Data Snooping in Equity Premium Prediction

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

This draft: November 2019

Abstract

We analyze the performance of a comprehensive set of equity premium forecasting strategies. All strategies were found to outperform the mean in previous academic publications. However, using a multiple testing framework to account for data snooping, our findings support Welch and Goyal (2008) in that almost all equity premium forecasts fail to beat the mean out-of-sample. Only few forecasting strategies that are based on Ferreira and Santa-Clara's (2011) "sum-of-the-parts" approach generate robust and statistically significant economic gains relative to the historical mean even after controlling for data snooping and accounting for transaction costs.

Keywords: Equity risk premium prediction; 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] Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556, USA.

[d] BlackRock Investment Management (UK) Limited, 12 Throgmorton Avenue, London EC2N 2DL, United Kingdom.

‡ *Acknowledgments*: We thank two anonymous referees, Amit Goyal, Michael Halling, Alexander Hillert, George Kapetanios (editor), Markus Leippold, Harald Lohre, Emanuel Mönch, and Tatjana Puhan for helpful comments. We also thank participants of the 2017 European Financial Management Association Meeting (Athens), the 2017 Financial Management Conference (Boston), the 2018 Southern Finance Association Conference (Asheville), the 2018 World Finance and Banking Symposium (Taichung), and the 2018 Paris Financial Management Conference for constructive suggestions. All remaining errors are our own.

1. Introduction

Does equity premium prediction pay off? While the in-sample predictability of the equity premium is largely undisputed (Campbell, 2000; Rapach and Zhou, 2013), investors are ultimately interested in whether there exists any forecasting strategy that delivers superior out-of-sample gains. Recognizing the controversial debate about the out-of-sample performance of established stock return prediction models, Spiegel (2008) poses the challenging question whether the standard approaches accurately forecast the equity premium any better than the historical mean.

While Welch and Goyal (2008) suggest that skepticism is appropriate when it comes to predicting the equity premium out-of-sample, a few subsequent studies indicate that some forecasting models deliver better results than the historical mean (Campbell and Thompson, 2008; Rapach, Strauss, and Zhou, 2010; Ferreira and Santa-Clara, 2011; Dangl and Halling, 2012; Neely et al., 2014; Huang et al., 2017). However, one challenge when testing for out-of-sample return predictability is that almost all forecasting strategies are analyzed on a single data set. When many models are evaluated individually on the same data set, some are bound to show superior performance by chance alone, even though they are not genuinely superior (Foster, Smith, and Whaley, 1997; Sullivan, Timmermann, and White, 1999). In any given sample, a forecasting model may produce a smaller average loss even though in expectation (i.e., across all samples we could have seen) the model would not be so good. This bias in statistical inference is referred to as 'data snooping'. Without adjusting for this bias in a multiple testing framework, one might commit a type I error, i.e., falsely assessing a forecasting strategy as being superior when it is not.

Multiple testing is critically important for assessing the success of equity premium prediction. 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 potential data snooping bias. We construct a set of 140 forecasting strategies that are based on both univariate predictive regression models 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., 2017). We use these strategies to predict the monthly U.S. equity premium strictly out-of-sample based on the most recent 180 months and track their performance for the subsequent month over the evaluation period from January 1966 to December 2018. As performance measures, we use the mean squared forecast error as well as mean absolute and risk-adjusted excess returns. Our analysis thus aims to answer Spiegel's (2008) question whether there exists any forecasting strategy that can provide a significantly higher economic value than the prevailing mean model.

Why is data snooping a concern in equity premium prediction? Suppose the 140 forecasting strategies are mutually independent, and we apply a *t*-test to each strategy with a significance level of 5%. The probability of falsely rejecting at least one correct null hypothesis is $1 - (1 - 5\%)^{140} \approx 0.999$. 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 appropriate methodology that controls for data snooping and avoids spurious inference when many models are examined together. In his AFA Presidential Address, Harvey (2017) summarizes the discussion by emphasizing 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 formally control for data snooping when testing for the possible superiority of a forecasting strategy in a multiple testing framework, we apply Hansen's (2005) test for superior predictive ability (SPA-test). His approach builds on White (2000) and examines how confident we can be that the best forecast, among a set of multiple forecasts, is genuinely better than the benchmark, given that the best forecast is selected from a potentially large pool of models. The SPA-test does not go beyond this first step in an attempt to identify which particular model is better than the benchmark, nor does it answer how many superior models exist. Therefore, to identify as many forecasting strategies that can outperform the benchmark as possible, we also use the stepwise extensions of the SPA-test recently proposed by Hsu, Hsu, and Kuan (2010) and Hsu, Kuan, and Yen (2014).

Confirming the earlier literature, our results show that many forecasting strategies outperform the historical mean when tested individually, i.e., pairwise against the benchmark. However, once we refrain from testing predictor variables in isolation and control for data snooping in a multiple testing setup, we find that no forecasting strategy is able to outperform the historical mean benchmark in purely statistical terms (using mean squared forecast errors). In contrast, we are able to detect some evidence for statistically significant economic gains for strategies based on Ferreira and Santa-Clara's (2011) sum-of-the-parts approach when exploiting our equity premium predictions in a standard mean-variance asset allocation, even after controlling for data snooping. Accounting for transaction costs, these forecasting strategies still provide superior risk-adjusted excess returns. Our results appear stronger when using recursive estimation methods, while varying model parameters negatively impacts our results. These findings caution against "mining" the overall estimation scheme. Taken together, *superior* predictive ability of any forecasting strategy relative to the prevailing mean benchmark model is hard to establish.

2. Literature review

In his recent 2017 AFA Presidential Address, Harvey (2017) provokingly argues that the competition for top journal space spurs the publication of "an embarrassing number of false positives." A search across multiple forecast models may result in the recovery of a genuinely good model, but it may also uncover a bad model that just happens to perform well in a given sample. Although data snooping and the lack of adjustment for multiple tests have been identified as major problems in financial economics, only few studies employ appropriate testing frameworks.

One of the first studies to address the problem of data snooping and correct for the search across multiple models was Sullivan, Timmermann, and White (1999), who use 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 wrongly identifying at least one forecasting model as superior (type I error) given the pre-specified significance level α. Subsequent studies extend the RC-method to reduce the influence of poor models (Hansen, 2005), account for the contribution of parameter estimation error (Corradi and Swanson, 2007), 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).

All of these testing methods have in common that they involve a composite null hypothesis, i.e., the benchmark is superior or equal to all alternative models. These types of tests are referred to as *superior* predictive ability (SPA) tests to distinguish them from *equal* predictive ability (EPA) tests. They address the question whether there is a *better* alternative to the benchmark, as opposed to whether the models are *equally* good as in the case of EPA-tests. Most EPA-tests build on the work of Diebold and Mariano (1995) and West (1996) and are used for stock return predictability when all competing models nest the benchmark (Inoue and Kilian, 2005; Rapach and Wohar, 2006a; Rapach, Strauss, and Zhou, 2013; Clark and McCracken, 2013; Neely et al., 2014). Given our research question, i.e., whether there is evidence of *superior* out-of-sample forecasting performance for at least one model, the composite hypothesis of the SPA-tests is pivotal in our setup (Elliott and Timmermann, 2016b).

There are a few studies that use SPA-tests in financial and economic applications. For example, Hansen and Lunde (2005) compare multiple volatility models and show that a GARCH(1,1) model is inferior to other volatility models for stock market returns, but is not outperformed in exchange rates data. Hsu and Kuan (2005) examine the performance of technical trading strategies and find significantly profitable strategies in relatively young markets. Similarly, Neuhierl and Schlusche (2011) test 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. Hsu, Lin, and Vincent (2017) analyze the performance of popular cross-sectional return predictors and infer that most predictor variables are no longer significant after adjusting for data snooping. Applying the multiple testing framework to evaluate the out-of-sample performance of asset allocation strategies, Hsu et al. (2018) conclude that only few 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 proportion (FDP), measured as the proportion of type I errors among all rejections, or the false discovery rate (FDR), defined as the expected FDP. Both FDP and FDR are less stringent than the FWER because they

account for the number of tested strategies. Earlier studies that implement the FDR-testing framework are 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.

3. Empirical procedure

3.1. 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, 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 possibly 'irrelevant' strategies with no hope of producing good results, thereby decreasing the power of the test, and including too few strategies, thereby overstating statistical significance (Hansen, 2005). Following Rapach and Zhou (2013), we survey forecasting strategies that have become popular in the literature, considering 28 univariate predictive regressions and 112 advanced forecasting techniques. Table 1 provides a brief overview of all 140 forecasting strategies.

[Insert Table 1 here]

Univariate predictive regressions: A simple univariate predictive regression is given as:

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

where r_{t+1} is the equity premium from period t to t+1, x_t a variable known at time t that is expected to predict the future equity premium, and ε_{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 index (including dividends) minus the log return on a one-month Treasury bill.[1]

[1] Given that predictability may increase with longer time horizons, we replicate our analyses using quarterly rather than monthly data. All results remain qualitatively unchanged and are available upon request.

While it is impossible to construct a set of all conceivable predictors and combinations, we aim to build a set that is representative of the return predictability literature. Using the updated monthly data set provided by Welch and Goyal (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).[2] Table 2 provides a description of our predictor variables.

[Insert Table 2 here]

Neely et al. (2014) highlight the predictive power of technical indicators that stems from information frictions, e.g., 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 technical indicators based on popular trend-following strategies, i.e., moving averages (Zhu and Zhou, 2009), time-series momentum (Moskowitz, Ooi, and Pedersen, 2012), and volume data (Blume, Easley, and O'Hara, 1994). In our empirical analysis, 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[3], where applicable. The construction of the technical indicators is detailed in Table 2. 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).

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 an out-of-sample equity risk premium forecast $\hat{r}_{t+1}$.[4] We employ a rolling scheme to derive the OLS parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ in order to capture uncertain model dynamics (Giacomini and White, 2006) and to account for possible breaks in the data-

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

[3] The volume data are available from http://finance.yahoo.com.

[4] Stambaugh (1999) shows that OLS coefficients of these predictive regression models are biased in small samples if the independent variables are highly persistent. Since the autoregressive coefficients of many of our predictor variables exceed 0.5, the Stambaugh (1999) bias might be a concern. Using several bias correction methods, Welch and Goyal (2008) conclude that out-of-sample prediction cannot be significantly improved. Therefore, we follow most prior studies on the out-of-sample predictability of stock returns and do not apply any bias-correction method.

generating process (Pesaran and Timmermann, 2002). Rolling windows in estimating the model parameters is appropriate given that Rapach and Wohar (2006b) show evidence for structural breaks in several predictive regression models of U.S. aggregate stock returns. In results not reported, we find a significant shift in the growth rate of the S&P 500 index itself (occurring around March 1979) when applying the structural break analysis method developed by Muggeo (2003).[5]

Forecast restrictions: Campbell and Thompson (2008) argue that the performance of univariate predictive regressions can be substantially improved by imposing restrictions on the signs of coefficients and return forecasts. Therefore, in one subset of our strategies, we impose both the restriction that the coefficient on the predictor must be of the correct sign (*positive slope*, otherwise we set the coefficient of the predictor to zero), and that the equity premium forecasts obtained from the regressions must be non-negative (*positive forecast*, otherwise we set the forecast to zero).

Regime shifts: As noted by Paye and Timmermann (2006) and Rapach and Wohar (2006b), the data-generating process for stock returns is likely subject to parameter instability because of structural breaks. Several methodological approaches were suggested to account for parameter instability. Building on Hamilton (1989), Guidolin and Timmermann (2007) estimate a multivariate Markov-switching model with four regimes – defined as crash, slow growth, bull, and recovery – and find that their model produces significant utility gains in asset allocation decisions. Exploiting the time-variation of fundamental variables, Henkel, Martin, and Nardari (2011) propose a regime-switching vector autoregressive framework with two states that closely resemble the business cycles dated by the National Bureau of Economic Research (NBER). They find that the historical average forecast is the best out-of-sample predictor in expansions, while fundamental variables provide useful information in recessions.

[5] See also Zakamulin (2015). The rolling estimation scheme is also preferable from a methodological point of view as it 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, 2016a).

In our empirical analysis, we use the state-dependent predictive regression approach of Huang et al. (2017) with the full set of fundamental variables and technical indicators.[6] As in Cooper, Gutierrez, and Hameed (2004), the market states are identified based on past return information

$$r_{t+1} = \alpha + \beta_{good} x_t I_{good,t} + \beta_{bad} x_t \left(1 - I_{good,t}\right) + \varepsilon_{t+1}, \quad (2)$$

where the indicator variable $I_{good,t}$ proxies for the market state and takes the value of one when the past six-month (log) equity premium is non-negative, and zero otherwise (Huang et al., 2017).

Shrinkage approach: As noted by Fama and French (1997), the coefficient estimates from rolling regressions can be improved by shrinking them towards a grand mean. Therefore, we use the simple shrinkage approach suggested by Connor (1997) to shrink the estimated slope coefficients from equation (1) toward zero and preserve the unconditional mean return

$$\beta^* = \frac{T}{T+s}\hat{\beta} \quad (3)$$

$$\alpha^* = \bar{r}_{t+1} - \beta^* \bar{x}_{t+1}, \quad (4)$$

where T is the length of the estimation period used to obtain $\hat{\beta}$, $\bar{r}_{t+1}$ and $\bar{x}_{t+1}$ are the means of the equity premium and the predictor variable over the estimation period, and s is the shrinkage intensity. The shrinkage factor $\frac{T}{T+s}$ drives the adjusted slope coefficient β^* toward zero, which is the prior expected value of the coefficient under the null of no predictability. The higher is s, the more weight is given to the prior of no predictability. Connor (1997) suggests that $s = 1/\rho$, where ρ is the expected explanatory power of the forecasting strategy. In our empirical analysis, we follow Ferreira and Santa-Clara (2011) and choose strong shrinkage by setting $s = 1{,}200$,[7] which can be interpreted as an R^2 of less than 0.1%.

[6] Huang et al. (2017) address Lettau and van Nieuwerburgh's (2008) critique that regime-shifting models often 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 introduced by Boyd, Hu, and Jagannathan (2005), they confirm that conventional predictive regressions are often misspecified, but show that their state-dependent approach is able to predict the equity premium in both bad and good times.

[7] Such strong shrinkage might effectively result in comparing the rolling window with the recursively estimated historical mean (benchmark model). However, the results of our SPA-tests are similar if we choose less strong shrinkage, e.g., by setting $s = 600$.

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 return forecasts captures different degrees of adaptability of forecasting strategies to structural breaks and mitigates the problem of potential model misspecification. Rapach, Strauss, and Zhou (2010) document that combinations of individual forecasts can deliver significant out-of-sample results due to reduced model uncertainty and parameter instability.

Combination forecasts are a function h of the N individual forecasts that are estimated using the predictive regression in equation (1):

$$\hat{r}_{combination,t+1} = h(\hat{r}_{1,t+1}, \hat{r}_{2,t+1}, \dots, \hat{r}_{N,t+1}). \quad (5)$$

We combine the individual forecasts based on either solely fundamental variables, solely technical indicators, or all predictors. Following Rapach and Zhou (2013), showing that simple combination forecasts work well, the mean forecast is $\hat{r}_{Mean,t+1} = \sum_{i=1}^{N} \omega_{i,t}\, \hat{r}_{i,t+1}$, with $\omega_{i,t} = \frac{1}{N}$; the median forecast is the median of $\{\hat{r}_{i,t+1}\}_{i=1}^{N}$; and the trimmed mean forecast is $\hat{r}_{Trimmed\ Mean,t+1} = \sum_{i=1}^{N} \omega_{i,t}\, \hat{r}_{i,t+1}$, with $\omega_{i,t} = 0$ 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, one could 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, the set of technical indicators, or all fundamental and technical indicators combined. Following Rapach and Zhou (2013), we use the first principal component of either only fundamental variables, only technical indicators, or all predictors. These estimated principal components then serve as independent variables in the predictive regression model in equation (1).

Kitchen sink forecast: One shortcoming of the univariate predictive regression models is that potential interdependencies between various predictor variables are not considered. Therefore, we also test a "kitchen sink" forecast (KSF), i.e., a multiple regression forecasting model that includes either solely fundamental variables, solely technical indicators, or all predictors together:

$$r_{t+1} = \alpha + \sum_{i=1}^{N} \beta_i x_{i,t} + \varepsilon_{t+1}. \tag{6}$$

LASSO: When estimating multivariate regression models such as the kitchen sink forecast, in-sample overfitting is a general concern. To alleviate this problem, Tibshirani (1996) introduces the least absolute shrinkage and selection operator (LASSO) to improve both prediction accuracy and model interpretation by shrinking OLS parameter estimates toward zero and permitting continuous shrinkage to exactly zero, i.e., performing variable selection.[8] The LASSO objective function is given by:

$$\min_{\alpha,\beta} \left(\sum_{t=1}^{T-1} \left(r_{t+1} - \alpha - \sum_{i=1}^{N} \beta_i x_{i,t} \right)^2 + \lambda \sum_{i=1}^{N} |\beta_i| \right), \tag{7}$$

where $\lambda \sum_{i=1}^{N} |\beta_i|$ is a penalty term (sum of absolute regression coefficients). If the tuning parameter $\lambda = 0$, the LASSO estimates are equivalent to the OLS estimates in equation (6). By increasing λ, the parameters are shrunk towards zero, i.e., LASSO performs both shrinkage and variable selection. To select the appropriate value for λ, we use ten-fold cross-validation and chose the value of λ that minimizes the mean cross-validated error. We apply LASSO to the multiple regression forecasting models that include either solely fundamental variables, solely technical indicators, or all predictors together.

Ridge regression: Similar to LASSO, ridge regressions attempt to alleviate the problems of in-sample overfitting and multicollinearity in linear regression models. While LASSO permits model selection that eventually yields sparse models, ridge regressions shrink all coefficients by the same factor

[8] See Rapach, Strauss, and Zhou (2013) and Rapach et al. (2015) for an application of LASSO in the context of stock return predictability as well as Freyberger, Neuhierl, and Weber (2017) on the cross-section of stock returns. Gu, Kelly, and Xiu (2019) provide a comparative analysis of machine learning techniques, e.g., LASSO and ridge regressions, for measuring asset risk premia.

but rules out variable selection, i.e., the ridge method is less flexible and does not eliminate any variables from the final regression model. The ridge regression objective function is given by:

$$\min_{\alpha,\beta}\left(\sum_{t=1}^{T-1}\left(r_{t+1}-\alpha-\sum_{i=1}^{N}\beta_i x_{i,t}\right)^2+\lambda\sum_{i=1}^{N}\beta_i^2\right), \tag{8}$$

where $\lambda\sum_{i=1}^{N}\beta_i^2$ is a penalty term (sum of squared regression coefficients). If $\lambda = 0$, the ridge regression estimates are equivalent to the OLS estimates in equation (6); by increasing λ, the parameters are shrunk towards zero. We again use ten-fold cross-validation to determine the value of λ that minimizes the mean cross-validated error. We estimate ridge regression forecasting models that include either solely fundamental variables, solely technical indicators, or all predictors together.

Elastic Net: The Elastic Net combines properties from both LASSO and ridge regression, providing parsimonious regression models through both model selection and shrinkage.[9] The Elastic Net objective function is given by:

$$\min_{\alpha,\beta}\left(\sum_{t=1}^{T-1}\left(r_{t+1}-\alpha-\sum_{i=1}^{N}\beta_i x_{i,t}\right)^2+\left(\rho\lambda\sum_{i=1}^{N}|\beta_i|+(1-\rho)\lambda\sum_{i=1}^{N}\beta_i^2\right)\right), \tag{9}$$

where $\left(\rho\lambda\sum_{i=1}^{N}|\beta_i|+(1-\rho)\lambda\sum_{i=1}^{N}\beta_i^2\right)$ is a penalty term, where $\rho = 1$ corresponds to LASSO and $\rho = 0$ to ridge regression. We set $\rho = 0.5$, use ten-fold cross-validation to select the appropriate value of λ, and estimate Elastic Nets that include either solely fundamental variables, solely technical indicators, or all predictors together.

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 stock market returns. The method offers one way of incorporating economic restrictions directly into the prediction. Returns are decomposed 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}):

[9] See Feng, Giglio, and Xiu (2019) and Kozak, Nagel, and Santosh (2019) for applications of Elastic Nets in explaining the cross-section of stock returns.

$$r_{t+1} = gm_{t+1} + ge_{t+1} + dp_{t+1} - r_{f,t+1}. \quad (10)$$

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:

$$\hat{r}_{t+1}^{SOP} = \overline{ge}_t + dp_t - r_{f,t}. \quad (11)$$

Building on Ferreira and Santa-Clara (2011), Bätje and Menkhoff (2016) develop an 'extended' sum-of-the-parts (ESOP) approach that combines the decomposition of stock market return forecasts 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 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 up the (combined) component forecasts, assuming that the dividend-price ratio follows a random walk:

$$\hat{r}_{t+1}^{ESOP} = \widehat{gm}_{t+1}^{combination,FUND} + \widehat{ge}_{t+1}^{combination,TECH} + dp_t - r_{f,t}. \quad (12)$$

Our empirical framework has two limitations. First, our choice of fundamental predictor variables is restricted to the data set provided by Welch and Goyal (2008), which is a standard and economically sound choice that has benefits in terms of cross-study comparison and replicability. However, there are other predictors in the literature that have been shown to perform well over monthly horizons. For example, Bollerslev, Tauchen, and Zhou (2009) show that the variance risk premium (VRP), i.e., the difference between implied and realized variance, has predictive power for stock market returns. Unfortunately, data for the VRP is only available starting January 1990 (which makes the evaluation period in our setup substantially shorter), and preliminary tests indicate that it performs poorly during our sample period when compared to the other fundamental predictors.

Second, we work within a class of models in which parameters are either not changing at all or only changing very slowly. Other models such as Bayesian approaches allow for time variation in both

betas and volatility (Dangl and Halling, 2012, Johannes, Korteweg, and Polson, 2014; Pettenuzzo and Ravazzolo, 2016; Bianchi and McAlinn, 2018).[10] However, the goal of our analysis is not to investigate the additional value of time-varying parameters (or stochastic volatility), but rather to analyze the consequences of multiple testing. The class of ("static") models we analyze is (still) the most frequently used in financial economics, and given that prior research claimed success within this class of models, our point is to caution researchers on this conclusion.

3.2. The multiple testing framework

When considering a large number of forecasting strategies, data snooping is a natural concern. Lo and MacKinlay (1990) show that tests of financial asset pricing models may yield misleading inferences when properties of the data are used to construct the test statistics. Data snooping occurs when a given set of data is used more than once for purposes of inference or model selection. Such a reuse leads to the possibility that any results obtained may be due to chance rather than to any merit inherent in the method yielding the results. Every time a new alternative model is considered, the yardstick for beating the benchmark must be adjusted, imposing a "tax" on mindless data mining. Hansen (2005) develops a test for superior predictive ability, or SPA-test, that allows for a comprehensive comparison of forecasting strategies, asking whether one particular forecast dominates the benchmark and all other forecasts, while ensuring that the results are robust to data snooping biases. 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 answers the question whether there is at least one superior forecasting strategy, if any, it is not able to identify all such strategies. Hsu, Hsu, and Kuan (2010) develop a stepwise extension of the SPA-test, the step-SPA-test, that is capable of identifying as many outperforming mod-

[10] Smith and Timmermann (2018) propose a new approach to return forecasting that exploits information in the cross-section of returns to rapidly detect and adapt to instability ("breaks") in return prediction models. Farmer, Schmidt, and Timmermann (2019) present an approach that lets the data determine both how large predictability is at a given point in time and how long it lasts. They provide evidence that short-horizon return predictability is quite concentrated or "local" in time and tends to fall in certain "pockets."

els 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 family-wise error rate (FWER), the probability of at least one false rejection given the pre-specified error rate α. In practice, when multiple testing involves a large number of hypotheses, incorrectly rejecting 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 probability 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 use these econometric techniques to control for data snooping when comparing the out-of-sample performance of our 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 (Welch and Goyal, 2008).

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:

$$H_0: \max_{j=1,\ldots,J} E\left(d_{j,t}\right) \equiv \mu_j \leq 0 \tag{13}$$

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 significantly outperforms the benchmark. Hansen (2005) proposes the studentized test statistic:

$$V_T^{SPA} = max\left(\max_{j=1,\ldots,J} \frac{\sqrt{T}\bar{d}_j}{\hat{\omega}_j}, 0\right) \tag{14}$$

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

To approximate the distribution of the test statistic, we follow Hansen (2005) and implement the stationary bootstrap of Politis and Romano (1994). In particular, for each strategy, we generate $b = 1, \ldots, B$ 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, which is defined as $g(\bar{d}_j) = \bar{d}_j \mathbb{1}_{\{\sqrt{T}\bar{d}_j/\widehat{\omega}_j \geq -\sqrt{2\,log\,log\,T}\}}$. The studentized test statistic under the bootstrap is computed as $V_{b,T}^{SPA^*} = max\left(\underset{j=1,\ldots,J}{max}\frac{\sqrt{T}\bar{Z}_{j,b}^*}{\widehat{\omega}_j}, 0\right)$, where $\bar{Z}_{j,b}^* = T^{-1}\sum_{t=1}^{T} Z_{j,b,t}^*$. A consistent estimate of the *p*-value is then given by:

$$\hat{p}_{SPA} = \sum_{b=1}^{B} \frac{\mathbb{1}_{\{V_{b,T}^{SPA^*} > V_T^{SPA}\}}}{B}. \tag{15}$$

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 apply the stepwise extension of the SPA-test (step-SPA-test) developed by Hsu, Hsu, and Kuan (2010) to identify all additional significant forecasting strategies, i.e., to determine if there is more than one model that beats the mean. 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.[11] Second, we remove $\bar{d}_j$ 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.

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 α. 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 α.

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

[11] In our empirical analysis, we determine the critical values for the pre-specified error rate $\alpha = 5\%$.

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. We compare the performance of the forecasting strategies with the performance of the historical mean using squared forecast errors $\left(r_t - \hat{r}_{j,t}\right)^2$, where r_t is the realized (log) equity premium, and $\hat{r}_{j,t}$ is the (log) equity premium forecast based on strategy *j*.

However, as shown by Leitch and Tanner (1991), there is only a weak association between statistical measures of forecasting performance such as the MSFE and economic forecast profitability; strategies that outperform the benchmark model in terms of MSFE often fail to outperform when using 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 based on the equity premium forecast of forecasting strategy *j*, $r_{j,t}^{abs}$, and risk-adjusted excess returns, $\frac{r_{j,t}^{abs} - r_{f,t}}{\sigma_j}$, where σ_j is the volatility of the excess return of strategy *j*, as adequate performance measures.

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*:

$$r_{j,t}^{abs} = w_{j,t} r_{j,t} + r_{f,t}, \tag{16}$$

where $w_{j,t}$ is the proportion of total wealth allocated to the stock market. For a given coefficient of relative risk aversion, γ, and a forecast of the equity premium variance, $\hat{\sigma}_t^2$, a mean-variance investor holds $w_{j,t} = \frac{\hat{r}_{j,t}}{\gamma \hat{\sigma}_t^2}$ of the risky asset. Following Neely et al. (2014), we set $\gamma = 5$ and estimate $\hat{\sigma}_t^2$ as a five-year rolling window of past monthly returns. Moreover, we impose portfolio constraints preventing investors from short-selling and levering more than 50%, so that $w_{j,t}$ is restricted between 0 and 1.5.

4. Empirical results

The sample period is from December 1950 to December 2018. To account for structural breaks in predictive relationship, 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 2018. The prevailing mean model, i.e., the recursive historical average equity premium since December 1950, serves as our benchmark model.[12]

4.1. Out-of-sample performance of forecasting strategies

Table 3 shows the out-of-sample performance of the historical mean benchmark and all forecasting strategies using the MSFE (panel A), the out-of-sample R^2 (panel B)[13], the mean monthly absolute return (panel C), and the mean monthly risk-adjusted excess return (panel D) as performance measures.

[Insert Table 3 here]

The results in panel A of Table 3 indicate that the historical mean generates lower mean squared forecast errors (MSFE of 18.83) than the average of all forecasting strategies (MSFE of 19.08). None of the univariate predictive regressions is able to outperform the historical mean, largely confirming the results of Welch and Goyal (2008) and Neely et al. (2014). Consistent with Campbell and Thompson (2008), forecast restrictions improve upon the univariate predictive regressions, but, on average, also do not outperform the historical mean. Similar findings are obtained for the shrinkage approach.

In contrast, state-dependent regressions tend to worsen the performance of univariate predictive regressions. Moreover, contrasting results of Neely et al. (2014), diffusion indices are not able to outperform the historical mean. Combination forecasts exhibit the lowest average MSFE of all forecasting strategies. Regarding the poor performance of the kitchen sink forecasts, our findings are in line with

[12] Following Simin (2008), we also consider a constant forecast as our benchmark. The constant return forecast is 0.6131% per month, estimated as the average monthly excess return over the five-year period prior to our sample period (December 1945 to November 1950). All results remain qualitatively similar and are available upon request.

[13] The out-of-sample R^2 gives the proportional reduction in MSFE for a forecasting strategy relative to the benchmark model and is computed as $R^2_{OOS} = 1 - MSFE_j/MSFE_{HistAvg}$, where $MSFE_j$ ($MSFE_{HistAvg}$) denotes the MSFE of forecasting strategy *j* (the historical mean benchmark).

Welch and Goyal (2008) and Rapach, Strauss, and Zhou (2010). Using LASSO regressions, Ridge regressions, and Elastic Nets improves upon the kitchen sink forecasts, but they still underperforms the historical mean. The sum-of-the-parts models do not outperform the historical mean, on average, but contain the best of all forecasting strategies, the SOP-approach developed by Ferreira and Santa-Clara (2011) with a MSFE of only 18.58. Overall, our results confirm that many forecasting strategies fail to outperform the historical mean when evaluated based on forecast errors.

To better understand why most forecasting strategies struggle to beat the historical mean in terms of forecast errors, we decompose the MSFE according to Theil (1971). He proposes the following MSFE decomposition: $MSFE = \left(\bar{\hat{r}} - \bar{r}\right)^2 + (\sigma_{\hat{r}} - \rho\sigma_r)^2 + (1-\rho^2)\sigma_r^2$, where $\bar{\hat{r}}$ ($\bar{r}$) is the average predicted (actual) return, $\sigma_{\hat{r}}$ (σ_r) is the standard deviation of predicted (actual) returns, and ρ is the correlation coefficient between predicted and actual returns. For our analysis, we follow Rapach, Strauss, and Zhou (2010) and assume that ρ is close to zero, such that $MSFE \approx \left(\bar{\hat{r}} - \bar{r}\right)^2 + \sigma_{\hat{r}}^2 + \sigma_r^2$.

Based on this this decomposition, Figure 1 plots the forecast variance $\left(\sigma_{\hat{r}}^2\right)$ of all strategies against their squared bias $\left(\left(\bar{\hat{r}} - \bar{r}\right)^2\right)$. The scatterplot reveals that most forecasting strategies produce relatively unbiased return predictions, many of them even better than the historical mean benchmark. However, their performance in terms of forecast error relative to the historical mean is negatively affected by their forecast variance, confirming the results of Rapach, Strauss, and Zhou (2010). None of the examined forecasting strategies exhibit a lower forecast variance than the historical average forecast.

[Insert Figure 1 here]

Turning to forecast profitability, the results in panels C and D of Table 3 indicate that an investor with mean-variance preferences can profoundly benefit from forecasting the equity premium. All forecasting strategies – except shrinkage regressions – outperform the historical mean both in terms of mean absolute returns and in terms of mean risk-adjusted excess returns, on average. Taking another look at Figure 1, we note that the most profitable forecasting strategies are not necessarily the ones producing

the lowest forecast variance. This observation lends support to Leitch and Tanner's (1991) conclusion that there is only a weak relationship between statistical performance measures and forecast profitability.

While Table 3 provides a first indication as to which forecasting strategies may offer an improvement upon the historical mean, the analysis does not account for data snooping. To address this concern, we apply Hansen's (2005) SPA-test and its extensions. Given that most advanced strategies are expected to improve upon univariate predictive regressions, in a first step we test each subset of advanced strategies separately, each time including the univariate predictive regressions in the test sample. However, testing only subsets of forecasting strategies is subject to data mining since the results do not incorporate the full set of strategies. To impose the most stringent test for superior predictive ability, we assess the performance of all strategies jointly against the historical benchmark forecast in a second step.

4.2. Test results based on mean squared forecast errors

In Table 4, we test whether any forecasting strategy can more accurately forecast the equity premium than the historical mean in terms of MSFE using Hansen's (2005) SPA-test. Column (1) gives the set of forecasting strategies we draw from. Column (2) identifies the 'most significant' strategy, i.e., the strategy with the lowest nominal *p*-value that results from 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) shows the consistent *p*-value together with 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., at least one strategy is better than the historical mean in terms of MSFE. Column (5) indicates the numbers of significant strategies identified by the step-SPA-test and the step-SPA(3)-test.

[Insert Table 4 here]

The first row in Table 4 shows the results of the SPA-test for the subset of univariate predictive regressions and forecast restrictions. The restricted forecast using the T-bill rate (TBL), denoted as TBL (rest.), is selected as the most significant strategy. However, the nominal *p*-value of 0.2387 indicates that this model is not able to outperform the historical mean when considered in isolation. The consistent

p-value of 0.9770 suggests 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 4, all state-dependent regressions are dominated by a univariate predictive regression based on the term spread (TMS), which is 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 significantly different from the MSFE of the historical mean benchmark (nominal *p*-value of 0.5285). Accordingly, the null hypothesis of the SPA-test also cannot be rejected for this subset (consistent *p*-value of 0.9975). Similar results apply for the subset of strategies that include the shrinkage approach (third row), combination forecasts (fourth row), the diffusion indices (fifth row), kitchen sink forecasts (sixth row), LASSO regressions (seventh row), Ridge regressions (eights row), and Elastic Nets (ninth row).

The results in the tenth row indicate that the SOP-approach performs better than the univariate predictive regressions in this subset. It significantly outperforms the historical mean when considered in isolation (nominal *p*-value of 0.0228), but the null hypothesis of the SPA-test cannot be rejected at the 5% level of significance (consistent *p*-value of 0.2201).

Finally, when we turn to the results in the last row of Table 4, 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 (step-SPA(3)-test), none of the forecasting strategies is identified as superior.

Taken together, the results in Table 4 indicate that many forecasting strategies do not outperform the historical mean once accounting for potential data snooping biases. Only Ferreira and Santa-Clara's (2011) SOP-approach shows a marginal superiority compared to both univariate predictive regressions and the historical mean. However, 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.

These findings support Torous and Valkanov (2000), showing that forecasts using predictive regression models will do no better than the simple unconditional mean when the signal-to-noise ratio is small even if a forecasting relationship is assumed to prevail between returns and some predictor. Using Monte Carlo simulations, they show that, once the signal-to-noise ratio drops below 0.1, forecasts from the 'true' model and from the unconditional mean produce similar results in terms of MSFE. Even for large sample sizes, the true model cannot be detected if the signal-to-noise ratio is low. In our sample, the average signal-to-noise ratio of predictors is very low at 0.03, i.e., failure to detect superior forecasting power might be attributable to 'noisy' predictors that do not generate sufficiently strong signals.[14]

4.3. Test results using recursive estimation scheme

One could argue that the mostly insignificant nominal *p*-values in the single setting in Table 4, i.e., when we test only one hypothesis at a time, suggest that we are not able to fully replicate the out-of-sample predictability shown in earlier research. As we are able to replicate the main findings in earlier studies when using the same time periods and research designs, these discrepancies might be attributable to either the divergent evaluation period or the use of a rolling instead of a recursive estimation scheme.

To ensure comparability with previous studies, we also estimate recursively. Although recursively estimated parameters are not compatible with the stationarity requirement of the SPA-test – making sure that the test statistics are asymptotically normally distributed, because the variance of prediction errors decreases with time (see Hansen (2005) for a discussion) – we show the results of pseudo-SPA-tests using recursive estimation in Appendix 1.[15] Our results are now "more statistically significant" when only one hypothesis is tested at a time. Of the ten tested forecasting strategy subsets, five exhibit a nominal *p*-value below 10%, and the nominal *p*-value when testing all strategies together is below 5%.

[14] Following Torous and Valkanov (2000), the signal-to-noise ratio is $\tau = \frac{\sigma_u \beta}{\sigma_\varepsilon}$, where σ_ε is the standard deviation of the disturbance term in equation (1), β the slope coefficient, and σ_u the standard deviation of the disturbance term from fitting an AR(1) model to the predictor variable x. We estimate the signal-to-noise ratio for each predictor using the sample analogues of the standard deviations and the OLS estimate of β over the full sample period.

[15] Granziera, Hubrich, and Moon (2014) also note that SPA-tests are appropriate for comparisons where at least one of the alternative models does not nest the benchmark, whereas they should not perform as well as EPA-tests when used in nested model comparisons.

However, our main result holds, i.e., we still find no statistically significant evidence that any forecasting strategy is better than the historical mean in terms of MSFE within the multiple testing framework.

We also use the reality check proposed by Clark and McCracken (2012) based on the *maxMSFE-F statistic*. This test is appropriate for comparing the forecasts from multiple strategies that all nest the benchmark model, as is the case in a recursive estimation scheme. The *maxMSFE-F statistic* does not rejects the null hypothesis that none of the forecasting strategies outperforms the historical mean at the 5% level of significance, which supports our previous results.

4.4. Test results based on economic forecast profitability

Table 5 shows the results of data snooping tests using mean monthly absolute (panel A) and risk-adjusted excess returns (panel B) that an investor with mean-variance preferences and risk aversion coefficient $\gamma = 5$ can generate over the out-of-sample period.[16] The results of panel A indicate regressions including shrinkage (third row), forecast combinations (fourth row), diffusion indices (fifth row), LASSO regressions (seventh row), and Elastic Nets (ninth row) do not show a better performance than the univariate predictive regression based on the term spread (TMS), which is selected as the most significant strategy in each subset (nominal *p*-value of 0.0489). Nevertheless, the null hypothesis of the SPA-test cannot be rejected for any of these subsets (all with consistent *p*-values above 5%).

The state-dependent regression based on the term spread, denoted as TMS (st.dep.), marginally improves upon its univariate predictive regression (nominal *p*-value of 0.0333). Similarly, the restricted forecast using the default yield spread as a predictor, DFY (rest.), the kitchen sink forecast based on all indicators, KSF (ALL), and Ridge regressions based on fundamental predictors, Ridge (FUND), exhibit more significant performance in isolation (nominal *p*-values of 0.0196, 0.0081, and 0.0166, respectively). However, we are not able to reject the null hypothesis of the SPA-test in either subset (consistent *p*-values of 0.3360, 0.2681, 0.0980, and 0.1610, respectively).

[16] The corresponding results for the forecasting strategies based on the recursive estimation scheme are presented in Appendix 2. The pseudo-SPA-test results are similar to those obtained under the rolling estimation scheme.

[Insert Table 5 here]

Turning to the subset including the sum-of-the-parts models (tenth row), the extended sum-of-the-parts strategy using median combination forecasts, ESOP (Median), is selected as the most significant strategy in a pairwise comparison against the historical mean (nominal *p*-value of 0.0031). Furthermore, we can reject the null hypothesis of the SPA-test at the 5% significance level (consistent *p*-value of 0.0500). If we allow up to three false rejections, the step-SPA(3)-test identifies all three extended sum-of-the-parts models as being superior to the historical mean in this subset. Finally, the null hypothesis of the SPA-test can be marginally rejected for the full set of forecasting strategies with a consistent *p*-value of 0.0934, with one forecasting strategy, ESOP (Median), being identified as superior based on the step-SPA(3)-test at the 5% level of significance.

As shown in panel B of Table 5, in most subsets, restricted, state-dependent, or simple forecasts using TBL are selected as the most significant strategies. While these strategies are able to outperform the historical mean when considered in isolation (nominal *p*-values between 0.0373 and 0.0531), their superiority is not robust to data snooping bias corrections, as indicated by large consistent *p*-values, all exceeding 5%. In the subset including combination forecasts (fourth row), the most significant strategy in a pairwise comparison against the historical mean is the Mean (FUND) strategy (nominal *p*-value of 0.0128). However, we again cannot reject the null hypothesis of the SPA-test for this subset (consistent *p*-value of 0.1231). Similar results are obtained for the subsets including the kitchen sink forecasts (sixth row), LASSO regressions (seventh row), Ridge regressions (eight row), and Elastic Net (ninth row).

The null hypothesis of the SPA-test can be rejected for the subset including the sum-of-the-parts models (tenth row; consistent *p*-value of 0.0130). The ESOP (Median) strategy is selected as the most significant strategy (nominal *p*-value of 0.0007), and the step-SPA-test identifies all three extended sum-of-the-parts strategies as superior to the historical mean. Finally, the null hypothesis of the SPA-test can also be rejected at the 5% level when accounting for all forecasting strategies (last row; consistent *p*-value of 0.0258). The step-SPA-test corroborates that at least the ESOP (Median) strategy significantly outperforms the historical mean, i.e., is superior even after accounting for data snooping biases.

Another issue when comparing the performance of multiple forecasting strategies relative to the historical mean are transaction costs. While the prevailing mean model dictates a relatively passive investment strategy, many of the forecasting strategies generate much more volatile return forecasts, such that the trading profits might be eroded by higher transaction costs (Pesaran and Timmermann, 1995). To assess the extent to which our results are influenced by the frequent trading associated with most forecasting strategies, we repeat our analyses of Table 5, while incorporating realistic transaction costs. Specifically, we follow the choice in Lynch and Balduzzi (2000) and assume 25 basis points as roundtrip transaction costs.[17] Table 6 summarizes the results.

[Insert Table 6 here]

The nominal *p*-values of the most significant strategies increase due to the smaller performance differences between the historical average forecast and the respective forecasting strategy as a result of incorporating transaction costs. Many strategies are no longer significant at the 5% level of significance even when considered in isolation. Only in the subsets including restricted forecasts (first row), kitchen sink forecasts (sixth row), and the sum-of-the-parts models (tenth row), we identify at least one strategy that is significantly superior relative to the historical mean in a pairwise comparison. When using mean absolute returns (panel A), this superiority is no longer robust to data snooping corrections, as indicated by the consistent *p*-value exceeding 5% when accounting for all strategies. However, on a risk-adjusted excess return basis (panel B), we are still able to identify all three extended sum-of-the-parts models as superior within the full set of forecasting strategies if allowing up to three false rejections.

Overall, our results suggest that the prevailing mean model is extremely hard to beat. We identify only very few forecasting strategies, the extended sum-of-the-parts models, that can offer superior performance to investors relative to the historical mean benchmark when used in a traditional mean-variance optimization that is robust to data snooping concerns.

[17] Given that our strategies can be implemented in practice by using, for example, highly liquid futures, transaction costs might be small (Solnik, 1993), and 25 basis points should be considered reasonably conservative. Therefore, we repeat our analysis using 10 basis points as roundtrip transaction costs. The results remain very similar.

5. Robustness checks

5.1. Variation of input parameters

So far, we prevent the investor from short selling and leveraging more than 50%. We repeat our baseline analysis of Table 6 without portfolio constraints. In results not reported, there is an increase of consistent *p*-values in the SPA-test in most subsets regardless of the performance measure used.[18]

Moreover, our analysis assumes a conservative investor with coefficient of relative risk aversion $\gamma = 5$. Repeating our tests for a more aggressive investor with $\gamma = 1$ (results not reported), we fail to reject the null hypothesis of the SPA-tests under the mean absolute return criterion. However, we identify the ESOP (Mean) strategy as being superior in terms of mean risk-adjusted excess returns within the full set of forecasting strategies even after accounting for the data snooping bias.

5.2. Rolling sub-sample period analysis

Although the bootstrap procedure implemented for the SPA-tests provides us with 'artificial' subsamples, it still relies on the original data series and is thus influenced by the choice of the evaluation period. Therefore, to verify whether the results are robust to the choice of the evaluation period, we repeat our analyses using shorter evaluation periods of ten years each. 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 results, using the mean absolute return over the respective ten-year sub-sample period for an investor with mean-variance preferences and a relative risk aversion coefficient $\gamma = 5$ as the performance measure (including transaction costs of 25 basis points).

[Insert Figure 2 here]

[18] These results can be interpreted in the spirt of Jagannathan and Ma (2003), arguing that no-short-sales constraints and upper bounds on portfolio weights can reduce sampling error and lead to improvements in the out-of-sample performance of the optimal portfolio even when the constraints are effectively wrong.

In panel A, we plot the mean absolute return over the respective ten-year sub-sample period of the benchmark model (black line) and each forecasting strategy *j* (grey crosses). Confirming our previous findings, there are many forecasting strategies that generate a higher mean absolute return than the historical mean model. We identify some forecasting strategies that substantially outperform all other strategies, especially in the mid-80s to mid-90s and towards the end of our sample period, as indicated by the outliers in the scatter plot. Moreover, the analysis reveals that forecasting the equity premium was especially beneficial during and after the global financial crisis of 2007-2009, when the historical mean model generated negative mean absolute returns. The performance of the forecasting strategies also became more dispersed during this period, but ex ante the correct choice of the outperforming strategies is not straightforward.

Panel B plots the time series of the consistent *p*-value over the associated ten-year sub-sample period. If we assume that equity premium predictability should be disappearing over time, we expect fewer or no forecasting strategies to outperform the benchmark model in the more recent past, leading to a gradual increase of the consistent *p*-value. However, as illustrated in panel B, the consistent *p*-value fluctuates strongly between 0.8011 (ten-year sub-sample period ending in 2005) and 0.0362 (ten-year sub-sample period ending in 2014). We only observe one sub-sample with a statistically significant out-of-sample predictability at the 5% level of significance, which is the ten-year period ending in 2014. The step-SPA(3)-tests identify six additional time periods that contain significant strategies (the ten-year periods ending in 1980, 1990, 1999, 2013, 2015, and 2016). However, this observed predictability is highly sensitive to the selected time window, i.e., predictability is no longer observable when we roll the ten-year window by only one year in either direction. In addition, the best (or most significant) strategy varies over both time windows. While the univariate predictive regression based on the default yield spread (DFY) was the best strategy during the ten-year period ending in 1980, the forecasting strategies based on the net equity expansion (NTIS) and the restricted forecast using a volume-based indicator that signals the crossing of the three-months moving average of the 'on balance' volume and the respective nine-months moving average (VOL 3-9 rest.), provided the highest performance measures during the periods ending in 1990 and 1999, respectively. Exceptions are the extended sum-of-the-parts

models, whose performances are more stable, dominating the ten-year windows ending in 2013, 2014, and 2015. During the period ending 2016, the kitchen sink forecast based on fundamentals variables, KSF (FUND), exhibits the best performance.

To further examine which type of forecasting models performs well over time, Figure 3 illustrates a 'heat map' of the average performance of each forecasting technique group in terms of mean absolute return over the ten-year sub-sample periods, i.e., we average the mean absolute return of each forecasting strategy within a forecasting technique group over the respective ten-year sub-sample period. To simplify comparison, we further standardize the average performance of each forecasting technique group within each sub-sample period (labelled 'Column *Z*-Score') and track the development of the *z*-score over the rolling sub-sample periods. Univariate predictive regressions and forecast restrictions rank relatively stable through time compared to other sets of forecasting strategies. Conversely, kitchen sink forecasts and the prevailing historical mean forecast exhibit distinctive periods of inferior performance. Only the sum-of-the-parts models dominate the other strategy sets over extended periods of time; however, even these models fall behind other forecasting strategies in the early 2000s.

5.3. How large does predictability need to be?

Given that only a few forecasting strategies are able to outperform the historical mean benchmark, there might be concerns that our results are due to a lack of power of the SPA-test in the specific research setup. While Hansen (2005), Hsu, Hsu, and Kuan (2010), and Hsu, Kuan, and Yen (2014) show that the SPA-test and its stepwise extensions exhibit good power properties using Monte Carlo simulations, we follow the 'post estimation' design in Hsu et al. (2018) to gain a better understanding of how large the predictability of a forecasting strategy, expressed as the return difference against the benchmark, needs to be to be identified as superior. In particular, we simulate 100 artificial forecasting strategies with excess monthly returns as follows:

$$r_{a,t}^{exc} = (H - \rho)\mu + \rho r_{HistAvg,t}^{exc} + \sqrt{1-\rho^2}\sigma\varepsilon_{a,t} \text{ for } a = 1,\ldots, 100, \quad (17)$$

where $r_{HistAvg,t}^{exc}$ are the excess returns of the benchmark model, i.e., the historical average forecast, over the out-of-sample period from January 1966 to December 2018 (including transaction costs of 25 basis points), μ and σ are the mean and standard deviation of those excess returns, ρ is the average correlation coefficient between all originally tested forecasting strategies (Table 6) and the benchmark model over the same period, and $\varepsilon_{a,t}$ are zero-mean disturbance terms. Based on sample data, the parameters are chosen as follows: $\mu = 0.25\%$, $\sigma = 3.64\%$, and $\rho = 0.82$. This simulation design ensures that the artificial forecasting strategies exhibit approximately the same mean return and standard deviation as the benchmark model and similar correlation with the originally tested forecasting strategies.

To examine by how much the excess return of a forecasting strategy has to exceed the mean return of the benchmark model μ to be identified as being superior, we use the scaling parameter H. In particular, we choose $H = \{1.1, 1.5, 2.0, 2.5, 3.0\}$, i.e., an increase by 10% up to 200% over the mean return of the benchmark model. Using the simulated excess returns, we compute the risk-adjusted excess returns of each artificial strategy a, merge the artificial strategies with our original set of forecasting strategies and conduct the SPA-tests using risk-adjusted excess returns as the performance measure. In accordance with Hsu et al. (2018), we repeat this simulation test 100 times. We are interested in the proportion of superior strategies rejected by the SPA-tests. The results are summarized in Table 7.

When H is set to 1.1, the mean consistent *p*-value of the SPA-test over all 100 trials is 0.1268. Therefore, contrasting the results in panel B of Table 6, we cannot reject the null hypothesis of the SPA-tests, on average, at the 5% level of significance. This highlights the danger of excessive data mining given that the SPA-test is not entirely immune to the inclusion of (a large number of) poor forecasts. When we further increase H, the mean consistent *p*-value of the SPA-test decreases, and the percentage of superior strategies that we identify increases accordingly. With $H = 3.0$, we are able to identify (almost) all superior strategies in the test set. While an increase by 200% over the mean return of the benchmark model may seem very large, note that the standard deviation of the benchmark model and the average standard deviation of the forecasting strategies are 3.70% and 3.42%, which are 13.8 and

12.8 times the mean of the benchmark model and larger than $H = 3.0$, respectively. We thus conclude that our testing methods are sufficiently powerful to identify superior forecasting strategies.

6. Conclusion

In this study, we jointly examine the out-of-sample performance of a comprehensive set of forecasting strategies relative to the historical mean as the benchmark. Most advanced forecasting strategies were documented in the literature to be superior against the historical mean and also against some of the simpler univariate predictive regressions. However, an immediate concern is that the best model may appear to beat the benchmark simply by luck or as a result of inspecting multiple models.

When we control for the potential data snooping bias within a multiple testing framework, we fail to establish superior predictive ability in terms of mean squared forecast errors, although identify several strategies based on Ferreira and Santa-Clara's (2011) sum-of-the-parts approach that significantly outperform the historical mean in terms of economic profitability. This result still holds after accounting for realistic transaction costs, at least on a risk-adjusted excess return basis. However, our rolling window analysis reveals that even when we detect a data snooping resistant superior model, this superiority is highly sub-sample dependent. Changing the estimation scheme, e.g., estimating parameters recursively rather than rolling, or assuming different model parameters, impact our results as well. Overall, by focusing on the application of equity premium prediction, our results support Harvey's (2017) more general concern that many of the published results in financial economics will fail to hold up.

We note that a caveat of the SPA-test and our testing framework is that it treats all forecasting strategies equally. However, some predictors are derived from economic theory, e.g., the dividend-price ratio within the present-value identity (Campbell and Shiller, 1988), and should thus be less susceptible to data snooping (Harvey, Liu, and Zhu, 2016). Finally, while we work within a class of models in which parameters are either not changing at all or only changing very slowly, an analysis of the value of time-varying parameters models is clearly important. We leave these issues as an avenue for further research.

References

Bajgrowicz, P., and Scaillet, O., 2012, Technical Trading Revisited: False Discoveries, Persistence Tests, and Transaction Costs, *Journal of Financial Economics* 106, 473–491.

Barras, L., Scaillet, O., and Wermers, R., 2010, False Discoveries in Mutual Fund Performance: Measuring Luck in Estimated Alphas, *Journal of Finance* 65, 179–216.

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

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

Bianchi, D., and McAlinn, K., 2018, Large-Scale Dynamic Predictive Regressions, Working Paper.

Bollerslev, T., Tauchen, G., and Zhou, H., 2009, Expected Stock Returns and Variance Risk Premia, *Review of Financial Studies* 22, 4464–4492.

Boyd, J.H., Hu, J., and Jagannathan, R., 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., and Shiller, R.J., 1988, The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors, *Review of Financial Studies* 1, 195–228.

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

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

Clark, T., and McCracken, M., 2013, *Advances in Forecast Evaluation*, in: Elliott, G. and A. Timmermann, Handbook of Economic Forecasting Volume 2B, Elsevier B.V., Amsterdam.

Connor, G., 1997, Sensible Return Forecasting for Portfolio Management, *Financial Analysts Journal* 53, 44–51.

Cooper, M.J., Gutierrez, R.C., and Hameed, A., 2004, Market States and Momentum, *Journal of Finance* 59, 1345–1365.

Corradi, V., and Swanson, N.R., 2007, Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes, *International Economic Review* 48, 67–109.

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

Diebold, F.X., and Mariano, R.S., 1995, Comparing Predictive Accuracy, *Journal of Business & Economic Statistics* 13, 253–263.

Elliott, G., and Timmermann, A., 2016a, *Economic Forecasting*, Princeton University Press, Princeton.

Elliott, G., and Timmermann, A., 2016b, Forecasting in Economics and Finance, *Annual Review of Economics* 8, 81–110.

Fama, E.F., and French, K., 1997, Industry Costs of Equity, *Journal of Financial Economics* 43, 153–193.

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

Farmer, L., Schmidt, L., and Timmermann, A., 2019, Pockets of Predictability, Working Paper.

Feng, G., S. Giglio, and D. Xiu, 2019, Taming the Factor Zoo: A Test of New Factors, Working paper.

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

Foster, F.D., Smith, T., and Whaley, R.E., 1997, Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2, *Journal of Finance* 52, 591–607.

Freyberger, J., Neuhierl, A., and Weber, M., 2017, Dissecting Characteristics Nonparametrically, Working paper.

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

Granziera, E., Hubrich, K., and Moon, H.R., 2014, A Predictability Test for a Small Number of Nested Models, *Journal of Econometrics* 182, 174–185.

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

Gu, S., B. Kelly, and D. Xiu, 2019, Empirical Asset Pricing via Machine Learning, Working paper.

Hamilton, J.D., 1989, A New Approach to the Economic Analysis of Non-Stationary 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.

Hansen, P.R., and Lunde, A., 2005, A Forecast Comparison of Volatility Models: Does Any-thing Beat a GARCH(1,1)?, *Journal of Applied Econometrics* 20, 873–889.

Harvey, C.R., 2017, Presidential Address: The Scientific Outlook in Financial Economics, *Journal of Finance* 72, 1399–1440.

Harvey, C.R., Liu, Y., and Zhu, H., 2016, …and the Cross-Section of Expected Returns, *Review of Financial Studies* 29, 5–68.

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

Hong, H., and Stein, J.C., 1999, A Unified Theory of Underreaction, Momentum Trading, and Overreaction in Asset Markets, *Journal of Finance* 54, 2143–2184.

Hsu, P.-H., Han, Q., and Wu, W., and Z. Cao, 2018, Asset Allocation Strategies, Data Snooping, and the 1/N Rule, *Journal of Banking and Finance* 97, 257–269.

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

Hsu, P.-H., and Kuan, C.-M., 2005, Reexamining the Profitability of Technical Analysis with Data Snooping Checks, *Journal of Financial Econometrics* 3, 606–628.

Hsu, Y.-C., Kuan, C.-M., and Yen, M.-G., 2014, A Generalized Stepwise Procedure with Improved Power for Multiple Inequalities Testing, *Journal of Financial Econometrics* 12, 730–755.

Hsu, Y.-C., Lin, H.-W., and Vincent, K., 2017, Do Cross-Sectional Stock Return Predictors Pass the Test without Data-Snooping Bias, Working paper.

Huang, D., Jiang, F., and Tu, J., and G. Zhou, 2017, Forecasting Stock Returns in Good and Bad Times: The Role of Market States, Working paper.

Inoue, A., and Kilian, L., 2005, In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?, *Econometric Reviews* 23, 371–402.

Jagannathan, R., and Ma, T., 2003, Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps, *Journal of Finance* 58, 1651–1683.

Johannes, M., Korteweg, A., and Polson, N., 2014, Sequential Learning, Predictability, and Optimal Portfolio Returns, *Journal of Finance* 69, 611–644.

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

Kozak, S., S. Nagel, and S. Santosh, 2019, Shrinking the Cross-Section, forthcoming in: *Journal of Financial Economics*.

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

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

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

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

Lynch, A.W., and P. Balduzzi, 2000, Predictability and Transaction Costs: The Impact on Rebalancing Rules and Behavior, *Journal of Finance* 55, 2285–2309.

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

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

Muggeo, V.M.R., 2003, Estimating Regression Models with Unknown Break-Points, *Statistics in Medicine* 22, 3055–3071.

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

Neuhierl, A., and Schlusche, B., 2011, Data Snooping and Market-Timing Rule Performance, *Journal of Financial Econometrics* 9, 550–587.

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

Pesaran, M.H., and Timmermann, A., 1995, Predictability of Stock Returns: Robustness and Economic Significance, Journal of Finance 50, 1201–1228.

Pesaran, M.H., and Timmermann, A., 2002, Market Timing and Return Prediction Under Model Instability, Journal of Empirical Finance 9, 495–510.

Pettenuzzo, D., and Ravazzolo, F., 2016, Optimal Portfolio Choice under Decision-Based Model Combinations, *Journal of Applied Econometrics* 31, 1312–1332.

Politis, D.N., and Romano, J.P., 1994, Large Sample Confidence Regions Based on Subsamples under Minimal Assumptions, *Annals of Statistics* 22, 2031–2050.

Rapach, D.E., Strauss, J., and Tu, J., and G. Zhou, 2015, Industry Interdependencies and Cross-Industry Return Predictability, Working paper.

Rapach, D.E., Strauss, J.K., and Zhou, G., 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., Strauss, J.K., and Zhou, G., 2013, International Stock Return Predictability: What Is the Role of the United States?, *Journal of Finance* 68, 1633–1662.

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

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

Romano, J.P., and Wolf, M., 2005, Stepwise Multiple Testing as Formalized Data Snooping, *Econometrica* 73, 1237–1282.

Romano, J.P., and Wolf, M., 2007, Control of Generalized Error Rates in Multiple Testing, *Annals of Statistics* 35, 1378–1408.

Simin, T., 2008, The Poor Predictive Performance of Asset Pricing Models, *Journal of Financial and Quantitative Analysis* 43, 355–380.

Smith, S., and Timmermann, A., 2018, Break Risk, Working Paper.

Solnik, B., 1993, The Performance of International Asset Allocation Strategies using Conditioning Information, *Journal of Empirical Finance* 1, 33–55.

Spiegel, M., 2008, Forecasting the Equity Premium: Where We Stand Today, *Review of Financial Studies* 21, 1453–1454.

Stambaugh, R.F., 1999, Predictive Regressions, Journal of Financial Economics 54, 375–421.

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

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

Theil, H., 1971, *Applied Economic Forecasting*, Elsevier B.V., Amsterdam.

Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* 58, 267–288.

Timmermann, A., 2006, *Forecast Combinations*, in: Elliott, G., C. Granger, and A. Timmermann, Handbook of Economic Forecasting Volume 1, Elsevier B.V., Amsterdam.

Torous, W., and Valkanov, R., 2000, Boundaries of Predictability: Noisy Predictive Regressions, Working paper.

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

West, K.D., 1996, Asymptotic Inference about Predictive Ability, *Econometrica* 64, 1067–1084.

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

Zakamulin, V., 2015, A Comprehensive Look at the Empirical Performance of Moving Average Trading Strategies, Working paper.

Zhu, Y., and Zhou, G., 2009, Technical Analysis: An Asset Allocation Perspective on the Use of Moving Averages, *Journal of Financial Economics* 92, 519–544.

Tables

Table 1
Overview of forecasting strategies

This table provides an overview of all forecasting strategies included in our empirical analysis. Column (1) lists the different forecasting techniques we use (see section 3.1 for more details). Columns (2) through (4) list the forecasting strategies within each forecasting technique group that rely solely on fundamental variables, solely on technical indicators, or both (all predictors). Column (5) shows the number of forecasting strategies within each forecast technique group.

Forecasting techniques	Based on fundamental variables	Based on technical indicators	Based on all predictors	Total
Univariate predictive regressions	14 fundamental variables following Welch and Goyal (2008)	14 technical indicators following Neely et al. (2014)		28
Forecast restrictions	14 fundamental variables following Welch and Goyal (2008)	14 technical indicators following Neely et al. (2014)		28
State-dependent predictive regressions	14 fundamental variables following Welch and Goyal (2008)	14 technical indicators following Neely et al. (2014)		28
Shrinkage approach	14 fundamental variables following Welch and Goyal (2008)	14 technical indicators following Neely et al. (2014)		28
Combination forecasts	Mean (FUND), Median (FUND), Trimmed Mean (FUND)	Mean (TECH), Median (TECH), Trimmed Mean (TECH)	Mean (ALL), Median (ALL), Trimmed Mean (ALL)	9
Diffusion indices	PC (FUND)	PC (TECH)	PC (ALL)	3
Kitchen sink forecasts	KSF (FUND)	KSF (TECH)	KSF (ALL)	3
LASSO regressions	LASSO (FUND)	LASSO (TECH)	LASSO (ALL)	3
Ridge regressions	Ridge (FUND)	Ridge (TECH)	Ridge (ALL)	3
Elastic Net	EN (FUND)	EN (TECH)	EN (ALL)	3
Sum-of-the-parts models			SOP, ESOP (Mean), ESOP (Median), ESOP (Trimmed Mean)	4
All forecasting strategies	**64**	**64**	**12**	**140**

Table 2

Definition of predictor variables

This table defines the predictor variables that are used in our forecasting strategies (as listed in Table 1).

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 stocks listed on the New York Stock Exchange (NYSE) to the total end-of-year market capitalization of NYSE-listed 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 $MAs\text{-}l_t = \begin{cases} 1 \; if \; MA_{s,t} \geq MA_{l,t} \\ 0 \; if \; MA_{s,t} < MA_{l,t} \end{cases}$ for $s = \{1,2,3\}$ and $l = \{9,12\}$ where $MA_{j,t} = \frac{1}{j}\sum_{i=0}^{j-1} P_{t-i}$ for $j = s, l$.
MOMm	Momentum-indicator, calculated as the difference between the current stock price and the stock price *m* months ago $MOMm = \begin{cases} 1 \; if \; P_t \geq P_{t-m} \\ 0 \; if \; P_t < P_{t-m} \end{cases}$ 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 $VOLs\text{-}l_t = \begin{cases} 1 \; if MA_{s,t}^{OBV} \geq MA_{l,t}^{OBV} \\ 0 \; if MA_{s,t}^{OBV} < MA_{l,t}^{OBV} \end{cases}$ for $s = \{1,2,3\}$ and $l = \{9,12\}$ where $MA_{j,t}^{OBV} = \frac{1}{j}\sum_{i=0}^{j-1} OBV_{t-i}$ for $j = s, l$ where $OBV_t = \sum_{k=1}^{t} VOL_k \, D_k$ where $D_k = \begin{cases} 1 \; if P_k \geq P_{k-1} \\ -1 if P_k < P_{k-1} \end{cases}$.

Table 3

Out-of-sample performance of forecasting strategies

This table reports the performance of the benchmark model and all forecasting strategies: univariate predictive regressions, forecast restrictions, state-dependent regressions, shrinkage approach, combination forecasts, diffusion indices, kitchen sink forecasts, LASSO regressions, Ridge regressions, Elastic Nets, sum-of-the-parts models, as well as all forecasting strategies combined over the out-of-sample period from January 1966 to December 2018. The benchmark model is the historical mean. The out-of-sample performance is evaluated based on mean squared forecast errors (panel A), out-of-sample R^2 (panel B), the mean absolute return (panel C, in % per month), and the mean risk-adjusted excess return (panel D, in % per month). We assume an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$. The "best" strategy is the forecasting strategy with the best performance measure, i.e., the lowest mean squared forecast error or the highest out-of-sample R^2, mean absolute and risk-adjusted excess return. All forecasting strategies are described in section 3.1.

	Panel A: MSFE			**Panel B**: Out-of-sample R^2			**Panel C**: Mean absolute return			**Panel D**: Mean risk-adj. excess return		
	Mean	**Min**	**Best**	**Mean**	**Min**	**Best**	**Mean**	**Max**	**Best**	**Mean**	**Max**	**Best**
Benchmark	18.83						0.67			0.07		
Univariate pred. regressions	19.04	18.84	TMS	-1.13	-0.08	TMS	0.71	0.83	TMS	0.09	0.12	TBL
Forecast restrictions	18.92	18.71	TBL (rest.)	-0.51	0.60	TBL (rest.)	0.72	0.85	DFY (rest.)	0.09	0.13	TBL (rest.)
State dep. regressions	19.17	18.98	TMS (st.dep.)	-1.85	-0.83	TMS (st.dep.)	0.72	0.89	TMS (st.dep.)	0.09	0.13	TMS (st.dep.)
Shrinkage approach	18.87	18.81	DFY (shrink.)	-0.23	0.11	DFY (shrink.)	0.62	0.65	DFY (shrink.)	0.07	0.08	DFY (shrink.)
Combination forecasts	18.82	18.72	Mean (FUND)	0.03	0.58	Mean (FUND)	0.72	0.75	Mean (FUND)	0.10	0.12	Mean (FUND)
Diffusion indices	19.12	19.00	PC (TECH)	-1.55	-0.92	PC (TECH)	0.68	0.73	PC (TECH)	0.09	0.10	PC (TECH)
Kitchen sink forecasts	21.89	20.56	KSF (TECH)	-16.26	-9.22	KSF (TECH)	0.93	1.02	KSF (ALL)	0.13	0.14	KSF (ALL)
LASSO regressions	19.63	19.20	LASSO (TECH)	-4.24	-1.97	LASSO (TECH)	0.78	0.86	LASSO (FUND)	0.11	0.14	LASSO (FUND)
Ridge regressions	19.22	19.00	Ridge (ALL)	-2.08	-0.94	Ridge (ALL)	0.84	0.96	Ridge (FUND)	0.13	0.17	Ridge (FUND)
Elastic Net	19.59	19.13	EN (TECH)	-4.06	-1.61	EN (TECH)	0.74	0.80	EN (ALL)	0.10	0.12	EN (ALL)
Sum-of-the-parts models	18.99	18.58	SOP	-0.87	1.30	SOP	0.90	0.98	ESOP (Median)	0.16	0.18	ESOP (Median)
All strategies	19.08	18.58	SOP	-1.37	1.30	SOP	0.71	1.02	KSF (ALL)	0.09	0.18	ESOP (Median)

Table 4

Results of the superior predictive ability tests based on mean squared forecast errors

This table reports the results of the superior predictive ability tests (SPA-tests) for the pre-specified error rate $\alpha = 5\%$ using mean squared forecast errors over the out-of-sample period from January 1966 to December 2018 as a performance measure. The benchmark model is the historical mean. Column (1) indicates the set of forecasting strategies the SPA-tests are applied to. All forecasting strategies are described in section 3.1. The "most significant" strategy in column (2) is the forecasting strategy with the lowest nominal *p*-value. The nominal *p*-values in column (3) 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, the table reports the consistent *p*-value of the SPA-test, as well as its lower and upper bounds, in column (4), and the number of superior forecasting strategies identified by the step-SPA- and step-SPA(3)-tests in column (5).

Forecasting strategies	**Most significant**	**Nominal *p*-value**	**SPA *p*-value [lower; upper]**	**# of strategies identified [step-SPA; step-SPA(3)]**
Univariate pred. regressions Forecast restrictions	TBL (rest.)	0.2387	0.9770 [0.8346; 0.9770]	[0; 0]
Univariate pred. regressions State-dependent pred. regressions	TMS	0.5285	0.9975 [0.9268; 0.9976]	[0; 0]
Univariate pred. regressions Shrinkage approach	DFY (shrink.)	0.4179	0.9824 [0.8579; 0.9824]	[0; 0]
Univariate pred. regressions Combination forecasts	Mean (FUND)	0.2769	0.9159 [0.6603; 0.9159]	[0; 0]
Univariate pred. regressions Diffusion indices	TMS	0.5285	0.9935 [0.9190; 0.9935]	[0; 0]
Univariate pred. regressions Kitchen sink forecasts	TMS	0.5285	0.9935 [0.9186; 0.9967]	[0; 0]
Univariate pred. regressions LASSO	TMS	0.5285	0.9941 [0.9189; 0.9941]	[0; 0]
Univariate pred. regressions Ridge	TMS	0.5285	0.9938 [0.9229; 0.9938]	[0; 0]
Univariate pred. regressions Elastic Net	TMS	0.5285	0.9941 [0.9187; 0.9941]	[0; 0]
Univariate pred. regressions Sum-of-the-parts models	SOP	0.0228	0.2201 [0.0843; 0.2201]	[0; 0]
All strategies	SOP	0.0228	0.3666 [0.1855; 0.3797]	[0; 0]

Table 5

Results of the superior predictive ability tests based on economic forecast profitability

This table reports the results of the superior predictive ability tests (SPA-tests) for the pre-specified error rate $\alpha = 5\%$ using the mean absolute returns (panel A) and mean risk-adjusted excess returns (panel B) over the out-of-sample period from January 1966 to December 2018 as performance measures. The benchmark model is the historical mean. We assume an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$. We impose portfolio constraints that prevent investors from short-selling and levering more than 50%. Column (1) indicates the set of forecasting strategies the SPA-tests are applied to. All forecasting strategies are described in section 3.1. The "most significant" strategy in column (2) is the forecasting strategy with the lowest nominal *p*-value. The nominal *p*-values in column (3) 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, the table reports the consistent *p*-value of the SPA-test, as well as its lower and upper bounds, in column (4), and the number of superior forecasting strategies identified by the step-SPA- and step-SPA(3)-tests in column (5).

Forecasting strategies	**Most significant**	**Nominal *p*-value**	**SPA *p*-value [lower; upper]**	**# of strategies identified [step-SPA; step-SPA(3)]**
Panel A: Mean absolute returns				
Univariate pred. regressions Forecast restrictions	DFY (rest.)	0.0196	0.2681 [0.2227; 0.2681]	[0; 0]
Univariate pred. regressions State-dependent pred. regressions	TMS (st.dep.)	0.0333	0.3360 [0.2851; 0.3360]	[0; 0]
Univariate pred. regressions Shrinkage approach	TMS	0.0489	0.3872 [0.3097; 0.3872]	[0; 0]
Univariate pred. regressions Combination forecasts	TMS	0.0489	0.3738 [0.3099; 0.3738]	[0; 0]
Univariate pred. regressions Diffusion indices	TMS	0.0489	0.3796 [0.3106; 0.3796]	[0; 0]
Univariate pred. regressions Kitchen sink forecasts	KSF (ALL)	0.0081	0.0980 [0.0810; 0.0980]	[0; 1]
Univariate pred. regressions LASSO	TMS	0.0489	0.3869 [0.3205; 0.3869]	[0; 0]
Univariate pred. regressions Ridge	Ridge (FUND)	0.0166	0.1610 [0.1317; 0.1610]	[0; 0]
Univariate pred. regressions Elastic Net	TMS	0.0489	0.3852 [0.3219; 0.3852]	[0; 0]
Univariate pred. regressions Sum-of-the-parts models	ESOP (Median)	0.0031	0.0500 [0.0415; 0.0500]	[1; 3]
All strategies	ESOP (Median)	0.0031	0.0934 [0.0776; 0.0934]	[0; 1]
Panel B: Mean risk-adjusted excess returns				
Univariate pred. regressions Forecast restrictions	TBL (rest.)	0.0373	0.4019 [0.3592; 0.4019]	[0; 0]
Univariate pred. regressions State-dependent pred. regressions	TMS (st.dep.)	0.0471	0.4429 [0.4032; 0.4429]	[0; 0]
Univariate pred. regressions Shrinkage approach	TBL	0.0531	0.4253 [0.3772; 0.4253]	[0; 0]
Univariate pred. regressions Combination forecasts	Mean (FUND)	0.0128	0.1231 [0.1090; 0.1231]	[0; 0]
Univariate pred. regressions Diffusion indices	TBL	0.0531	0.4147 [0.3745; 0.4147]	[0; 0]
Univariate pred. regressions Kitchen sink forecasts	KSF (ALL)	0.0275	0.2664 [0.2425; 0.2664]	[0; 0]
Univariate pred. regressions LASSO	LASSO (ALL)	0.0491	0.3922 [0.3511; 0.3922]	[0; 0]
Univariate pred. regressions Ridge	Ridge (FUND)	0.0059	0.0703 [0.0628; 0.0703]	[0; 0]
Univariate pred. regressions Elastic Net	TBL	0.0531	0.4186 [0.3789; 0.4186]	[0; 0]
Univariate pred. regressions Sum-of-the-parts models	ESOP (Median)	0.0007	0.0130 [0.0119; 0.0130]	[3; 3]
All strategies	ESOP (Median)	0.0007	0.0258 [0.0236; 0.0258]	[1; 3]

Table 6

Influence of transaction costs

This table reports the results of the superior predictive ability tests (SPA-tests) for the pre-specified error rate $\alpha = 5\%$ using mean absolute returns (panel A) and mean risk-adjusted excess returns (panel B) over the out-of-sample period from January 1966 to December 2018 as performance measures. The benchmark model is the historical mean. We assume an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$. We impose portfolio constraints that prevent investors from short-selling and levering more than 50% and roundtrip transaction costs of 25 basis points. Column (1) indicates the set of forecasting strategies the SPA-tests are applied to (all forecasting strategies are described in section 3.1). The "most significant" strategy in column (2) is the forecasting strategy with the lowest nominal *p*-value. The nominal *p*-values in column (3) 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, the table reports the consistent *p*-value of the SPA-test, as well as its lower and upper bounds, in column (4), and the number of outperforming strategies identified by the step-SPA- and step-SPA(3)-tests in column (5).

Forecasting strategies	**Most significant**	**Nominal *p*-value**	**SPA *p*-value [lower; upper]**	**# of strategies identified [step-SPA; step-SPA(3)]**
Panel A: Mean absolute returns				
Univariate pred. regressions Forecast restrictions	DFY (rest.)	0.0312	0.3416 [0.2617; 0.3416]	[0; 0]
Univariate pred. regressions State-dependent pred. regressions	TMS (st.dep.)	0.0556	0.4698 [0.3665; 0.4698]	[0; 0]
Univariate pred. regressions Shrinkage approach	TMS	0.0698	0.4797 [0.3584; 0.4797]	[0; 0]
Univariate pred. regressions Combination forecasts	TMS	0.0698	0.4676 [0.3611; 0.4676]	[0; 0]
Univariate pred. regressions Diffusion indices	TMS	0.0698	0.4726 [0.3595; 0.4726]	[0; 0]
Univariate pred. regressions Kitchen sink forecasts	KSF (ALL)	0.0353	0.3235 [0.2534; 0.3235]	[0; 0]
Univariate pred. regressions LASSO	TMS	0.0698	0.4767 [0.3724; 0.4767]	[0; 0]
Univariate pred. regressions Ridge	Ridge (FUND)	0.0720	0.4711 [0.3686; 0.4711]	[0; 0]
Univariate pred. regressions Elastic Net	TMS	0.0698	0.4777 [0.3718; 0.4777]	[0; 0]
Univariate pred. regressions Sum-of-the-parts models	ESOP (Median)	0.0068	0.0830 [0.0646; 0.0830]	[0; 3]
All strategies	ESOP (Median)	0.0068	0.1491 [0.1112; 0.1491]	[0; 0]
Panel B: Mean risk-adjusted excess returns				
Univariate pred. regressions Forecast restrictions	TBL (rest.)	0.0436	0.4377 [0.3565; 0.4377]	[0; 0]
Univariate pred. regressions State-dependent pred. regressions	TBL	0.0607	0.5139 [0.4063; 0.5139]	[0; 0]
Univariate pred. regressions Shrinkage approach	TBL	0.0607	0.4619 [0.3703; 0.4619]	[0; 0]
Univariate pred. regressions Combination forecasts	Mean (FUND)	0.0220	0.2131 [0.1691; 0.2131]	[0; 0]
Univariate pred. regressions Diffusion indices	TBL	0.0607	0.4525 [0.3670; 0.4525]	[0; 0]
Univariate pred. regressions Kitchen sink forecasts	TBL	0.0607	0.4711 [0.3896; 0.4711]	[0; 0]
Univariate pred. regressions LASSO	TBL	0.0607	0.4546 [0.3704; 0.4546]	[0; 0]
Univariate pred. regressions Ridge	Ridge (FUND)	0.0381	0.3152 [0.2532; 0.3152]	[0; 0]
Univariate pred. regressions Elastic Net	TBL	0.0607	0.4551 [0.3680; 0.4551]	[0; 0]
Univariate pred. regressions Sum-of-the-parts models	ESOP (Median)	0.0017	0.0266 [0.0216; 0.0266]	[1; 3]
All strategies	ESOP (Median)	0.0017	0.0509 [0.0404; 0.0509]	[0; 3]

Table 7

Robustness checks: Simulated strategies

This table reports the average results of superior predictive ability tests (SPA-tests) for the pre-specified error rate $\alpha = 5\%$ over 100 trials using the mean risk-adjusted excess returns over the out-of-sample period from January 1966 to December 2018 as the performance measure. The benchmark model is the historical mean. The set of forecasting strategies the SPA-tests are applied to include all forecasting strategies considered in Table 6 (Panel B) and the 100 artificial forecasting strategies simulated according to equation (17). We report the average consistent *p*-value of the SPA-tests as well as the average percentage of superior strategies identified by the step-SPA- and step-SPA(3)-tests over 100 trials.

Scaling parameter	SPA *p*-value	% of superior strategies identified by step-SPA	% of superior strategies identified by step-SPA(3)
H = 1.1	0.1268	0.05%	1.32%
H = 1.5	0.0147	2.89%	11.04%
H = 2.0	0.0001	43.24%	69.15%
H = 2.5	0.0000	94.41%	98.29%
H = 3.0	0.0000	99.47%	99.98%

Figures

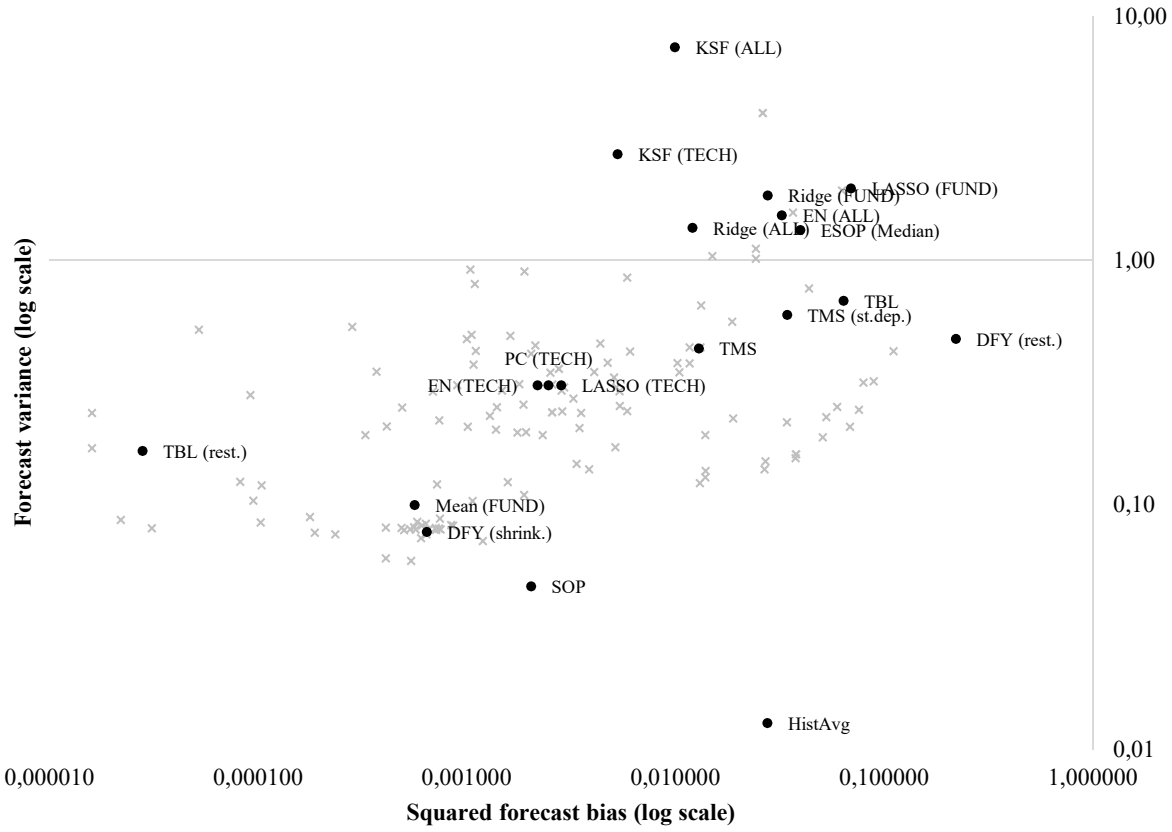

Figure 1. Forecast variances and squared forecast bias. For each forecasting strategy *j* the squared forecast bias $\left(\bar{\hat{r}} - \bar{r}\right)^2$ over the out-of-sample period from January 1966 to December 2018 is plotted against its forecast variance $\sigma^2_{\hat{r}}$ (depicted by the grey crosses). The "best" strategies, as identified in Table 3, as well as the historical average (HistAvg) are highlighted by black dots. To avoid cluttering, the values are depicted on a logarithmic scale.

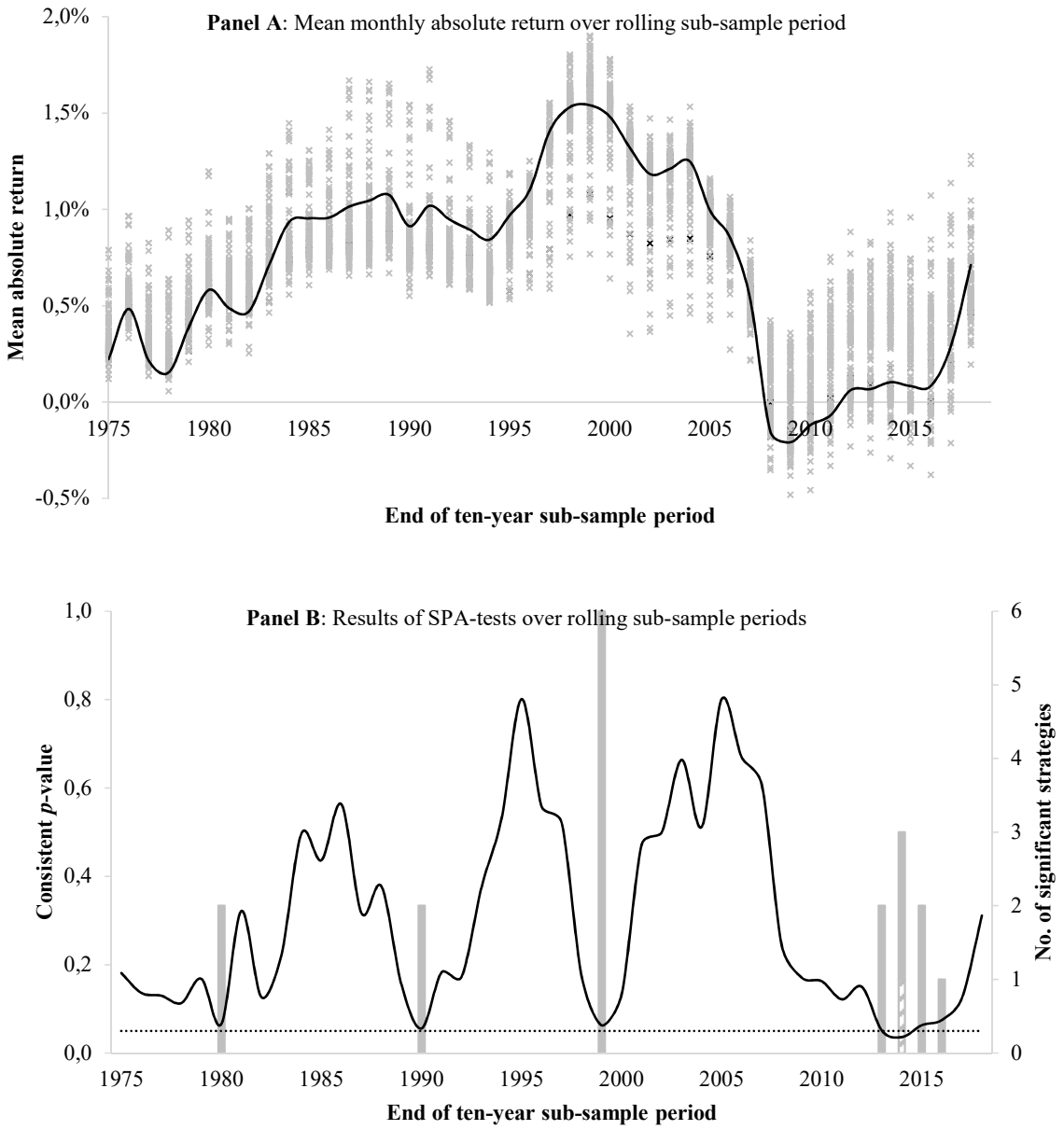

Figure 2. Rolling sub-sample analysis. In panel A, the mean absolute return of each forecasting strategy *j* over the respective ten-year sub-sample period for an investor with mean-variance preferences and relative risk aversion coefficient $\boldsymbol{\gamma = 5}$ (including transaction costs of 25 basis points) is depicted as a grey cross. For comparison, the mean absolute return of the benchmark model over the same sub-sample period is plotted as a black 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 black line identifies the 5% level of significance. The striped bars indicate the number of significant strategies identified by the step-SPA-tests, the grey bars identify the number of significant strategies additionally identified by the step-SPA(3)-tests.

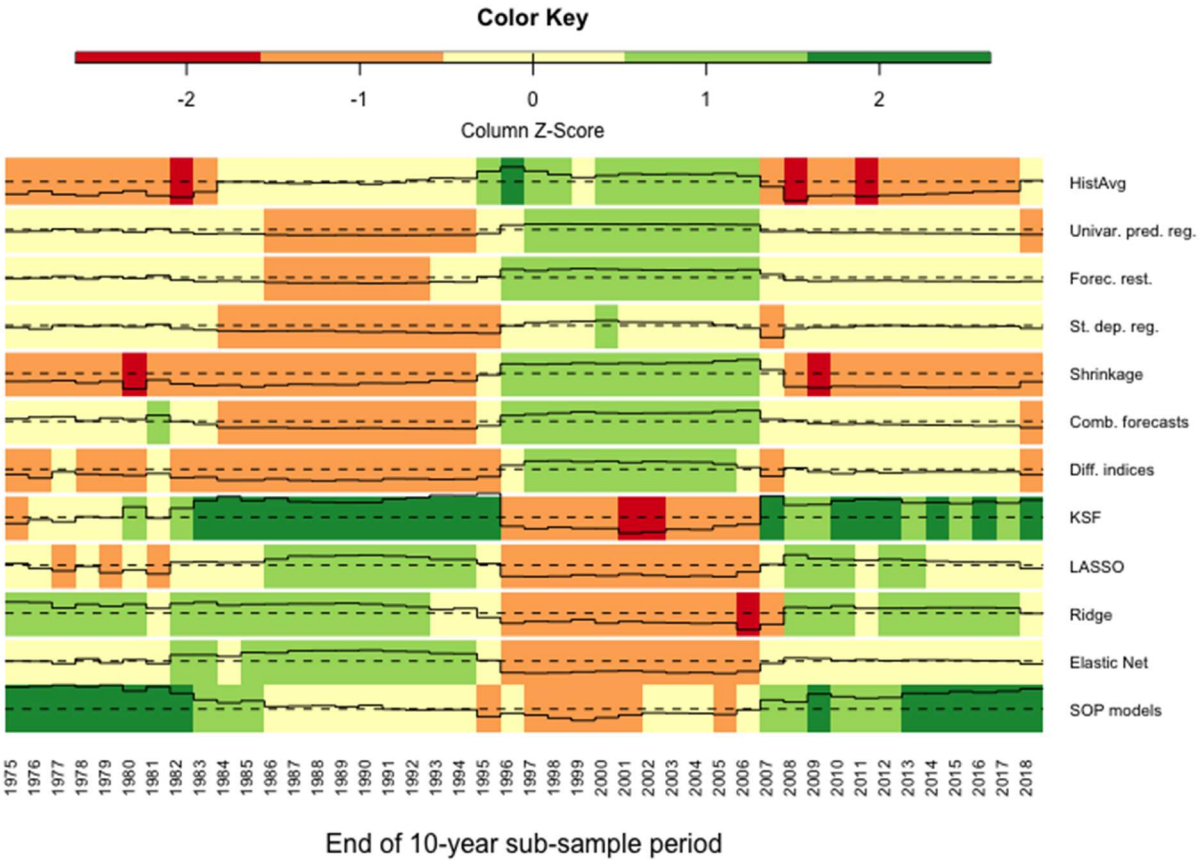

Figure 3. Heat map. This figure visualizes the relative performance of each group of forecasting techniques over the respective ten-year sub-sample period using the average monthly absolute return over the sub-sample period as the performance measure. "HistAvg" denotes the benchmark model. We assume an investor with mean-variance preferences and relative risk aversion coefficient $\boldsymbol{\gamma = 5}$. We impose portfolio constraints that prevent investors from short-selling and levering more than 50% and roundtrip transaction costs of 25 basis points. The shades indicate the *z*-score within the sub-sample period (see color key). The solid black line traces the development of the *z*-score over the rolling sub-sample periods, while the dashed black line indicates a *z*-score of zero.

Appendix

Appendix 1

Results of the pseudo-SPA-tests based on mean squared forecast errors (recursive estimation)

This table reports the results of the superior predictive ability tests (SPA-tests) for the pre-specified error rate $\alpha = 5\%$ using mean squared forecast errors over the out-of-sample period from January 1966 to December 2018 as a performance measure. The benchmark model is the historical mean. Column (1) indicates the set of forecasting strategies the SPA-tests are applied to. All forecasting strategies are described in section 3.1 and based on a recursive, instead of a rolling, estimation scheme. The "most significant" strategy in column (2) is the forecasting strategy with the lowest nominal *p*-value. The nominal *p*-values in column (3) 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, the table reports the consistent *p*-value of the SPA-test, as well as its lower and upper bounds, in column (4), and the number of superior forecasting strategies identified by the step-SPA- and step-SPA(3)-tests in column (5).

Forecasting strategies	**Most significant**	**Nominal *p*-value**	**SPA *p*-value [lower; upper]**	**# of strategies identified [step-SPA; step-SPA(3)]**
Univariate pred. regressions Forecast restrictions	MA2-12 (rest.)	0.0813	0.7259 [0.5725; 0.7259]	[0; 0]
Univariate pred. regressions State-dependent pred. regressions	MA2-12	0.1875	0.9490 [0.7998; 0.9490]	[0; 0]
Univariate pred. regressions Shrinkage approach	MA2-12 (shrink.)	0.0695	0.6368 [0.5421; 0.6368]	[0; 0]
Univariate pred. regressions Combination forecasts	Trim. Mean (FUND)	0.0049	0.0544 [0.0365; 0.0544]	[0; 2]
Univariate pred. regressions Diffusion indices	PC (FUND)	0.0323	0.3205 [0.2108; 0.3205]	[0; 0]
Univariate pred. regressions Kitchen sink forecasts	MA2-12	0.1875	0.9335 [0.7772; 0.9335]	[0; 0]
Univariate pred. regressions LASSO	MA2-12	0.1875	0.9286 [0.7784; 0.9286]	[0; 0]
Univariate pred. regressions Ridge	MA2-12	0.1875	0.9273 [0.7890; 0.9273]	[0; 0]
Univariate pred. regressions Elastic Net	MA2-12	0.1875	0.9278 [0.7803; 0.9278]	[0; 0]
Univariate pred. regressions Sum-of-the-parts models	ESOP (Median)	0.0045	0.0762 [0.0494; 0.0762]	[0; 3]
All strategies	ESOP (Median)	0.0045	0.0990 [0.0735; 0.0990]	[0; 2]

Appendix 2

Results of the pseudo-SPA-tests based on economic forecast profitability (recursive estimation)

This table reports the results of the superior predictive ability tests (SPA-tests) for the pre-specified error rate $\alpha = 5\%$ using the mean absolute returns (panel A) and mean risk-adjusted excess returns (panel B) over the out-of-sample period from January 1966 to December 2018 as performance measures. The benchmark model is the historical mean. We assume an investor with mean-variance preferences and relative risk aversion coefficient $\gamma = 5$. We impose portfolio constraints that prevent investors from short-selling and levering more than 50%. Column (1) indicates the set of forecasting strategies the SPA-tests are applied to. All forecasting strategies are described in section 3.1 and based on a recursive, instead of a rolling, estimation scheme. The "most significant" strategy in column (2) is the forecasting strategy with the lowest nominal *p*-value. The nominal *p*-values in column (3) 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, the table reports the consistent *p*-value of the SPA-test, as well as its lower and upper bounds, in column (4), and the number of superior forecasting strategies identified by the step-SPA- and step-SPA(3)-tests in column (5).

Forecasting strategies	**Most significant**	**Nominal *p*-value**	**SPA *p*-value [lower; upper]**	**# of strategies identified [step-SPA; step-SPA(3)]**
Panel A: Mean absolute returns				
Univariate pred. regressions Forecast restrictions	TMS	0.0032	0.0565 [0.0500; 0.0565]	[0; 2]
Univariate pred. regressions State-dependent pred. regressions	TMS	0.0032	0.0701 [0.0609; 0.0701]	[0; 2]
Univariate pred. regressions Shrinkage approach	TMS	0.0032	0.0700 [0.0583; 0.0713]	[0; 1]
Univariate pred. regressions Combination forecasts	TMS	0.0032	0.0556 [0.0502; 0.0556]	[0; 1]
Univariate pred. regressions Diffusion indices	TMS	0.0032	0.0543 [0.0484; 0.0543]	[0; 1]
Univariate pred. regressions Kitchen sink forecasts	TMS	0.0032	0.0558 [0.0496; 0.0558]	[0; 2]
Univariate pred. regressions LASSO	TMS	0.0032	0.0553 [0.0496; 0.0553]	[0; 1]
Univariate pred. regressions Ridge	TMS	0.0032	0.0540 [0.0480; 0.0540]	[0; 1]
Univariate pred. regressions Elastic Net	TMS	0.0032	0.0547 [0.0486; 0.0547]	[0; 1]
Univariate pred. regressions Sum-of-the-parts models	TMS	0.0032	0.0558 [0.0499; 0.0558]	[0; 2]
All strategies	TMS	0.0032	0.1010 [0.0880; 0.1020]	[0; 0]
Panel B: Mean risk-adjusted excess returns				
Univariate pred. regressions Forecast restrictions	MA2-12	0.0105	0.1238 [0.1060; 0.1238]	[0; 0]
Univariate pred. regressions State-dependent pred. regressions	MA2-12	0.0105	0.1487 [0.1277; 0.1487]	[0; 0]
Univariate pred. regressions Shrinkage approach	MA2-12	0.0105	0.1592 [0.1289; 0.1611]	[0; 0]
Univariate pred. regressions Combination forecasts	Mean (ALL)	0.0032	0.0542 [0.0472; 0.0542]	[0; 5]
Univariate pred. regressions Diffusion indices	MA2-12	0.0105	0.1209 [0.1020; 0.1209]	[0; 0]
Univariate pred. regressions Kitchen sink forecasts	MA2-12	0.0105	0.1249 [0.1063; 0.1249]	[0; 0]
Univariate pred. regressions LASSO	MA2-12	0.0105	0.1221 [0.1030; 0.1221]	[0; 0]
Univariate pred. regressions Ridge	MA2-12	0.0105	0.1230 [0.1043; 0.1230]	[0; 0]
Univariate pred. regressions Elastic Net	MA2-12	0.0105	0.1220 [0.1030; 0.1220]	[0; 0]
Univariate pred. regressions Sum-of-the-parts models	ESOP (Median)	0.0003	0.0094 [0.0084; 0.0094]	[1; 3]
All strategies	ESOP (Median)	0.0003	0.0205 [0.0185; 0.0205]	[1; 2]