 'nominal' *p*-value of 0.0233). However, we note that the predictive superiority of the ESOP models when applied in the traditional mean-variance asset allocation framework is not robust to the split of the out-of-sample period after controlling for data snooping biases; the null hypothesis of the SPA-test can only be rejected in the second half of the out-of-sample period when based on mean absolute returns (with consistent *p*-value of 0.0704) or the first half of the out-of-sample period when based on risk-adjusted excess returns (with consistent *p*-value of 0.0296).

Finally, panel C of Table V shows the results for the market timer. The results are consistent with those above to the extent that the null hypothesis of the SPA-tests cannot be rejected for any sub-sample or specification of the loss function. Moreover, the ESOP (Mean) model that was selected as

the most significant and/or best model in the full out-of-sample period (see the eighth row in panel A and B of Table IV), only exhibits predictive superiority over the remaining forecasting strategies in the first half of the out-of-sample period.

Since the null hypothesis of the SPA-test is constructed relative to the performance of a benchmark model, our results depend critically on the assumed benchmark model. So far, in line with existing studies, we have considered the historical average forecast as the appropriate benchmark. However, in practice, often a buy-and-hold or a 50:50 constant-mix (rebalancing) strategy may be deemed more appropriate when evaluating the economic value of forecasting strategies.[7] Therefore, we repeat our analyses of section 4.2 for the full set of forecasting strategies using these alternative benchmark models. Table VI summarizes the results.

[Insert Table VI here]

Panel A shows the results for a mean-variance investor. We note that both the buy-and-hold strategy and the constant mix strategy yield higher mean absolute returns (0.8747% and 0.6403% per month, respectively) and higher risk-adjusted excess returns (0.1055% and 0.1035% per month, respectively) than the historical average forecast. Therefore, we expect that these alternatives serve as more stringent benchmark models when testing for statistical significance in a traditional mean-variance asset allocation. Our results confirm this hypothesis.

While none of the forecasting strategies is able to outperform the buy-and-hold strategy in terms of mean absolute return even in a pairwise comparison, as indicated by a comparatively large 'nominal' *p*-value, both the ESOP (Median) model and the ESOP (Mean) model deliver significantly higher mean absolute returns than the constant mix strategy of 0.9290% per month (with 'nominal' *p*-value

[7] For a 50:50 constant mix strategy, we set $w_{k,t} = 0.5$, i.e., an investor is 50% invested in the S&P 500 index and 50% in cash. A constant mix strategy requires monthly rebalancing to the target weights (subject to transaction costs of 50 basis points).

of 0.0021) and 0.9305% per month (with 'nominal' *p*-value of 0.0020), respectively. The null hypothesis of the SPA-test when applying a loss function based on mean absolute returns cannot be rejected when using the buy-and-hold strategy as a benchmark model. However, in line with the results presented in panel A of Table III above, the null hypothesis of the SPA-test can still be rejected when considering a constant mix (rebalancing) strategy as the benchmark model (with consistent *p*-value of 0.0306). The step-SPA-test confirms that all three ESOP models significantly outperform the constant mix strategy in terms of mean absolute returns.

The results for risk-adjusted excess returns indicate that the ESOP (Median) model significantly outperforms both alternative benchmark models in isolation (with 'nominal' *p*-values of 0.0444 and 0.0408, respectively). However, in contrast to the results in panel B of Table III, the null hypothesis of the SPA-test cannot be rejected when incorporating alternative benchmark models (with consistent *p*-values of 0.2883 and 0.2633, respectively).

Finally, panel B of Table VI summarizes the results for a market timer. As already noted, when applied to time the market, the historical average forecast is approximately identical in performance to a buy-and-hold strategy. The results of the respective SPA tests are thus nearly indistinguishable; we omit a further discussion. When considering the constant-mix strategy as an alternative benchmark, however, we observe a significant outperformance of the sum-of-the-part models when considered in isolation, both in terms of mean absolute return and risk-adjusted excess return. Contrary to the results in panel B of Table IV above, the null hypothesis of the SPA-test for a loss function based on the mean absolute return can now be rejected (with consistent *p*-value of 0.0016).

6. Conclusions

In this study, we jointly examine the out-of-sample performance of conventional predictive regressions and a comprehensive set of advanced forecasting strategies. Statistical inference might be

biased due to data snooping in comparisons of the out-of-sample performance of different forecasting strategies using the same set of data. We address this challenge by applying Hansen's (2005) SPA-test and its stepwise extension by Hsu, Hsu, and Kuan (2010), which allows us to infer whether any of the forecasting strategies under investigation exhibits truly superior performance against a given benchmark model.

Our results indicate that, when controlling for data snooping biases, only the sum-of-the parts approach proposed by Ferreira and Santa-Clara (2011) outperforms the historical average forecast out-of-sample in terms of mean squared forecast error. We further confirm the results of Leitch and Tanner (1991) by documenting that several forecasting strategies, based on combination forecasts, regime shifts, and the sum-of-the-parts approach, significantly outperform the historical average forecast in a mean-variance framework when considered in isolation. Extensions of the sum-of-the-parts approach proposed by Bätje and Menkhoff (2016) exhibit predictive superiority both on a mean absolute return and a risk-adjusted excess return basis even when controlling for data snooping biases. In contrast, none of the forecasting strategies under investigation, with exemption of the sum-of-the-parts models when assessed against a constant mix benchmark, can help an investor who wishes to engage in market timing. Overall, our results provide evidence that mean-variance investors will benefit from using more advanced forecasting strategies rather than conventional predictive regressions.

References

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

Balduzzi, P., and A.W. Lynch, 1999, Transaction Costs and Predictability: Some Utility Cost Calculations, *Journal of Financial Economics* 52, 47-78.

Blume, L., D. Easley, and M. O'Hara M, 1994, Market Statistics and Technical Analysis: The Role of Volume, *Journal of Finance* 49,153–181.

Boyd, J.H., J. Hu, and R. Jagannathan, 2005, The Stock Market's Reaction to Unemployment News: Why Bad News Is Usually Good for Stocks, *Journal of Finance* 60, 649-672.

Campbell, J.Y., 2000, Asset Pricing at the Millennium, *Journal of Finance* 55, 1515-1567.

Campbell, J.Y., 2008, Viewpoint: Estimating the Equity Premium, *Canadian Journal of Economics* 41, 1-21.

Campbell, J.Y., and R.J. Shiller, 1989, The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors, *Review of Financial Studies* 1, 195-228.

Campbell, J.Y., and S.B. Thompson, 2008, Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?, *Review of Financial Studies* 21, 1509-1531.

Clark, T.E., and M.W. McCracken, 2012, Reality Checks and Comparisons of Nested Predictive Models, *Journal of Business and Economic Statistics* 30, 53-66.

Clark, T.E., and K.D. West, 2007, Approximately Normal Tests for Equal Predictive Accuracy in Nested Models, *Journal of Econometrics* 138, 291-311.

Cooper, M.J., R.C. Gutierrez Jr., and A. Hameed, 2004, Market States and Momentum, *Journal of Finance* 54, 1345-1365.

Dangl, T., and M. Halling, 2012, Predictive Regressions with Time-Varying Coefficients, *Journal of Financial Economics* 106, 157-181.

Fama, E.F., and G.W. Schwert, 1977, Asset Returns and Inflation, *Journal of Financial Economics* 5, 115-146.

Ferreira, M.A., and P. Santa-Clara, 2011, Forecasting Stock Market Returns: The Sum of the Parts is more than the Whole, *Journal of Financial Economics* 100, 514-537.

Giacomini, R., and H. White, 2006, Tests of Conditional Predictive Ability, *Econometrica* 74, 1545-1578.

Goyal, A., and I. Welch, 2008, A Comprehensive Look at The Empirical Performance of Equity Premium Prediction, *Review of Financial Studies* 21, 1455-1508.

Granville, J.E., 1963, *Granville's New Key to Stock Market Profits*, Prentice-Hall, New York.

Guidolin, M., and A. Timmermann, 2007, Asset Allocation under Multivariate Regime Switching, *Journal of Economic Dynamics & Control* 31, 3503-3544.

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

Hansen, P.R., 2005, A Test for Superior Predictive Ability, *Journal of Business and Economic Statistics* 23, 365-380.

Henkel, S.J., J.S. Martin, and F. Nardari, 2011, Time-Varying Short-Horizon Predictability, *Journal of Financial Economics* 99, 560-580.

Hong, H., W. Torous, and R. Valkanov, 2007, Do Industries Lead Stock Markets?, *Journal of Financial Economics* 83, 367-396.

Hsu, P., Y. Hsu, and C. Kuan, 2010, Testing the Predictive Ability of Technical Analysis Using a New Stepwise Test without Data Snooping Bias, *Journal of Empirical Finance* 17, 471-484.

Huang, D., Jiang, F., Tu, J., and G. Zhou, 2016, Forecasting Stock Returns in Good and Bad Times: The Role of Market States, Working Paper.

Inoue, A., and L. Kilian, 2004. In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use? *Econometric Reviews* 23, 371–402.

Kothari, S.P., and J. Shanken, 1997, Book-to-Market, Dividend Yield, and Expected Market Returns: A Time-Series Analysis, *Journal of Financial Economics* 44, 169-203.

Leitch, G., and J.E. Tanner, 1991, Economic Forecast Evaluation: Profits versus the Conventional Error Measures, *American Economic Review* 81, 580-590.

Lettau, M., and S. Van Nieuwerburgh, 2008, Reconciling the Return Predictability Evidence, *Review of Financial Studies* 21, 1607-1652.

Lo, A.W., and A.C. MacKinlay, 1990, Data-Snooping Biases in Tests of Financial Asset Pricing Models, *Review of Financial Studies* 3, 431-467.

Ludvigson, S.C., and S. Ng, 2007, The Empirical Risk–Return Relation: A Factor Analysis Approach, *Journal of Financial Economics* 83, 171-222.

McCracken, M.W., 2007, Asymptotics for Out of Sample Tests of Granger Causality, *Journal of Econometrics* 140, 719–752.

Mele, A., 2007, Asymmetric Stock Market Volatility and the Cyclical Behavior of Expected Returns, *Journal of Financial Economics* 86, 446-478.

Moskowitz, T.J., Y.H. Ooi, and L.H. Pedersen, 2012, Time Series Momentum, *Journal of Financial Economics* 104, 228-250.

Neely, C.J., D.E. Rapach, J. Tu, and G. Zhou, 2014, Forecasting the Equity Risk Premium: The Role of Technical Indicators, *Management Science* 60, 1772-1791.

Paye, B.S., and A. Timmermann, 2006, Instability of Return Prediction Models, *Journal of Empirical Finance* 13, 274-315.

Politis, D.N., and J.P. Romano, 1994, Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions, *Annals of Statistics* 4, 2031-2050.

Rapach, D.E., J.K. Strauss, and G. Zhou, 2010, Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy, *Review of Financial Studies* 23, 821-862.

Rapach, D.E., and M.E. Wohar, 2006a, In-sample vs. Out-of-sample Tests of Stock Return Predictability in the Context of Data Mining, *Journal of Empirical Finance* 13, 231-247.

Rapach, D.E., and M.E. Wohar, 2006b, Structural Breaks and Predictive Regression Models of Aggregate U.S. Stock Returns, *Journal of Financial Econometrics* 4, 238-274.

Rapach, D.E., and G. Zhou, 2013, *Forecasting Stock Returns*, in G. Elliot and A. Timmermann, eds.: Handbook of Economic Forecasting, Volume 2, Elsevier, Amsterdam.

Stock, J.H., and M.W. Watson, 2006, *Forecasting with Many Predictors*, in G. Elliot, C.W.J. Granger and A. Timmermann, eds.: Handbook of Economic Forecasting, Volume 1, Elsevier, Amsterdam.

Sullivan, R., A. Timmermann, and H. White, 1999, Data-Snooping, Technical Trading Rule Performance, and the Bootstrap, *Journal of Finance* 54, 1647-1691.

Timmermann, A., 2006, *Forecast Combinations*, in G. Elliot, C.W.J. Granger and A. Timmermann, eds.: Handbook of Economic Forecasting, Volume 1, Elsevier, Amsterdam.

White, H., 2000, A Reality Check for Data Snooping, *Econometrica* 68, 1097-1126.

Tables

Table I
Out-of-sample forecasting results

This table reports out-of-sample forecasting results for conventional predictive regressions based on fundamental variables and advanced forecasting strategies in the out-of-sample period from January 1966 to December 2015. Definitions of fundamental variables are shown in appendix A ***, **, and * indicate significance at the 1%, 5% and 10% levels, respectively.

Forecasting strategies	Predictor	R^2_{OOS} (%)	MSFE-adj.	
Panel A: Conventional predictive regressions				
Fundamental variables	NTIS	-0.54	1.28	*
	TBL	-0.81	1.58	*
	LTR	0.08	1.73	**
	TMS	0.24	2.26	**
	DFY	-1.22	1.43	*
Panel B: Advanced forecasting strategies				
Lagged industry returns	TV	0.30	1.36	*
Combination forecasts	Mean	0.85	1.57	*
	Median	0.61	1.70	**
	Trimmed mean	0.76	1.54	*
Forecast restrictions	DE	0.01	0.79	
	RVOL	0.30	1.51	*
	NTIS	0.25	1.81	**
	TBL	0.76	2.55	***
	LTY	0.30	2.13	**
	LTR	0.37	1.72	**
	TMS	0.42	2.15	**
	DFY	0.13	1.89	**
	INFL	0.26	1.21	
State-dependent predictive regressions	TBL	-1.53	1.33	*
	TMS	-0.56	1.99	**
	DFY	-2.22	1.31	*
Sum-of-the-parts models	SOP	1.62	3.42	***
	ESOP (Mean)	0.04	2.72	***
	ESOP (Median)	-0.34	2.98	***
	ESOP (Trimmed mean)	-0.30	2.69	***

Table II

SPA-tests based on MSFE loss function

This table reports the results of Hansen's (2005) SPA-test for conventional predictive regressions based on fundamental variables and advanced forecasting strategies compared to the historical average forecast in the out-of-sample period from January 1966 to December 2015 for a loss function based on mean squared forecast errors. The table reports the sample loss for the benchmark and the two forecasting strategies that have the smallest sample loss value and the largest *t*-statistic for average relative performance $(\bar{d}_k)$. These two strategies are referred to as the "best" and "most significant" model, respectively. The loss value is shown in column (3). Column (4) reports the 'nominal' *p*-values from the pairwise comparisons of the best and the most significant model with the benchmark. These p-values (unlike the SPA *p*-value) ignore the search over all strategies that preceded the selection of the strategy being compared to the benchmark, i.e., they do not account for the entire set of forecasting strategies. Finally, column (5) shows the consistent *p*-value of the SPA test, as well as the lower and upper bounds for *p*-values. Results that are significant at the 10% level of significance are printed in bold.

	(1)	(2)	(3)	(4)	(5)
	Set of forecasting strategies	Benchmark model Most significant model Best model	Loss value	*p*-value	Consistent *p*-value Lower *p*-value Upper *p*-value
(1)	14 conventional pred. regressions 33 lagged industry returns	Historical average TV TV	19.4044 19.3467 19.3467	 0.3594 0.3594	0.9988 0.9449 0.9988
(2)	14 conventional pred. regressions 14 technical indicators	Historical average TMS TMS	19.4044 19.3586 19.3586	 0.4268 0.4268	0.9963 0.9544 0.9963
(3)	14 conventional pred. regressions 3 diffusion indices	Historical average TMS TMS	19.4044 19.3586 19.3586	 0.4268 0.4268	0.9928 0.9266 0.9928
(4)	14 conventional pred. regressions 3 combination forecasts	Historical average Median Mean	19.4044 19.2852 19.2403	 **0.0761** 0.1241	0.5468 0.3223 0.5468
(5)	14 conventional pred. regressions 14 forecast restrictions	Historical average TBL (restricted) TBL (restricted)	19.4044 19.2562 19.2562	 **0.0969** **0.0969**	0.7473 0.5761 0.7473
(6)	14 conventional pred. regressions 14 state-dependent regressions	Historical average TMS TMS	19.4044 19.3586 19.3586	 0.4268 0.4268	0.9974 0.9295 0.9974
(7)	14 conventional pred. regressions 5 sum-of-the-parts models	Historical average SOP SOP	19.4044 19.0901 19.0901	 **0.0058** **0.0058**	**0.0556** **0.0291** **0.0556**
(8)	14 conventional pred. regressions All advanced forecasting strategies	Historical average SOP SOP	19.4044 19.0901 19.0901	 **0.0058** **0.0058**	0.1660 **0.0085** 0.1660

Table III

SPA-tests based on profit-based loss function (mean-variance investor)

This table reports the results of Hansen's (2005) SPA-test for conventional predictive regressions based on fundamental variables and advanced forecasting strategies compared to the historical average forecast in the out-of-sample period from January 1966 to December 2015 for a loss function based on the (negative) mean absolute return (panel A) or the risk-adjusted excess return (panel B) assuming a mean-variance investor. The table reports the sample loss for the benchmark and the two forecasting strategies that have the smallest sample loss value and the largest *t*-statistic for average relative performance $(\bar{d}_k)$. These two strategies are referred to as the "best" and "most significant" model, respectively. The loss value is shown in column (3). Column (4) reports the 'nominal' *p*-values from the pairwise comparisons of the best and the most significant model with the benchmark. These *p*-values (unlike the SPA *p*-value) ignore the search over all strategies that preceded the selection of the strategy being compared to the benchmark, i.e., they do not account for the entire set of forecasting strategies. Finally, column (5) shows the consistent *p*-value of the SPA-test, as well as the lower and upper bounds for *p*-values. Results that are significant at the 10% level of significance are printed in bold.

	(1)	(2)	(3)	(4)	(5)
	Set of forecasting strategies	Benchmark model Most significant model Best model	Loss value	*p*-value	Consistent *p*-value Lower *p*-value Upper *p*-value
Panel A: Loss function based on mean absolute return					
(1)	14 conventional pred. regressions 33 lagged industry returns	Historical average TMS TMS	-0.5767 -0.7824 -0.7824	 **0.0283** **0.0283**	0.3859 0.2046 0.4633
(2)	14 conventional pred. regressions 14 technical indicators	Historical average TMS TMS	-0.5767 -0.7824 -0.7824	 **0.0283** **0.0283**	0.2820 0.2384 0.2820
(3)	14 conventional pred. regressions 3 diffusion indices	Historical average TMS TMS	-0.5767 -0.7824 -0.7824	 **0.0283** **0.0283**	0.2586 0.2139 0.2586
(4)	14 conventional pred. regressions 3 combination forecasts	Historical average Mean TMS	-0.5767 -0.7021 -0.7824	 **0.0109** **0.0283**	0.1005 **0.0830** 0.1005
(5)	14 conventional pred. regressions 14 forecast restrictions	Historical average TMS TMS	-0.5767 -0.7824 -0.7824	 **0.0283** **0.0283**	0.2422 0.1965 0.2422
(6)	14 conventional pred. regressions 14 state-dependent regressions	Historical average TMS TMS (state-dependent)	-0.5767 -0.7824 -0.8108	 **0.0283** **0.0302**	0.3136 0.2445 0.3136
(7)	14 conventional pred. regressions 5 sum-of-the-parts models	Historical average ESOP (Median) ESOP (Mean)	-0.5767 -0.9280 -0.9305	 **0.0011** **0.0018**	**0.0124** **0.0104** **0.0124**
(8)	14 conventional pred. regressions All advanced forecasting strategies	Historical average ESOP (Median) ESOP (Mean)	-0.5767 -0.9280 -0.9305	 **0.0011** **0.0018**	**0.0346** **0.0215** **0.0396**

(continued)

Table III– *Continued*

Panel B: Loss function based on risk-adjusted excess return					
(1)	14 conventional pred. regressions 33 lagged industry returns	Historical average	-0.0503		0.3221
		TBL	-0.1071	**0.0234**	0.1610
		TBL	-0.1071	**0.0234**	0.4066
(2)	14 conventional pred. regressions 14 technical indicators	Historical average	-0.0503		0.2368
		TBL	-0.1071	**0.0234**	0.1893
		TBL	-0.1071	**0.0234**	0.2368
(3)	14 conventional pred. regressions 3 diffusion indices	Historical average	-0.0503		0.2178
		TBL	-0.1071	**0.0234**	0.1699
		TBL	-0.1071	**0.0234**	0.2178
(4)	14 conventional pred. regressions 3 combination forecasts	Historical average	-0.0503		**0.0195**
		Mean	-0.0993	**0.0019**	**0.0158**
		TBL	-0.1071	**0.0234**	**0.0195**
(5)	14 conventional pred. regressions 14 forecast restrictions	Historical average	-0.0503		0.2022
		TBL	-0.1071	**0.0234**	0.1528
		TBL	-0.1071	**0.0234**	0.2022
(6)	14 conventional pred. regressions 14 state-dependent regressions	Historical average	-0.0503		0.2627
		TBL	-0.1071	**0.0234**	0.1919
		TBL	-0.1071	**0.0234**	0.2627
(7)	14 conventional pred. regressions 5 sum-of-the-parts models	Historical average	-0.0503		**0.0099**
		ESOP (Median)	-0.1617	**0.0008**	**0.0083**
		ESOP (Median)	-0.1617	**0.0008**	**0.0099**
(8)	14 conventional pred. regressions All advanced forecasting strategies	Historical average	-0.0503		**0.0265**
		ESOP (Median)	-0.1617	**0.0008**	**0.0169**
		ESOP (Median)	-0.1617	**0.0008**	**0.0318**

Table IV

SPA-tests based on profit-based loss function (market timer)

This table reports the results of Hansen's (2005) SPA-test for conventional predictive regressions based on fundamental variables and advanced forecasting strategies compared to the historical average forecast in the out-of-sample period from January 1966 to December 2015 for a loss function based on the (negative) mean absolute return (panel A) or the risk-adjusted excess return (panel B) assuming a market timer. The table reports the sample loss for the benchmark and the two forecasting strategies that have the smallest sample loss value and the largest t-statistic for average relative performance $(\bar{d}_k)$. These two strategies are referred to as the "best" and "most significant" model, respectively. The loss value is shown in column (3). Column (4) reports the 'nominal' *p*-values from the pairwise comparisons of the best and the most significant model with the benchmark. These *p*-values (unlike the SPA *p*-value) ignore the search over all strategies that preceded the selection of the strategy being compared to the benchmark, i.e., they do not account for the entire set of forecasting strategies. Finally, column (5) shows the consistent *p*-value of the SPA-test, as well as the lower and upper bounds for *p*-values. Results that are significant at the 10% level of significance are printed in bold.

	(1)	(2)	(3)	(4)	(5)
	Set of forecasting strategies	Benchmark model Most significant model Best model	Loss value	*p*-value	Consistent p-value Lower *p*-value Upper *p*-value
Panel A: Loss function based on mean absolute return					
(1)	14 conventional pred. regressions 33 lagged industry returns	Historical average TMS TMS	-0.8459 -0.8884 -0.8884	 0.3024 0.3024	0.9395 0.7251 0.9576
(2)	14 conventional pred. regressions 14 technical indicators	Historical average TMS TMS	-0.8459 -0.8884 -0.8884	 0.3024 0.3024	0.8733 0.7046 0.8929
(3)	14 conventional pred. regressions 3 diffusion indices	Historical average TMS TMS	-0.8459 -0.8884 -0.8884	 0.3024 0.3024	0.8598 0.6917 0.8824
(4)	14 conventional pred. regressions 3 combination forecasts	Historical average TMS TMS	-0.8459 -0.8884 -0.8884	 0.3024 0.3024	0.8539 0.6826 0.8777
(5)	14 conventional pred. regressions 14 forecast restrictions	Historical average TMS TMS	-0.8459 -0.8884 -0.8884	 0.3024 0.3024	0.8432 0.6585 0.8688
(6)	14 conventional pred. regressions 14 state-dependent regressions	Historical average TMS TMS	-0.8459 -0.8884 -0.8884	 0.3024 0.3024	0.8822 0.7030 0.9081
(7)	14 conventional pred. regressions 5 sum-of-the-parts models	Historical average ESOP (Mean) ESOP (Mean)	-0.8459 -0.9909 -0.9909	 **0.0873** **0.0873**	0.4607 0.3202 0.4857
(8)	14 conventional pred. regressions All advanced forecasting strategies	Historical average ESOP (Mean) ESOP (Mean)	-0.8459 -0.9909 -0.9909	 **0.0873** **0.0873**	0.6539 0.4011 0.7033

(continued)

Table IV– *Continued*

Panel B: Loss function based on risk-adjusted excess return					
(1)	14 conventional pred. regressions 33 lagged industry returns	Historical average TMS DE	-0.1001 -0.1225 -0.1232	 0.1266 0.1797	0.8186 0.5992 0.8450
(2)	14 conventional pred. regressions 14 technical indicators	Historical average TMS DE	-0.1001 -0.1225 -0.1232	 0.1266 0.1797	0.6747 0.5577 0.7017
(3)	14 conventional pred. regressions 3 diffusion indices	Historical average TMS DE	-0.1001 -0.1225 -0.1232	 0.1266 0.1797	0.6423 0.5159 0.6722
(4)	14 conventional pred. regressions 3 combination forecasts	Historical average TMS DE	-0.1001 -0.1225 -0.1232	 0.1266 0.1797	0.6282 0.4960 0.6591
(5)	14 conventional pred. regressions 14 forecast restrictions	Historical average TMS DE	-0.1001 -0.1225 -0.1232	 0.1266 0.1797	0.6137 0.4782 0.6462
(6)	14 conventional pred. regressions 14 state-dependent regressions	Historical average TMS TBL (state-dependent)	-0.1001 -0.1225 -0.1262	 0.1266 0.1824	0.7171 0.5396 0.7370
(7)	14 conventional pred. regressions 5 sum-of-the-parts models	Historical average ESOP (Mean) ESOP (Median)	-0.1001 -0.1560 -0.1565	 **0.0165** **0.0165**	0.1408 0.1021 0.1408
(8)	14 conventional pred. regressions All advanced forecasting strategies	Historical average ESOP (Mean) ESOP (Median)	-0.1001 -0.1560 -0.1565	 **0.0165** **0.0165**	0.2841 0.1647 0.3004

Table V

Robustness check: Sub-sample analyses

This table reports the results of Hansen's (2005) SPA-test for conventional predictive regressions based on fundamental variables and advanced forecasting strategies compared to the historical average forecast in the out-of-sample sub-sample periods from January 1966 to December 1990 and from January 1991 to December 2015, respectively, for a loss function based on mean squared forecast errors (panel A), the (negative) mean absolute return or the risk-adjusted excess return assuming a mean-variance investor (panel B) or a market timer (panel C). The table reports the sample loss for the benchmark and the two forecasting strategies that have the smallest sample loss value and the largest *t*-statistic for average relative performance $(\bar{d}_k)$. These two strategies are referred to as the "best" and "most significant" model, respectively. The loss value is shown in column (3). Column (4) reports the 'nominal' p-values from the pairwise comparisons of the best and the most significant model with the benchmark. These *p*-values (unlike the SPA *p*-value) ignore the search over all strategies that preceded the selection of the strategy being compared to the benchmark, i.e., they do not account for the entire set of forecasting strategies. Finally, column (5) shows the consistent p-value of the SPA-test, as well as the lower and upper bounds for *p*-values. Results that are significant at the 10% level of significance are printed in bold.

(1)	(2)	(3)	(4)	(5)
Set of forecasting strategies	Benchmark model Most significant model Best model	Loss value	*p*-value	Consistent *p*-value Lower *p*-value Upper *p*-value
Panel A: SPA tests based on MSFE loss function				
1966-1990	Historical average	21.2376		**0.0219**
14 conventional pred. regressions	SOP	20.7880	**0.0009**	**0.0164**
All advanced forecasting strategies	ESOP (Mean)	20.4293	**0.0031**	**0.0219**
1991-2015	Historical average	17.5712		0.9950
14 conventional pred. regressions	SOP	17.3922	0.1840	0.8123
All advanced forecasting strategies	SOP	17.3922	0.1840	0.9964
Panel B: SPA tests based on profit-based loss function (mean-variance investor)				
I. Loss function based on mean absolute return				
1966-1990	Historical average	-0.6164		0.2318
14 conventional pred. regressions	ESOP (Median)	-1.0150	**0.0139**	0.1200
All advanced forecasting strategies	ESOP (Median)	-1.0150	**0.0139**	0.3120
1991-2015	Historical average	-0.5370		**0.0704**
14 conventional pred. regressions	ESOP (Mean)	-0.9461	**0.0022**	**0.0485**
All advanced forecasting strategies	ESOP (Mean)	-0.9461	**0.0022**	**0.0704**
II. Loss function based on risk-adjusted excess return				
1966-1990	Historical average	0.0241		**0.0296**
14 conventional pred. regressions	ESOP (Mean)	-0.1687	**0.0009**	**0.0196**
All advanced forecasting strategies	ESOP (Mean)	-0.1687	**0.0009**	**0.0296**
1991-2015	Historical average	-0.1019		0.3104
14 conventional pred. regressions	EP	-0.1645	**0.0229**	0.1752
All advanced forecasting strategies	ESOP (Median)	-0.1940	**0.0233**	0.4367

Table V – *Continued*

Panel C: SPA tests based on profit-based loss function (market timer)				
I. Loss function based on mean absolute return				
1966-1990	Historical average	-0.8091		0.3595
14 conventional pred. regressions	ESOP (Mean)	-1.1126	**0.0306**	0.2181
All advanced forecasting strategies	ESOP (Mean)	-1.1126	**0.0306**	0.3800
1991-2015	Historical average	-0.8827		0.9729
14 conventional pred. regressions	TMS	-0.9746	0.2312	0.6812
All advanced forecasting strategies	TMS	-0.9746	0.2312	0.9838
II. Loss function based on risk-adjusted excess return				
1966-1990	Historical average	-0.0471		0.2101
14 conventional pred. regressions	ESOP (Mean)	-0.1349	**0.0100**	0.1310
All advanced forecasting strategies	ESOP (Mean)	-0.1349	**0.0100**	0.2213
1991-2015	Historical average	-0.1576		0.6385
14 conventional pred. regressions	DFY	-0.2083	**0.0715**	0.3920
All advanced forecasting strategies	DFY	-0.2083	**0.0715**	0.6989

Table VI

Robustness check: Alternative benchmark models

This table reports the results of Hansen's (2005) SPA-test for conventional predictive regressions based on fundamental variables and advanced forecasting strategies compared to a buy-and-hold or 50:50 constant mix strategy in the out-of-sample period from January 1966 to December 2015 for a loss function based on the negative absolute return or the risk-adjusted excess return assuming a mean-variance investor (panel A) or a market timer (panel B). The table reports the sample loss for the benchmark and the two forecasting strategies that have the smallest sample loss value and the largest *t*-statistic for average relative performance $(\bar{d}_k)$. These two strategies are referred to as the "best" and "most significant" model, respectively. The loss value is shown in column (3). Column (4) reports the 'nominal' *p*-values from the pairwise comparisons of the best and the most significant model with the benchmark. These *p*-values (unlike the SPA *p*-value) ignore the search over all strategies that preceded the selection of the strategy being compared to the benchmark, i.e., they do not account for the entire set of forecasting strategies. Finally, column (5) shows the consistent *p*-value of the SPA test, as well as the lower and upper bounds for *p*-values. Results that are significant at the 10% level of significance are printed in bold.

(1)	(2)	(3)	(4)	(5)
Set of forecasting strategies	Benchmark model Most significant model Best model	Loss value	*p*-value	Consistent *p*-value Lower *p*-value Upper *p*-value
Panel A: SPA tests based on profit-based loss function (mean-variance investor)				
I. Loss function based on mean return				
14 conventional pred. regressions All advanced forecasting strategies	Buy-and-hold ESOP (Mean) ESOP (Mean)	-0.8747 -0.9305 -0.9305	 0.3340 0.3340	0.7678 0.4452 0.8597
14 conventional pred. regressions All advanced forecasting strategies	Constant mix ESOP (Median) ESOP (Mean)	-0.6403 -0.9280 -0.9305	 **0.0021** **0.0020**	**0.0306** **0.0196** **0.0316**
II. Loss function based on risk-adjusted excess return				
14 conventional pred. regressions All advanced forecasting strategies	Buy-and-hold ESOP (Median) ESOP (Median)	-0.1055 -0.1617 -0.1617	 **0.0444** **0.0444**	0.2883 0.1624 0.3881
14 conventional pred. regressions All advanced forecasting strategies	Constant mix ESOP (Median) ESOP (Median)	-0.1035 -0.1617 -0.1617	 **0.0408** **0.0408**	0.2633 0.1579 0.3580
Panel B: SPA tests based on profit-based loss function (market timer)				
I. Loss function based on mean return				
14 conventional pred. regressions All advanced forecasting strategies	Buy-and-hold ESOP (Mean) ESOP (Mean)	-0.8747 -0.9909 -0.9909	 0.1269 0.1269	0.7638 0.4414 0.8252
14 conventional pred. regressions All advanced forecasting strategies	Constant mix ESOP (Mean) ESOP (Mean)	-0.6403 -0.9909 -0.9909	 **0.0000** **0.0000**	**0.0016** **0.0016** **0.0016**
II. Loss function based on risk-adjusted excess return				
14 conventional pred. regressions All advanced forecasting strategies	Buy-and-hold ESOP (Mean) ESOP (Median)	-0.1055 -0.1560 -0.1565	 **0.0211** **0.0216**	0.3342 0.1685 0.3648
14 conventional pred. regressions All advanced forecasting strategies	Constant mix ESOP (Mean) ESOP (Median)	-0.1035 -0.1560 -0.1565	 **0.0164** **0.0184**	0.2971 0.1559 0.3219

Appendix A: Definition of fundamental 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