EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

University of Tübingen
Working Papers in
Business and Economics

No. 130

Diverging Roads: Theory-based vs. machine learning-implied stock risk premia

by

Joachim Grammig, Constantin Hanenberg,
Christian Schlag, Jantje Sönksen

Faculty of Economics and Social Sciences
www.wiwi.uni-tuebingen.de

Diverging Roads: Theory-based vs. machine learning-implied stock risk premia

Joachim Grammig[*] Constantin Hanenberg[†]
Christian Schlag[‡] Jantje Sönksen[§¶]

February 12, 2020

Abstract

We assess financial theory-based and machine learning-implied measurements of stock risk premia by comparing the quality of their return forecasts. In the low signal-to-noise environment of a one month horizon, we find that it is preferable to rely on a theory-based approach instead of engaging in the computer-intensive hyper-parameter tuning of statistical models. The theory-based approach also delivers a solid performance at the one year horizon, at which only one machine learning methodology (random forest) performs substantially better. We also consider ways to combine the opposing modeling philosophies, and identify the use of random forests to account for the approximation residuals of the theory-based approach as a promising hybrid strategy. It combines the advantages of the two diverging paths in the finance world.

Key words: stock risk premia, return forecasts, machine learning, theory-based return prediction

JEL: C53, C58, G12, G17

[*] University of Tübingen, Department of Statistics and Econometrics and Center for Financial Studies (CFS), Frankfurt. joachim.grammig@uni-tuebingen.de

[†] University of Tübingen, Department of Statistics and Econometrics. constantin.hanenberg@uni-tuebingen.de

[‡] University of Frankfurt, CFS, and Leibniz Institute for Financial Research SAFE, Frankfurt. schlag@finance.uni-tuebingen.de

[§] University of Tübingen, Department of Statistics and Econometrics. jantje.soenksen@uni-tuebingen.de

[¶] We thank Michael Lechner, Yarema Okhrin, Eric Renault as well as participants of an invited session on Advances of Financial Econometrics at the 13th International Conference on Computational and Financial Econometrics for helpful comments. We are indebted to Amit Goyal and Jeremiah Green for their provision of public goods (macro predictor data and SAS code) and to Jonathan Bosoni and Volker Rogg who provided outstanding research assistance. We acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1134-1 FUGG.

1 Introduction

When it comes to measuring stock risk premia, two roads diverge in the finance world. Or so it may seem to a student of the recent literature on empirical asset pricing. From Martin and Wagner (2019) that reader could learn how to quantify the conditional expected return of a stock by exploiting the forward-looking information embedded in option prices. The contribution by Gu et al. (2019b) pursues the same end, but by completely different means. Reflecting the surge of data science applications in economics and finance, and benefiting from advances in computer technology, they advocate the use of machine learning techniques for the measurement of stock risk premia. While Martin and Wagner (2019) derive their results from asset pricing paradigms, Gu et al. (2019b) do not provide deeper references to financial economic theory.

These stunningly different ways to address the same issue motivate us to provide a level playing field for a comprehensive performance comparison of the theory-based and machine learning approaches towards measuring stock risk premia. Because the conditional expected value is the best predictor in terms of the mean squared error (MSE), it is natural to compare the opposing modeling philosophies by gauging the quality of their stock return forecasts at short and longer horizons. Such an analysis tests whether the use of forward-looking information embedded in current market prices is superior to sophisticated statistical analyses of historical data, or whether it is the other way around. Precisely because of their different vantage points and use of complementary data, we find it intriguing to combine the two methodologies. In particular, we employ machine learning techniques to alleviate the approximation errors of the theory-based approach towards measuring stock risk premia.

The present study thereby aims to connect two strands of literature, of which the aforementioned papers are conspicuous examples. The first draws on the basic asset pricing equation and uses option data to quantify risk premia of financial assets. It originates in Martin's (2017) derivation of the lower bound for the conditional expected return of a market index. Kadan and Tang (2019) take up this idea and argue that it can also be applied to quantify risk premia for a certain type of stocks. Assuming that the lower bound of the market index is binding, Martin and Wagner (2019) extend Martin's (2017) approach to the quantification of stock risk premia. The distinctive feature of the first strand approaches is that the econometric estimation of unknown model parameters – traditionally the focus of empirical asset pricing – can be dispensed with.

The second strand consists of applications that address issues in empirical asset pricing and return predictions with the help of machine learning. For example, Light et al. (2017) predict returns of NYSE-, AMEX-, and NASDAQ-traded stocks by using the partial least squares method to aggregate information from 26 observable firm characteristics. Using an adaptive group LASSO, Freyberger et al. (2019) perform a rolling estimation to assess which characteristics contain incremental information on the cross-section of expected returns and how this set of characteristics changes over time. Gu et al. (2019a) note that the prediction problems considered in Gu et al. (2019b) do not constitute asset pricing models and propose an autoencoder model that allows to address risk-return trade offs directly. Kelly et al. (2019) apply instrumented principal component analysis to construct a five-factor model that spans the cross-section of average returns and allows testing for anomalies. Focusing on risk prices instead of premia, Kozak et al. (2019) use penalized regressions for the purpose of shrinking coefficients on risk factors in the pricing kernel. Acknowledging the importance of interactions among firm characteristics, Bryzgalova et al. (2019) generalize the approach by Kozak et al. (2019) and use decision trees for the purpose of constructing a set of base assets that span the efficient frontier. In a Bayesian study, Martin and Nagel (2019) apply shrinkage and selection techniques to a high-dimensional vector of firm characteristics for the purpose of return prediction. They find that standard tests of market efficiency are not applicable in such a setting and stress the importance of out-of-sample tests. Addressing the fact that there are hundreds of competing factors described in recent asset pricing literature, Feng et al. (2019) combine two-pass regressions with regularization methods to assess the marginal contribution of an individual factor in pricing the cross-section of expected returns. Bianchi et al. (2019) study different regularization techniques in the context of a large set of firm characteristics and allow for a time-varying degree of sparsity. They identify a strong relationship between their sparsity measure and the VIX. Two recent papers share our intent to combine theory-based and machine-learning approaches. Chen et al. (2019) use Gu et al.'s (2019b) study as a benchmark and find that the inclusion of the no-arbitrage constraints improves the prediction results. Gu et al. (2019a) note that the prediction problems considered in Gu et al. (2019b) do not constitute asset pricing models and propose an autoencoder model that allows to address risk-return trade-offs directly. Finally, Avramov et al. (2020) take a practitioner's perspective and assess advantages and limitations of the approaches by Kozak et al. (2019), Kelly et al. (2019), Gu et al. (2019a), Gu et al.'s (2019b), and Chen et al. (2019) in a broad reality check.

The main results of our study can be summarized as follows. At the one month forecast horizon, a theory-based approach outperforms the machine learning methods that we consider: elastic net, neural network, boosted trees, and random forest. Of the two theory-based procedures, Martin and Wagner's (2019) take, which is more costly in terms of data input, is preferable to Kadan and Tang's (2019). This conclusion also holds true at the one year horizon, where the expedient theory-based approach surpasses two of the four machine learning methods. Boosted trees offer a comparable performance, but the random forest delivers a notably better out-of-sample forecast. An approach that combines the superior theory-based alternative with this effective machine learning technique is identified as a promising hybrid strategy. Although it has to rely on fewer data, due to the late availability of information required for the theory-based approaches, it performs at least as good as the best (pure) machine learning method. This hybrid approach might counter the critique of "measurement without theory" regarding the use of machine learning techniques by relying on financial economic paradigms and using statistical assistance only for the components that are left unexplained by theory.

The remainder of the paper is structured as follows. Section 2 contrasts the theory-based and the data science methodologies towards measuring stock risk premia, and outlines ideas to combine them. Section 3 describes the preparation of the database and the implementations of the respective risk premia formulas. Section 4 reports the results of comparisons of the forecast quality of the theory-based and machine learning approaches at short and longer horizons, and provides an assessment of the potential of hybrid approaches. Section 5 concludes. Sections A.1–A.4 of the Appendix can be consulted for more details on methodology, data, and implementation.

2 Methodological considerations

2.1 Two diverging roads

This section outlines the basic concepts and key equations associated with the theory-based and the machine learning approaches that we focus on in our study. We explain how, from a common starting point, the methodologies to measure stock risk premia diverge. For the sake of a concise exposition in the main text, we outline details of the respective approaches in the Appendix in Sections A.1–A.3 (theory-based) and in Appendix A.4 (machine learning). The implementation aspects that are specific to the data that we use for our study are discussed in Section 2.4.

The theory-based approach (explicitly) and the machine learning approach (implicitly) take as a point of reference the basic asset pricing equation applied to a gross return of asset i from period t to T ($R^i_{t,T}$) in excess of the respective gross risk-free rate ($R^f_{t,T}$),

$$\mathbb{E}_t(R^{ei}_{t,T}) = \mathbb{E}_t(R^i_{t,T}) - R^f_{t,T} = -R^f_{t,T} \cdot \mathrm{cov}_t(m_{t,T}, R^i_{t,T}), \tag{2.1}$$

where the t subindex indicates that expected values are computed conditional on time t information. In preference-based asset pricing, the stochastic discount factor (SDF) $m_{t,T}$ represents the marginal rate of substitution between consumption in t and T. The sign and size of the risk premia, reflected in the conditional expected excess return on asset i, is determined by the conditional covariance on the right-hand side of Equation (2.1). In the absence of arbitrage, a positive SDF exists, such that $R^f_{t,T} = \mathbb{E}_t(m_{t,T})^{-1} > 0$.

Let us first take a look at the theory-based route. Using (2.1) as a starting point, Appendix A.1 shows how Martin and Wagner (2019) derive the following expression for the right-hand side of (2.1),

$$\mathbb{E}_t(R^{ei}_{t,T}) = R^f_{t,T} \cdot \left\{ \mathrm{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2} \cdot \left[\mathrm{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \mathrm{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right]\right\} + a^i_{t,T}, \tag{2.2}$$

where R^m denotes the return of a market index proxy and w^j_t is the time-varying value weight of index constituent j. var^*_t denotes a conditional variance that is computed under the risk-neutral measure. $a^i_{t,T}$ is a time-varying, asset-specific component, which is, as shown in Appendix A.1, a function of conditional moments either calculated under the risk-neutral or the physical measure.

In a similar vein, Kadan and Tang (2019) advocate the following, even more succinct theory-consistent formula for stock risk premia:

$$\mathbb{E}_t(R^{ei}_{t,T}) = \frac{1}{R^f_{t,T}} \cdot \mathrm{var}^*_t(R^i_{t,T}) - \xi^i_{t,T}, \tag{2.3}$$

where $\xi^i_{t,T} = \mathrm{cov}_t(m_{t,T} \cdot R^i_{t,T}, R^i_{t,T})$. In Appendix A.1 we show how Kadan and Tang (2019) draw on Martin's (2017) derivation of a lower bound for the market equity premium. They argue that depending on the level of risk aversion that one is willing to assume, $\xi^i_{t,T} < 0$ holds for a large fraction of stocks, such that $1/R^f_{t,T} \cdot \mathrm{var}^*_t(R^i_{t,T})$ represents a lower bound for the risk premium of these stocks.

As shown by Martin (2017), the risk-neutral variances in (2.2) and (2.3) can be obtained as follows (suppressing the asset index i for notational brevity):

$$\mathrm{var}^*_t\left(\frac{R_{t,T}}{R^f_{t,T}}\right) = \frac{\int_0^{F_{t,T}} \mathrm{put}_{t,T}(K)dK + \int_{F_{t,T}}^{\infty} \mathrm{call}_{t,T}(K)dK}{0.5 \cdot S_t^2 \cdot R^f_{t,T}}, \tag{2.4}$$

where $\mathrm{call}_{t,T}(K)$ and $\mathrm{put}_{t,T}(K)$ denote the time t prices of European call and put options with strike price K and time to maturity T. S_t is the spot price and $F_{t,T}$ the forward price of the underlying asset. The components of the right-hand sides of (2.2) and (2.3) except the "residuals" $a^i_{t,T}$ and $\xi^i_{t,T}$ can thus be approximated using data on risk-free rate proxies and current option prices for a sufficient number of strike prices. For (2.3), these data are only required for asset i. Equation (2.2) is more demanding in that the option data, along with the time-varying index weights, must also be provided for the constituent stocks of the market index, as well as for the index itself. Martin and Wagner (2019) argue that the consequences of setting $a^i_{t,T} = 0$ should be benign, such that stock risk premia can be quantified without the need to estimate any unknown parameters using:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx R^f_{t,T}\left\{\text{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2}\cdot\left[\text{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \text{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right]\right\}. \tag{2.5}$$

Similarly, assuming that the NCC holds and that the lower bound in (2.3) is binding, a parsimonious theory-consistent approximative formula for the risk premium on stock i is given by:

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx \frac{1}{R^f_{t,T}}\cdot\text{var}^*_t(R^i_{t,T}). \tag{2.6}$$

Now let us sketch the alternative route that the data science approach takes. They may not spell it out explicitly, but Gu et al. (2019b) also use Equation (2.1) as a point of reference. However, the road taken from there is very different. While the theory-based approach gives weight to the risk premium aspect of the basic asset pricing equation, which naturally entails the change from the physical to the risk-neutral measure, the data science perspective emphasizes the forecast implications of (2.1). Recalling that the conditional expected value is the best predictor in terms of MSE, Equation (2.1) states that the MSE-optimal prediction of $R^{ei}_{t,T}$ is given by $-R^f_{t,T}\cdot\text{cov}_t(m_{t,T}, R^i_{t,T})$. Because the exact functional form of the conditional covariance is not known (least not under the physical measure, we have seen that the theory-based approach can characterize it under the risk-neutral measure), it may be conceived as a function that depends on state variables $z^i_t \in \mathcal{F}_t$, such that:

$$\mathbb{E}_t(R^{ei}_{t,T}) = g^0_T(z^i_t), \tag{2.7}$$

where the subindex T indicates that the functional form is assumed to depend on the horizon of interest. The data science approach then proceeds to approximate

$g_T^0(z_t^i)$ by $g_T(z_t^i, \theta_T)$, a parametric function implied by some statistical model with parameter vector θ_T to be estimated. The estimation of θ_T using machine learning procedures (henceforth MLPs) instead of standard econometric methods is advocated because of the following reasons.

First, the number of candidates for state variables z_t^i is large. A myriad of correlated stock-level and macro-economic return predictors (named "features" in machine learning terms) appear in the empirical finance literature, and dimension reduction and feature selection is the very domain of MLPs. Second, the suite of statistical models employed for MLPs trade analytical tractability and rigorous statistical inference for flexible functional form and predictive performance (artificial neural networks are a prominent example). The prediction aspect implied by the basic asset pricing equation naturally provides the "learning" objective, the minimization of the forecast MSE. However, the combination of these two issues – large number of features and desire for flexibility – entails the risk of over-fitting. MLPs deal with this caveat by dividing the data into three parts, a training sample, a validation sample, and a forecast sample. The training sample is used to estimate θ_T by pursuing the learning objective to minimize the forecast MSE at the horizon of interest.

The hallmark of MLPs is to introduce regularization in this process, that is, measures to mitigate the risk of over-fitting. Regularization is controlled via the tuning of so-called hyper-parameters. Such a hyper-parameter can be a penalty parameter that is applied to the learning objective, early stopping rules applied to its optimization, or, more generally, coefficients that determine the complexity of the statistical model, for example, the number of layers in a neural network. Using a given combination of hyper-parameters, the parameter vector θ_T is estimated on the training sample, and the model performance is evaluated, in terms of forecast MSE, on the validation sample. A search across hyper-parameter combinations will ultimately point to a specification that delivers the best performance. Using this hyper-parameter combination, θ_T is re-estimated on the merged training/validation sample. The result is the final estimated model, $g_T(z_t^i, \hat{\theta}_T)$, that is used for out-of-sample evaluation on the forecast sample and the machine learning implied approximative risk premium,

$$\mathbb{E}_t(R_{t,T}^{ei}) \approx g_T(z_t^i, \hat{\theta}_T). \tag{2.8}$$

Another hallmark of MLPs is to employ a variety of statistical models that offer flexible approximations of $g_T^0(z_t^i)$. For the purpose of quantifying stock risk premia, Gu et al. (2019b) advocate the use of artificial neural networks and regressions trees, as well as elastic net regression. Because the number of potential combinations of hyper-parameter values in these models is so large that probing all of them for tuning is infeasible, efficient search algorithms are required to estimate θ_T. The statistical models and associated hyper-parameter value combinations that we consider for the present study are described in detail in Section 3.3.

2.2 Pros and cons

We have seen that the theory-based and data science approaches towards quantifying stock market premia have a common starting point, the basic asset pricing equation (2.1), from which the two modeling philosophies depart in opposite directions. In terms of empirical implementations, the two opposing approaches have, concerning the following aspects, their pros and cons.

Parameter estimation and approximation errors

Using the theory-based formulas (2.5) or (2.6) one can, a result of working under the risk-neutral measure, dispense with the estimation of unknown parameters altogether. However, this parsimony of the theory-based approach comes at the cost of approximation errors, the practical consequences of which are not quite clear. By contrast, the machine learning approach deals with a huge number of unknown parameters, which have to be estimated while balancing the risk of overfitting. On the other hand, the flexibility of machine learning models helps mitigate approximation errors.

Time-varying distributions and parameters

A conspicuous feature of the theory-based approach is that it can naturally deal with changing conditional distributions and even non-stationarity of the data generating process. The machine learning approach, like any statistical/econometric method, struggles much more with ensuing problems like the incidental parameter problem that occurs if the parameters in θ_T were time-varying. Gu et al. (2019b) account for this caveat within their model validation scheme. Starting from an initial sample split, the model is re-trained on updated splits. The updated training sample receives an additional year of data previously included in the validation sample, and the validation and training samples are also shifted one year forward. The optimal hyperparameter combination, and thus the statistical models' complexity, can change with every new split. Instead of (2.8) it is notationally more precise, albeit more cluttered, to write

$$\mathbb{E}_t(R^{ei}_{t,T}) \approx g_{s,T}(z^i_t, \hat{\theta}_{s,T}), \tag{2.9}$$

indicating the dependence of the functional form and estimates on the sample split s as well as on the horizon T.

Demand on data quality and computational resources

The demands on data quality and quantity in both the theory-based and the machine learning approach are considerable, and they are, not surprisingly, different and complementary. For the machine learning approach, one has to supply historical data on stock-level predictors for every asset of interest. While macro predictors are publicly available, the stock-level predictors are not and access to Compustat and CRSP is mandatory. The theory-consistent approaches have no need for these data, but they require high quality option data. In particular, they require option prices with times-to-maturity that match the horizons of interest, and a sufficiently large number of strike prices K to provide a good approximation to the integrals on the right-hand side of (2.4). Equation (2.5) reveals that these data are not only required for the stocks of interest, but also for every current member of the market index, as well as the index itself. Index membership changes, and these changes must be tracked, which is, as outlined below, not straightforward using the available data. The advantage of the theory-based approach by Martin and Wagner (2019) is that, provided the index membership can be tracked, the computational resources that are needed to provide quantifications of stock risk premia are moderate. The same holds true for Kadan and Tang's (2019) alternative approach. By contrast, the computational resources required for the machine learning approaches are considerable and access to a high performance computer cluster is mandatory. Training and hyper-parameter tuning are required for each statistical model considered, for each horizon of interest, and for every new sample split.

2.3 Roads not traveled: Hybrid approaches

"And sorry I could not travel both," writes Frost (1916), contemplating which road to take. In the present case, it is precisely because of the diversity of their respective pros and cons that it is intriguing to actually travel both roads, by combining the theory-based and machine learning philosophies. We consider two such hybrid strategies. The first and obvious is motivated by the observation that while Gu et al.'s (2019b) machine learning approach includes a plethora of stock-level and macro features (and interactions), it does not use the information provided by the theory-based risk premium measure, or any other conditional time t moment computed under the risk-neutral measure. By augmenting the set of features accordingly, we can assess whether the theory-based measurements enhance the explanatory power of the data science approach or even render the other predictor variables obsolete.

Our second hybrid approach starts from (2.2) and the approximative formula (2.5), and then employs MLPs to account for the approximation residuals $a^i_{t,T}$. This strategy may be described as a theory-based, machine learning assisted approach towards measuring stock risk premia.[1] For that purpose, let us denote by $\widetilde{\mathbb{E}}_t(R^{ei}_{t,T})$ the right-hand side of (2.5). Then $\widetilde{R}^{ei}_{t,T} = R^{ei}_{t,T} - \widetilde{\mathbb{E}}_t(R^{ei}_{t,T})$ gives the component of the excess return that is left unexplained by the theory-based approximation of the stock risk premium. Provided that the aforementioned data requirements are met, $\widetilde{R}^{ei}_{t,T}$ can be computed for every i, t, and T. Emphasizing the prediction aspect of the basic asset pricing equation, consider the following decomposition,

$$\widetilde{R}^{ei}_{t,T} = a^i_{t,T} + \varepsilon^i_{t,T}, \tag{2.10}$$

where $\varepsilon^i_{t,T} = R^{ei}_{t,T} - \mathbb{E}_t(R^{ei}_{t,T})$ can be conceived of as the irreducible idiosyncratic forecast error (the smallest forecast MSE would be $\mathbb{E}([\varepsilon^i_{t,T}]^2)$). Now consider applying the machine learning procedures described above instead of to $R^{ei}_{t,T}$ and $\mathbb{E}_t(R^{ei}_{t,T})$ to $\widetilde{R}^{ei}_{t,T}$ and the approximation residuals $a^i_{t,T}$. This is a sensible approach because as Appendix A.1 shows, $a^i_{t,T}$ is a function of time t conditional moments. Similarly to (2.7), we may therefore represent $a^i_{t,T}$ as a function of the time t state variables z^i_t, such that $a^i_{t,T} = h^0_T(z^i_t)$, and use a parametric statistical model with parameters ϑ_T

[1] Alternatively, we could also take Kadan and Tang's (2019) approximation (2.6) as a starting point, but (2.2) is arguably more appropriate for a larger number of stocks. However, we will consider the alternative in a later version of the paper.

to approximate $h_T^0(z_t^i) \approx h_T(z_t^i, \vartheta_T)$.

Machine learning-style estimation of the parameters ϑ_T employs the learning objective to minimize the forecast MSE of $\widetilde{R}_{t,T}^{ei} - h_T(z_t^i, \vartheta_T)$ instead of $R_{t,T}^{ei} - g_T(z_t^i, \theta_T)$, applying the hyper-parameter tuning procedures described supra. A hybrid risk premia quantification/excess return forecast is then given by:

$$\mathbb{E}_t(R_{t,T}^{ei}) \approx \widetilde{\mathbb{E}}_t(R_{t,T}^{ei}) + h_T(z_t^i, \hat{\vartheta}_T). \tag{2.11}$$

The residual of this composite forecast can be decomposed as

$$R_{t,T}^{ei} - \underbrace{(\widetilde{\mathbb{E}}_t(R_{t,T}^{ei}) + h_T(z_t^i, \hat{\vartheta}_T))}_{\text{hybrid risk premium/ forecast}} = \underbrace{(a_{t,T}^i - h_T(z_t^i, \vartheta_T))}_{\text{approximation error}} + \underbrace{(h_T(z_t^i, \vartheta_T) - h_T(z_t^i, \hat{\vartheta}_T))}_{\text{estimation error}} + \varepsilon_{t,T}^i. \tag{2.12}$$

Diligent hyper-parameter tuning is mandated to avoid over-fitting, i.e. acknowledge that $\varepsilon_{t,T}^i$ represents the inherently unforecastable part of the excess return $R_{t,T}^i$.

To account for time-varying model parameters and complexity, the dynamic hyper-parameter tuning described in Section 2.3 can be applied in the same way as described supra, which yields the hybrid approximative formula for the stock risk premium,

$$\mathbb{E}_t(R_{t,T}^{ei}) \approx \widetilde{\mathbb{E}}_t(R_{t,T}^{ei}) + h_{s,T}(z_t^i, \hat{\vartheta}_{s,T}). \tag{2.13}$$

Neither the theory-based (“Econ”) approach nor the machine learning (“Metrics”) approach would be described as Econometrics, the discipline once founded to connect economic theory and statistics. Yet, the hybrid formula (2.13) may be seen as a novel way of combining Econ and Metrics in the age of data science.

2.4 A level playing field for comparison and evaluation

The approximative risk premium formulas – (2.5) for the theory-based forecast, (2.6) for the theory-based alternative, (2.9) for the machine learning-based prediction, and (2.13) for the hybrid approach – all have a common root in the basic asset pricing equation (2.1). And they all provide an approximation of the MSE optimal excess return forecast, $\mathbb{E}_t(R^{ei}_{t,T})$. It is therefore quite natural to assess the alternative approaches based on their out-of-sample forecast quality at various horizons.

A central tenet of financial economics, derived from the basic asset pricing equation, states that marginal utility weighted prices follow martingales. This statement implies that return predictability should be a long-horizon phenomenon. Short-run prices should behave like martingales, such that the MSE-optimal return forecast at short horizons should be close to the zero forecast (cf. Cochrane (2005), Section 2.4). Put differently, the signal $(\mathbb{E}_t(R^{ei}_{t,T}))$ to noise $(\varepsilon^i_{t,T})$ ratio should be considerably higher at long horizons than at short horizons. So the question is, which of the two approaches delivers the better approximation to $\mathbb{E}_t(R^{ei}_{t,T})$ at given horizons, and by how much the hybrid approaches can enhance the performance. These are empirical questions that we address in our study. To answer them we have to set up a comprehensive data base.

3 Data and implementations

3.1 Data base: Making of

The selection of stocks for which we compare the alternative risk premium measurements is defined by a firm's membership in the S&P 500 index.[2] One reason to choose

[2] There can be multiple securities associated with an S&P 500 firm, e.g. Apple. An S&P 500 constituent is a specific company-security combination, but we refer to them, as is common in the literature, interchangeably as "stocks" or "firms."

this criterion is that if we want to compute theory-based risk premia according to Equation (2.5), we have to provide information about the index constituents. Using the S&P 500 as the market index proxy, S&P 500 membership provides the obvious selection criterion for our analysis.

The identification of S&P 500 constituents works as follows.[3] Using a procedure proposed by Wharton Research Data Services (WRDS), we first retrieve information about a firm's S&P 500 membership status from Compustat. We thereby obtain for every month between March 1964 to December 2018 a list of S&P 500 constituents. In total, we identify 1,697 firms that have been a member of the S&P 500 at least for one month.[4] For these stocks of interest, we retrieve price and return data from CRSP. The option data, which are required to compute the theory-based measures, come from OptionMetrics. From Compustat and CRSP we obtain feature data used for the machine learning approaches. Linking the information across these WRDS data bases is hampered, because the security identifiers are not unified. A perfect match of securities across data bases is infeasible, although the WRDS linkage tables are a great help. Notwithstanding these challenges, and as shown in Panel A of Figure 1, we are quite successful in recovering the Compustat-identified S&P 500 members also in CRSP. Moreover, Panel B of Figure 1 shows that the true S&P 500 market capitalization is closely tracked by the aggregated market capitalization of the S&P 500 constituents that we identify with our procedure.

[Insert Figure 1 about here]

The matching with OptionMetrics is notoriously less precise. OptionMetrics, which provides daily data from January 1996 onward, uses its own security identifier. Panel A in Figure 1 shows that our procedure can nevertheless recover a large fraction

[3] A more detailed description is provided in Appendix A.2.

[4] More precisely, these are 1,697 distinct GVKEY+IID combinations.

of the S&P 500 constituents in OptionMetrics, too. The approximation formula (2.5) shows that the higher the coverage of index stocks, the better the theory-based approach can be expected to perform, while a poor match adds another source of approximation error. Comparing descriptive statistics, we note that our coverage rate is notably higher than that reported by Martin and Wagner (2019). Averaged over the respective sample periods, we succeed in recovering 483/500 constituents; Martin and Wagner's (2019) ratio is 451/500.

We perform our comparative analysis, the out-of-sample forecast comparison, for the period from January 1996, the starting date of OptionMetrics, until December 2018, the most recent CRSP date available at the time of writing. This is arguably a challenging playground, because this time interval is cluttered with financial crises (Asian, LTCM, Subprime), new economy euphoria, and bursts of (alleged) price bubbles. Both theory-based and machine learning approaches are able to provide stock risk premia measurements for this 22 years interval, during which 1,145 of the initially identified 1,697 S&P 500 constituents still appear and can be recovered in the OptionMetrics data. These stocks represent the cross section of assets that we are focusing on in our analysis. We note that data on these 1,145 stocks are required also before 1996, for purpose of parameter estimation with the help of MLPs.

Adopting the procedure used by Gu et al. (2019b), who draw on Green et al.'s (2017) prior work, 93 stock-level predictor variables (collected in a vector c_t^i) and 74 dummy variables that identify a firm's industry are obtained from Compustat and CRSP.[5]

[5] For that purpose, we adapt the SAS program from Jeremiah Green's website https://sites.google.com/site/jeremiahrgreenacctg/home, accessed January 20, 2020. We are one feature short of the 94 firm characteristics extracted by Gu et al. (2019b), because the industry-adjusted firm size was not implemented in the SAS program. The industry dummies are based on the first two digits of the SIC code. In line with Gu et al. (2019b), we use cross-sectional median-based imputation to deal with missing observations. Note that Gu et al. (2019b) additionally (and differently from us) rescale their features to the interval [-1,1] before using them. Appendix A.4 provides details on these issues.

The stock-level characteristics are augmented by 8 macro predictor variables at the monthly level, obtained from Amit Goyal's website.[6] The stock-level and macro features have a mixed frequency: monthly (20 stock level + 8 macro variables), quarterly (13 stock level variables) or annual (60 stock level variables). Extracting from CRSP the date of the last trading day of each month as a point of reference, the stock-level and macro features are aligned according to Green et al.'s (2017) assumptions about delayed availability to avoid the forward-looking bias.[7] Moreover, we match CRSP returns at horizons of one month (30 calendar days) and one year (365 calendar days). These returns are forward-looking from the vantage point of the end-of-month alignment day.

When proceeding as described, we obtain an unbalanced panel data set at the monthly frequency that contains information about the 1,145 stocks of interest, with a varying number of observations per stock. From October 1974, which is the first month used for training the machine learning algorithms, until the end of 2018, we obtain 362,306 stock/month observations. Table 1 reports descriptive information about this data set, which provides the basis for our analysis.

[Insert Table 1 about here]

To compute excess returns and all of the theory-based measures, we need risk-free rate proxies that match the return horizon. Conveniently, OptionMetrics provides time series of the zero curve, from which risk-free rate proxies at different horizons can be computed at the daily frequency. These risk-free rate proxies are used by

[6] http://www.hec.unil.ch/agoyal, accessed January 20, 2020. These variables are the dividend-price ratio, earnings-price ratio, book-to-market ratio, net equity expansion, stock variance (all non-stock-specific), as well as the Treasury-bill rate, term spread, and the default spread. Their definitions are given by Welch and Goyal (2008). The macro variables are collected in a vector denoted x_t.

[7] Following Green et al.'s (2017), variables at the monthly frequency are delayed at most one month, quarterly variables with at least four month lag, and annual variables with at least six months lag.

Martin and Wagner (2019) and Kadan and Tang (2019). However, like any data supplied by OptionMetrics, the zero curve is not available before January 1996. We therefore employ the Treasury-bill rate as a risk-free rate proxy for earlier periods. The monthly Treasury-bill rate is used to construct risk-free rate proxies at longer horizons. Figure 2 shows that the two risk-free rate proxies behave similarly.

[Insert Figure 2 about here]

3.2 Theory-consistent approaches: Empirical implementations

This section addresses problems and solutions associated with the implementation of the approximative risk premia formulas (2.5) and (2.6).[8] The obstacles are the following: First, options on S&P 500 stocks are American options, while the computation of risk-neutral variances according to (2.4) relies on European options. Second, as a continuum of strike prices is not available, the integrals in (2.4) must be approximated, using a grid of discrete strikes. As pointed out by Martin (2017), a lack of a sufficient number of strikes may severely downward bias the computation of risk-neutral variances and thus that of the theory-based stock-risk premia. In economic terms, the option market must be liquid enough to provide option prices for a large number of strikes and maturities.[9]

Martin and Wagner (2019) advocate the use of the OptionMetrics volatility surface to address these issues such that the computation of risk-neutral variances according to (2.4) can be performed. A detailed exposition of how exactly we use the volatility surface for this study is provided in Appendix A.3. A favorable innovation is

[8] The main text provides a succinct discussion, Appendix A.3 outlines technical details.

[9] Relatedly, we also note that the out-of-the money call options needed for the computation in (2.4) are notoriously illiquid assets, but that problem that can be alleviated by exploiting the put-call parity and using more liquid out-of-the money put options instead.

that the latest OptionMetrics release (including data until December 2018) provides the volatility surface for an increased number of standardized strike prices. Previous releases provided 13 strikes, the recent release has 17. The wider range of strikes is available also for the historical data back to January 1996. As reasoned supra, the theory-based approaches should benefit from that modification; more strikes should facilitate a more accurate approximation of risk-neutral variances. Although European options are traded on the S&P 500 index, and their prices are available in OptionMetrics, we also rely on the volatility surface to compute risk-neutral index variances. We find that it is more convenient to approximate $\mathrm{var}_t^*\left(R_{t,T}^m / R_{t,T}^f\right)$ in (2.5) in that way (see Appendix A.3). Albeit they are not explicit about it, our calculations suggest that Martin and Wagner (2019) pursue the same strategy.

Using the most recent release of the OptionMetrics volatility surface, we compute the theory-based risk premia measures for the selected stocks and the two horizons of interest. These data are matched by security identifier and end-of-month date to the unbalanced panel data described in the previous section; Table 1 contains these additional descriptive statistics. Because the theory-consistent measures can be computed on a daily frequency, we also construct a data set that contains these measures along with the corresponding forward-looking returns at the horizons of interest.

3.3 Machine learning and hybrid approaches: Empirical implementations

This section provides information about the implementation of the machine learning approaches: the statistical models and their hyper-parameters, the selection of features used in $g_T(z_t^i, \hat{\theta}_T)$, and model validation strategies.[10] The statistical models that we focus on are those identified by Gu et al. (2019b) as the most appropriate for the task of predicting excess returns and stock risk premia: random forests, artificial neural networks, and gradient boosted regression trees. We also include elastic net as an instance of penalized regression.[11] The hyper-parameters of the respective models are presented in Table 2.

[Insert Table 2 about here]

The selection of features z_t^i follows Gu et al. (2019b), such that we include 91 stock-level characteristics as well as their interactions with the 8 macro predictors.[12] Formally, z_t^i is comprised of the vector $(1, x_t')' \otimes c_t^i$, which is augmented with the industry dummies. Altogether we have $91 \times 9 + 72 = 891$ features. As in Gu et al. (2019b), there are no interactions across stocks or lagged variables.

There is a considerable number of missing values for stock-level predictors dating further back. To avoid negative consequences from massively imputing missing values, we start the estimation when the problem is somewhat alleviated. Deviating from Gu et al. (2019b), who start in 1957, our first training period is October 1974. Our concrete implementation of the sequential validation procedure mentioned in Section 2.1 is illustrated in Figure 3. Adapting the procedure by Gu et al. (2019b), the length of the training period increases from initially 10 years to 31, the 12 year validation period is shifted forward by one year with every new split, and there are

[10] The succinct exposition in the main text is augmented by Appendix A.4, where we explain technical details.

[11] We assume that the reader has some familiarity with these standard models and their application. These statistical models are covered in Hastie et al. (2017), and Gu et al. (2019b) can be consulted for a succinct but useful overview. Gu et al. (2019b) also consider linear regression, partial least squares, and principal component regression. We do not consider these approaches, because we only focus on the most flexible and thus promising competitors. Elastic net is included mainly because of its relative computational ease and popularity.

[12] Compared to Gu et al. (2019b), we exclude two features from the set of firm characteristics, because they contain an excessive amount of missing values. The two features are “real estate holdings” and “secured debt.” Also, we only find 72 SIC codes instead of 74.

22 out-of-sample years with the final one year predictions made in December 2017 for December 2018.[13] For each of the sample splits, hyper-parameter tuning for each of the statistical models is performed at the one month and one year forecast horizons.

[Insert Figure 3 about here]

While the basic setup remains the same when considering hybrid approaches, the validation procedure changes because of the delayed availability of the OptionMetrics data that are required for the theory-based approaches. Augmenting z_t^i by the two theory-consistent measures in (2.5) and (2.6) becomes possible beginning January 1996, which is also when the implementation of the machine learning-assisted theory-based approach becomes feasible. We therefore consider the alternative "short" validation procedure that is depicted in Figure 4.

[Insert Figure 4 about here]

The alternative validation procedure is applied using the theory-augmented feature set and for the training on the residuals of the theory-based forecasts. Moreover, to provide a benchmark for how these hybrid approaches compare to the base case, we also re-train the models with the original feature set, but using the alternative validation procedure. There, the training and validation periods are notably shorter.[14] The initial 10 year training period is reduced to one year until it increases to 20 years; the validation periods comprise one year instead of 12. The reason for this configuration is that we want to retain a sufficiently large number of out-of-sample years comparable to the base case: At the one month forecast horizon, the alternative

[13] Note that this out-of-sample forecast period only relates to the one year forecast horizon. When considering the one month horizon, the number of splits increases to 23, because we can then also make forecasts during the year 2018.

[14] The calculations for the theory-based approaches remain unchanged, because no estimation is required. However, two years of out-of-sample evaluation for the base case (1996 and 1997) are lost.

has 20 years instead of 22. Comparing the results with the base case is interesting for another reason. It allows to study how important the length of the training period is and to assess the effect of the length of the validation period.

4 Empirical Results

4.1 Comparison by out-of-sample forecast performance

This section compares the theory-based and machine learning approaches to measure stock risk premia by their out-of-sample forecast performance. We have argued supra that this is a sensible criterion for comparison, because the different methodologies deliver quantifications of the conditional expected excess return, the MSE-optimal prediction. We consider forecast horizons of one month (30 calendar days) and one year (365 calendar days), for which both Gu et al. (2019b) and Martin and Wagner (2019) argue that their approaches are most suitable. As is standard in the literature, and following Welch and Goyal (2008), we use a performance measure that relates the MSE of a model's out-of-sample forecast to that of a benchmark. We use the zero forecast for that purpose, which is, as argued in Section 2.4, the natural choice at short horizons.[15] Using the zero forecast as a benchmark, the performance criterion is given by:

$$R^2_{oos} = 1 - \frac{\sum_i \sum_t \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2}{\sum_i \sum_t \left(R^{ei}_{t,T}\right)^2}, \tag{4.1}$$

[15] Of course, there are alternative choices. Martin and Wagner (2019) also consider stock-specific historical mean excess returns. However, because the goal of this study is a juxtaposition of theory-based and machine learning approaches, the benchmark forecast is of lesser importance. The zero forecast has the appeal of providing a theory-consistent, parameter-free benchmark. Moreover, as pointed out by Gu et al. (2019b), the zero forecast is not a scapegoat. They report that using the stock-specific historical mean excess returns instead of the zero forecast considerably improves the relative performance of the competitor forecasts that they consider. Another advantage of the zero forecast is that results can be better compared across studies.

where the horizon index is dropped for notational brevity. The calculation of R^2_{oos} uses only observations included in the forecast samples, data of which are not used for training or validation. There are S=22 forecast sample years for the “long training/validation” scheme in Figure 3 and S=20 years for the “short training/validation scheme” in Figure 4. One of the advantages of the theory-based approaches is that the risk premia approximations according to (2.2) and (2.6) can be computed on a daily frequency. We therefore compute R^2_{oos} and the subsequent statistics for the theory-based approaches in two variants. The base version uses the end-of-month forecasts, the alternative the daily forecasts.

To study the performance over time, we also compute the out-of-sample R^2 for each of the forecast samples $s = 1, 2, \ldots, S$ separately, viz:

$$R^2_{oos,s} = 1 - \frac{\sum_i \sum_t \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2 \cdot \mathbb{1}[t \in \mathcal{S}(s)]}{\sum_i \sum_t \left(R^{ei}_{t,T}\right)^2 \cdot \mathbb{1}[t \in \mathcal{S}(s)]} \qquad s = 1, 2, \ldots, S, \tag{4.2}$$

where $\mathcal{S}(s)$ denotes the set of time indices of forecast sample s such that $\mathbb{1}[t \in \mathcal{S}(s)]$ returns 1 if the observation at period t belongs to the sample year s, and 0 else.

For assessments of the statistical significance of differences in forecast performance, we report p-values associated with Gu et al.’s (2019b) adapted Diebold-Mariano test. The test statistic is based on the MSEs computed for the forecasts samples,

$$MSE_s = \frac{1}{N_s} \sum_i \sum_t \left(R^{ei}_{t,T} - \hat{R}^{ei}_{t,T}\right)^2 \cdot \mathbb{1}[t \in \mathcal{S}(s)] \qquad s = 1, 2, \ldots, S, \tag{4.3}$$

where N_s is the number of forecasts issued in forecast sample s. Denoting the difference of (4.3) implied by two models by $d_s = MSE_s^{(1)} - MSE_s^{(2)}$, one can compute the mean over the S forecast samples, $\overline{d} = \frac{1}{S} \sum_{s=1}^{S} d_s$, as well as its Newey-West (NW) standard error $\hat{\sigma}(\overline{d})$.[16] Gu et al. (2019b) assume that a central limit theorem (CLT)

[16] The NW-correction accounts for serial correlation in d_s.

can be applied, such that the test statistic $DM = \overline{d}/\hat{\sigma}(\overline{d})$ is approximately distributed $\mathcal{N}(0,1)$ under the null hypothesis that the population MSEs implied by the two model forecasts are identical. When using the DM-statistic for our purposes, we keep the theory-based forecast implied by (2.5) as the first of the two forecasts (using the end-of-month variant).

We also report p-values associated with a test of the null hypothesis that a model's forecast has no explanatory power over the zero forecast, formally phrased as $\mathbb{E}(R^2_{oos,s}) < 0$. The construction of the test statistic is motivated in the same vein as the DM-statistic. Take the mean of $R^2_{oos,s}$ across the testing samples, $\overline{R^2_{oos}} = \frac{1}{S}\sum_{s=1}^{S} R^2_{oos,s}$, and compute its NW-standard error $\hat{\sigma}(\overline{R^2_{oos}})$. Then, provided that a CLT can be applied, and assuming that $\mathbb{E}(R^2_{oos,s}) = 0$, the test statistic $\overline{R^2_{oos}}/\hat{\sigma}(R^2_{oos})$ is approximately standard normally distributed, such that a one-sided p-value associated with the null hypothesis that $\mathbb{E}(R^2_{oos,s}) < 0$ can be computed.[17]

[Insert Table 3 about here]

Comparing the results reported in Panel A of Table 3 with Panel B shows that the one month horizon during the years 1996-2018 is a harsh environment for forecasting. None of the machine learning approaches achieves a positive $R^2_{oos,s}$. With an R^2_{oos} of 0.9% (daily variant), the theory-consistent forecast implied by Martin and Wagner's approximation formula (2.5) stands out, also against the alternative theory-based forecast based on (2.6). An R^2_{oos} of about 1% may appear small, but compared to the numbers reported by Gu et al. (2019b), it is notably high.[18] Figure 5 illustrates the advantage of the preferred theory-based approach. The idea behind this graphical representation is to sort stocks into decile portfolios formed according to the excess

[17] Because of the small number of observations S used to compute the means, the power of the two tests is inevitably limited.

[18] The neural networks that Gu et al. (2019b) train, arguably the best performing approach, achieve R^2_{oos} between 0.3% and 0.7%, depending on stock selection and network architecture.

return forecast, and plotting the average predicted excess returns against the average realized excess returns, which should align along the 45-degree line. It is evident that the theory-based approach does a better job.

[Insert Figure 5 about here]

These results suggest that at the one month horizon not much may be gained by investing in computer-intensive data science methods. Relying on the (superior) theory-consistent approach seems to be the prudent choice in this low signal-to noise environment. The favorable conclusions regarding the use of machine learning approaches at the one month horizon reported by Gu et al. (2019b) should therefore be taken with a grain of salt.

One may express reservations about this conclusion. It may be argued that the results are based on a different sample period and a different selection of stocks, for which the forecasting task is more difficult for machine learning (yet arguably not so much for the theory-based approach). One could also point out that we use fewer stocks for training and validation, and that our training begins in a later year, all of which may prevent the machine learning approaches to unfold their full potential. Moreover, our choice of hyper-parameters may be to blame, because it is well known that highly non-linear models, like artificial neural networks, are very sensitive towards the hyper-parameter tuning procedure.[19]

Most of these concerns are alleviated when taking a look at Panel B of Table 3, which presents the results for the one year forecast horizon. Compared to the Panel A results, the R^2_{oos} increase by an order of magnitude. In line with notions of financial theory, the prediction job becomes easier as the signal-to-noise ratio is more favorable. The results reported in Panel B of Table 3 refute the notion that our selection of

[19] Unfortunately, Gu et al. (2019b) do not provide sufficient details on their hyper-parameter combinations for a comprehensive comparison.

stocks constitutes a more difficult environment for machine learning approaches, because their performance also considerably improves. Both neural network and elastic net deliver R^2_{oos} that are comparable to those reported by Gu et al. (2019b).[20] Notwithstanding, the theory-based forecast using Martin/Wagner's approximation in Equation (2.5) delivers a notably higher R^2_{oos} than the neural network. Among the two theory-consistent alternatives, the more sophisticated approach by Martin/Wagner outperforms Kadan/Tang's, which delivers the smallest R^2_{oos} among the competitors. The performance of the regression-tree methods, identified by Gu et al. (2019b) as promising models for the one month horizon, is remarkably good. In our study, these approaches show their potential at the one year horizon. While the performance of gradient-boosted trees is comparable to the preferred theory-based method, the random forest offers a considerable improvement. The prediction deciles plots in Figure 6 corroborate these conclusions.

[Insert Figure 6 about here]

The favorable results for the regression tree methods (and for the other machine learning approaches, too) mitigates the caveat that our hyper-parameter tuning may be completely ill-advised.

A issue that has not been explicitly addressed in previous literature is the variation of the forecast performance over time. This phenomenon is reflected in the high $R^2_{oos,s}$ standard deviations and p-values reported in Table 3. While Gu et al. (2019b) do not discuss the time series variation of $R^2_{oos,s}$, it also translates into the p-values of the DM-statistic that they report.[21]

The time series variation of the $R^2_{oos,s}$ is illustrated in Figures 7 and 8. The three

[20] Depending on the selection of stocks, the R^2_{oos} of their best trained neural network range from 3.4% to 5.2%, for elastic net from 1.8% to 3.9%.

[21] As Table 3 in Gu et al. (2019b) reveals, significant differences of the forecasting performance can only be detected if a very poorly performing model is compared to a well performing alternative.

panels take Martin/Wagner's theory-consistent approach as a point of reference and compare it in Panel A against the best-performing machine learning approach, in Panel B against the least successful competitor, and against all the others in Panel C. The volatility of the $R^2_{oos,s}$ during our forecast sample years 1996-2018 is not surprising, because it is a period that is rife with crises and crashes.[22] Their impacts are conspicuous in the $R^2_{oos,s}$ time series plots in Figures 7 and 8.

[Insert Figures 7 and 8 about here]

Figure 8 shows that on the one year horizon the impact of the build-up and burst of the so-called dot-com bubble is more pronounced than that of the financial crisis in 2008. The yearly forecasts of all models (safe one) issued in the years 2000 and 2001 yield negative R^2_{oos}.[23] Panel A illustrates that the random forest forecast offers an overall improvement over the theory-consistent forecast, albeit the negative R^2_{oos} associated with 2015 forecasts indicates some erratic behavior. The year 2016 may not be considered a particularly conspicuous period. An interesting observation in Panel C of Figure 8 is that the neural network is the only model that weathered the dot-com bubble well, but it loses it at the end of the sample period when it exhibits very erratic behavior. A poor forecast performance during black swan-like crashes is explainable, a sudden drop in forecast performance during calm times is not.

[22] Gu et al.'s (2019b) forecast sample contains the more tranquil years between 1986 and 1995. Because of the late availability of the Optionmetrics data required for the theory-based approaches, we have to start in 1996.

[23] Note that the R^2_{oos} in Figures 7 and 8 refer to the year at which the forecast was issued. For example, the R^2_{oos} for the one year horizon forecasts for the year 2008 have been issued from January 2007 to December 2007, thus the effect of the financial crises is visible in 2007. Except for the end-of-month December forecast, all monthly forecasts are issued in the same year as the realization.

4.2 Short training and hybrid approaches

So far, we have treated the theory-based and the machine learning methodologies as competitors, but as outlined in Section 2.3, we also want to assess the potential of hybrid approaches that combine these opposing philosophies. The results reported in Table 4 indicate that this may be a promising idea. We observe that although the theory-based and machine learning forecasts covary positively, the correlations are not strong, such that the two approaches appear to account for different aspects.

[Insert Table 4 about here]

Any hybrid methodology must accommodate the late availability of the Optionmetrics data, which are required for the implementation of the theory-based formulas. The training of hybrid models therefore can not begin before 1996. Accordingly, training and/or validation samples must be shortened to retain a sufficient number of out-of-sample years for comparison with the "long-training" results. As outlined supra, we deal with this issue by applying the alternative validation scheme in Figure 4. Tables 5 (one month horizon) and 6 (one year horizon) report the results thus obtained.[24] These tables display two sets of machine learning results, one using the same feature input as before, but applying the short training scheme. The other set, referred to as "ML with theory features" is obtained by adding to the feature set the two theory-based stock risk premia measures (Martin/Wagner's and Kadan/-Tang's) as well as Martin's (2017) lower bound of the expected market return. The following discussion of the results intertwines the assessment of the incremental effect of including these theory-based features with that of applying the short validation scheme.

[Insert Tables 5 and 6 about here]

[24] Comparing Tables 5 and 6 with Table 3, it should be noted that theory-based results only change because the out-of-sample evaluation period is shortened: The years 1996 and 1997 are not used.

We have seen that at the one month horizon the machine learning approaches do not perform well. As expected, the results worsen when applying the short training scheme. Panel A of Table 5 shows that this deterioration is only mildly mitigated by the inclusion of the theory-consistent features. This result corroborates the conclusion that pursuing a (pure) theory-based approach is the preferable strategy at the short horizon.

Table 6 shows that the detrimental effect of shortened training is also observed on the one year horizon, but that it is different across machine learning approaches. The neural network performances deteriorate so substantially that the conclusion is supported that the short training scheme should be avoided altogether. By contrast, boosted trees and random forests are affected to a much lesser extent: As a result of the shorter training, the R^2_{oos} decreases and its standard deviation increases for random forests, too; but Panel A of Figure 9 suggests that the effect is mitigated as the training sample grows. At the beginning of the sequential validation procedure, there are only a few years of observations available for training. Accordingly, the dot-com turmoil hits a relatively "untrained" random forest, which results in a deterioration of the R^2_{oos}. This can be seen by comparing the year 2000 $R^2_{oos,s}$ of the random forest in Panel A of Figure 8 with the counterpart in Figure 9. As the training sample grows, the random forest performance picks up and reaches, towards the end of the sample period, the performance level of the "long training" variant. These observations suggests that the shortening of the validation period (from 12 years to one) is of a much lesser importance than the shortening of the training period.

[Insert Figure 9 about here]

Table 6 indicates that the augmentation of the feature set by theory-based measures has a positive effect only for random forests. The effect is not large, but it is interesting to see how the augmentation helps the "short-trained" random forest when it is needed, namely for the 2008 forecast (cf. Panel A of Figure 9). These results support the notion that random forest are the most useful machine learning approach for the job at hand.

An epistemological concern regarding the use of machine learning techniques is that although they may perform well empirically, it is often unclear why. Take the random forest results at the one year horizon. Our training and validation scheme can obviously detect non-linearities that translate into a superior forecast performance. As a next step, one may investigate the reasons by (metaphorically speaking) exploring the depths of the forest. And as it sometimes happens, when relying on data science methods, one would engage in trying to understand the complexity of a trained model, instead of economic reality itself.

The hybrid approach proposed in Section 2.3 is based on a different philosophy. It relies on Martin/Wagner's theory-based approach, which starts from the basic asset pricing equation, the keystone of financial economics. We have seen that this approach is empirically not unsuccessful. Our idea for a hybrid strategy is to take it as a basis, and to model what theory can not account for, the approximation errors, with the help of machine learning techniques.

In the segment labeled "Theory assisted by ML" in Table 6, we report the results obtained when taking this idea to the data.[25] We observe that an improvement of the theory-consistent approach through machine learning assistance is not a given. In fact, elastic net and neural network drive the originally positive R^2_{oos} of the theory-consistent approach into the negative domain. While the assistance of boosted

[25] We do not report the one month horizon results, because we have seen supra that machine learning approaches do not perform well here. We did not expect more when explaining the residual of the theory-based forecast, and indeed we did not find an improvement. These additional results are available upon request. Moreover, we only use Martin/Wagner's formula (2.5) as a basis for the hybrid approach, because the previous results indicate that it is the more promising of the two theory-based approaches.

trees offer just a mild improvement, it is again the random forest that stands out and increases the theory-based R^2_{oos} of 9.1% by 7 percentage points. The standard deviation of $R^2_{oos,s}$ also grows, but as can be seen in Figure 10, this increase is mainly due to the adverse effect of short training. We again observe the harsh drop of the $R^2_{oos,s}$ in the year 2000, when an insufficiently trained random forest is asked to perform in a turbulent time. However, Figure 10 also illustrates the machine learning-assisted improvement of the theory-based forecast as the training sample size increases towards the recent past.

[Insert Figure 10 about here]

Due to the "short-training" effect, the 16.1% out-of-sample R^2 delivered by the random forest-assisted Martin/Wagner approach is not directly comparable with the 19.1% of the "long-trained" pure random forest reported in Table 3. However, by zooming in on more recent forecast samples, we observe in Figure 11 that with increasing training sample size, the performance of the ML-assisted, theory-based forecast aligns with that of the long-trained random forest, and at the very end of the sample period even surpasses it.

[Insert Figures 11 and 12 about here]

The prediction deciles plots in Figure 12 further corroborate the conclusion that a hybrid approach that combines Martin/Wagner's theory-consistent approach with random forest machine learning is, not only from an epistemological point of view, a promising alternative for the task of quantifying stock risk premia.

5 Conclusion

In this study, we have followed and compared two diverging paths towards measuring stock risk premia and attempted a reconciliation of the opposing philosophies.

Exploiting the alleged predictive abilities of theory-based and machine learning methodologies, we compare them at the one month and the one year forecast horizon. We find that the theory-consistent approach offers advantages at the one month horizon, where machine learning approaches do not perform well. Recommendations regarding the use of data science methods at short horizons should therefore be taken with a grain of salt. At the one year horizon, the picture is more complex. Of the four machine learning methods that are considered in this study, two deliver worse performances than the theory-based approach, one is comparable, and one, the random forest approach, is superior. Its out-of-sample R^2 is notably higher than what is reported in previous work. Neural networks, which prior literature characterizes as both extremely flexible, but also somewhat unstable, are outperformed by the theory-based approach.

When considering hybrid approaches that aim for a combination of the theory-consistent and the machine learning methodologies, we have to acknowledge restrictions on data availability. The computation of the theory-consistent measures is not possible before 1996. Machine learning techniques that attempt to make use of theory-consistent features must therefore rely on shorter training samples, which adversely affect the initial forecast performance. Fortunately, the effect is mitigated as the training sample size increases, when a dynamic training procedure is applied.

Acknowledging epistemological concerns regarding the use of agnostic machine learning procedures in a well-developed field like finance, we propose a hybrid methodology that takes the theory-based approach towards stock risk premia as its basis and then applies machine learning techniques to a residual component unexplained by theory. The empirical performance of this combined approach is encouraging. Of its overall explanatory power in excess of a zero forecast, which matches that of the best agnostic machine learning approach, 57 percent comes from the theory-based part, 43 percent are attributable to machine learning assistance. We view this hybrid model as a promising alternative to unite the diverging paths in finance.

A Appendix

A.1 Details on the theory-based stock risk premia formulas

This section provides details behind the stock risk premia formulas (2.2) and (2.3) and the nature of the approximation residuals $a^i_{t,T}$ and $\xi^i_{t,T}$. We delineate the assumptions and rationales behind their omission, which provides the theory-based approximation formulas in Equations (2.5) and (2.6).

Martin and Wagner's (2019) derivations originate from basic asset pricing equation by focusing on the gross return of a portfolio with maximal expected log return ($R^g_{t,T}$). This growth-optimal return has the unique property among gross returns that its reciprocal is an SDF, such that $m_{t,T} = 1/R^g_{t,T}$. Using this SDF to price the payoff $X^i_{t,T} = R^i_{t,T} \cdot R^g_{t,T}$ gives:

$$\mathbb{E}_t\Big(m_{t,T} \cdot X^i_{t,T}\Big) = \mathbb{E}_t\Big(R^i_{t,T}\Big) = \frac{1}{R^f_{t,T}} \mathbb{E}^*_t(R^i_{t,T} \cdot R^g_{t,T}), \tag{A-1}$$

where the $*$ notation again (and henceforth) indicates that the expected value is computed with respect to the risk-neutral measure. Division by $R^f_{t,T}$ and subtracting $\mathbb{E}^*_t\Big(R^i_{t,T}/R^f_{t,T}\Big) \times \mathbb{E}^*_t\Big(R^g_{t,T}/R^f_{t,T}\Big) = 1$ (the price of any gross return is 1) yields:

$$\mathbb{E}_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) = 1 + \operatorname{cov}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}, \frac{R^g_{t,T}}{R^f_{t,T}}\right), \tag{A-2}$$

An orthogonal projection under the risk-neutral measure of $R^i_{t,T}/R^f_{t,T}$ on $R^g_{t,T}/R^f_{t,T}$ and a constant gives:

$$\frac{R^i_{t,T}}{R^f_{t,T}} = \alpha^i_{t,T} + \beta^i_{t,T} \cdot \frac{R^g_{t,T}}{R^f_{t,T}} + u^i_{t,T}, \tag{A-3}$$

where the moment conditions $\mathbb{E}^*_t(u^i_{t,T}) = 0$ and $\mathbb{E}^*_t(u^i_{t,T} \cdot R^g_{t,T}) = 0$ define the projection coefficient

$$\beta^i_{t,T} = \frac{\operatorname{cov}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}, \frac{R^g_{t,T}}{R^f_{t,T}}\right)}{\operatorname{var}^*_t\left(\frac{R^g_{t,T}}{R^f_{t,T}}\right)}$$

and $\alpha^i_{t,T} = 1 - \beta^i_{t,T}$. Insertion in (A-2) gives:

$$\mathbb{E}_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) = 1 + \beta^i_{t,T} \cdot \operatorname{var}^*_t\left(\frac{R^g_{t,T}}{R^f_{t,T}}\right). \tag{A-4}$$

Moreover, (A-3) implies:

$$\operatorname{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) = (\beta^i_{t,T})^2 \cdot \operatorname{var}^*_t\left(\frac{R^g_{t,T}}{R^f_{t,T}}\right) + \operatorname{var}^*_t(u^i_{t,T}). \tag{A-5}$$

To make these results practically usable, Martin and Wagner (2019) propose to linearize $(\beta^i_{t,T})^2 \approx 2\beta^i_{t,T} - k$, which for $k = 1$ amounts to a first-order Taylor approximation at $\beta^i_{t,T} = 1$. Using this approximation and inserting in (A-4) (for $k = 1$) removes the dependence on $\beta^i_{t,T}$, viz:

$$\mathbb{E}_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) \approx 1 + \frac{1}{2}\operatorname{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2}\operatorname{var}^*_t\left(\frac{R^g_{t,T}}{R^f_{t,T}}\right) - \frac{1}{2}\operatorname{var}^*_t(u^i_{t,T}). \tag{A-6}$$

We note that the term that is neglected on the right-hand side due to the linearization is $-\operatorname{var}^*_t(R^g_{t,T}/R^f_{t,T})(\beta^i_{t,T}-1)^2$. The approximation should thus be reasonable for stocks whose $\beta^i_{t,T}$ is close to one.

Using w_t^j, the weight of stock j in a market index with gross return $R_{t,T}^m$, Martin and Wagner (2019) perform a value-weighting of (A-6) to obtain:

$$\mathbb{E}_t\left(\frac{R_{t,T}^m}{R_{t,T}^f}\right) \approx 1 + \frac{1}{2}\sum_j w_t^j \text{var}_t^*\left(\frac{R_{t,T}^j}{R_{t,T}^f}\right) + \frac{1}{2}\text{var}_t^*\left(\frac{R_{t,T}^g}{R_{t,T}^f}\right) - \frac{1}{2}\sum_j w_t^j \cdot \text{var}_t^*(u_{t,T}^i). \tag{A-7}$$

Subtracting (A-7) from (A-6) removes the dependence on the unobservable optimal growth portfolio, such that:

$$\begin{aligned}\mathbb{E}_t\big(R_{t,T}^i\big) \approx \mathbb{E}_t\big(R_{t,T}^m\big) + \frac{R_{t,T}^f}{2}\left[\text{var}_t^*\left(\frac{R_{t,T}^i}{R_{t,T}^f}\right) - \sum_j w_t^j \cdot \text{var}_t^*\left(\frac{R_{t,T}^j}{R_{t,T}^f}\right)\right] \\ - \frac{R_{t,T}^f}{2}\left(\text{var}_t^*(u_{t,T}^i) - \sum_j w_t^j \cdot \text{var}_t^*(u_{t,T}^j)\right).\end{aligned} \tag{A-8}$$

Keeping track of the approximation error due to the linearization, we note that the term that is omitted on the right-hand side of (A-8) is

$$\kappa_{t,T}^i = -\frac{1}{2R_{t,T}^f}\text{var}_t^*\big(R_{t,T}^g\big) \cdot \left[(\beta_{t,T}^i - 1)^2 - \sum_j w_t^j \cdot (\beta_{t,T}^i - 1)^2\right].$$

To account for the first term on the right-hand side of (A-8), Martin and Wagner (2019) draw on a result by Martin (2017) who derives a lower bound for the expected return of a market index. His starting point is again the basic asset pricing equation (2.1), which can be written in terms of the price of the payoff $(R_{t,T}^i)^2$ using an add and subtract strategy:

$$\mathbb{E}_t(R_{t,T}^i) - R_{t,T}^f = \big(\mathbb{E}_t[m_{t,T} \cdot (R_{t,T}^i)^2] - R_{t,T}^f\big) - \big(\mathbb{E}_t[m_{t,T} \cdot (R_{t,T}^i)^2] - \mathbb{E}_t(R_{t,T}^i)\big). \tag{A-9}$$

The first term on the right-hand side of (A-9) can be related to a risk-neutral variance and the second term to a covariance under the physical measure, viz:

$$\mathbb{E}_t(R^i_{t,T}) - R^f_{t,T} = \frac{1}{R^f_{t,T}} \text{var}^*_t(R^i_{t,T}) - \text{cov}_t(m_{t,T} \cdot R^i_{t,T}, R^i_{t,T}). \tag{A-10}$$

As noted in the main text, Kadan and Tang (2019) use (A-10) for their quantification and approximation of stock risk premia.

Martin (2017) argues that for an asset return that qualifies as a market return proxy (denoted $R^m_{t,T}$), it should be the case that

$$\xi_{t,T} = \text{cov}_t(m_{t,T} \cdot R^m_{t,T}, R^m_{t,T}) < 0, \tag{A-11}$$

which is referred to as the negative correlation condition (NCC). Intuitively, an investor's marginal rate of intertemporal substitution should be negatively correlated with any portfolio that qualifies as a market index. Accordingly,

$$\mathbb{E}_t(R^m_{t,T}) - R^f_{t,T} \geq \frac{1}{R^f_{t,T}} \text{var}^*_t(R^m_{t,T}). \tag{A-12}$$

Assuming that the inequality (A-12) is binding, and using it for (A-8) yields:

$$\begin{aligned} \mathbb{E}_t\left(R^i_{t,T}\right) - R^f_{t,T} \approx R^f_{t,T} \cdot \left[\text{var}^*_t\left(\frac{R^m_{t,T}}{R^f_{t,T}}\right) + \frac{1}{2}\left\{ \text{var}^*_t\left(\frac{R^i_{t,T}}{R^f_{t,T}}\right) - \sum_j w^j_t \cdot \text{var}^*_t\left(\frac{R^j_{t,T}}{R^f_{t,T}}\right)\right\}\right] \\ - \frac{R^f_{t,T}}{2} \cdot \left[\text{var}^*_t(u^i_{t,T}) - \sum_j w^j_t \cdot \text{var}^*_t(u^j_{t,T}) \right], \end{aligned} \tag{A-13}$$

where the approximative formula (A-13) omits the term $\kappa^i_{t,T} - \xi_{t,T}$ on the right-hand side. Equation (2.2) in the main text thus obtains using

$$a^i_{t,T} = \kappa^i_{t,T} - \xi_{t,T} - \zeta^i_{t,T} \tag{A-14}$$

where

$$\zeta^i_{t,T} = \frac{1}{2} R^f_{t,T} \cdot \left[\mathrm{var}^*_t(u^i_{t,T}) - \sum_j w^j_t \cdot \mathrm{var}^*_t(u^j_{t,T}) \right]. \tag{A-15}$$

Working with the abbreviated formula in (2.5) in the main text thus entails three approximations. First, the linearization of $(\beta^i_{t,T})^2$, second, the assumption that Martin's (2017) lower bound for the expected return of the market is binding, and third, that the residual variances $\mathrm{var}^*_t(u^i_{t,T})$ are very similar across stocks, such that $\zeta^i_{t,T}$ is negligibly small in absolute terms.

A.2 Details on the construction of the data base

Identifying historical S&P 500 constituents

In our analysis we focus on firms that appear at least once as an S&P 500 constituent during March 1964 and December 2019. For that purpose, we must identify the set of active S&P 500 constituents for each date of the sample period. WRDS allows for different ways to obtain the desired list of historical S&P 500 consituents (HSPC). Due to user-specific WRDS data access constraints, not all paths are feasible for us. In particular, we cannot access the HSPC tables that are, according to WRDS, available from CRSP.[26] The access to these tables is desirable because if we had it, the identification of historical S&P 500 constituents and the extraction of security level information from CRSP would be based on the same security identifier. Instead, we have to take a detour via Compustat to connect securities across data bases. To our knowledge, there are two different paths that help achieve this goal.

The first is based on a monthly security query from Compustat's `SECM` file. In the `SECM` file, the variable `SPMIM` (S&P Major Index Code - Historical) allows us to

[26] The (for us) inaccessible SAS data files in question are named `dsp500list` and `msp500list`. They contain the starting and ending dates for each security identified by PERMNO.

identify the S&P 500 constituents. As advised by WRDS, from March 1964 until November 1994 we select `SPMIM` $\in \{10, 40, 49, 60\}$ to identify the S&P 500 constituents. According to WRDS, S&P 500 constituents are the union of S&P Transportation (`SPMIM`=40), Utilities (`SPMIM`=49), Financial (`SPMIM`=60), and Industrial (`SPMIM` 10).[27] From December 1994 onwards, WRDS advises to just select `SPMIM`=10 to identify S&P 500 constituents. The resulting table contains information on each security's identity providing permanent company (named `GVKEY`) and security (named `IID`) identifiers. Moreover, it reports the dates at which a given security was part of the S&P 500.

The alternative path to generate the HSPC list is based on the Compustat table `IDXCST_HIS`, which collects securities that are identified by the variable `GVKEYX`, which indicates the membership of a company in the S&P 500.

The HSPC lists resulting from the two approaches differ only slightly, and we believe that both methods yield reliable data on S&P 500 firm membership over time. The first of the two approaches identifies 1,697 S&P 500 constituents between January 1962 and December 2019, whereas the second identifies 1,713 S&P 500 constituents between March 1964 and December 2019. Both agree on 1,691 of the union of 1,719 identified S&P 500 constituents. There are six S&P 500 constituents that are identified exclusively by the first approach, and 22 S&P 500 constituents that are identified exclusively by the second approach. We choose the first approach because it provides a more consistent coverage of HSPCs during the 1970s.

Linking Compustat and CRSP

The security identifier `CUSIP`, just as ticker symbols, can change over time. `CUSIPs`

[27] This information is taken from https://wrds-www.wharton.upenn.edu/pages/support/applications/programming-examples-and-other-topics/sp-500-datasets-and-constituents/

are not permanently linked to a firm. As noted by WRDS,[28] “[a] change in CUSIP [...] could be triggered by any change in the security, including non-fundamental events such as splits and company name changes.” Compustat therefore also supplies a permanent security identifier, which results from the combination of the variables GVKEY and IID, which we retrieve along with CUSIP when constructing the list of HSPC. A permanent security identifier is important in order to keep track of any legal or structural changes that may occur during the lifetime of a company.

Even though Compustat, CRSP and OptionMetrics have CUSIP identification in common, their permanent identifiers differ. Moreover, despite the common identification via CUSIP, it is still possible that the tags of one and the same security differ. The reason for this is often a different assessment of whether a change in the corporate structure should actually be recorded in CUSIP. First we will examine the possibilities of combining Compustat and CRSP and then we will take a closer look at the connection possibilities of Compustat and OptionMetrics.

Establishing a connection between CRSP and Compustat, and merging the respective data for a security, is a common task in empirical finance. For that purpose, WRDS provides a linkage table that enables the cross-database identification of securities using each database’s permanent identifiers. As mentioned before, Compustat uses a combination of GVKEY and IID to track securities whereas CRSP relies on a permanent security identifier called PERMNO.[29]

We find that using the linkage table, it is not possible to find a one-to-one assignment of the permanent identifiers across CRSP and Compustat. Instead, there are several CRSP PERMNO identifiers that can be assigned to a unique combination of Compus-

[28] For a detailed description of the cross-database identification problem see https://wrds-www.wharton.upenn.edu/pages/support/applications/linking-databases/linking-crsp-and-compustat/

[29] CRSP additionally provides company identification via PERMCO. However, since securities are unambiguously identified by PERMNO, PERMCO is not of particular importance in our context.

tat's GVKEY+IID. As a result, when we use the S&P 500 constituents identified with the help of Compustat, the list of matched CRSP CUSIPs is longer than the list of Compustat CUSIPs. This suggests that it may sometimes be necessary to merge the price information of multiple CRSP CUSIPs with a single Compustat CUSIP.

However, a particular connection between the permanent identifiers in CRSP and Compustat must be one-to-one only at a given point in time. For this purpose we use the variables LINKDT and LINKENDDT, which contain information about the validity of a connection of the permanent CRSP and Compustat identifiers at a certain point in time. The good news is that the connection of the permanent identifiers in CRSP and Compustat indeed is one-to-one at corresponding dates.

Using the list of S&P 500 constituents obtained from Compustat, we extract security level information from CRSP. The crspa library provides price information and the number of outstanding shares on a daily frequency for each index constituent. The CRSP index price data is obtained from the library crsp via the table dsi.

Linking Compustat and OptionMetrics

We obtain the OptionMetrics volatility surface data from the library optionm where there is a separate volatility surface table for each available year. WRDS offers a beta version of a linkage table, named opcrsphist, to connect CRSP and OptionMetrics. In this table there are scores that indicate the quality of a match between SECID and PERMNO. The highest score for the most reliable link is given to 8-digit CUSIP identification. To the best of our knowledge, there is currently no better way to link either Compustat or CRSP to OptionMetrics. Putting aside all the shortcomings that are attached to CUSIP identification, we search OptionMetrics for the list of HSCP that we derived from Compustat. Clearly, this approach does not yield a 100% coverage of S&P 500 constituents in the OptionMetrics data, however, the average per day coverage of constituents is still quite satisfactory and improves on

Martin and Wagner (2019).

The calculation of the theory-consistent excess return forecasts requires the price of the underlying at the day the forecast is made. There are several ways to obtain these prices. Firstly, CRSP provides daily security price data for the period from December 1925 until December 2019. Furthermore, the OptionMetrics database provides prices for all stocks for which options are traded between January 1996 and December 2018. It seems natural to take the prices from OptionMetrics and calculate expected excess returns. This would ensure that we get both the volatility surface and the associated price data from the same data source, so that we do not face any cross-database security identification problems. However, at the latest when comparing the theory-consistent forecasts to those of machine learning models, our study requires that we have prices available for all S&P 500 constituents at times before 1996. These are needed to train the machine learning models on a period prior to the out-of-sample testing period. Thus, we decide to get the price data from CRSP instead of OptionMetrics.

A.3 Details on the implementation of the theory-based approaches

In the following, we describe how to approximate the formula for risk-neutral variances in (2.4) using volatility surface data from OptionMetrics. The right-hand side of (2.4) depends on observable time t information only. We require the price of the underlying, a proxy for the risk-free interest rate, the price of the forward contract and the prices of European options at different strikes, each of the latter with maturity in T. Since the formula in (2.4) is based on the put-call parity, it exclusively applies to European option contracts. In Martin (2017) this is not an issue since options on the S&P 500 index are traded European style and the observed option price data from OptionMetrics can be used to approximate the risk-neutral variances. However, options on the constituents of the S&P 500 are traded American style. As is well known, there is no put-call parity for American options and thus prices of American options are not directly useful. However, we find the prices of equivalent European options using OptionMetrics' volatility surface. This volatilty surface is constructed using the price of the underlying, the risk-free interest rate and various times to maturity and strike prices of American options. With these implied volatilities one can calculate the price of an equivalent European option using the Black-Scholes-Merton (BSM) formula. This is possible, because the implied volatility is, apart from the risk-free interest rate, the only ingredient of the BSM model which is not directly observable. Further, the implied volatility depends on the nature of the underlying but is independent from the terms and conditions of the option contract. Thus, it may be conceived as a conversion factor for translating between otherwise equivalent European and American style options.

Besides the issue with the exercise style of options on S&P 500 constituents, we have to approximate the integrals in (2.4) because we do not observe option prices at a continuum of strikes. Martin's (2017) strategy amounts to using for every strike K_j from the ascending list of available strikes $K_1, K_2, \dots K_n$[30]

$$\Omega_j(K_j) = \begin{cases} \text{put}_j(K_j) & \text{if} \quad K_j < F_j \\ \text{call}_j(K_j) & \text{if} \quad K_j \geq F_j \end{cases} \tag{A-16}$$

to approximate the sum of the two integrals on the right-hand side of (2.4) via

$$\sum_j \Omega_j(K_j) \cdot \Delta K_j, \tag{A-17}$$

[30] For notational convenience, we drop the security, time, and maturity indices.

where

$$
\begin{aligned}
\Delta K_j &= \frac{K_{j+1} - K_{j-1}}{2} \qquad j = 2, \dots, n-1, \\
\Delta K_1 &= K_2 - K_1, \\
\Delta K_n &= K_n - K_{n-1}.
\end{aligned}
$$

Put differently, we sum over discrete strikes, weighting observations with the distance between previous and next strike divided by two. Hence, we center the rectangles, of which we add their areas, around the observed strikes. This approach ensures that the approximation error, that is due to the discreteness of strikes, is limited from above.

It is important to mention that Martin (2017) uses observed European option prices to calculate risk-neutral variances for the S&P 500 index, whereas Martin and Wagner (2019) base their calculations on the volatility surface with its standardized implied strikes, which has the advantage that all calculations are based on the same standardized number of strikes, both for the index and its constituents. We decide to follow the latter approach.

A.4 Details on the implementation of the machine learning approaches

A.4.1 Software and computing resources

We implement our machine learning procedures using Python's scikit-learn ecosystem, which provides a considerable number of popular machine learning models. For the training of neural networks we rely on Python's deep learning library Keras with the Tensorflow backend. Although scikit-learn also allows the training of neural networks, it is less flexible than Keras and lacks some degrees of freedom in the construction of network architectures. In order to achieve maximum parallelization during our extensive hyper-parameter search, we further combine scikit-learn with the parallel computing environment Dask. The hyper-parameter optimization for the Keras networks is conducted using Talos, an efficient hyper-parameter optimization toolbox which delivers Tensorflow's native parallelism out of the box.

The computing infrastructure that we use for our experiments is provided by the state of Baden Württemberg in the shape of a High Performance Computing (HPC) cluster named MLS&WISO, which is short for Molecular Life Science (MLS) and Economics and Social Science (WISO). On this cluster, we primarily use the "standard" compute nodes, of which there are 476 in total, each equipped with 2 x Intel Xeon E5-2630v3 Haswell Processors (2.4 GHz), 16 Cores and 64 (128) GB working memory (RAM).

A.4.2 Comparison with Gu et al. (2019b)

Gu et al. (2019b) set a benchmark for machine learning based excess return predictions with a forecast horizon of both one month and one year. Since in our experiment we rely on the same source of data (Green et al.'s (2017) SAS code), a natural check is whether the machine learning models can live up to their promises. Even though we believe that the long-term forecasts should be more promising, our out-of-sample results for the one month horizon should ideally be in vicinity of what is reported in Gu et al. (2019b). We find, however, that the one month performance of our machine learning models lag behind the performance of the models presented in Gu et al. (2019b). This observation may be attributed to several reasons.

Universe of stocks

We believe that the deviations in out-of-sample performance partially stem from differences in the universes of stocks being considered. Since the theory-consistent approach allows us to compute excess return forecasts exclusively for S&P 500 constituents, we feed our machine learning models with features of S&P 500 consituents only. In this way we are not only able to establish a fair comparison between the two approaches, but further make it possible to combine them in a hybrid manner, which for non-S&P 500 constituents is impossible. Gu et al. (2019b) instead rely on a much broader set of NYSE-, AMEX- and NASDAQ-traded firms and also include penny stocks, yielding an average number of stocks per month that exceeds 6,200.[31]

Sample period

Our overall sample period deviates from the one reported in Gu et al. (2019b). They start their training in 1957, the beginning of the S&P 500. However, taking the problem of missing values into consideration, we find that we are unable to replace all of the missing values that occur between 1957 and October 1974. The reason is that we follow Gu et al. (2019b) in filling a feature's missing values with the cross-sectional median at a given point in time. A problem that is attached to any cross-section based imputation method is, that if the first observation of some variable is dated October 1974 (as is the case for "cash flow volatility"), these methods break down since there is no cross-section from which the missing data prior to October 1974 could be inferred. One might be inclined to resort to more sophisticated imputation

[31] In case of the long training/validation scheme described in Figure 3 we do not use any of the theory-consistent measures. Hence, in principle, we could have used the extended set of NYSE-, AMEX- and NASDAQ-traded firms which is used in Gu et al. (2019b) to train our machine learning models. However, we want to be able to compare the models with long training/validation scheme to the (hybrid) machine learning models with the short training/validation scheme described in Figure 4. Hence, for a true ceteris paribus assessment of the effect of reducing the size of the training/validation windows, we must exclude all non-S&P 500 stocks. What speaks in favor of an extended universe of stocks is that with a higher number of observations, the machine learning procedures are less likely to overfit. However, this advantage may be compensated by an (allegedly) higher signal-to-noise ratio, which we believe is due to the fact that S&P 500 stocks are larger than the average stock listed on NYSE, AMEX and NASDAQ. Ultimately, it is unclear whether additional stocks with a lower market capitalization add information for a prediction of excess returns on S&P 500 stocks. The figures in Figure 1 Gu et al. (2019b), however, suggest that there is quite some variation in a model's performance across stocks with different market capitalizations.

methods, but we refrain from doing so because we fear that our results could be biased by overly restrictive assumptions about the structure of the missing data.

Out-of-sample testing period

We need to adjust our out-of-sample testing period in order to enable a comparison with the theory-consistent forecasts proposed by Martin and Wagner (2019) and Kadan and Tang (2019). Gu et al. (2019b) report their out-of-sample results on a testing period that ranges from 1987 to 2016. The option data that are used to construct the theory-consistent forecasts are, however, only available from 1996 onwards. We therefore extend our training/validation period until 1995 and begin with out-of-sample testing not before 1996.

Feature transformations

In the construction of our feature set, we follow an earlier version of Gu et al. (2019b) which is dated April 9, 2018. In this version they seem to take their set of firm characteristics as they are, without any transformations except for the Kronecker product with the macroeconomic variables. In the recently published version of their paper, however, it is mentioned in footnote 29, that they map their firm characteristics based on a period-by-period ranking to the interval $[-1, 1]$, an approach that is used, among others, by Freyberger et al. (2019). Thereby they account for the fact, that "[one is] typically not interested in the value of a characteristic in isolation, but rather in the rank of the characteristic in the cross section." (cf. Freyberger et al. (2019)) The effect of switching from the one approach to the other will be discussed in a future version of our paper.

Hyper-parameter tuning

We adapt the search space for the hyper-parameters of each machine learning model to the requirements of our restricted sample. Any differences in the out-of-sample forecast performance may be due to these changes. For example, Gu et al. (2019b) set the maximum depth of each tree in their random forest to 6, which in our case appears to be too restrictive. Thus, we increase the upper boundary of the interval from 6 to 30, which improves our results especially with a forecast horizon of one year. Also we extend the search space for the elastic net's L1-ratio, which in Gu et al. (2019b) is fixed at 0.5, to allow for a more flexible combination of L1- and L2-penalization. For the gradient boosted regression trees we limit the number of trees to the interval $[2, 100]$, increase the maximum tree depth to 3 and extend the interval for the learning rate to $[0.005, 0.12]$. In case of the neural networks, we switch from the seed-value based ensemble approach that is propagated in Gu et al. (2019b) to dropout regularization, which is a more efficient way of regularizing neural networks. Admittedly, ensemble methods have proven to be the gold standard in many machine learning applications since they allow the different aspects learned by each individual model to be subsumed in a single prediction. However, creating ensembles can become prohibitively expensive if the number of sample observations is large and/or each individual model is highly complex. Srivastava et al. (2014) address this issue by proposing dropout regularization which retains the capability of neural networks to learn different aspects of the data while being computationally more efficient than the standard ensemble approach. As proposed in the original paper, we also introduce a maximum weight norm for each hidden layer. Compared to Gu et al. (2019b), we also reduce the batch size, both due to the fact that a smaller batch size typically improves the generalization capabilities of a model that is trained with stochastic gradient descent (cf. Keskar et al. (2016)) and also due to our reduced sample size which is restricted to S&P 500 constituents only. For a detailed comparison of the hyper-parameter search spaces the reader may refer to our Table 2 and Table A.5 in Gu et al. (2019b).

References

Avramov, D., S. Cheng, and L. Metzker (2020): "Machine Learning versus Economic Restrictions: Evidence from Stock Return Predictability," Working Paper.

Bianchi, D., M. Büchner, and A. Tamoni (2019): "What Matters When? Time-Varying Sparsity in Expected Returns," Working Paper.

Bryzgalova, S., M. Pelger, and J. Zhu (2019): "Forest Through the Trees: Building Cross-Sections of Stock Returns," Working Paper.

Chen, L., M. Pelger, and J. Zhu (2019): "Deep Learning in Asset Pricing," Working Paper.

Cochrane, J. H. (2005): *Asset Pricing.* Princeton University Press, Princeton, NJ.

Feng, G., S. Giglio, and D. Xiu (2019): "Taming the Factor Zoo: A Test for New Factors," *forthcoming: The Journal of Finance.*

Freyberger, J., A. Neuhierl, and M. Weber (2019): "Dissecting Characteristics Nonparametrically," *forthcoming: The Review of Financial Studies.*

Frost, D. (1916): *Mountain Interval.* Henry Hold and Company., New York, NY, USA.

Green, J., J. R. M. Hand, and X. F. Zhang (2017): "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns," *The Review of Financial Studies*, 30(12), 4389–4436.

Gu, S., B. T. Kelly, and D. Xiu (2019a): "Autoencoder Asset Pricing Models," *forthcoming: Journal of Econometrics.*

——— (2019b): "Empirical Asset Pricing via Machine Learning," *forthcoming: The Review of Financial Studies.*

HASTIE, T., R. TIBSHIRANI, AND J. FRIEDMAN (2017): *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

KADAN, O., AND X. TANG (2019): "A Bound on Expected Stock Returns," *forthcoming: The Review of Financial Studies*.

KELLY, B. T., S. PRUITT, AND Y. SU (2019): "Characteristics are Covariances: A Unified Model of Risk and Return," *Journal of Financial Economics*, 134(3), 501–524.

KESKAR, N. S., D. MUDIGERE, J. NOCEDAL, M. SMELYANSKIY, AND P. T. P. TANG (2016): "On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima," *arXiv preprint arXiv:1609.04836*.

KOZAK, S., S. NAGEL, AND S. SANTOSH (2019): "Shrinking the Cross-Section," *forthcoming: Journal of Financial Economics*.

LIGHT, N., D. MASLOV, AND O. RYTCHKOV (2017): "Aggregation of Information About the Cross Section of Stock Returns: A Latent Variable Approach," *The Review of Financial Studies*, 30(4), 1339–1381.

MARTIN, I. (2017): "What is the Expected Return on the Market?," *The Quarterly Journal of Economics*, 132(1), 367–433.

MARTIN, I., AND S. NAGEL (2019): "Market Efficiency in the Age of Big Data," NBER Working Papers 26586, National Bureau of Economic Research.

MARTIN, I. W. R., AND C. WAGNER (2019): "What Is the Expected Return on a Stock?," *The Journal of Finance*, 74(4), 1887–1929.

SRIVASTAVA, N., G. HINTON, A. KRIZHEVSKY, I. SUTSKEVER, AND R. SALAKHUTDINOV (2014): "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *The Journal of Machine Learning Research*, 15(1), 1929–1958.

Welch, I., and A. Goyal (2008): “A Comprehensive Look at The Empirical Performance of Equity Premium Prediction,” *The Review of Financial Studies*, 21(4), 1455–1508.

Figures and Tables

Table 1: Variable descriptives. The table reports descriptive statistics and general information on the variables used for the empirical analyses. Panel A1 is based on Table A.6 by Gu et al. (2019b) and contains information on the firm characteristics used for the machine learning approach. Panel A2 does the same for the macro features and Panels B and C refer to the realized excess returns and the theory-based forecasts that are proposed by Martin and Wagner (2019), Kadan and Tang (2019) and Martin (2017), respectively. For each measure, we report its debut in finance literature (author(s), year, journal), from which database it can be obtained (source), at which frequency it is reported (freq.), and the mean and standard deviation in our sample (avg. and std., respectively).

Panel A1: Firm characteristics	**Source**	**Freq.**	**Avg.**	**Std.**	**Author(s)**	**Year**	**Journal**
1-month momentum	CRSP	Monthly	0.015	0.106	Jegadeesh & Titman	1993	JF
12-month momentum	CRSP	Monthly	0.175	0.429	Jegadeesh	1990	JF
36-month momentum	CRSP	Monthly	0.397	0.766	Jegadeesh & Titman	1993	JF
6-month momentum	CRSP	Monthly	0.077	0.257	Jegadeesh & Titman	1993	JF
Abnormal earnings announcement volume	Compustat/CRSP	Quarterly	0.631	1.173	Lerman, Livnat & Mendenhall	2007	WP
Absolute accruals	Compustat	Annual	0.065	0.060	Bandyopadhyay, Huang & Wirjanto	2010	WP
Accrual volatility	Compustat	Quarterly	0.561	13.997	Bandyopadhyay, Huang & Wirjanto	2010	WP
Asset growth	Compustat	Annual	0.145	0.324	Cooper, Gulen & Schill	2008	JF
Beta	CRSP	Monthly	1.022	0.533	Fama & MacBeth	1973	JPE
Beta squared	CRSP	Monthly	1.329	1.399	Fama & MacBeth	1973	JPE
Bid-ask spread	CRSP	Monthly	0.028	0.020	Amihud & Mendelson	1989	JF
Book-to-market	Compustat/CRSP	Annual	0.582	0.507	Rosenberg, Reid & Lanstein	1985	JPM
Capital expenditures and inventory	Compustat	Annual	0.074	0.138	Chen & Zhang	2010	JF
Cash flow to debt	Compustat	Annual	0.212	0.930	Ou & Penman	1989	JAE
Cash flow to price ratio	Compustat	Annual	0.078	0.193	Desai, Rajgopal & Venkatachalam	2004	TAR
Cash flow volatility	Compustat	Quarterly	0.998	26.096	Huang	2009	JEF
Cash holdings	Compustat	Quarterly	0.107	0.136	Palazzo	2012	JFE
Cash productivity	Compustat	Annual	1.774	57.792	Chandrashekar & Rao	2009	WP
Change in 6-month momentum	CRSP	Monthly	0.001	0.404	Gettleman & Marks	2006	WP
Change in inventory	Compustat	Annual	0.011	0.037	Thomas & Zhang	2002	RAS
Change in shares outstanding	Compustat	Annual	0.125	0.347	Pontiff & Woodgate	2008	JF
Change in tax expense	Compustat	Quarterly	0.001	0.010	Thomas & Zhang	2011	JAR
Convertible debt indicator	Compustat	Annual	0.178	0.382	Valta	2016	JFQA
Corporate investment	Compustat	Quarterly	-0.002	0.429	Titman, Wei & Xie	2004	JFQA
Current ratio	Compustat	Annual	2.344	3.795	Ou & Penman	1989	JAE
Debt capacity/firm tangibility	Compustat	Annual	0.483	0.136	Almeida & Campello	2007	RFS
Depreciation/PP&E	Compustat	Annual	0.191	0.233	Holthausen & Larcker	1992	JAE
Dividend initiation	Compustat	Annual	0.017	0.130	Michaely, Thaler & Womack	1995	JF
Dividend omission	Compustat	Annual	0.014	0.117	Michaely, Thaler & Womack	1995	JF
Dividend to price	Compustat	Annual	0.027	0.032	Litzenberger & Ramaswamy	1982	JF

Table 1 continued . . .

Table 1 continued . . .

. . .	Source	Freq.	Avg.	Std.	Author(s)	Year	Journal
Dollar trading volume	CRSP	Monthly	14.595	2.307	Chordia, Subrahmanyam & Anshuman	2001	JFE
Earnings announcement return	Compustat/CRSP	Quarterly	0.005	0.061	Kishore, Brandt, Santa-Clara & Venkatachalam	2008	WP
Earnings to price	Compustat	Annual	0.044	0.217	Basu	1977	JF
Earnings volatility	Compustat	Quarterly	0.013	0.020	Francis, LaFond, Olsson & Schipper	2004	TAR
Employee growth rate	Compustat	Annual	0.069	0.249	Bazdresch, Belo & Lin	2014	JPE
Financial statement score	Compustat	Quarterly	4.427	1.675	Mohanram	2005	RAS
Financial statements score	Compustat	Annual	4.746	1.607	Piotroski	2000	JAR
Gross profitability	Compustat	Annual	0.357	0.293	Novy-Marx	2013	JFE
Growth in capital expenditures	Compustat	Annual	0.515	2.023	Anderson & Garcia-Feijoo	2006	JF
Growth in common shareholder equity	Compustat	Annual	0.142	0.514	Richardson, Sloan, Soliman & Tuna	2005	JAE
Growth in long term net operating assets	Compustat	Annual	0.087	0.114	Fairfield, Whisenant & Yohn	2003	TAR
Growth in long-term debt	Compustat	Annual	0.181	0.523	Richardson, Sloan, Soliman & Tuna	2005	JAE
Idiosyncratic return volatility	CRSP	Monthly	0.042	0.020	Ali, Hwang & Trombley	2003	JFE
Illiquidity	CRSP	Monthly	0.000	0.000	Amihud	2002	JFM
Industry momentum	CRSP	Monthly	0.153	0.281	Moskowitz & Grinblatt	1999	JF
Industry sales concentration	Compustat	Annual	0.075	0.078	Hou & Robinson	2006	JF
Industry-adjusted book to market	Compustat/CRSP	Annual	15.292	656.543	Asness, Porter & Stevens	2000	WP
Industry-adjusted cash flow to price ratio	Compustat	Annual	8.657	271.989	Asness, Porter & Stevens	2000	WP
Industry-adjusted change in asset turnover	Compustat	Annual	-0.007	0.168	Soliman	2008	TAR
Industry-adjusted change in employees	Compustat	Annual	-0.120	0.648	Asness, Porter & Stevens	1994	WP
Industry-adjusted change in profit margin	Compustat	Annual	-0.073	19.993	Soliman	2008	TAR
Industry-adjusted % change in capital exp.	Compustat	Annual	0.247	13.836	Abarbanell & Bushee	1998	TAR
Leverage	Compustat	Annual	2.097	4.584	Bhandari	1988	JF
Maximum daily return	CRSP	Monthly	0.045	0.036	Bali, Cakici & Whitelaw	2011	JFE
Number of earnings increases	Compustat	Quarterly	1.098	1.483	Barth, Elliott & Finn	1999	JAR
Number of years since first Compustat coverage	Compustat	Annual	23.113	13.623	Jiang, Lee & Zhang	2005	RAS
Operating profitability	Compustat	Annual	0.358	0.472	Fama & French	2015	JFE
Organizational capital	Compustat	Annual	0.009	0.009	Eisfeldt & Papanikolaou	2013	JF
% change in current ratio	Compustat	Annual	0.033	0.357	Ou & Penman	1989	JAE
% change in depreciation	Compustat	Annual	0.038	0.276	Holthausen & Larcker	1992	JAE
% change in gross margin - % change in sales	Compustat	Annual	-0.004	0.515	Abarbanell & Bushee	1998	TAR
% change in quick ratio	Compustat	Annual	0.050	0.424	Ou & Penman	1989	JAE
% change in sales - % change in A/R	Compustat	Annual	-0.025	0.399	Abarbanell & Bushee	1998	TAR
% change in sales - % change in inventory	Compustat	Annual	-0.021	0.503	Abarbanell & Bushee	1998	TAR

Table 1 continued . . .

Table 1 continued . . .

. . .	Source	Freq.	Avg.	Std.	Author(s)	Year	Journal
% change in sales - % change in SG&A	Compustat	Annual	0.003	0.160	Abarbanell & Bushee	1998	TAR
% change sales-to-inventory	Compustat	Annual	0.065	0.621	Ou & Penman	1989	JAE
Percent accruals	Compustat	Annual	-1.068	4.907	Hafzalla, Lundholm & Van Winkle	2011	TAR
Price delay	CRSP	Monthly	0.098	0.609	Hou & Moskowitz	2005	RFS
Quick ratio	Compustat	Annual	1.785	3.521	Ou & Penman	1989	JAE
R&D increase	Compustat	Annual	0.094	0.292	Eberhart, Maxwell & Siddique	2004	JF
R&D to market capitalization	Compustat	Annual	0.031	0.049	Guo, Lev & Shi	2006	JBFA
R&D to sales	Compustat	Annual	0.057	0.554	Guo, Lev & Shi	2006	JBFA
Return on assets	Compustat	Quarterly	0.014	0.025	Balakrishnan, Bartov & Faurel	2010	JAE
Return on equity	Compustat	Quarterly	0.034	0.102	Hou, Xue & Zhang	2015	RFS
Return on invested capital	Compustat	Annual	0.105	0.199	Brown & Rowe	2007	WP
Return volatility	CRSP	Monthly	0.021	0.014	Ang, Hodrick, Xing & Zhang	2006	JF
Revenue surprise	Compustat	Quarterly	0.024	0.130	Kama	2009	JBFA
Sales growth	Compustat	Annual	0.138	0.319	Lakonishok, Shleifer & Vishny	1994	JF
Sales to cash	Compustat	Annual	49.249	135.423	Ou & Penman	1989	JAE
Sales to inventory	Compustat	Annual	22.766	57.214	Ou & Penman	1989	JAE
Sales to price	Compustat	Annual	1.602	2.383	Barbee, Mukherji, & Raines	1996	FAJ
Sales to receivables	Compustat	Annual	11.823	23.007	Ou & Penman	1989	JAE
Secured debt indicator	Compustat	Annual	0.368	0.482	Valta	2016	JFQA
Share turnover	CRSP	Monthly	1.359	1.511	Datar, Naik & Radcliffe	1998	JFM
Sin stocks	Compustat	Annual	0.013	0.115	Hong & Kacperczyk	2009	JFE
Size	CRSP	Monthly	14.742	1.735	Banz	1981	JFE
Tax income to book income	Compustat	Annual	0.064	1.632	Lev & Nissim	2004	TAR
Volatility of liquidity (dollar trading vol.)	CRSP	Monthly	0.542	0.250	Chordia, Subrahmanyam & Anshuman	2001	JFE
Volatility of liquidity (share turnover)	CRSP	Monthly	3.622	5.271	Chordia, Subrahmanyam, &Anshuman	2001	JFE
Working capital accruals	Compustat	Annual	-0.021	0.085	Sloan	1996	TAR
Zero trading days	CRSP	Monthly	0.041	0.477	Liu	2006	JFE

Panel A2: Macro features	Source	Freq.	Avg.	Std.	Author(s)	Year	Journal
Book-to-market ratio	Goyal	Monthly	0.421	0.262	Goyal & Welch	2008	RFS
Default yield spread	Goyal	Monthly	0.011	0.004	Goyal & Welch	2008	RFS
Dividend price ratio	Goyal	Monthly	-3.713	0.424	Goyal & Welch	2008	RFS

Table 1 continued . . .

Table 1 continued . . .

. . .	Source	Freq.	Avg.	Std.	Author(s)	Year	Journal
Earnings price ratio	Goyal	Monthly	-2.913	0.464	Goyal & Welch	2008	RFS
Net equity expansion	Goyal	Monthly	0.006	0.020	Goyal & Welch	2008	RFS
Stock variance	Goyal	Monthly	0.002	0.005	Goyal & Welch	2008	RFS
Term spread	Goyal	Monthly	0.022	0.014	Goyal & Welch	2008	RFS
Treasury bill rate	Goyal	Monthly	0.043	0.033	Goyal & Welch	2008	RFS

Panel B: Excess returns	Source	Freq.	Avg.	Std.	Author(s)	Year	Journal
Excess return 1 month	CRSP/OptionM./Goyal	Daily	0.010	0.110			
Excess return 1 year	CRSP/OptionM./Goyal	Daily	0.150	0.494			

Panel C: Theory-based measures	Source	Freq.	Avg.	Std.	Author(s)	Year	Journal
Martin/Wagner 1 month	Compu./CRSP/OptionM.	Daily	0.006	0.010	Martin & Wagner	2019	JF
Kadan/Tang 1 month	Compu./CRSP/OptionM.	Daily	0.015	0.019	Kadan & Tang	2019	forth. RFS
Martin 1 month	Compu./CRSP/OptionM.	Daily	0.003	0.003	Martin	2017	QJE
Martin/Wagner 1 year	Compu./CRSP/OptionM.	Daily	0.071	0.112	Martin & Wagner	2019	JF
Kadan/Tang 1 year	Compu./CRSP/OptionM.	Daily	0.159	0.230	Kadan & Tang	2019	forth. RFS
Martin 1 year	Compu./CRSP/OptionM.	Daily	0.040	0.022	Martin	2017	QJE

Table 2: Hyper-parameter combinations. Panels A to D present the hyper-parameter combinations that we search during our training/validation procedure. Any parameter configurations that are not listed here correspond to the respective default settings of the software packages used.

Panel A: Elastic Net	**Panel B: Random Forest**
Software: Scikit-learn (SGDRegressor)	*Software:* Scikit-learn (RandomForestRegressor)
Feature transformation: Standard & robust scaling Selection by variance threshold	*Feature transformation:* Standard & robust scaling Selection by variance threshold
Model parameters: L1-L2-penalty: $\{x \in \mathbb{R} : 10^{-5} \leq x \leq 10^{-1}\}$ L1-ratio: $\{x \in \mathbb{R} : 0 \leq x \leq 1\}$	*Model parameters:* Number of trees: 300 Max. depth: $\{x \in \mathbb{N} : 2 \leq x \leq 30\}$ Max. features: $\{x \in \mathbb{N} : 2 \leq x \leq 150\}$
Optimization: Stochastic gradient descent Tolerance: 10^{-4} Max. epochs: $1,000$ Learning rate: $10^{-4}/t^{0.1}$	
Random search: Number of combinations: $1,000$	*Random search:* Number of combinations: 500

Panel C: Boosted Trees	**Panel D: Neural Network**
Software: Scikit-learn (GradientBoostingRegressor)	*Software:* Keras/Tensorflow (Sequential)
Feature transformation: Standard & robust scaling Selection by variance threshold	*Feature transformation:* Standard & robust scaling Selection by variance threshold
Model parameters: Number of trees: $\{x \in \mathbb{N} : 2 \leq x \leq 100\}$ Max. depth: $\{x \in \mathbb{N} : 1 \leq x \leq 3\}$ Max. features: $\{20, 50, \text{All}\}$ Learning rate: $\{x \in \mathbb{R} : 5 \times 10^{-3} \leq x \leq 1.2 \times 10^{-1}\}$	*Model parameters:* Activation: TanH, ReLU Hidden layers: $\{1, 2, 3, 4, 5\}$ First hidden layer nodes: $\{32, 64, 128, 256\}$ Network architecture: Rectangle, Pyramid Max. weight norm: $\{3, 4, 5\}$ Dropout rate: $\{x \in \mathbb{R} : 0 \leq x \leq 0.5\}$ L1-penalty: $\{x \in \mathbb{R} : 10^{-5} \leq x \leq 10^{-2}\}$
	Optimization: Adaptive moment estimation Batch size: $\{10, 20, 50, 100, 200, 500, 1,000\}$ Learning rate: $\{x \in \mathbb{R} : 0.01 \leq x \leq 0.1\}$ Early stopping Epochs w/o change: $\{2, 4, 6, 8, 10, 12\}$ Max. epochs: 150 Batch normalization
Random search: Number of combinations: 300	*Random search:* Number of combinations: 500

Table 3: Performance comparison: Theory-based vs. machine learning forecasts. The table presents information on the out-of-sample forecast performance for variants of the theory-based forecasts (Martin and Wagner (2019) and Kadan and Tang (2019)) and the machine learning models with the long training/validation scheme. For each competitor, we report the associated $R^2_{oos}\times100$ and the standard deviations of $R^2_{oos,s}$, calculated on annual splits. The last two columns contain the p-values associated with testing the null hypothesis that a model's forecast has no explanatory power over the zero forecast, i.e. $\mathbb{E}(R^2_{oos}) \leq 0$ and the p-values of the Diebold-Mariano test (using end-of-month theory-based forecast of Martin and Wagner (2019) as a basecase), respectively. Panel A refers to the one month forecast horizon and Panel B to the one year forecast horizon. For Panel A, the out-of-sample testing period starts in January 1996 and ends in November 2018. For Panel B, the out-of-sample testing period ends in December 2017. Both panels consider the theory-based forecasts on a monthly and on a daily frequency.

Panel A: One month horizon

				p-values	
		$R^2_{oos}\times100$	Std Dev	$\mathbb{E}(R^2_{oos}) \leq 0$	DM
Theory-Based	Martin/Wagner	0.2	3.2	0.154	
	Martin/Wagner (daily)	0.9	2.3	0.008	0.612
	Kadan/Tang	−1.8	6.9	0.704	0.089
	Kadan/Tang (daily)	−0.5	5.3	0.502	0.115
Machine Learning	Elastic Net	−0.3	3.5	0.161	0.479
	Neural Network	−68.7	121.9	1.000	0.022
	Boosted Trees	−0.6	4.2	0.248	0.353
	Random Forest	−1.6	5.2	0.435	0.301

Panel B: One year horizon

				p-values	
		$R^2_{oos}\times100$	Std Dev	$\mathbb{E}(R^2_{oos}) \leq 0$	DM
Theory-Based	Martin/Wagner	8.8	16.3	0.051	
	Martin/Wagner (daily)	9.0	16.2	0.046	0.132
	Kadan/Tang	3.1	47.6	0.694	0.295
	Kadan/Tang (daily)	3.5	48.2	0.677	0.217
Machine Learning	Elastic Net	5.5	18.5	0.125	0.259
	Neural Network	5.6	28.6	0.512	0.324
	Boosted Trees	10.6	20.5	0.035	0.195
	Random Forest	19.5	23.6	0.002	0.003

Table 4: Forecast correlations. The table reports pearson correlation coefficients for the out-of-sample forecasts of the theory-based approaches (Martin and Wagner (2019) and Kadan and Tang (2019)) and the machine learning models with the long training/validation scheme. Panel A refers to a forecast horizon of one month with a testing period from January 1996 to November 2018 and Panel B refers to a forecast horizon of one year and a testing period from January 1996 to December 2017.

Panel A: One month horizon

	Neural Network	Random Forest	Boosted Trees	Elastic Net	Kadan/Tang
Martin/Wagner	0.01	0.25	0.32	−0.06	0.98
Kadan/Tang	0.00	0.25	0.31	−0.04	
Elastic Net	0.01	0.70	0.45		
Boosted Trees	0.07	0.82			
Random Forest	0.09				

Panel B: One year horizon

	Neural Network	Random Forest	Boosted Trees	Elastic Net	Kadan/Tang
Martin/Wagner	0.12	0.33	0.34	0.00	0.98
Kadan/Tang	0.08	0.32	0.35	0.02	
Elastic Net	−0.02	0.49	0.57		
Boosted Trees	0.06	0.72			
Random Forest	0.15				

Table 5: Performance comparison: Pure vs. hybrid forecasts (one month horizon). The table presents information on the out-of-sample forecast performance for variants of the theory-based forecasts (*Theory-based*), the machine learning models (*Machine Learning*), and a hybrid approach in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*). The general outline is as in Table 3. All results refer to a one month forecast horizon and use the period January 1998 to November 2018 for out-of-sample forecasting.

				p-values	
		$R^2_{oos}\times100$	Std Dev	$\mathbb{E}(R^2_{oos}) \leq 0$	DM
Theory-Based	Martin/Wagner	0.1	3.4	0.206	
	Martin/Wagner (daily)	0.9	2.4	0.016	0.756
	Kadan/Tang	−2.0	7.2	0.739	0.086
	Kadan/Tang (daily)	−0.6	5.6	0.575	0.160
Machine Learning	Elastic Net	−4.0	8.6	0.840	0.130
	Neural Network	−43.1	57.3	0.996	0.087
	Boosted Trees	−29.5	57.7	0.860	0.245
	Random Forest	−8.4	15.1	0.869	0.173
ML with theory features	Elastic Net	−3.2	7.1	0.790	0.130
	Neural Network	−29.5	38.3	1.000	0.002
	Boosted Trees	−25.6	53.1	0.855	0.253
	Random Forest	−7.6	13.3	0.871	0.157

Table 6: Performance comparison: Pure vs. hybrid forecasts (one year horizon). The table presents information on the out-of-sample forecast performance for variants of the theory-based forecasts (*Theory-based*), the machine learning models (*Machine Learning*), a hybrid approach in which the theory-consistent forecasts serve as additional features in the machine learning models (*ML with theory features*), and a second hybrid attempt in which ML models are trained to deliver predictions of the forecast errors of the theory-based approaches (*Theory assisted by ML*). The general outline is as in Table 3. All results refer to a one year forecast horizon and use the period January 1998 to December 2017 for out-of-sample forecasting.

				p-values	
		R^2_{oos}×100	Std Dev	$\mathbb{E}(R^2_{oos}) \leq 0$	DM
Theory-Based	Martin/Wagner	9.1	17.1	0.072	
	Martin/Wagner (daily)	9.3	17.0	0.066	0.158
	Kadan/Tang	3.1	49.9	0.706	0.315
	Kadan/Tang (daily)	3.5	50.5	0.692	0.231
Machine Learning	Elastic Net	−31.6	153.6	0.873	0.131
	Neural Network	−346.2	398.1	0.859	0.263
	Boosted Trees	10.3	36.6	0.308	0.849
	Random Forest	12.4	45.1	0.329	0.645
ML with theory features	Elastic Net	−32.6	160.3	0.868	0.139
	Neural Network	−2.4	22.6	0.604	0.117
	Boosted Trees	9.7	39.7	0.356	0.973
	Random Forest	14.6	42.3	0.244	0.387
Theory assisted by ML	Elastic Net	−38.2	192.9	0.885	0.168
	Neural Network	−1.9	42.6	0.633	0.308
	Boosted Trees	9.2	45.2	0.440	0.955
	Random Forest	16.1	50.6	0.259	0.367

Figure 1: Identification of S&P 500 constituents. The figure illustrates the ability to detect historical S&P 500 constituents using our identification strategy. Panel A presents the coverages of S&P 500 constituents that we achieve at different stages of our data preprocessing. The line in light grey refers to the historical S&P 500 constituents in Compustat. The blue line shows for how many of these constituents we can find stock price information on CRSP when combining the permanent stock identifiers of Compustat and CRSP. The red line starting in 1996 illustrates those constituents for which we can furthermore find information on OptionMetrics and thus are able to compute theory-based forecasts. Panel B visualizes the aggregate market capitalizations for each of these three groups.

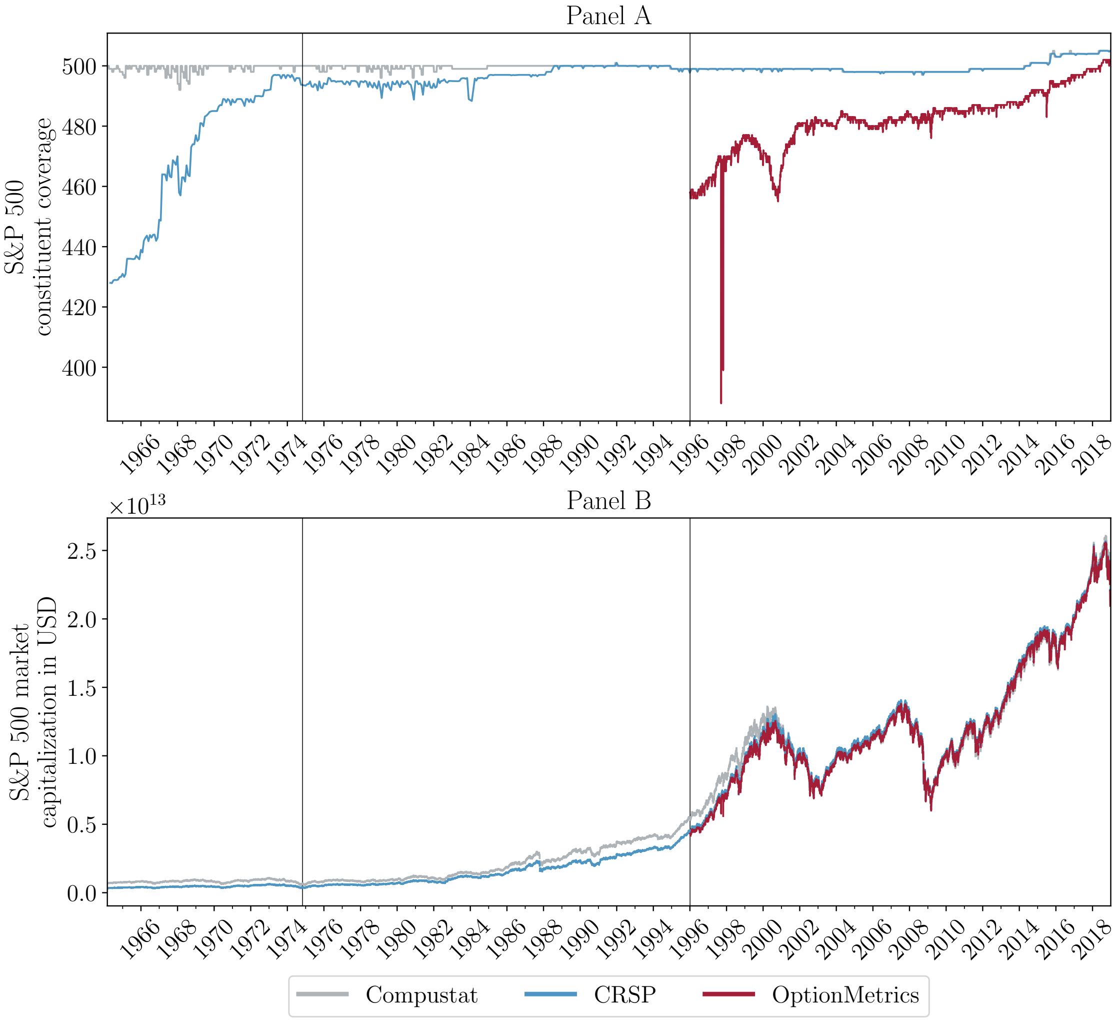


Figure 2: Comparison of risk-free rate proxies. This figure presents a comparison between the two annual risk-free rate proxies that we use in our study. The zero curve obtained from OptionMetrics (dashed blue) is available from January 1996 until December 2018. The Treasury-bill rate (solid red) is taken from Amit Goyal's webpage and is available from March 1964 until December 2018.

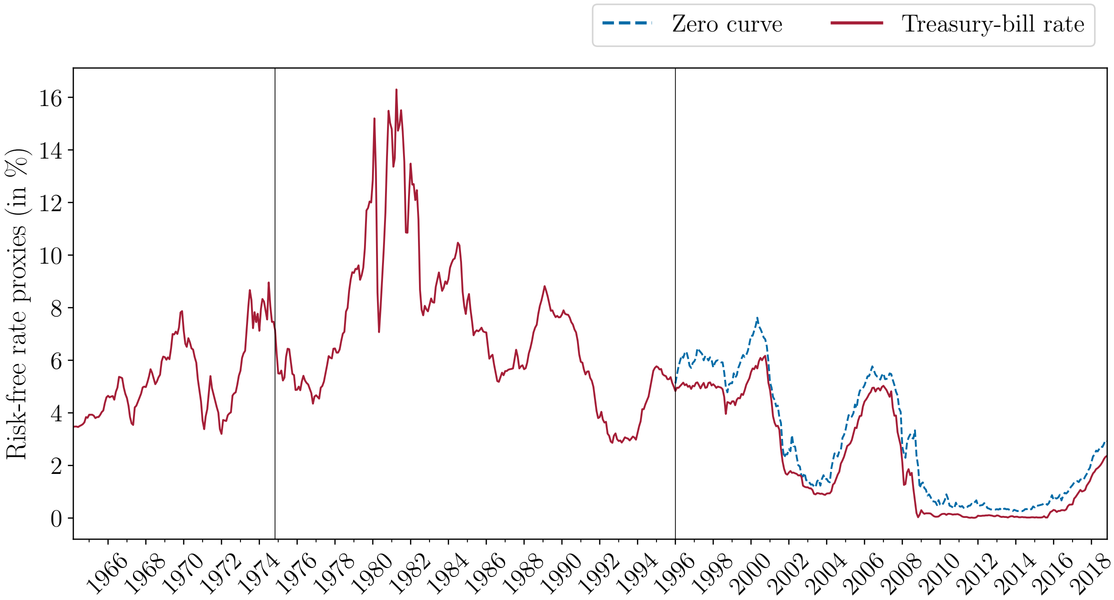


Figure 3: Long training/validation scheme with a forecast horizon of one year. The data range from October 1974 to December 2017. The training period (red/dark grey) initially spans 10 years and increases by one year after each validation step. Each of the 22 validation steps delivers a new set of parameter estimates. Each validation window (gold/light grey) covers 12 years and is rolled forward with a fixed width. After each validation step, there is one year of out-of-sample testing (checkered blue/grey).

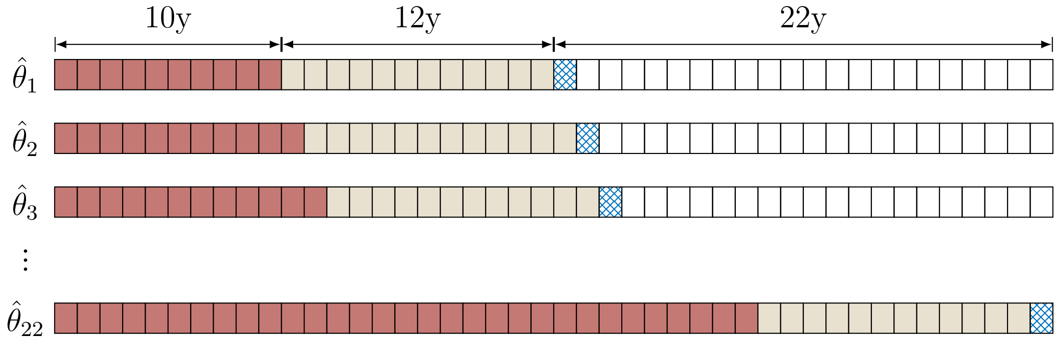


Figure 4: Short training/validation scheme with a forecast horizon of one year. The data range from January 1996 to December 2017. The training period (red/dark grey) initially spans 1 year and increases by one year after each validation step. Each of the 20 validation steps delivers a new set of parameter estimates. Each validation window (gold/light grey) covers 1 year. After each validation step, there is one year of out-of-sample testing (checkered blue/grey).

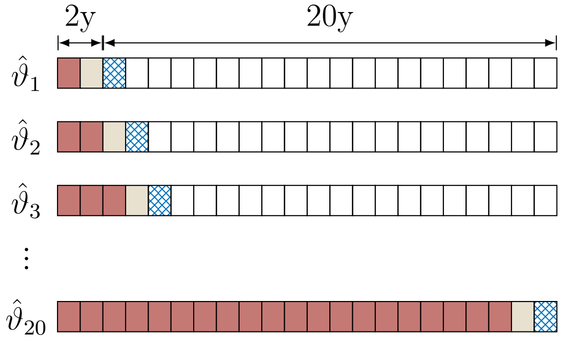


Figure 5: Decile portfolios (one month horizon): Theory-based vs. machine learning forecasts. Panels A to D show average realized excess returns of stocks which are, each month, sorted into deciles based on the models' predictions. The models being considered are the theory-based approach of Martin and Wagner (2019) and its machine learning competitors based on the long training/validation scheme described in Figure 3. Panels A and B cover the theory-based approach with a monthly (A) and a daily (B) frequency. Panels C and D contain the best (C) and the worst (D) machine learning competitor. The forecast horizon is one month and the sample period ranges from January 1996 to November 2018 for all panels. Each of the symbols represents a combination of average predicted decile excess returns and average realized decile excess returns (in %). The numbers that are associated with the symbols indicate the rank of the prediction deciles. We also include a 45-degree line for reference.

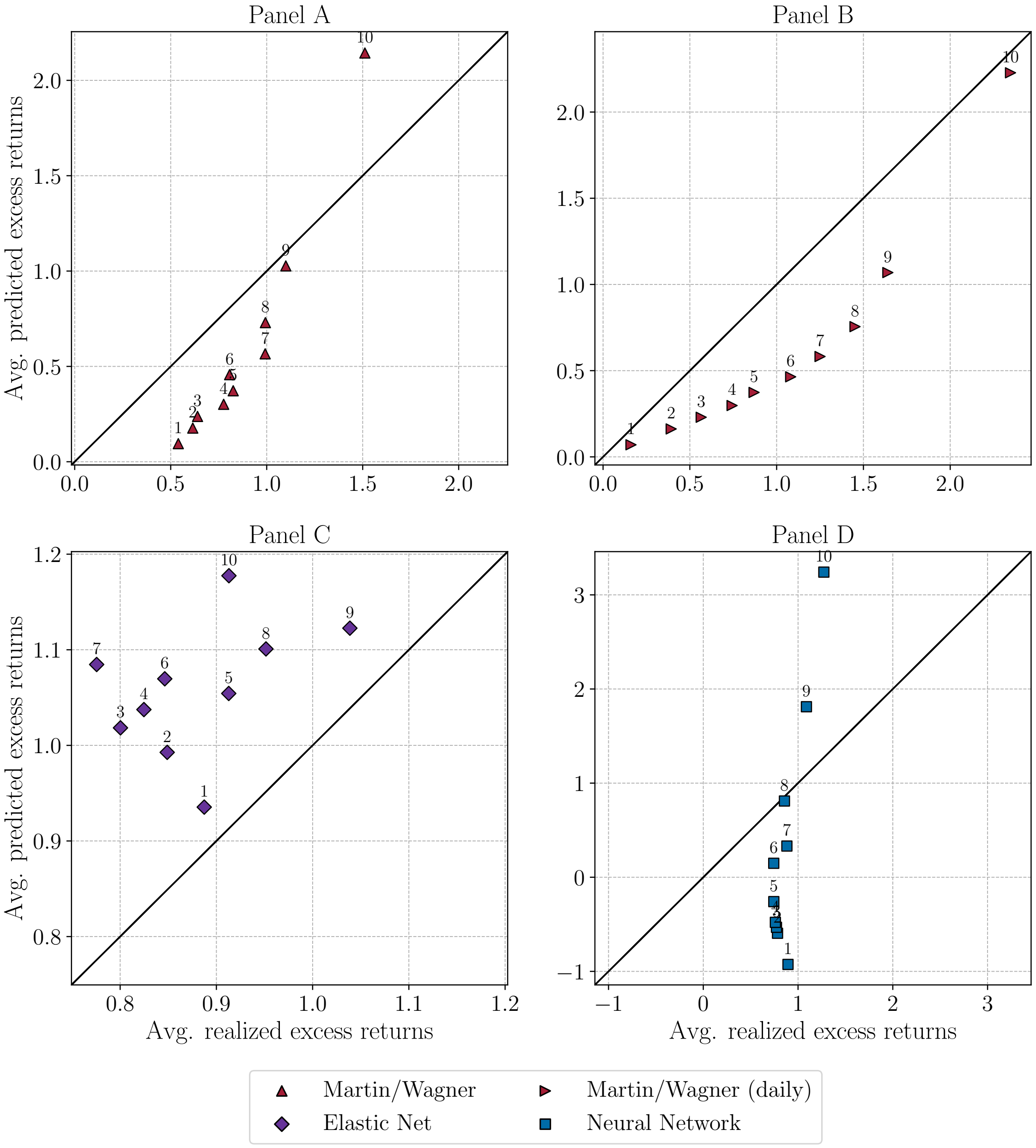


Figure 6: Decile portfolios (one year horizon): Theory-based vs. machine learning forecasts. Panels A to D show average realized excess returns of stocks which are, each month, sorted into deciles based on the models' predictions. The models being considered are the theory-based approach of Martin and Wagner (2019) and its machine learning competitors based on the long training/validation scheme described in Figure 3. Panels A and B cover the theory-based approach with a monthly (A) and a daily (B) frequency. Panels C and D contain the best (C) and the worst (D) machine learning competitor. The forecast horizon is one year and the sample period ranges from January 1996 to December 2017 for all panels. Each of the symbols represents a combination of average predicted decile excess returns and average realized decile excess returns (in %). The numbers that are associated with the symbols indicate the rank of the prediction deciles. We also include a 45-degree line for reference.

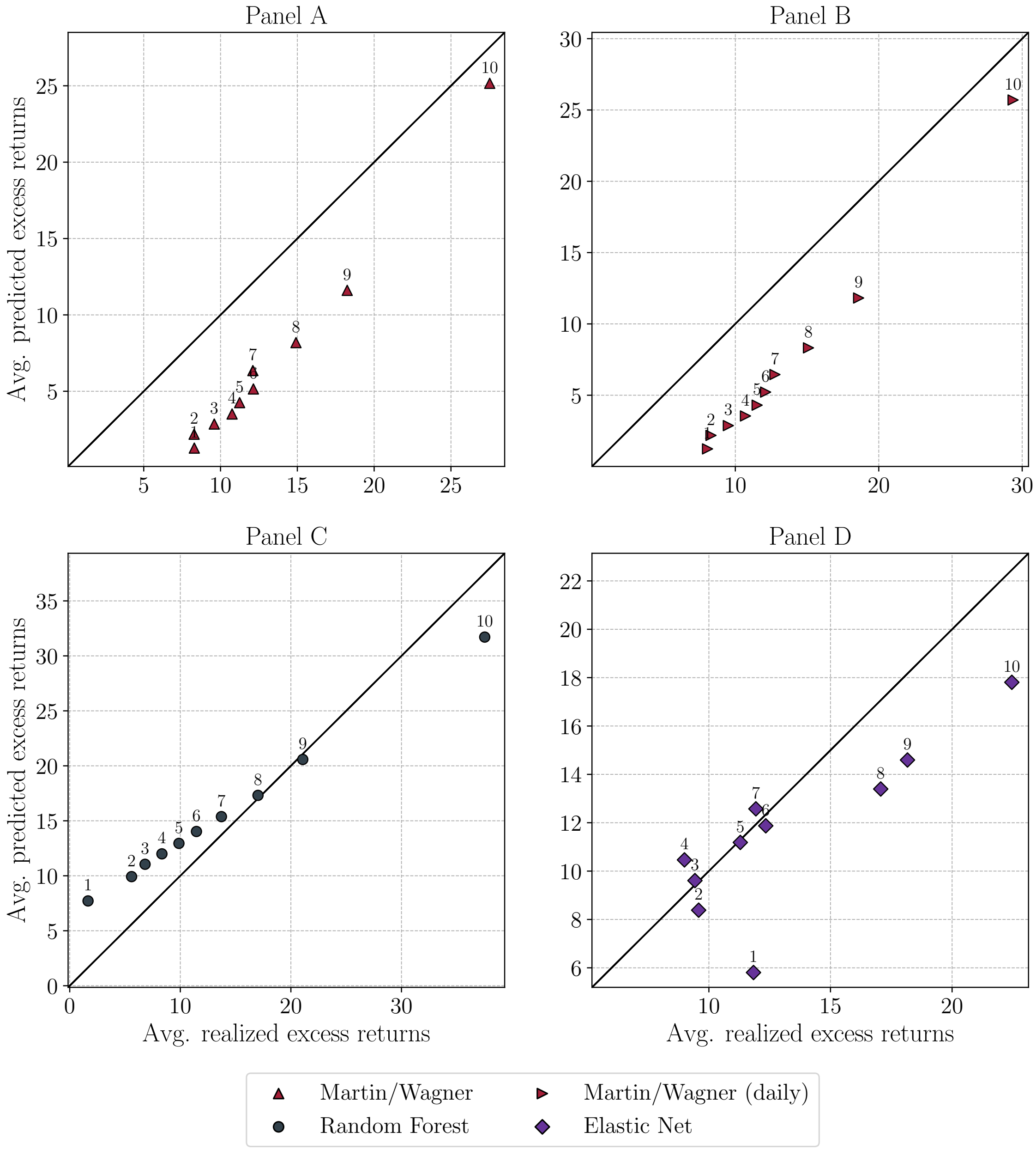


Figure 7: Performance comparison (one month horizon): Theory-based vs. machine learning forecasts. The figure depicts the one month out-of-sample forecast performance of competing theory-based and machine learning approaches as a time series of $R^2_{oos,s}$ on annual splits. The out-of-sample period ranges from January 1996 to November 2018. Each of the three panels contains the theory-based forecast as proposed in Martin and Wagner (2019) (upward pointing triangle/red) along with a set of competitor models. These competitor models include both the alternative theory-based forecast by Kadan and Tang (2019) (downward pointing triangle/yellow) and machine learning models similar to those used in Gu et al. (2019b). In Panel A, we present the best among the machine learning competitors, whereas in Panel B, we contrast the forecast by Martin and Wagner (2019) to the weakest machine learning competitor. In Panel C we collect the $R^2_{oos,s}$ for all remaining models. All machine learning results are obtained using the validation scheme outlined in Figure 3.

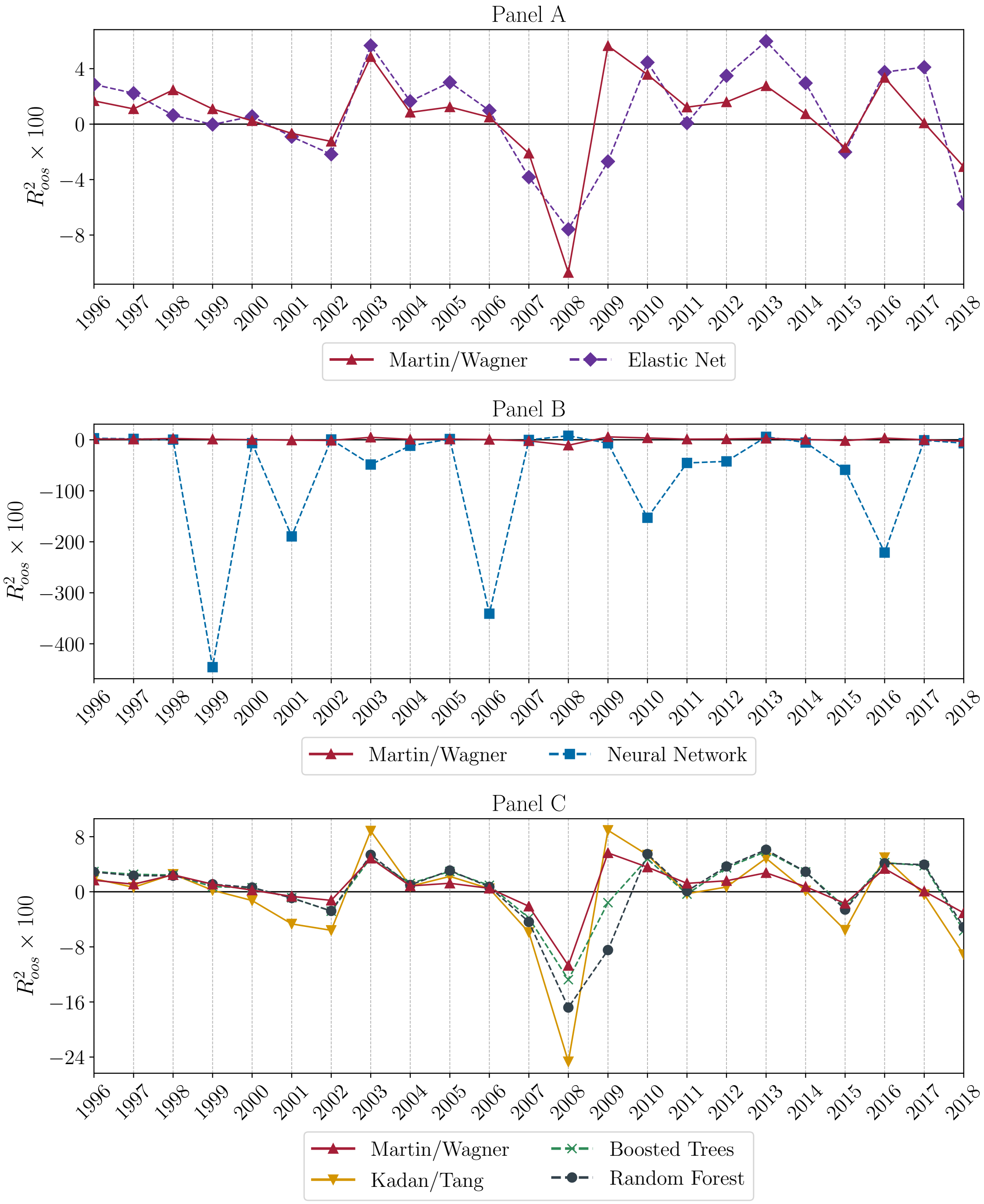


Figure 8: Performance comparison (one year horizon): Theory-based vs. machine learning forecasts. The figure depicts the one year out-of-sample forecast performance of competing theory-based and machine learning approaches as a time series of $R^2_{oos,s}$ on annual splits. The out-of-sample period ranges from January 1996 to December 2017. Each of the three panels contains the theory-based forecast as proposed in Martin and Wagner (2019) (upward pointing triangle/red) along with a set of competitor models. These competitor models include both the alternative theory-based forecast by Kadan and Tang (2019) (downward pointing triangle/yellow) and machine learning models similar to those used in Gu et al. (2019b). In Panel A, we contrast the forecast by Martin and Wagner (2019) with the best among the machine learning competitors, whereas in Panel B, we choose the weakest machine learning competitor. In Panel C, we collect the $R^2_{oos,s}$ for all remaining models. All machine learning results are obtained using the validation scheme outlined in Figure 3.

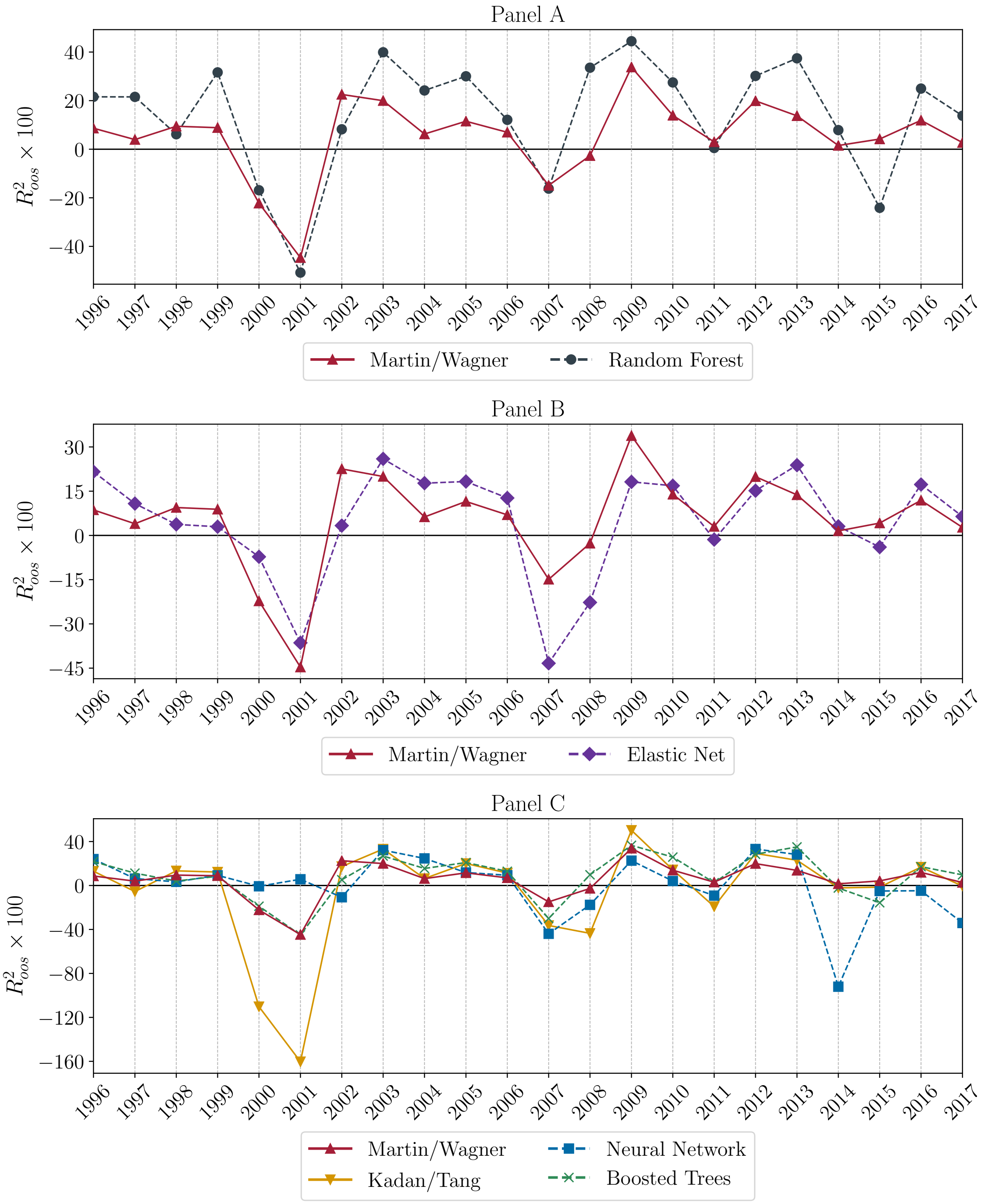


Figure 9: Performance comparison (one year horizon): Theory-based vs. augmented machine learning forecasts. The figure depicts the one year out-of-sample forecast performance of competing theory-based approaches and machine learning models that include the theory-consistent forecasts in their feature set. That performance is visualized as a time series of $R^2_{oos,s}$ on annual splits. The out-of-sample period ranges from January 1998 to December 2017. Each of the three panels contains the theory-based forecast as proposed in Martin and Wagner (2019) (upward pointing triangle/red) along with a set of competitor models. These competitor models include both the alternative theory-based forecast by Kadan and Tang (2019) (downward pointing triangle/yellow) and augmented machine learning models. In Panel A, we contrast the forecast by Martin and Wagner (2019) with the best among the augmented machine learning competitors, whereas in Panel B, we choose the weakest augmented machine learning competitor. In Panel C, we collect the $R^2_{oos,s}$ for all remaining models. All machine learning results are obtained using the validation scheme outlined in Figure 4.

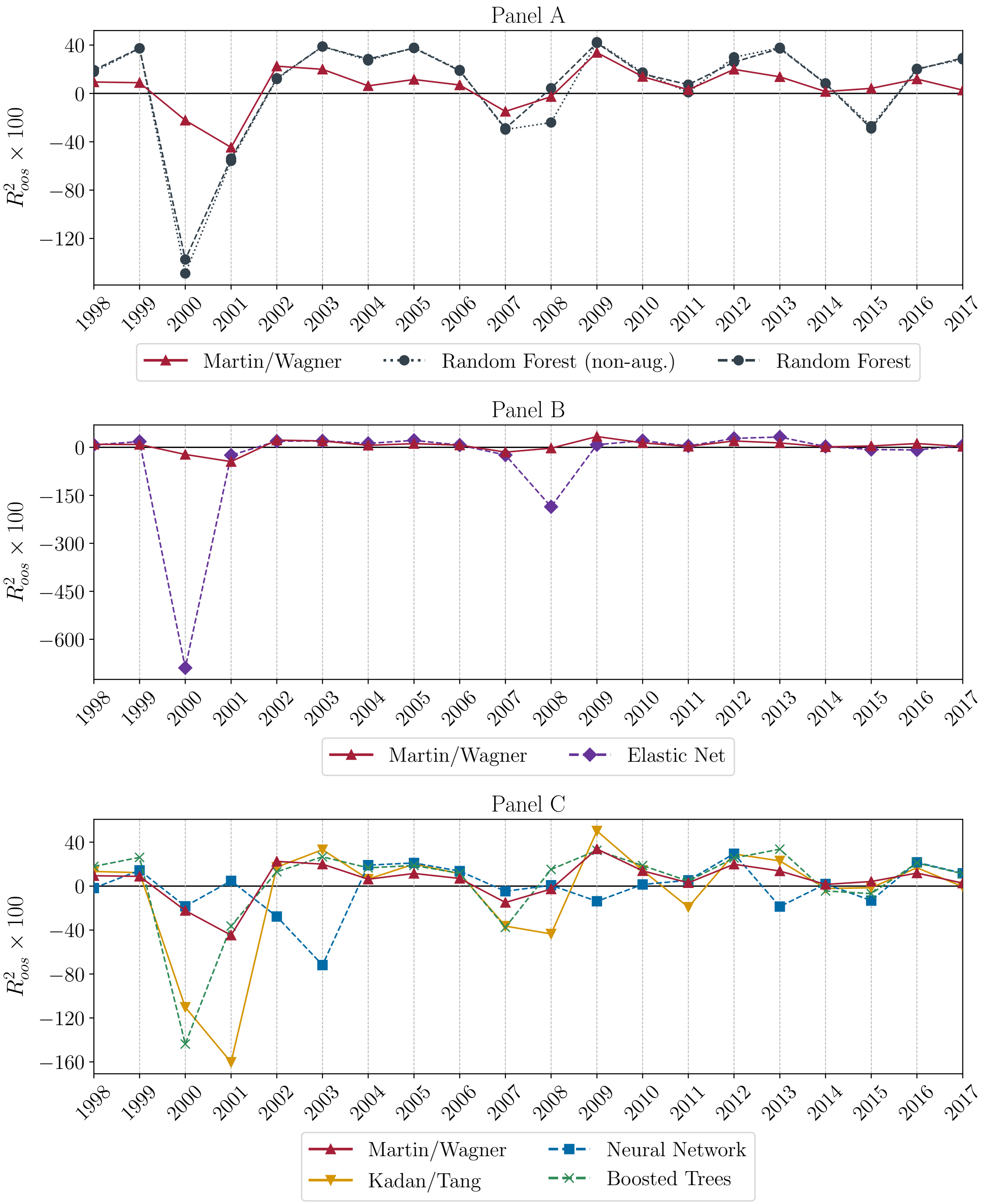


Figure 10: Performance comparison (one year horizon): Theory-based vs. machine learning assisted forecasts. This figure shows time series of annual out-of-sample $R^2_{oos,s}$ for the theory-based forecast proposed in Martin and Wagner (2019) (upward-pointing triangle/light red), the random forest forecast that is based on the short training/validation scheme as described in Figure 4 (circle/light grey) and the machine learning assisted, theory-based approach (Martin/Wagner+random forest; hexagon/dark grey). The results are presented for a one year forecast horizon. The out-of-sample period ranges from January 1998 to December 2017.

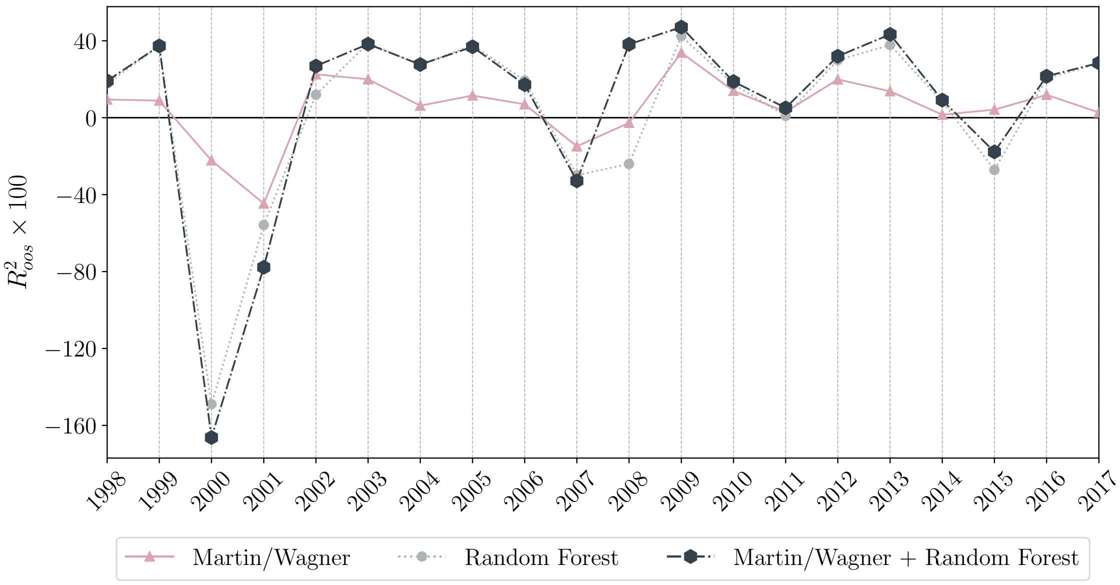


Figure 11: Performance comparison (one year horizon): Random forests vs. hybrid approach. This figure contrasts the out-of-sample performance of a random forest trained using the long validation scheme depicted in Figure 3 (circle/light grey) with that of a theory-based, random forests assisted forecast obtained from the short validation scheme illustrated in Figure 4 (hexagon/black). In both cases, the forecast horizon is one year and the forecast performance is measured by time series of $R^2_{oos,s}$ on annual splits

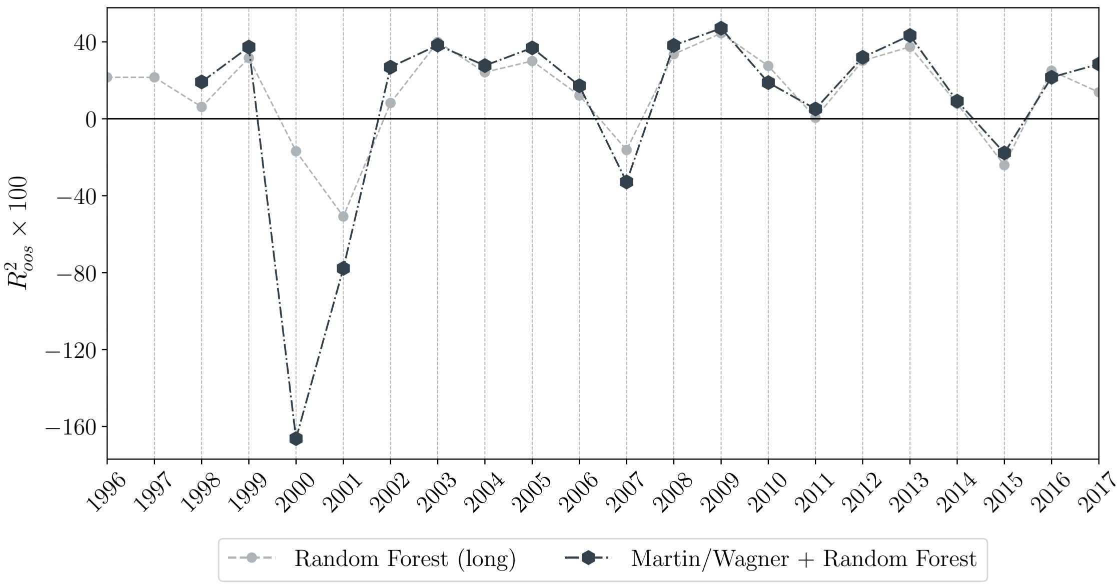
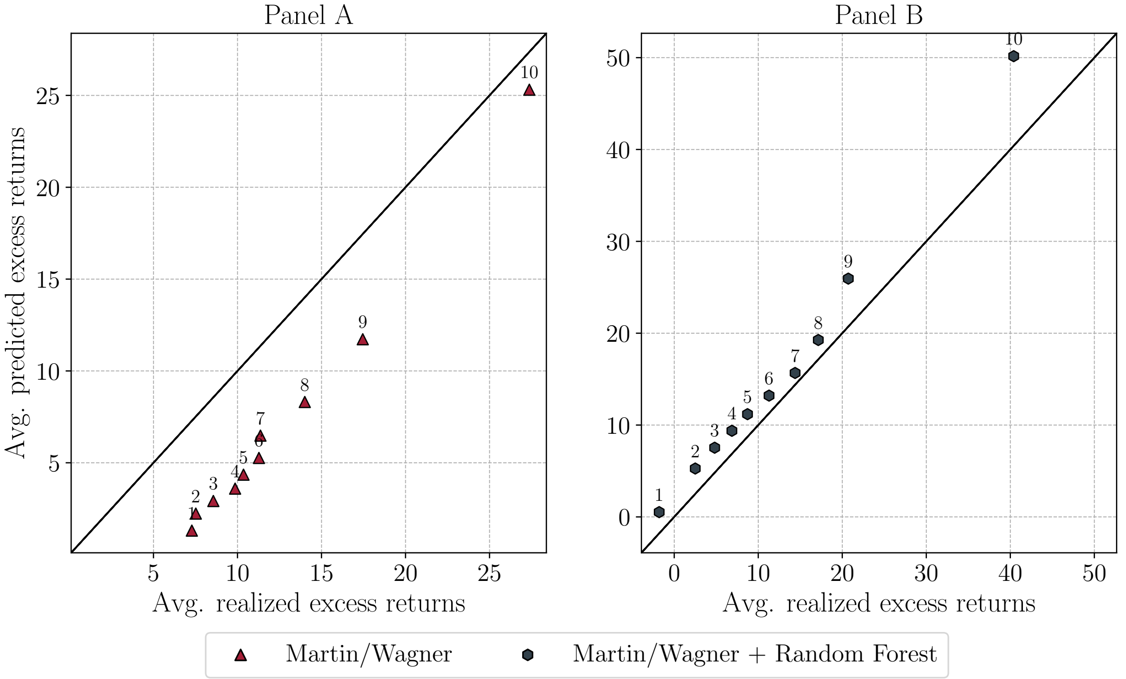


Figure 12: Decile portfolios (one year horizon): Theory-based vs. theory-assisted machine learning forecasts. Panels A and B show average realized excess returns of stocks which are, each month, sorted into deciles based on the models' predictions. The models being considered are the theory-consistent approach of Martin and Wagner (2019) and the theory-assisted random forest based on the short training/validation scheme described in Figure 4. The forecast horizon is one year and the sample period ranges from January 1998 to December 2017 for both panels. Each of the symbols represents a combination of average predicted decile excess returns and average realized decile excess returns (in %). We also include a 45 degree line for reference.


Panel A
Avg. predicted excess returns
Avg. realized excess returns
Panel B
Avg. realized excess returns
▲ Martin/Wagner
⬢ Martin/Wagner + Random Forest