University of Oxford
Department of Statistics
Master of Science in Statistical Science

The Mosaic of Predictability for Corporate Bonds

Twan Richard Mulder

Supervisors:	Mihai Cucuringu
	Maria Grith
	Stefan Zohren
Date final version:	15th September 2025

The Mosaic of Predictability for Corporate Bonds

Twan R. Mulder

Abstract

We study the heterogeneity in return predictability across corporate bonds using a tree-based clustering procedure that groups bonds based on cross-sectional asset characteristics and time-varying macroeconomic variables. First, we show that corporate bonds with both low distance-to-default *and* low prior-month returns are substantially less predictable, both in- and out-of-sample. Second, we find that the *predictability disagreement anomaly* documented in the equity market is not present in the corporate bond market. Third, we document that bond returns are less predictable during periods of high inflation.

Keywords: Machine learning, corporate bonds, clustering, return predictability

JEL classification: C38, C55, G12

Acknowledgements

I would like to thank Mihai Cucuringu, Maria Grith and Stefan Zohren, for their guidance and support.

Contents

List of Tables

List of Figures

1 Introduction

Corporate bonds represent a particularly challenging asset class for empirical research. Unlike equities, which trade through centralized limit order books, corporate bonds are primarily traded over the counter. Researchers typically rely on transaction-based data, which are affected by market microstructure noise such as bid-ask spread bias (Dickerson, Robotti & Rossetti, 2023).[1] Additionally, recent studies highlight that the absence of error-free datasets and the lack of a standardized data-cleaning framework have contributed to a replication and credibility crisis in the corporate bond literature (Dick-Nielsen, Feldhütter, Pedersen & Stolborg, 2023; Dickerson, Mueller & Robotti, 2023). These challenges likely explain why many empirical asset pricing studies test novel theories across a broad range of markets while largely excluding the corporate bond market (e.g., Asness, Moskowitz & Pedersen, 2013).[2] As a result, our understanding of the risk-return trade-off in the corporate bond market remains relatively underdeveloped compared to equities. The corporate bond market is, however, similar in size and economic importance to the equity market (Kelly, Palhares & Pruitt, 2023). In this paper, we contribute to the literature by being the first to investigate both the cross-sectional and time-series variation in corporate bond return predictability.

The predictability of asset returns has already been extensively studied in the finance literature.[3] Most of these studies propose novel predictors or models that are claimed to advance return predictability significantly. However, they typically assume that all assets are equally predictable or that predictability is constant over time. Instead, in this paper, we regard return predictability as an unobservable asset characteristic—such as return volatility, which varies over time and across individual assets. Our objective is to group a panel of monthly bond returns into clusters with heterogeneous levels of predictability, using cross-sectional asset characteristics and time-varying macroeconomic variables. In this way, we can determine which corporate bonds—as identified by cross-sectional characteristics—are more predictable and under what regimes—as identified by macroeconomic variables.

In practice, predictability is typically measured by the R^2 statistic. If return predictability is indeed an asset characteristic, then low-predictable assets (with low signal-to-noise ratios) have low R^2 estimates even when the true conditional expectation is used as return forecast (Cong, Feng, He & Wang, 2024). Thus, the R^2 statistic is an informative

[1] Although researchers could use quote-based data (e.g., from Intercontinental Exchange)—which are also used by hedge funds, pension funds, and other market participants—this data is typically expensive.

[2] Asness et al. (2013) study value and momentum across eight different markets and asset classes: commodity futures, currencies, equity index futures, government bonds and individual stocks in Europe, Japan, the UK, and the US. However, they do not consider corporate bonds.

[3] Return predictability has mainly been studied in the equity market (e.g., Campbell & Thompson, 2008; Ang & Bekaert, 2007; Rapach, Strauss & Zhou, 2013; Golez & Koudijs, 2018), but recently also in the corporate bond market (e.g., Lin, Wang & Wu, 2014; Lin, Wu & Zhou, 2018; Feng, He, Wang & Wu, 2025)

proxy for the level of return predictability of an asset. However, obtaining a sensible R^2 estimate for each corporate bond i and each month t separately (i.e., $R^2_{i,t}$) is infeasible, given that we only have a single observation for each bond at each month. This implies that a conventional two-step approach—where first the level of return predictability for each asset-return observation is determined, and then a clustering algorithm is applied—does not work in practice. Therefore, we use an algorithm that *endogenously* clusters bond-return observations with similar levels of predictability.

More specifically, we extend the clustering framework of Cong et al. (2024) to the corporate bond market. This tree-based clustering procedure groups assets by their degree of return predictability using time-varying macroeconomic variables and cross-sectional asset characteristics as potential splitting variables. These so-called *panel trees*—as introduced by Cong, Feng, He and He (2025)—are grown by iteratively splitting one of the clusters (i.e., leaf nodes) into two novel clusters according to the optimal split rule that maximizes the *global* heterogeneity in *clusterwise* return predictability, starting with all observations grouped in a single cluster (i.e., root node). Each cluster is associated with a separate predictive model, which is fitted to only those observations assigned to the same cluster and used to obtain a clusterwise measure of return predictability. This means that all observations in the same cluster are assigned the same level of predictability.

The replication and credibility crisis in the corporate bond literature is largely attributed to the lack of reliable datasets and to incorrect data preprocessing procedures (Dick-Nielsen et al., 2023; Dickerson, Mueller & Robotti, 2023; Dickerson, Robotti & Rossetti, 2023). For this reason, we apply the data-cleaning procedures proposed by the same literature in order to obtain robust results. More specifically, we use monthly corporate bond returns from the Wharton Research Data Services (WRDS) Bond Database and apply the filters of Dick-Nielsen et al. (2023), among others. Furthermore, we consider a panel dataset of 53 bond and equity characteristics constructed by Dickerson, Robotti and Nozawa (2024), which serve both as inputs for our return-predictive models and as cross-sectional splitting variables. These variables include popular characteristics shown to predict bond returns by prior studies (e.g., Feng, He, Wang & Wu, 2025; Kelly et al., 2023). Additionally, we consider eight macroeconomic indicators measuring the status of the equity market (e.g., stock issuance), the corporate bond market (e.g., default spread) and the US economy (e.g., inflation), which are used to construct *regime-switching* panel trees.

Our main finding is that return predictability varies significantly across corporate bonds. In particular, we show that corporate bonds with both low previous-month returns and low distance-to-default are substantially less predictable, both in- and out-of-sample. We provide a liquidity-based interpretation to this finding. More specifically, corporate bonds with poor past performance and high default risk tend to be illiquid due to dealer balance-sheet constraints and trading frictions, which implies a weak relationship between

cross-sectional asset characteristics and next month's return. Besides that, we further find that corporate bonds with different credit ratings have, nevertheless, relatively similar levels of return predictability over time. Interestingly, most studies evaluate their novel predictors or models on rating-sorted groups of corporate bonds, in order to show that they forecast the returns of both low- and high-rated bonds (e.g., Feng, He, Wang & Wu, 2025; Lin et al., 2018). Our findings, however, question the effectiveness of this approach. Instead, we suggest that future research should evaluate their predictors and models on groups with genuine heterogeneous levels of predictability.

Moreover, we find that the *predictability disagreement anomaly* documented in the US equity market by Cong et al. (2024) is not present in the corporate bond market. More specifically, Cong et al. (2024) construct long-short portfolios that go long the assets with the largest disagreement between clusterwise and global models and short the assets with the smallest disagreement. They find that these portfolios earn statistically significant abnormal returns after accounting for common risk factors. We show that similar portfolios also earn economically significant average excess returns in the corporate bond market—with annualized Sharpe ratios more than 3.5 times higher than the value-weighted bond market portfolio, and even larger than the value-weighted equity market portfolio. However, the abnormal returns (i.e., alphas) are typically not statistically significant after accounting for common bond and equity risk factors. This finding is consistent with recent evidence that return-based anomalies—such as momentum and short- and long-term reversal—do not yield significant abnormal returns in the corporate bond market. More specifically, while past literature claimed that these anomalies, which were first documented in the equity market, were also present in the corporate bond market,[4] Dickerson, Robotti and Rossetti (2023) show that these findings are completely attributable to the use transaction-based TRACE data, which are subject to market microstructure noise. This paper contributes to this literature, in the sense that we show that the *return-predictability disagreement* anomaly is not present in the corporate bond market.

Finally, we document that corporate bond returns are less predictable when inflation is high. In particular, when inflation exceeds, or is close to exceeding, its 2% target, return predictability declines substantially for most corporate bonds. Following the literature, we attribute this to heightened uncertainty about future monetary policy. Investors expect central banks to raise interest rates in response to higher inflation, but remain uncertain about the magnitude of these interventions.

The remainder of this paper is structured as follows. In Section 2, we compare our work with related literature. In Section 3, we introduce the tree-based clustering procedure that groups assets by their level of return predictability, as measured by R^2. Furthermore,

[4]See, for instance, Bali, Subrahmanyam and Wen (2017); Jostova, Nikolova, Philipov and Stahel (2013) and Bali, Subrahmanyam and Wen (2021).

in Section 4, we describe the data-collection procedure for corporate bond returns, asset characteristics and macroeconomic predictors. Finally, we present our results in Section 5, and Section 6 concludes.

2 Literature review

This paper contributes to the literature on return predictability. Early studies in this area investigated whether aggregate stock returns are predictable, particularly using valuation ratios (Campbell & Thompson, 2008). These include the dividend yield (Rozeff, 1984; Fama & French, 1988), the price-earnings ratio (Basu, 1977), and the term structure of interest rates (Campbell, 1987). Subsequent research, however, questioned the evidence of stock return predictability documented by these studies. First, predictive regressions were found to have limited out-of-sample forecasting power (Goyal & Welch, 2003; Welch & Goyal, 2008). Second, the high persistence of commonly used predictors was shown to bias coefficient estimates and generate spurious evidence of predictability (Nelson & Kim, 1993; Stambaugh, 1999). Third, concerns were raised about data mining, the use of large predictor sets, and the selective reporting of statistically significant results (Foster, Smith & Whaley, 1997; Ferson, Sarkissian & Simin, 2003). Nevertheless, the recent adoption of machine learning (ML) methods in finance has helped address some of these issues. In particular, ML models have been shown to deliver robust forecasting performance, to accommodate flexible model selection in the presence of many predictors, and to mitigate data-mining concerns (Bali, Goyal, Huang, Jiang & Wen, 2020; Feng, Giglio & Xiu, 2020; Gu, Kelly & Xiu, 2020; Giglio, Liao & Xiu, 2021; Feng, He, Wang & Wu, 2025). In this paper, we build on this literature by taking advantage of predictive ML models while also addressing their limitations, especially with respect to interpretability.

While the literature mentioned above is mainly concerned with the predictability of equity returns, there are also studies investigating predictability in the corporate bond market. For example, Fama and French (1989) document that dividend yields, default spreads, and term spreads contain predictive power for corporate bond returns. Furthermore, Greenwood and Hanson (2013) show that low issuer quality is associated with low future excess corporate bond returns. Lin et al. (2014) highlight the role of liquidity and forward rate factors in predicting bond returns, and conclude that corporate bond returns exhibit greater predictability than equity returns. Lin et al. (2018) construct an iterated combination model to forecast corporate bond returns that substantially outperforms the traditional predictive models mentioned above. More recently, Bao, Hou and Zhang (2023) introduce a global measure of systematic default risk and demonstrate its ability to predict future equity and corporate bond index returns, both in- and out-of-sample.

In contrast to the aforementioned literature, we study the *heterogeneity* in the predictability of corporate bond returns. Specifically, we treat return predictability as an

asset characteristic, extending the framework of Cong et al. (2024) to the corporate bond market. We investigate which bonds are more predictable and how this varies across market regimes by clustering bonds according to cross-sectional characteristics and macroeconomic variables. Cong et al. (2024) show that stocks with high earnings-to-price ratios, high earnings surprises, and low trading volumes are the most predictable. Avramov (2002) find that small-cap value stocks are more predictable than large-cap growth stocks. Additionally, Dangl and Halling (2012) show that return predictability is stronger during recessions than during expansions. While these studies are related to stocks, others have presented evidence of time-varying return predictability in fixed-income markets. For instance, Borup, Eriksen, Kjær and Thyrsgaard (2024) find that excess returns on US Treasury bonds are more predictable during periods of high economic activity and low uncertainty. In contrast, Feng, He, Wang and Wu (2025) show that corporate bonds are more predictable during periods of slow economic growth and high risk aversion. To our knowledge, we are the first to investigate which cross-sectional asset characteristics identify variations in corporate bond return predictability.

Moreover, we contribute to the ML literature by employing a tree-based clustering procedure to classify bonds into groups that differ in their degree of return predictability. Recent literature shows that ML models can successfully forecast corporate bond returns. For instance, Bali et al. (2020) find that ML models significantly improve the out-of-sample performance of bond and equity characteristics in forecasting corporate bond returns compared to ordinary least squares. Kelly et al. (2023) develop a conditional five-factor model for corporate bonds using instrumented principal component analysis (IPCA), which accommodates time-varying factor loadings by specifying them as functions of asset characteristics. Additionally, Feng, He, Wang and Wu (2025) apply ML models to predict the monthly excess returns of individual corporate bonds. They document that short-term bond reversal, downside risk, credit spread, and return skewness are among the most important characteristics enhancing predictive performance.

More specifically, we form clusters using the panel tree framework introduced by Cong et al. (2025)—a novel extension of conventional Classification and Regression Tree (CART) models. The key innovation of these so-called *P-Trees* is that they optimize a global objective function—evaluated over all observations in the tree—rather than a sequence of local objectives at each split. Besides that, as the name suggests, this framework is specifically designed to accommodate (unbalanced) panel data. In their asset pricing application, Cong et al. (2025) use P-Trees to construct optimal test assets and latent factors within a mean-variance efficiency framework. Specifically, asset-return observations are clustered into value-weighted portfolios through sequential tree splits, where the splitting criterion is chosen to maximize the Sharpe ratio of the mean-variance efficient (MVE) portfolio constructed from these individual *leaf portfolios*. However, panel trees have also been applied in other contexts. For instance, P-Trees have been used to construct

a dynamic Nelson-Siegel yield-curve model (Bie, Diebold, He & Li, 2024), to construct uncommon factor models for regime-switching and characteristics-based clusters of assets (Cong, Feng, He & Li, 2023), to investigate the time-varying influence of US macroeconomic regimes on risk factors in the currency market (Feng, He, Li, Sarno & Zhang, 2025), and to group equity returns with heterogeneous levels of predictability (Cong et al., 2024).

In addition, we contribute to the literature that documents a link between portfolio performance and return predictability. Lo and MacKinlay (1997) introduce maximally predictable portfolios (MPPs), in which portfolio weights are chosen to maximize the R^2 of the resulting portfolio returns. However, a limitation of this framework is that the predictive model for portfolio returns is often specified as a linear function of a handful of regressors for computational reasons—thus, the resulting portfolios are maximally *linearly*-predictable portfolios. Goulet Coulombe and Göbel (2023) address this limitation by introducing MACE, a multivariate extension of the Alternating Conditional Expectations (ACE) algorithm of Breiman and Friedman (1985). Analogous to canonical correlation analysis, MACE tries to maximize the association between actual portfolio returns and their forecasts produced by a random forest (RF), by iteratively updating the portfolio weights and return forecasts while keeping the other fixed. The resulting maximally machine-learnable portfolios (MMLPs) exhibit strong out-of-sample predictability and good investment performance even under unfavorable market conditions. In this paper, we construct separate portfolios for each cluster, and show that those for the highest-predictable cluster obtain larger excess returns and Sharpe ratios than those for the lowest-predictable cluster.

We further contribute to the recent literature documenting a relationship between forecast disagreement and future returns. Bali, Kelly, Mörke and Rahman (2023) represent heterogeneous investor beliefs with distinct ML models, each trained on the same dataset. Furthermore, the dispersion in return forecasts across ML models represents the level of disagreement between investors' forecasts. The authors show that a value-weighted portfolio that buys low-disagreement stocks and sells high-disagreements stocks delivers an annual return of 14%. Similarly, Cong et al. (2024) construct an anomaly tied to predictability disagreement and model misspecification. Specifically, they measure predictability disagreement as the difference between in-sample R^2 estimates obtained from cluster-specific models and a global predictive model (i.e., a single model fitted to all observations grouped together). They find that a value-weighted portfolio that buys high-disagreement stocks and sells low-disagreement stocks earns statistically-significant monthly abnormal returns exceeding 1%. Although the findings in these two papers appear contradictory—one buys low-disagreement stocks while the other buys high, but both strategies earn large positive returns—the underlying frameworks differ fundamentally. In addition to employing different disagreement measures and datasets, the ML models are

estimated in two distinct ways: Bali et al. (2023) estimate different ML models on the same dataset, whereas Cong et al. (2024) fit the same ML model to different subsamples (i.e., clusterwise models are fitted to only the observations in the same cluster). Systematically comparing these two forms of disagreement (i.e., model class vs. dataset) is a promising avenue for future research. In this paper, we construct similar long-short portfolios as Cong et al. (2024) to form an anomaly related to predictability disagreement.

Finally, we contribute to the literature documenting that return-based anomalies—while significant in the equity market (e.g., Jegadeesh & Titman, 1993)—generate negligible abnormal returns in the corporate bond market. In particular, we show that the predictability-disagreement anomaly of Cong et al. (2024) yields economically significant, but statistically insignificant, abnormal returns. Recently Dickerson, Robotti and Rossetti (2023) show that return-based anomalies, such as momentum and short- and long-term reversal, produce average returns close to zero in the corporate bond market once market microstructure noise (MMN) is accounted for in transaction-based data from the Trade Reporting and Compliance Engine (TRACE) database. For example, while Jostova et al. (2013) document momentum effects—primarily in noninvestment-grade bonds—Dickerson, Robotti and Rossetti (2023) find that momentum delivers negative average returns after proper data cleaning and without relying on ex-post return winsorization. Even more interestingly, they show that bond winners perform worse than bond losers.

3 Methodology

In this section, we explain the tree-based clustering procedure introduced by Cong et al. (2024), which groups assets according to their degree of return predictability. First, we work under the assumption that the cluster-assignment of each asset-return observation is known, and specify the models used to predict returns within each cluster in Section 3.1. In Section 3.2, we introduce our measure of return predictability, the R^2 statistic. Finally, in Section 3.3, we relax the assumption of predetermined cluster assignments and show how panel trees learn to partition assets into groups with homogeneous levels of predictability directly from the data.

3.1 Clusterwise predictive model

Let $r_{i,t}$ denote the excess return of asset i at time t, defined as the return in excess of the one-month US Treasury bill rate. We consider a potentially unbalanced panel of asset returns, such that assets are indexed by $i = 1, \ldots, N_t$ and time periods by $t = 1, \ldots, T$. Additionally, for each asset i at time t, we observe a vector of C asset characteristics, denoted by $\mathbf{z}_{i,t}$, as well as an M-dimensional vector of macroeconomic predictors, $\mathbf{x}_t$. We

construct these variables using the data introduced in Section 4.

Consider a general predictive model for asset returns given by

$$r_{i,t+1} = \mathbb{E}_t[r_{i,t+1}] + \varepsilon_{i,t+1}, \tag{1}$$

where $\mathbb{E}[\varepsilon_{i,t+1}] = 0$. It is common practice in the literature to estimate $\mathbb{E}_t[r_{i,t+1}]$ using a—potentially highly nonlinear—function of predictor variables, $g(\mathbf{z}_{i,t}, \mathbf{x}_t)$. Some studies restrict this specification to be constant over time and cross-section (e.g., Gu et al. (2020)), while others estimate a time-varying function, $g_t(\cdot)$, using rolling- or expanding-window models (e.g., Feng, He, Wang and Wu (2025)). However, recent literature document that both approaches exhibit important shortcomings. First, they neglect the heterogeneous predictive power of features across assets (Evgeniou, Gueciouer & Prieto, 2023; Feng & He, 2022). Second, they implicitly assume that return predictability is constant across assets (Avramov, Cheng & Metzker, 2023).

Instead, our framework allows return predictability to vary along both the cross-sectional and time-series dimensions. We achieve this by clustering observations with similar levels of predictability using the *panel trees* introduced by Cong et al. (2025). These panel trees consider asset characteristics and macroeconomic predictors as potential splitting variables, and their structure is learned directly from the data. Specifically, macroeconomic variables are used to group assets along the time dimension, while standardized asset characteristics partition observations along the cross-sectional dimension. Within each resulting cluster, we estimate a separate predictive model, which yields a cluster-specific measure of return predictability and accounts for the heterogeneous predictive power of features across assets. Section 3.3 provides more details on this endogenous clustering procedure, but here we take as given a set of J non-overlapping clusters. Figure 1 illustrates how a panel tree partitions asset-return observations into clusters when $J = 3$.

Thus, instead of fitting a single predictive model on all observations, as in Gu et al. (2020), we end up with a set of J clusterwise predictive models. More precisely, for each cluster $j \in \{1, \ldots, J\}$, we estimate a separate model, such that

$$\mathbb{E}_t[r_{i,t+1}] = g_j(\mathbf{z}_{i,t}, \mathbf{x}_t), \tag{2}$$

where all observations assigned to cluster j share the same predictive model. We want to highlight that this predictive modeling is fully integrated into the clustering procedure. Namely, our tree-based approach operates iteratively: at each step, new groups of asset-return observations are formed together with their corresponding predictive models, which are then used to update the cluster assignments. This distinguishes our method from the common two-step approach, where first clusters are formed and then predictive models are estimated on these clusters (Evgeniou et al., 2023).

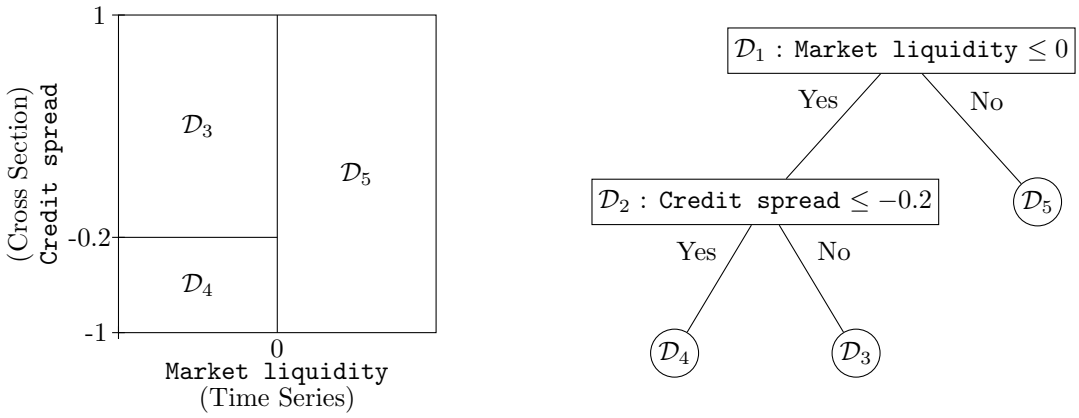


Figure 1: This figure illustrates our clustering procedure. We use macroeconomic variables to cluster assets along the time dimension and asset characteristics to cluster them along the cross-sectional dimension. All asset characteristics are cross-sectionally ranked between $[-1, 1]$ each month, and the macroeconomic variables are standardized to the interval $[-0.5, 0.5]$ based on their values over the last 10 years. In the left panel we observe three clusters: $\mathcal{D}_3$, $\mathcal{D}_4$ and $\mathcal{D}_5$. In each cluster, we fit a separate model for corporate bond returns. In the right panel we observe the panel tree that creates these clusters.

In theory, any predictive model could be used for $g_j(\cdot)$ in Eq. (2). In practice, however, this flexibility is limited. As we discuss in Section 3.2, the model should yield a generalizable measure of predictability that remains valid out-of-sample, at least approximately. This means that we should prevent the model from overfitting the training data. Thus, while complex deep learning methods combined with bootstrapping methods could in theory be applied, this would impose a substantial computational burden in the context of our tree-based clustering framework. For this reason, we use Ridge regressions and random forests in this paper.

First, ridge regressions are well-suited for low signal-to-noise environments and, thus, most empirical applications in finance. Additionally, ridge regressions have previously been applied by Cong et al. (2024) to estimate the *mosaic of predictability* for US equities. We investigate whether this ML model can also be used to construct the mosaic of predictability for corporate bonds. Not only are corporate bonds exposed to different market dynamics than equities (e.g., corporate bonds trade over-the-counter), but reliable data on corporate bonds are also in much shorter supply than for equities—typically beginning only in the early 2000s. In order to prevent the ML models from overfitting the training data, we apply careful hyperparameter tuning. For the empirical applications in Section 5, we use two-fold cross-validation to determine the $L2$-penalization parameter, λ. Following Feng, He, Wang and Wu (2025), we choose λ from the interval (e^{-40}, e^{40}).[5]

Moreover, random forests (RFs) are non-parametric and often deliver superior out-of-sample performance compared to linear Ridge regressions. For example, Feng, He, Wang and Wu (2025) show that RFs are among the most accurate models for predicting the

[5]Specifically, we consider the range `np.exp(np.linspace(-40, 40, num=100))` in `NumPy` (Python).

returns of individual corporate bonds. An additional advantage is that RF performance tends to be relatively insensitive to hyperparameter choices, which is particularly convenient for our tree-based clustering procedure (Goulet Coulombe, 2024). In our empirical applications, estimating a single panel tree with RFs as clusterwise predictive models requires approximately 12 hours of computation. To balance performance and computational feasibility, we therefore fix the hyperparameters of our RFs. Specifically, following the results in Feng, He, Wang and Wu (2025), we consider RFs with 500 trees, each having a minimum leaf size of 5% of the sample, and $\sqrt{\texttt{num\_features}}$ possible candidates at each split. While we also considered alternative hyperparameter settings, we found that they had little impact on predictive accuracy but substantially increased computational costs.[6]

3.2 A measure of return predictability

Before we can even cluster assets based on their degree of return predictability, we must first define a measure of predictability. Following Cong et al. (2024), we motivate for the use of the in-sample R^2. Specifically, we interpret a higher in-sample R^2 as indicating greater predictability of an asset's returns.

Under the model in Eq. (1), the *coefficient of determination* for asset i at time t is defined as

$$R^2_{i,t} = 1 - \frac{\mathbb{E}\big[(r_{i,t} - \hat{r}_{i,t})^2\big]}{\mathbb{E}\big[(r_{i,t} - \bar{r})^2\big]}, \tag{3}$$

where $\bar{r}$ is the average return over all stocks and time periods, and $\hat{r}_{i,t}$ is an estimate of $\mathbb{E}_{t-1}[r_{i,t}]$ in Eq. (1) (i.e., a one-step-ahead return forecast). A large $R^2_{i,t}$ indicates that our forecast, $\hat{r}_{i,t}$, aligns better with the conditional expectation $\mathbb{E}_{t-1}[r_{i,t}]$ than the average return, $\bar{r}$, does. Furthermore, for an unpredictable asset—characterized with a low signal-to-noise ratio (SNR)—this $R^2_{i,t}$ will be small even if we know the true conditional expectation (i.e., $\hat{r}_{i,t} = \mathbb{E}_{t-1}[r_{i,t}]$). This can be seen by looking at the numerator of $R^2_{i,t}$ in Eq. (3), which is essentially the variance of $\varepsilon_{i,t}$ in Eq. (1) when we know $\mathbb{E}_{t-1}[r_{i,t}]$. For an unpredictable asset, with low SNR, the variance of $\varepsilon_{i,t}$ is large, and so will be the numerator in Eq. (3). As a consequence, the $R^2_{i,t}$ of an unpredictable asset is small. In other words, $R^2_{i,t}$ provides a proxy for the SNR of assets, and is thus a reasonable measure for unobservable return predictability.

We use the clusterwise predictive models from Section 3.1 to estimate the R^2 values. We emphasize that estimating $R^2_{i,t}$ for each asset i and time period t is infeasible, since we only have a single observation for each asset at each time period. While most studies impose strict homogeneity assumptions in order to make the R^2-calculations feasible again,

[6] All models are estimated on a MacBook Air M1 with an 8-core processor and 8 GB of memory.

such as by fitting a single model on all panel data together (R^2; e.g., Gu et al., 2020) or by fitting models with time-varying parameters but which are assumed homogeneous across assets (R^2_t; e.g., Feng, He, Wang & Wu, 2025), we estimate clusterwise R^2 values. More specifically, we form clusters of asset-return observations with different levels of predictability, and where cluster-membership is determined by the values of standardized asset characteristics and macroeconomic variables. Then, for each cluster $j \in \{1, \ldots, J\}$, we estimate R^2_j using the predictive models detailed in Section 3.1 and all observations in cluster j, which is defined by

$$R^2_j = 1 - \frac{\sum_{\{i,t\} \in \text{cluster}_j} (r_{i,t} - \hat{r}_{i,t})^2}{\sum_{\{i,t\} \in \text{cluster}_j} r^2_{i,t}}. \tag{4}$$

The denominator in Eq. (4) is different from the definition of the coefficient of determination usually used in literature. Following Gu et al. (2020) and Bali et al. (2020), we use the sum of squared returns *without demeaning* as the denominator in Eq. (4). The reason is that the sample average, $\bar{r}$, is usually a noisy proxy, which lowers the bar to achieve a high R^2_j. Instead, a zero-return forecast is stable and typically outperforms a historical average return forecast in practice. Therefore, we compare our forecasts $\hat{r}_{i,t}$ against a zero-return forecast instead of the historical average, $\bar{r}$, in Eq. (4).

Moreover, since we use in-sample data to measure return predictability and fit our ML models, we conduct careful hyperparameter tuning to prevent our models from overfitting the data and, thus, artificially increasing R^2_j. More details on hyperparameter tuning are provided in Section 3.1, but we want to pay again particular attention to RF here. Namely, when calculating R^2_j in Eq. (4), we should not use RF's *fitted values*. These are prone to overfitting and will almost always deliver R^2_j values close to 1 (Goulet Coulombe & Göbel, 2023).[7] Therefore, instead of using RF's fitted values for $\hat{r}_{i,t}$ in Eq. (4), we use RF's out-of-bag (OOB) predictions (Breiman, 2001). The accuracy of these predictions is aligned with the out-of-sample accuracy of RF's predictions. More specifically, when using RF, we measure return predictability by

$$R^2_{OOB,j} = 1 - \frac{\sum_{\{i,t\} \in \text{cluster}_j} (r_{i,t} - \hat{r}^{OOB}_{i,t})^2}{\sum_{\{i,t\} \in \text{cluster}_j} r^2_{i,t}}, \tag{5}$$

where

$$\hat{r}^{OOB}_{i,t} = \frac{1}{|\mathcal{B}_{i,t}|} \sum_{b \in \mathcal{B}_{i,t}} h_b(\mathbf{z}_{i,t-1}, \mathbf{x}_{t-1}),$$

where h_b is the predictive function represented by the b-th tree in the RF, and $\mathcal{B}_{i,t}$ is the set of trees that do not use the observation for asset i at time t during estimation (i.e., it

[7] Goulet Coulombe (2024) illustrates the immense overfitting of RF when applied to some common empirical applications and popular datasets, and provides an explanation for this phenomenon.

is not in the bootstrapped sample).

Finally, the main reason why we use the in-sample R^2 instead of its out-of-sample equivalent (R^2_{OOS}) is that R^2_{OOS} would introduce test-sample information into our clustering procedure. Ultimately, we want to evaluate our clusters, and their corresponding predictive models, out-of-sample. Using R^2_{OOS} to form clusters would, therefore, lead to data-snooping bias (Cong et al., 2024).

3.3 Clustering using panel trees

As mentioned before, we group corporate bonds using the tree-based clustering procedure introduced by Cong et al. (2025). In Section 3.3.1 we introduce these panel trees (i.e., *P-Trees*) by comparing them with CART (Breiman, Friedman, Olshen & Stone, 2017). After that, we detail how the first, second and subsequent splitting points in these binary P-Trees are chosen, and specify under which conditions this iterative clustering procedure is terminated. The pseudocode for our algorithm is provided in Appendix B.

3.3.1 Panel trees versus CART

CART is a popular ML model for both classification and regression problems, and forms the basis for other popular ML techniques such as RF and bagging. CART is essentially a binary decision tree. It partitions observations into disjoint clusters, each represented by a leaf node in the tree, as illustrated in Figure 1. In each cluster a constant parameter is estimated using observations in the training data, which is then used to make predictions for new test observations. For instance, in the case of regression problems and a quadratic loss function, the tree would predict the outcome of a test observation ending up in leaf node $j \in \{1, \ldots, J\}$ by the average outcome of all training observations in the j-th leaf node. For classification problems, majority voting is often used. Tree-based models are particularly popular because of their interpretability, which is also one of the main reasons why we use them in this paper.

These conventional decision trees (e.g., CART) are grown using a heuristic procedure. Namely, each node is iteratively split into two child nodes, where the split point is chosen to optimize some local objective function. This means that the objective is a deterministic function of solely the training observations that end up in the child nodes, and does not take the other observations in the tree into account. As the tree grows deeper and deeper, the number of observations that ends up in each node declines monotonically, which—in combination with the local splitting objective—makes it sensitive to outliers and increases the chances of overfitting the training data (Cong et al., 2025). Furthermore, the local splitting procedure of CART makes it impossible to impose *global* economic restrictions on the tree growing procedure.

P-Trees are an extension on CART, and address the two problems mentioned above

by using global instead of local split criteria. More specifically, a P-Tree uses a global objective function that incorporates all observations in the tree to determine the best splitting point and, thus, the best way to split an already existing cluster (i.e., leaf node) into two separate clusters. This does not only make panel trees more robust to low SNR environments, but also allows us to incorporate economic restrictions. For instance, one could use P-Trees to cluster assets into disjoint groups which form portfolios that are chosen to maximize some financial objective (e.g., mean-variance optimization), while imposing restrictions on overall transaction costs (Cong et al., 2025).

Moreover, panel trees are also time-invariant and specifically designed for (financial) panel data. CART, on the other hand, assumes that all observations in the same leaf node (i.e., cluster) are independently and identically distributed; and, thus, not suitable for panel data. Instead, P-Trees address this issue by standardizing all features. For instance, we cross-sectionally rank each asset characteristics, $\mathbf{z}$, to the interval $[-1, 1]$ each month in our empirical applications in Section 5. This makes P-Trees robust to outliers and time-invariant. By time-invariance, we mean that we do not need to re-estimate the tree structure over time as the levels of asset characteristics change. While, in practice, we could also use cross-sectionally standardized features with CART, this is not considered a characteristic inherent to CART in contrast to P-Trees, and for that reason worth to emphasize here.[8]

Finally, we want to highlight that k-means and other conventional clustering techniques are not helpful in our context (Cong et al., 2024). Our goal is to cluster assets by their level of return predictability. The objective of k-means, however, is to cluster observations in order to maximize the between-cluster variation and minimize the within-cluster variation, by measuring the distance between observations in a potentially high-dimensional feature space. As discussed in Section 3.2, we use R^2 to measure return predictability, which is unobservable and depends on the predictive model used. Computing the R^2 for each asset-return observation individually (i.e., $R^2_{i,t}$) is also infeasible. Therefore, we cannot compute the distance between observations using their endogenous R^2 values, which makes k-means clustering infeasible. Furthermore, while in practice it would be possible to first compute R^2 values for each asset individually (i.e., R^2_i) and use these together with k-means to form clusters, this would assume that asset predictability stays constant over time—both during economic business cycles (time series) as well as across the business cycle of companies that, for instance, develop from growth to value firms (cross-section). Recent literature, however, finds that return predictability varies over the time-series and cross-sectional dimension (Avramov et al., 2023; Cong et al., 2024). Thus, this work-around to use k-means instead of P-Trees places strict assumptions on the resulting clusters. We relax these assumption with P-Trees, by integrating predictive modeling with clustering based on cross-sectional asset characteristics

[8] See e.g., Bie et al. (2024); Cong et al. (2023, 2024, 2025); Feng, He, Li et al. (2025).

and time-varying macroeconomic variables. Lastly, P-Trees are also more interpretable than k-means clusters, since we can better understand how cluster assignment changes—and, thus, return predictability—when characteristics or market conditions change.

3.3.2 First split

Next, we discuss the iterative clustering procedure of P-Trees to form groups with heterogeneous levels of return predictability, by illustrating how the first split is determined. Without loss of generality, we generate a P-Tree using only the asset characteristics, $\mathbf{z}$, which are cross-sectionally ranked, demeaned, and standardized each month to the interval $[-1, 1]$. We do so because our first empirical application in Section 5 follows this approach and does not consider macroeconomic splitting variables, $\mathbf{x}$. Later, in Section 5.4, we consider asset characteristics and macroeconomic predictors together, and that extension to the methodology outlined here follows trivially. We denote the k-th asset characteristic by $z_{\cdot,\cdot,k}$, and let c_m be a split threshold value. In our empirical applications, we consider $c_m \in \{-0.2, 0.2\}$.[9]

We start the clustering procedure with a single cluster that contains all observations, which is represented by a P-Tree with a single (root) node. Then, in order to determine the next split point we consider all candidate splits. Each candidate is a combination of an asset characteristic, $z_{\cdot,\cdot,k}$, and a threshold value, c_m, denoted by $\tilde{\mathbf{c}}_{k,m} = (z_{\cdot,\cdot,k}, c_m)$. Then, when the k-th characteristic of asset i at time t is less than or equal to the threshold value (i.e., $z_{i,t,k} \leq c_m$), observation $\{r_{i,t}, \mathbf{z}_{i,t}\}$ is assigned to the left child node, and otherwise the right child node. In this way, the original cluster is split into two non-overlapping groups of observations.

For each candidate split $\tilde{\mathbf{c}}_{k,m}$, we fit two predictive model—one for each resulting cluster (leaf node), that uses only the training data allocated to that cluster. These predictive models are functions of lagged asset characteristics, $\mathbf{z}_{i,t-1}$, and we consider RF and Ridge regressions separately in this paper. Let the fitted model for the j-th leaf node be denoted by $\widehat{g}_j(\cdot)$. Then, the clusterwise return forecasts are used to calculate R^2_j as defined in Eq. (4) (or $R^2_{OOB,j}$ in Eq. (5) when RF is used).

From all these split rule candidates, we choose the one that maximizes the split criterion given by

$$S_{\{\text{leaf}_l, \text{leaf}_r\}}(\tilde{\mathbf{c}}_{k,m}) = |R^2_{\text{leaf}_l} - R^2_{\text{leaf}_r}|. \tag{6}$$

This split criterion is motivated by our objective to separate highly predictable returns

[9] For the regime-switching panel tree in Section 5.4, we consider only threshold values of 0, both for the standardized asset characteristics as well as the macroeconomic predictors. We make this decision in order to prevent these P-Trees from overfitting the training data, taking into account the relatively small sample period we have for corporate bonds, which only gets smaller when we also consider time-series splits.

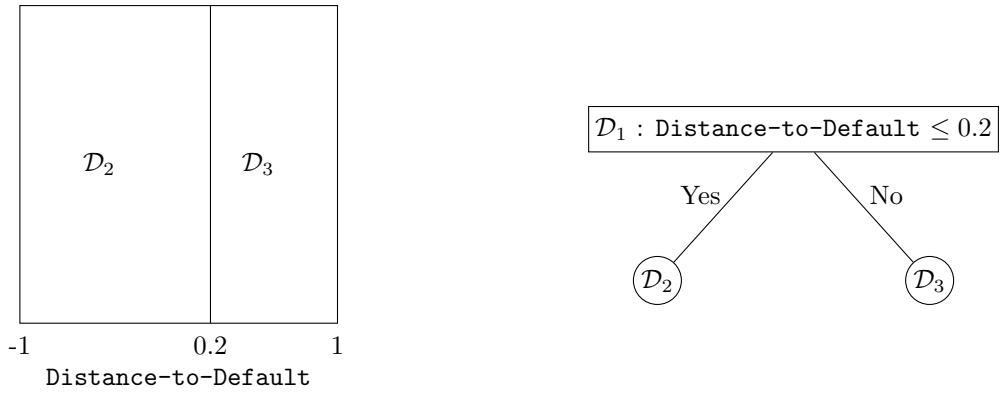


Figure 2: This figure illustrates the first split in our tree-based clustering procedure. The root node ($\mathcal{D}_1$) containing all asset observations, is split into two non-overlapping clusters (i.e., leaf nodes) $\mathcal{D}_2$ and $\mathcal{D}_3$. Those asset-returns observations that satisfy the split criterion: `Distance-to-Default` ≤ 0.2, are assigned to $\mathcal{D}_2$, while those that do not are placed in cluster $\mathcal{D}_3$. In order to determine the optimal first split, all candidate splits (i.e., combinations of split variables and thresholds) are considered, and used to evaluate the split criterion in Eq. (6). The one that maximizes this criterion is chosen as the first split.

from those that are less predictable. It maximizes the heterogeneity in return predictability in the clusters created by the split. As long as the stopping criteria detailed in Section 3.3.4 are not activated by the optimal split rule candidate, we perform the split and look for the optimal second split.

Figure 2 illustrates how two non-overlapping clusters are created, $\mathcal{D}_2$ and $\mathcal{D}_3$, by splitting the complete sample $\mathcal{D}_1$ according to a potential first split rule. For each candidate split, two predictive models are estimated, $\hat{g}_2(\cdot)$ and $\hat{g}_3(\cdot)$, using only training data in clusters $\mathcal{D}_2$ and $\mathcal{D}_3$, respectively. Then, the corresponding return forecasts for each cluster are used to calculate R_2^2 and R_3^2 as defined in Eq. (4). These estimates are, in turn, used to evaluate the split criterion in Eq. (6). Doing this for all combinations of split variables and threshold values, we choose the candidate that maximizes the split criterion as first split.

3.3.3 Second and subsequent splits

After the first split is determined, we repeat a similar procedure as outlined in Section 3.3.2 to determine the second split. More specifically, we consider again all candidate splits, fit separate predictive models for each cluster, and choose the candidate that maximizes the split criterion in Eq. (6). However, instead of a single node to split, there are now two potential nodes that can be split into two new leaf nodes, namely D_2 and D_3 in Figure 2. Thus in both of the current leaf nodes, we consider all candidate splits. Then, we choose the candidate split that optimize the global criterion (i.e., over the whole tree), which means splitting either D_2 or D_3. Cong et al. (2024) term this procedure as a *local-global framework*: the candidates for each leaf node are considered locally, but the optimal

candidate is chosen globally (i.e., over both leaf nodes). Figure 3 illustrates how the second split is determined, where short-term bond reversal is considered as a candidate split with a threshold value of -0.2.

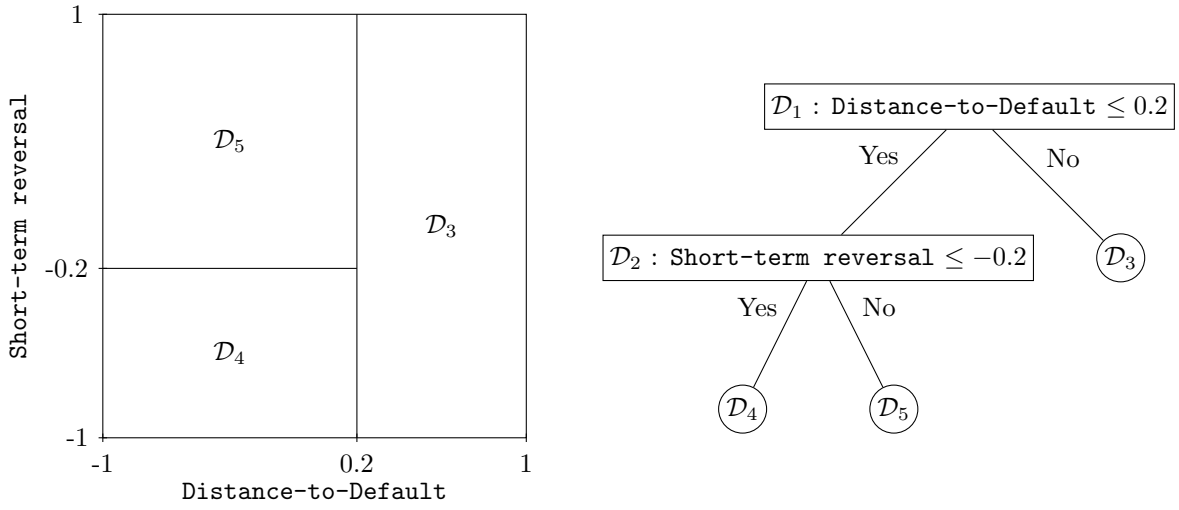
All subsequent splits are determined following a similar procedure as the first two splits. First, we evaluate the split criterion in Eq. (6) for all candidate splits and each leaf node. Then, we choose the candidate split that maximizes the split criterion as the optimal subsequent split. This means that at each iteration, only a single node is split into two new leaf nodes (i.e., clusters). After the split is performed, we check whether one of the stopping conditions detailed in Section 3.3.4 is activated. If so, we stop our clustering algorithm. The pseudocode for this algorithm is provided in Appendix B.

As discussed in Section 3.3.1, P-Trees are time-invariant, in the sense that the estimated tree structure does not need to be re-estimated over time. Nevertheless, this does not imply that cluster membership of asset i is time-invariant. Since cluster membership is determined by asset characteristics, changes in these characteristics over time naturally induce changes in cluster assignment. For example, imagine that the tree in the right panel of Figure 2 represents our fully grown P-Tree. Furthermore, suppose that asset i has a low distance-to-default at time t (i.e., $\mathbf{z}_{i,t,k} \leq 0.2$, where the k-th feature is distance-to-default). In Section 5, we show that such bonds exhibit substantially lower return predictability than corporate bonds with higher distance-to-default. If, over time, bond i becomes more stable and its standardized distance-to-default rises above 0.2 at time $t+1$, then its cluster membership changes accordingly. This shift reflects a change in how investors value the bond, which translates into different return predictability. Our tree-based clustering framework is specifically designed to capture such dynamics by making cluster assignment a function of asset characteristics (and macroeconomic variables). In this specific example, asset i transitions from cluster D_2 to D_3 in Figure 2 at time $t+2$ (taking into account the use of lagged predictors), and thus its level of return predictability is now estimated by R^2_3 instead of R^2_2 in Eq. (4).

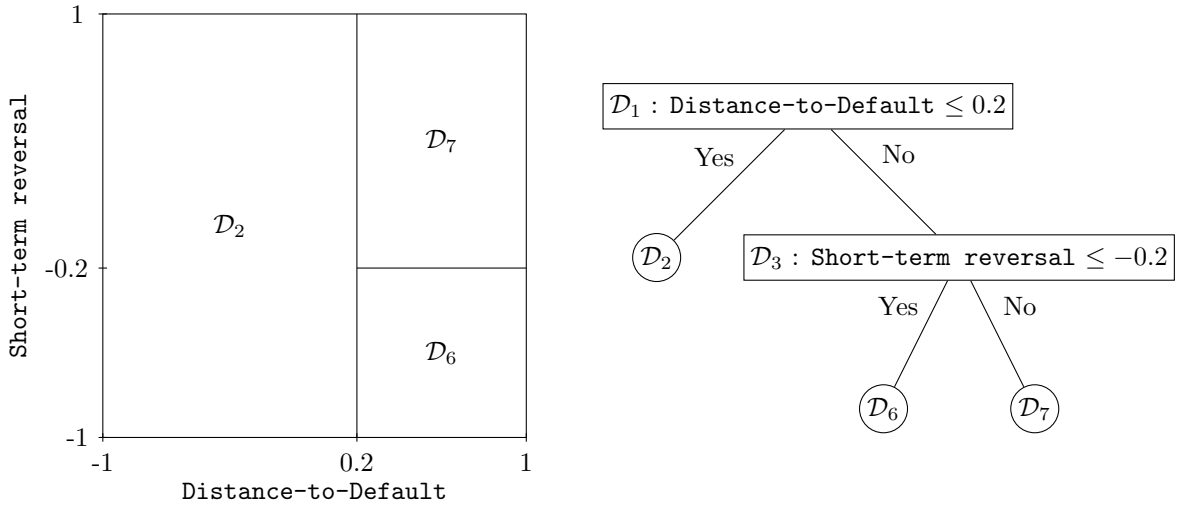
3.3.4 Stopping criteria

Furthermore, in order to prevent the tree clustering procedure from overfitting our data, we impose a set of stopping criteria. First, we set a maximum tree-depth, $d_{\max}$. This means we can generate at most $2^{d_{\max}}$ separate clusters. Following the ML literature, we define the depth of the tree as the length of the longest path in the tree. This means that a tree with only a single (root) node has a depth of zero.[10] In our applications, we set $d_{\max} = 4$. Second, we require a minimum number of monthly observations in each node, and eliminate candidate splits that do not meet this requirement. In order to balance robustness of our results and the specificity of our clustering procedure, we set

[10] Our definition of tree-depth differs from Cong et al. (2024). They assume the depth of a tree is 1 plus the length longest path, such that a tree with a single (root) node has a depth of 1.



(a) Split D_2 with "Short-term reversal ≤ -0.2"



(b) Split D_3 with "Short-term reversal ≤ -0.2"

Figure 3: This figure illustrates the second split in our tree-based clustering procedure. It follows the first split in Figure 2, and shows how either the left (D_2) or right (D_3) child node can be split into two new leaf nodes (right panels) and corresponding clusters (left panels). It does so by using the split rule: "Short-term reversal ≤ -0.2". In practice, all candidate splits (i.e., combinations of split variables and thresholds) are considered in both leaf nodes D_2 and D_3. Then, the candidate split that maximizes the split criterion in Eq. (6) is chosen as the second split. This means, thus, that either the left or right child node is split, but not both at the same time.

this minimum to 30 monthly observations. Finally, we stop a node from splitting further, when both child nodes have a smaller R^2 than their parent node.

4 Data

In this section, we describe the data used to grow the panel trees and estimate the predictive models. First, we detail the construction of monthly excess corporate bond returns in Section 4.1. Then, we present the equity and bond characteristics in Section 4.2. Finally, we introduce the set of macroeconomic indicators used as time-varying splitting variables in Section 4.3.

4.1 Corporate bond returns

Unlike equities, error-free databases for corporate bonds are not readily available. Recent studies document a credibility and replication crisis in the corporate bond literature, and attribute this to the use of unreliable data sources and the absence of a standardized framework for cleaning bond data (Dick-Nielsen et al., 2023; Dickerson, Robotti & Rossetti, 2023). For example, inconsistencies in the data-cleaning approach of Bai, Bali and Wen (2019) recently led to the article's retraction from the *Journal of Financial Economics* (Dickerson, Mueller & Robotti, 2023). Moreover, Dickerson, Robotti and Rossetti (2023) show that all return-based anomalies documented in prior corporate bond literature—such as return reversals and momentum—are largely explained by market microstructure noise (MMN) and ad-hoc return winsorization.[11]

For this reason, we use the dataset on monthly bond returns constructed by Dickerson, Robotti and Rossetti (2023), which we obtain from the *Open Source Bond Asset Pricing* website.[12] This dataset is based on pre-processed monthly bond data from the Wharton Research Data Services (WRDS) Bond Database.[13] The WRDS Bond Database itself is constructed from intraday transaction data in the enhanced Trade Reporting and Compliance Engine (TRACE) database and bond characteristics from the Mergent Fixed Income Securities Database (FISD). In Appendix D, we provide additional details on these databases and describe the construction of our dataset. After all, the monthly return for bond i in month t is defined as

$$R_{i,t} = \frac{P_{i,t} + AI_{i,t} + C_{i,t}}{P_{i,t-1} + AI_{i,t-1}} - 1,$$

[11] Examples of studies that document return-based anomalies in the corporate bond market include Jostova et al. (2013) and Bali et al. (2021).

[12] https://openbondassetpricing.com

[13] This database is available via https://wrds-www.wharton.upenn.edu/pages/get-data/wrds-bond-returns/wrds-bond-returns/.

where $P_{i,t}$ is the trade price of bond i in month t, $AI_{i,t}$ is the accrued interest, and $C_{i,t}$ is the coupon payment. The excess return is then defined as $r_{i,t} = R_{i,t} - R_{f,t}$, where $R_{f,t}$ is the risk-free rate (i.e., the one-month US Treasury bill rate).

However, while WRDS/TRACE is still widely used in academic research, it is not the ideal database for corporate bond returns, as discussed by Kelly et al. (2023). The most common source used by hedge funds, asset managers, and banks is the Intercontinental Exchange (ICE) database. The primary reason for this distinction in popularity is that WRDS/TRACE data is transaction-based, which means that we can only calculate the returns for a bond when it trades sufficiently close to both the start and end of a month. ICE, however, is a quote-based database, which means that we can always calculate the *exact* monthly returns. This difference also has implications for return dynamics. For instance, Andreani, Palhares and Richardson (2024) show that the first-order autocorrelation for monthly returns from ICE is closer to zero than those from WRDS/TRACE. Specifically, the autocorrelation of monthly WRDS/TRACE returns is -0.09 for both investment-grade (IG) and high-yield (HY) bonds, whereas ICE yields values of -0.06 for IG and -0.01 for HY bonds. We take this into account when interpreting our results in Section 5. Ideally, we would collect monthly bond returns from the ICE database, but this is not publicly available and expensive. For more details on the differences between ICE and WRDS/TRACE, we refer to Kelly and Pruitt (2022) and Dickerson, Robotti and Rossetti (2023).

4.2 Corporate bond predictors

To train the ML models in the leaves of the panel tree, we use a set of 53 bond and equity characteristics constructed by Dickerson, Robotti and Nozawa (2024), which spans a sample period from August 2002 to August 2022.[14] The predictors are defined in Table 5 of Appendix C, which also includes references to the literature motivating their use in our models. The 37 bond characteristics are constructed from daily bond data for the constituent bonds of the Bank of America (BAML) Investment Grade (C0A0) and High Yield (H0A0) indices available via ICE. The remaining 16 equity characteristics are constructed using data from the Center for Research in Security Prices (CRSP) and COMPUSTAT (COMP), both available via WRDS. Furthermore, our dataset has $855,729$ observations in total. The average and median number of monthly observations from August 2002 to August 2015 (i.e., the training sample in Section 5) are $2,887$ and $2,778$, respectively. For the test sample from September 2015 to August 2022, the average and median number of monthly observations are $4,792$ and $4,713$, respectively.

Our dataset includes those asset characteristics shown to have predictive power for corporate bonds. For instance, Feng, He, Wang and Wu (2025) find that short-term bond

[14] We download this dataset from the author's website: https://openbondassetpricing.com/machine-learning-data/.

reversal (STREVB; i.e., previous month's bond return), return skewness (SKEW), downside risk (VaR) and credit spreads (SPREAD) are important predictors for corporate bond returns. Besides that, our dataset includes all characteristics used by Kelly et al. (2023) to construct a conditional factor model for corporate bond with IPCA (Kelly, Pruitt & Su, 2019).

Finally, we emphasize that these asset characteristics are cross-sectionally ranked each month, such that they lie within the interval $[-1, 1]$. Furthermore, we set any missing values equal to 0 (i.e., the median). Following the same motivation as Cong et al. (2024) and Kelly et al. (2023), this standardization mitigates the influence of outliers and enhances the robustness of our results.

4.3 Macroeconomic splitting variables

To construct regime-switching panel trees, we incorporate time-varying, marketwide predictors as potential splitting variables. We follow the literature on panel trees and (regime-switching) return predictability by considering a set of nine macroeconomic variables, which are defined in Table 1 (e.g., Welch & Goyal, 2008; Cong et al., 2025, 2024). This set contains indicators for the bond market (e.g., default yield spread), the equity market (e.g., dividend yield of S&P 500), and the real economy (e.g., inflation). We construct all predictors, except the liquidity variable (LIQ), using data from Amit Goyal's website.[15] Furthermore, liquidity is defined as the aggregate liquidity measure of Pástor and Stambaugh (2003), which is collected from Robert Stambaugh's website.[16] We provide a more detailed discussion on the construction of this dataset in Appendix D.[17]

Additionally, we standardize each macroeconomic variable to the interval $[-0.5, 0.5]$. More specifically, for each variable j (e.g., INFL) in month t, we compute the fraction of values over the preceding 120 months that are smaller than the current observation, which maps the variable to the interval $[0, 1]$. Then, since we consider only splitting thresholds of 0 when constructing the regime-switching panel trees in Section 5, we subtract 0.5 from each variable. In this way, the tree partitions year-month observations by considering whether a marketwide splitting variable is currently above or below its median value over the past 120 months. Overall, this standardization procedure ensures that our panel tree is able to detect robust macroeconomic regimes. In Figure 8 of Appendix D.2, we plot the macroeconomic variables.

[15] https://sites.google.com/view/agoyal145

[16] https://finance.wharton.upenn.edu/~stambaug/

[17] In Appendix D, we also compare our dataset with the macroeconomic variables of Cong et al. (2025), which are publicly available through their replication package (https://data.mendeley.com/datasets/k7d7xmdy4y/2). While we consider the same macroeconomic variables, we cannot directly use their dataset because their sample ends in 2021M12, whereas ours extends to 2022M12. Although we construct our dataset following the same steps as Cong et al. (2024), the two do not align perfectly. Nonetheless, the correlations exceed 0.8 for each series (except for NI).

Table 1
This table presents the macroeconomic splitting variables used to grow regime-switching panel trees.

Variable	Description
DFY	Default yield spread.
TMS	Term spread.
SVAR	Stock variance on daily S&P 500 returns.[a]
LIQ	Pástor-Stambaugh aggregate liquidity.[a]
INFL	CPI from Bureau of Labor Statistics.[a]
DY	Dividend yield of S&P 500.
EP	Earnings-to-price ratio of S&P 500.
NI	Net equity expansion of NYSE listed stocks.
TBL	3-month US Treasury bill rate.

[a] This variable is defined by its 12-month moving average.

5 Results

In this section, we analyze the heterogeneity in return predictability in the corporate bond market.[18] First, we present the panel tree used to cluster bond-return observations, and interpret the resulting *mosaic of predictability*, in Section 5.1. This tree is constructed using only asset characteristics as splitting variables. Then, in Section 5.2, we evaluate the predictive performance of clusterwise models against a global predictive model. In Section 5.3, we investigate whether the *predictability disagreement anomaly* of Cong et al. (2024) is also present in the corporate bond market. Finally, we examine time-varying predictability in bond returns by estimating a regime-switching panel tree in Section 5.4, which is allowed to split on both time-varying macroeconomic variables and cross-sectional asset characteristics.

Moreover, in this section, we focus exclusively on Ridge regression as the predictive model. The results obtained with RFs are reported in Appendix G. We found that RFs tend to overfit the training data, even when OOB forecasts are used. Increasing the number of trees beyond 500 could yield more stable R^2_{OOB} estimates, but this comes at the cost of sharply higher computation time and memory requirements, making it impractical to run the code on a standard computer. While we considered relying on external computing resources, the potential gains were not sufficient given time constraints. Instead, we devoted our efforts to obtaining additional insights with Ridge regressions. Our goal is not to identify the single best predictive model, but rather to cluster assets based on their return predictability, with the expectation that the most predictable cluster in-sample also exhibits the strongest out-of-sample predictability. We demonstrate in this section that Ridge regressions are well suited for this purpose.

[18] Our code is publicly available at https://github.com/trmulder/MosaicsOfPredictability.

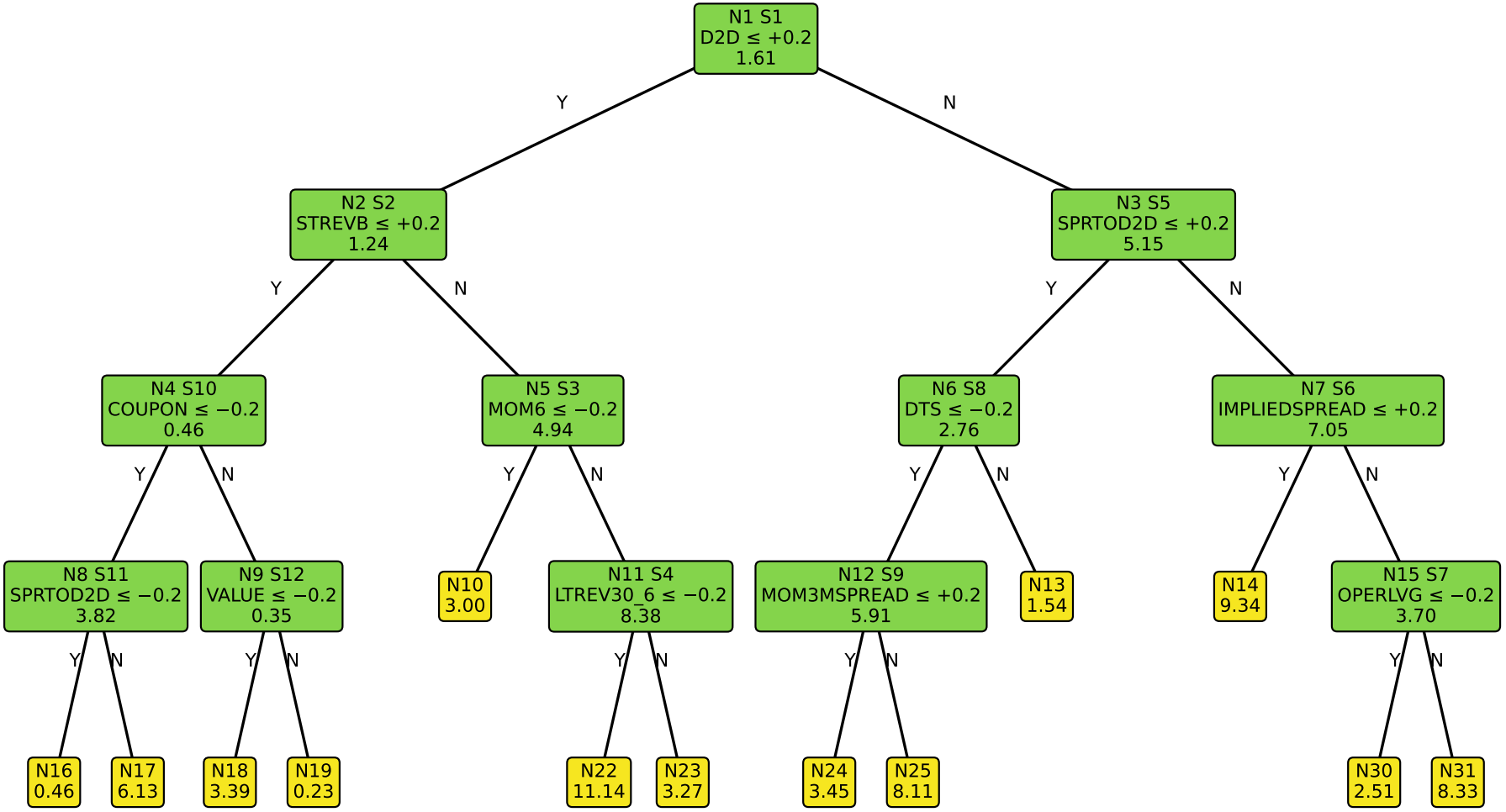


Figure 4: This figure illustrates the panel tree used to form clusters of bond-return observations with heterogeneous levels of predictability, measured by R^2. The tree is estimated using monthly data from 2002M8 to 2015M8, and splits on cross-sectional asset characteristics that are rank-standardized to the interval $[-1, 1]$ (see Appendix C). Each intermediate node (green) and leaf node (yellow) displays the node's ID, denoted by $N\#$, which is used to identify the node in this paper. Intermediate nodes additionally report the split rule (second line), which sends observations satisfying the rule to the left child node and the rest to the right child node, as well as the node's split order, denoted by $S\#$. Finally, all nodes display the in-sample R^2 statistic (in percentages; last line), estimated using the clusterwise Ridge regressions introduced in Section 3.

5.1 The mosaic of return predictability

Figure 4 presents the panel tree that clusters observations based on cross-sectional asset characteristics, forming groups with heterogeneous levels of return predictability, as measured by in-sample R^2. The tree is estimated using the 53 characteristics defined in Appendix C over a sample period from 2002M8 to 2015M8. Within each cluster, Ridge regression is used as the predictive model. Furthermore, by splitting the data into a *training* set (2002M8–2015M8) and *test* set (2015M9–2022M8), we are able to evaluate the predictive performance of our cluster-specific models *out-of-sample* in Section 5.2. Nevertheless, a full sample analysis is also provided in Appendix H, whose results are consistent with the findings reported here.

Each node in Figure 4 has a unique identifier, denoted by $N\#$, which is used to identify the node throughout this paper.[19] Intermediate nodes (green) additionally display the split order, $S\#$, along with the splitting rule that partitions the observations into a left and right child node. Both leaf nodes (yellow) and intermediate nodes report the in-sample R^2 (in percentages) estimated by a clusterwise predictive model fitted to all observations assigned to that node.

First, we observe that a global predictive model (i.e., a single model fitted to all observations grouped together) achieves an in-sample R^2 of 1.61%, as shown at the root node (N1) in Figure 4. Then, to determine the optimal first split, we evaluate the split criterion in Eq. (6) for all 53×2 candidate splits. The split rule maximizing this criterion is: $\texttt{D2D} \leq 0.2$, which is thus chosen as the first split. Observations with a high distance-to-default (i.e., $\texttt{D2D} > 0.2$) are more predictable, as the estimated R^2 statistic for this partition increases to 5.15% (node N3). For the group of remaining observations, assigned to node N2, the R^2 declines slightly to 1.24%. While this result suggests a positive relationship between distance-to-default and return predictability, subsequent splits in the fully grown tree could attenuate or reverse this relationship. Therefore, to assess the overall effect, we use the heatmap in the top-right panel of Figure 5. This heatmap depicts the average R^2 from the panel tree in Figure 4 for observations sorted into D2D deciles each year. The figure confirms that observations with higher distance-to-default tend to be more predictable, and that this pattern has persisted over time.

Distance-to-default (i.e., D2D) is a measure of default risk estimated with observed stock prices and book leverage under the credit risk model of Merton (1974).[20] It can be interpreted as the number of standard deviations a firm's asset value exceeds the face value of its debt. Prior studies mainly investigated how well distance-to-default forecasts

[19] Node IDs are assigned as follows: the root node has ID N1, and the left and right child of node Ni are N$\{2i\}$ and N$\{2i+1\}$, respectively.

[20] The conventional distance-to-default measure solves a structural option-pricing model for unobservable firm asset values and volatilities. However, we use the *naive* measure proposed by Bharath and Shumway (2008), which approximates these inputs directly with equity plus debt and observable volatility. This naive distance-to-default is easier to compute, and even outperforms the original measure.

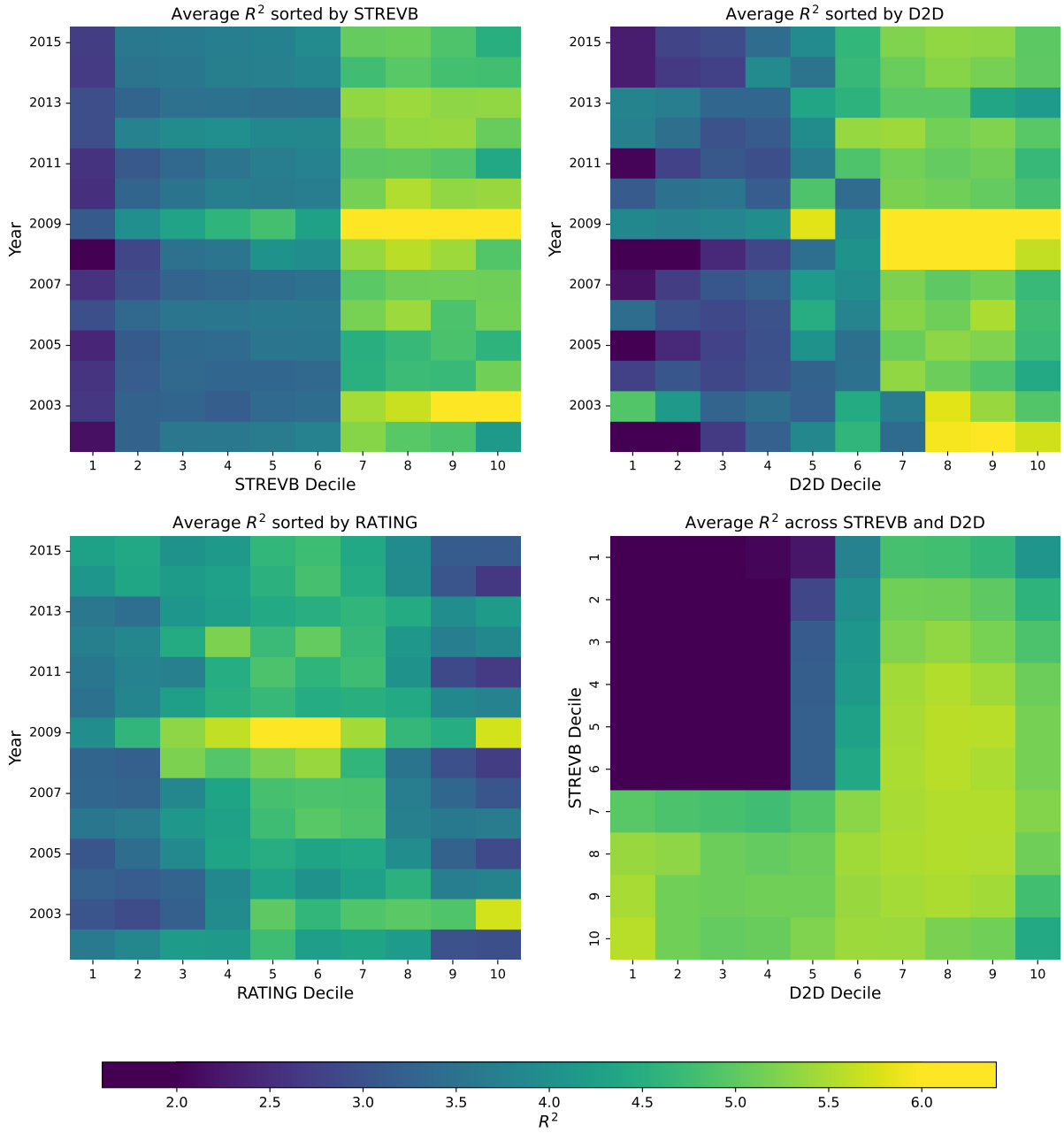


Figure 5: This figure presents four heatmaps of average in-sample R^2 under the panel tree in Figure 4 and when sorting observations into deciles of bond characteristics. The top panels show average R^2 for asset-returns observations sorted annually into deciles of short-term bond reversal (`STREVB`, left) and distance-to-default (`D2D`, right). The bottom left panel shows results for annual rating deciles, while the bottom right panel shows R^2 for 10×10 bivariate-sorted groups based on the first two splitting variables in Figure 4: `D2D` and `STREVB`. Lighter colors correspond to higher average R^2 values, as shown by the colorbar. Note that higher rating deciles are associated to lower-rated bonds, as can be seen from the definition of `RATING` in Table 5 of Appendix C.

corporate failure compared to other models, and the robustness of this measure (e.g., Bharath & Shumway, 2008; Jessen & Lando, 2015). However, to our knowledge, we are the first to document the relationship between distance-to-default and the predictability of corporate bond returns.[21]

We find several reasons that justify this positive relationship between return predictability and distance-to-default. First, from an economic perspective, we can argue that bonds with a low distance-to-default are dominated by idiosyncratic credit events and liquidity shocks, which creates an instable return-characteristic relationship and hence results in low in-sample R^2 estimates. In contrast, high distance-to-default bonds are far from distress and their returns are driven primarily by systematic risk factors—such as interest rate exposure, market betas, and persistent mispricing patterns—that can be forecasted from historical characteristics. Second, from a statistical perspective, distance-to-default (i.e., `D2D`) is shown to be robust to model misspecification (Jessen & Lando, 2015). More specifically, it is able to rank firms according to their default risk in empirical applications (Hillegeist, Keating, Cram & Lundstedt, 2004; Duffie, Saita & Wang, 2007). For this reason, distance-to-default is often used as the preferred proxy for default risk compared to other measures such as bond rating (e.g., Vassalou & Xing, 2004; Chava & Purnanandam, 2010).

Next, after the first split, our framework determines the optimal second split. This procedure begins by evaluating the split criterion in Eq. (6) for each candidate split. In total there are $53 \times 2 \times 2$ candidates—namely, 53×2 possible splits for both leaf nodes N2 and N3 in Figure 4. The candidate that maximizes the split criterion is chosen as the second split. In Figure 4, we observe that this corresponds to node N2, which is split on previous month's bond return (`STREVB`; short-term bond reversal) according to the rule: `STREVB` ≤ 0.2. Observations with high lagged returns (i.e., `STREVB` > 0.2) are more predictable, as the resulting child node (N5) achieves an R^2 of 4.94%. All other observations in node N2 are assigned to the left child node (N4), for which the clusterwise predictive model estimates an R^2 of 0.46%.

Moreover, to illustrate the overall heterogeneity between `STREVB` and return predictability estimated by our fully grown panel tree, we examine the heatmap in the top-left panel of Figure 5. We observe a strong positive relationship between return predictability and previous month's bond return. Furthermore, the bottom-right panel of Figure 5 illustrates the differences in average R^2 estimates when in-sample observations are grouped into bivariately sorted `STREVB`–`D2D` deciles (i.e., the first two splitting variables). We find that corporate bonds with both a low distance-to-default *and* a low previous-month return have a substantially lower R^2, indicating that the returns of these bonds are more difficult

[21] Recently, Bao et al. (2023) construct a global measure of systematic default risk, and show that it is able to predict future excess corporate bond and equity index returns, both in- and out-of-sample. We contribute to this study by showing that corporate bonds with higher distance-to-default are more predictable.

to predict. This highlights that heterogeneity in bond return predictability depends not only on individual characteristics but also on their interactions.

These findings can be motivated from both theoretical perspectives and related literature. Corporate bonds are traded over-the-counter and are therefore typically illiquid. Low previous-month returns increase the holding costs of corporate bonds, making them more sensitive to shocks in liquidity provision. For instance, a temporary negative liquidity shock raises holding costs further, which may induce investors to sell bonds at a discount relative to their fundamental value based on lagged asset characteristics. This mechanism can explain the weaker return predictability observed in lower `STREVB` deciles. Additionally, it particularly explains the why corporate bonds with both low distance-to-default *and* low previous-month returns are substantially less predictable. More specifically, Chen, Cui, He and Milbradt (2018) introduce the *liquidity-default spiral*: when default risk is high, even a small liquidity shock raises holding costs, which in turn increases default risk and again reduces liquidity. Therefore, especially for corporate bonds with a high default risk (i.e,. low distance-to-default) and low returns—which therefore already have a high holding cost—a relatively small liquidity shock can lead to a fire sale and, thus, a weak association between bond returns and lagged asset characteristics. For instance, this scenario could be motivated from a regulatory perspective: distressed bonds with low returns are often linked to credit rating downgrades, and insurance companies facing regulatory constraints are forced to sell downgraded bonds. Such fire sales influence bond returns independently of fundamentals, reducing return predictability (Ellul, Jotikasthira & Lundblad, 2011; Feldhütter, 2012; Wang, Zhang & Zhang, 2020).

Furthermore, the bottom-right panel of Figure 5 would also be the perfect explanation for why prior literature only found short-term reversals in illiquid bonds, and not in liquid corporate bonds, if it were not that these findings are questioned by recent literature. More specifically, Bali et al. (2017) show that a strategy that goes long corporate bonds with low previous-month returns and short bonds with high previous-month returns earns economically and statistically significant average excess returns in the sample of illiquid bonds, but these trading profits are economically and statistically insignificant in the sample of very liquid bonds. The bottom-right panel of Figure 5 could be used to motivate this result from a risk-return based perspective. Namely, the figure shows that the returns of *illiquid* corporate bonds, with low distance-to-default, and low previous-month returns are less predictable than illiquid bonds with high previous-month returns. Therefore, an investors that holds a long (short) position in these less (more) predictable bonds is subject to a higher level of risk, and should therefore command a higher return. On the other hand, the *liquid* bonds, with high distance-to-default, and low previous-month returns are not riskier than those with high previous-month returns, since their level of return predictability is the same. This means that an investor does not command an positive return when following a short-term reversal strategy. Nevertheless, in a recent

working paper, Dickerson, Robotti and Rossetti (2023) show that these return-based anomalies documented in the corporate bond literature are completely attributable to the incorrect handling of MMN in transaction-based TRACE data and ex-post return winsorization. Also we find that corporate bonds with lower levels of return predictability do not command per se higher return (see Section 5.2). If there is even a relationship between investment outcomes and return predictability, then this relationship seems to be positive, which is consistent with the findings of Dickerson, Robotti and Rossetti (2023).

Finally, the bottom-left panel of Figure 5 depicts the average R^2 values for observations sorted into `RATING` deciles each year. We observe no clear relationship between return predictability and bond rating. This finding may appear unexpected given the relationship between distance-to-default and predictability discussed above. A likely explation, however, is that bond ratings are updated only infrequently, so bonds with the same rating can nevertheless have very different *current* default risk (Hilscher & Wilson, 2017). In contrast, distance-to-default is a market-implied measure that reacts to immediate changes in the firm's capital structure and return volatility.

Besides that, the lack of a relationship between bond ratings and return predictability raises questions about the effectiveness of evaluating novel predictors and models on rating-sorted groups of corporate bonds. It is common in the literature to assess new predictive models on such groups as a robustness check, to demonstrate that predictors can improve return forecasts for both low-rated as well as high-rated corporate bonds (e.g., Lin, Wu & Zhou, 2013; Lin et al., 2018; Feng, He, Wang & Wu, 2025). However, the heatmap in the bottom-left panel of Figure 5 suggests that this may not be a sufficiently stringent test for robustness, since our Ridge regressions predict the returns of low- and high-rated bonds with similar accuracy. This means that we do not expect return predictability to vary over rating-sorted groups in the first place. In this respect, our results complement the findings of Elton, Gruber, Agrawal and Mann (2004), who show that grouping bonds by rating is not equivalent to grouping them by risk, leading to biased bond pricing models when researchers do make this assumption. We extend this argument by showing that grouping bonds by rating is also not equivalent to grouping them by return predictability.

5.2 Predictive performance of clusterwise versus global models

In addition to the clusters produced by the panel tree in Figure 4, our framework also yields a set of predictive models—one for each cluster. In this section, we compare the predictive performance of these clusterwise models against the global model fitted to all observations grouped together. These results are presented in Table 2. Panel A reports some summary statistics for each cluster produced by the panel tree in Figure 4, which include: (i) the number of observations; (ii) the R^2 statistic estimated by the correspond-

Table 2
This table report summary statistics for each cluster identified by the panel tree in Figure 4. The sample period is from 2002M8 to 2015M8, and Ridge regressions are used as predictive models in the leaves of the tree. Panel A shows the number of observations in each leaf, the return predictability (R^2, in percentages) from cluster-specific (R^2_C) and global (R^2_G) predictive models, and their difference ($R^2_{CMG} = R^2_C - R^2_G$). Panel B reports the average monthly excess return (in percentages) and the annualized Sharpe ratio for the equal-weighted (EW) and value-weighted (VW) portfolios of clustered asset-return observations. Rows are sorted in descending order of R^2_C.

	Panel A				Panel B			
Leaf	Obs.	R^2_C	R^2_G	R^2_{CMG}	Avg_{EW}	SR_{EW}	Avg_{VW}	SR_{VW}
N22	26,603	11.14	4.85	6.29	0.64	0.99	0.52	0.80
N14	31,338	9.34	3.50	5.85	0.53	0.97	0.45	0.74
N31	18,493	8.33	4.40	3.93	0.64	1.42	0.59	1.24
N25	40,778	8.11	4.04	4.07	0.31	0.93	0.30	0.89
N17	26,223	6.13	1.55	4.57	0.58	0.80	0.46	0.82
N24	48,610	3.45	2.39	1.06	0.28	0.95	0.28	0.92
N18	26,588	3.39	1.66	1.74	0.44	0.74	0.36	0.63
N23	37,630	3.27	2.47	0.80	0.53	1.18	0.48	0.97
N10	40,309	3.00	2.19	0.81	0.79	1.03	0.61	0.87
N30	21,377	2.51	2.45	0.06	0.56	1.05	0.56	0.98
N13	40,955	1.54	1.43	0.10	0.48	0.63	0.47	0.60
N16	31,220	0.46	0.55	-0.09	0.27	0.53	0.29	0.57
N19	63,091	0.23	0.06	0.17	0.38	0.47	0.37	0.55

ing clusterwise model in percentages, denoted R^2_C; (iii) the R^2 statistic estimated by the global predictive model, denoted R^2_G; and (iv) the difference between R^2_C and R^2_G, denoted R^2_{CMG}.[22]

We find substantial heterogeneity in bond return predictability across clusters. The most predictable cluster (N22) attains an R^2_C of 11.14%, whereas the least predictable cluster (N19) has an R^2_C of only 0.23%. Furthermore, as the rows in Table 2 are sorted in descending order of R^2_C, we can conclude that, on average, the levels of return predictability under the global predictive model (R^2_G) align with the clusterwise estimates (R^2_C). This alignment suggests that our clustering procedure effectively groups corporate bonds by their degree of predictability, and supports the view that return predictability itself can be regarded as an asset characteristic.

In order to examine the relationship between predictability and investment outcomes, we construct equal- and value-weighted portfolios for each cluster. Panel B of Table 2 reports the average monthly excess returns (in percentages) and annualized Sharpe ratios for these portfolios. The most predictable cluster (N22) achieves average monthly excess

[22]In other words, $R^2_{C,j}$ is the R^2 statistic estimated by a model fitted to all observations in cluster j, whereas $R^2_{G,j}$ is estimated by a model fitted to all observations but evaluated only on cluster j.

returns of 0.64% and 0.52% for the equal- and value-weighted portfolio, respectively, with corresponding Sharpe ratios of 0.99 and 0.80. These outcomes exceed those of the least predictable cluster (N19), which delivers average excess returns of 0.38% and 0.37% for the equal- and value-weighted portfolios, respectively, with Sharpe ratios of 0.47 and 0.55. Nevertheless, higher return predictability does not always translate into superior investment performance. More specifically, we observe that more predictable clusters do not consistently generate higher returns or Sharpe ratios. Instead, Table 2 suggests that the three least predictable clusters—those with R^2 values below the global estimate of 1.61% for the root node (N1) in Figure 4—stand out for their substantially weaker investment outcomes.

Moreover, Panel A of Table 2 reports the difference between the clusterwise and global predictability measures, R^2_{CMG}, for each cluster. This statistic captures the extent of disagreement between a cluster-specific and a global predictive model. Comparing these values with the investment results in Panel B, we find that the average monthly excess returns and annualized Sharpe ratios are higher for the cluster with the largest predictability disagreement than for the cluster with the smallest R^2_{CMG}. In the equity market, Cong et al. (2024) document a novel anomaly associated with such return predictability disagreement. Specifically, they construct long-short portfolios that go long (short) the cluster of stocks with the largest (smallest) disagreement. These portfolios deliver statistically significant abnormal returns even after controlling for common risk factors. In this paper, we examine whether a similar anomaly exists in the corporate bond market. The corresponding results are presented in Section 5.3.

5.2.1 Out-of-sample predictive performance

So far, we have focused on heterogeneity in *in-sample* return predictability. Ideally, however, the tree-based clusters should also capture heterogeneity *out-of-sample*—that is, the most predictable clusters in-sample should remain the most predictable when evaluated on new data. To test this, we evaluate the out-of-sample performance of our clustering procedure in this section. We use the panel tree in Figure 4, estimated on the training sample (2002M8–2015M8), to assign bond-return observations in the test sample (2015M9–2022M8) to clusters. Each cluster is associated with a cluster-specific model fitted on the training data, and we also estimate a global Ridge regression on all in-sample observations (i.e., without clustering). Both the global and clusterwise models are then used separately to produce forecasts, which allow us to compute clusterwise levels of return predictability using Eq. (4).

Table 3 presents the clusterwise R^2 estimates (in percentages) for both the in- and out-of-sample periods. We evaluate the predictive performance of our models by grouping the clusters in Figure 4 according to their level of in-sample return predictability. Specifically, the *High* group contains the three clusters with the highest R^2_C, the *Low* group contains

Table 3
This table presents R^2 values (in percentages; Eq. (4)) estimated from the global predictive model and clusterwise models from the cross-sectional panel tree in Figure 4, both in- and out-of-sample. More specifically, we define four different samples of observations: (i) *Aggregate*, all forecasts combined; (ii) *High*, the forecasts for observations ending up in clusters N22, N14 or N31 in Figure 4 (i.e., with highest in-sample R^2_C); (iii) *Low*, the forecasts for observations ending up in the clusters N19, N16 or N13 (i.e., with lowest in-sample R^2_C); and (iv) *Medium*, the forecasts for all observations not ending up in either sample *High* or *Low*. Panel A presents the R^2 values estimated by a Ridge regression fitted to all in-sample observations grouped together (i.e., a global model), but evaluated on all samples separately to produce the forecasts. Panel B reports similar results, but now the clusterwise models—Ridge regressions fitted to all observations in the same cluster—are used to produce the forecasts.

	In-sample	Out-of-sample		
	2002M8–2015M8	2015M9–2022M8	2015M9–2021M8	2021M9–2022M8
Panel A: Global forecasts				
Aggregate	1.61	-0.71	1.75	-13.80
High	4.30	0.36	4.39	-14.70
Medium	2.13	-1.15	2.58	-16.31
Low	0.26	-0.70	0.62	-10.55
Panel B: Clusterwise forecasts				
Aggregate	3.16	-0.08	2.29	-12.67
High	10.07	1.48	5.73	-14.42
Medium	3.73	-0.34	3.79	-17.16
Low	0.39	-0.32	0.55	-6.78

the three clusters with the lowest R^2_C, and the remaining clusters form the *Medium* group. The *Aggregate* sample includes all observations grouped together. Furthermore, Panel A presents the results when forecasts are generated by the global Ridge regression, evaluated separately on each group. Similarly, Panel B shows the results when forecasts are generated by clusterwise Ridge regressions.

First, we observe that the in-sample results confirm that predictability is aligned with the cluster groupings—a higher predictable cluster has a higher in-sample R^2—both under the global predictive model (Panel A) and clusterwise models (Panel B). Moreover, we observe that the aggregate R^2 is higher for the clusterwise models (3.16%) than for the global model (1.61%). This result is expected since we use 13 different clusterwise models to estimate the in-sample R^2 in Panel B, while only a single Ridge regression in Panel A.

Besides that, also out-of-sample, the clusterwise predictive models forecast returns better than the global model (-0.08% vs. -0.71%; second column). However, these out-of-sample R^2 estimates are close to zero, and even negative—indicating that zero-return forecasts are more accurate. In Table 3, we further observe that the ordering on predictability does not align anymore with the in-sample results. While the most predictable cluster remains most predictable out-of-sample, the relative rankings of the *Medium* and

Low groups is reversed. This result would imply that our tree-based clustering procedure has low accuracy out-of-sample, and would question the results documented in other section. Therefore, we examine the source of the poor performance and find that it is entirely attributable to the final year of our 7-year out-of-sample period.[23]

More specifically, in the last two columns of Table 3, we split the out-of-sample period into two subperiods: 2015M9–2021M8 and 2021M9–2020M8. For the longer subperiod up to 2021M8, we observe that the aggregate R^2 estimated by the clusterwise predictive models (2.29%) is higher than that estimated by the global Ridge regression (1.75%), and now both of these measures are also positive. We further document that the levels of return predictability align with the in-sample estimates—the most (least) predictable cluster in-sample is also the most (least) predictable cluster out-of-sample. Therefore, we conclude that the deterioration in out-of-sample performance is entirely driven by the final year (2021M9–2022M8), during which all R^2 values become sharply negative (last column). This indicates that our clustering procedure performs well under normal conditions, but fails during the most recent year.

There are several reasons for the poor out-of-sample performance of our tree-based clustering procedure over the period from 2021M9 to 2022M8. First, the Chinese property crisis began in 2021, as Evergrande started missing bond payments in September (Altman, Hu & Yu, 2022). Second, inflation surged in 2021 and 2022, prompting the Federal Reserve to raise interest rates for the first time since 2018, beginning in March 2022. This subsequently led to rising yields and financing costs for corporates, which heightened investors' concerns about defaults (OECD, 2022). For instance, Bao et al. (2023) show that their measure of systematic default risk—the probability that at least 2% of firms default—peaked during 2021 and 2022, at levels unseen since the start of their sample in 1961M3.[24] All in all, these two developments created an unusually adverse environment for corporate bonds, which explains the breakdown in predictability. At the same time, it shows that our tree-based clustering procedure is not able to account for regime-switches in return predictability, and motivates the construction of a regime-switching panel tree in Section 5.4.

5.3 An anomaly related to predictability disagreement

As mentioned in Section 2, most studies forecast returns using a global predictive model estimated on the pooled sample, rather than clusterwise models that allow for heterogeneity in return predictability across bonds. Table 2 shows, however, that clusterwise models deliver substantially higher in-sample R^2, suggesting that global models are misspecified.

[23]Note that we do not re-estimate the predictive models in our out-of-sample period, in order to highlight the good out-of-sample performance, even without re-estimation.

[24]We contacted Bao et al. (2023) to request their data but did not receive a response. We recommend consulting Figure IA.1 in the online appendix of their paper.

Because global models are the dominant approach among researchers and practitioners, investors may face the risk that information is not efficiently incorporated into bond prices. In particular, we expect that corporate bonds for which clusterwise and global models disagree most strongly are subject to greater model-misspecification risk, and therefore demand higher returns. In this section, we investigate whether this *predictability disagreement anomaly* is present in the corporate bond market, by constructing long-short portfolios in the spirit of Cong et al. (2024).

We measure predictability disagreement by R^2_{CMG}, defined as the difference between the R^2 of a clusterwise model (R^2_C) and a global model (R^2_G). To compute this measure, we cluster bond-return observations using the panel tree in Figure 4, and calculate R^2_C, R^2_G, and R^2_{CMG} for each cluster. All bonds within the same cluster are assigned the same predictability disagreement, but a bond's R^2_{CMG} can vary over time as changes in its characteristics could lead to different cluster assignments (see Section 3).

Table 2 shows that clusters with higher R^2_{CMG} earn higher returns. For example, the average monthly excess return and annualized Sharpe ratio of the cluster with the highest disagreement (N22) are nearly twice as large as those of the cluster with the lowest disagreement (N16). This finding suggests that investors are compensated for bearing model-misspecification risk. To study this more formally, and also out-of-sample, we construct long-short portfolios similar to Cong et al. (2024), which takes an equal-weighted long position in the top i clusters with the highest R^2_{CMG} and a short position in the bottom i clusters with the lowest disagreement, for $i \in \{1, 2, 3, 5\}$. The resulting portfolios are denoted by Ti–Bi.

Panel A of Table 4 reports some summary statistics for these portfolios, which include: (i) the average monthly excess return (Avg, in percentages), (ii) the standard deviation of returns (Std, in percentages), and (iii) the corresponding annualized Sharpe ratios (SR). These statistics are shown for both the in-sample period (2002M8–2015M8) and the out-of-sample period (2015M9–2022M8). The out-of-sample results are obtained using the panel tree in Figure 4, which is fitted to the in-sample data. This tree partitions the out-of-sample observations into clusters, from which value-weighted returns and the corresponding long-short portfolios are constructed.

We observe that portfolios with a larger spread between the long and short position (e.g., T1–B1) outperform those with smaller spreads (e.g., T5–B5), both in- and out-of-sample.[25] For instance, the out-of-sample average monthly excess return (annualized SR) of T1–B1 is 0.265% (0.705), compared to 0.091% (0.470) for T2–B2. Furthermore, we highlight that the investment outcomes are economically significant—in particular the outcomes for T1–B1. For instance, the out-of-sample Sharpe ratio for T1–B1 (0.705) is more than 3.5 times that of the value-weighted bond market portfolio (0.186), and even

[25] There is one exception for the in-sample period, where T5–B5 has a slightly higher annualized Sharpe ratio (0.198) than T3–B3 (0.185).

Table 4

This table reports summary statistics (Panel A) and abnormal returns (Panel B) for long–short portfolios based on the return-predictability disagreement measure R^2_{CMG} from Table 2. Clusters are first sorted in descending order of R^2_{CMG}, and value-weighted portfolios are constructed for each cluster. For each $i \in \{1, 2, 3, 5\}$, we form an equal-weighted long position in the top i clusters (highest predictability disagreement) and an equal-weighted short position in the bottom i clusters (lowest predictability disagreement), yielding the Ti–Bi long–short portfolio. Clusters are estimated using in-sample data from 2002M8 to 2015M8 and correspond to the panel tree in Figure 4. Results are shown separately for the in-sample and out-of-sample periods. Panel A reports the average monthly excess return (Avg, in percentages), standard deviation (Std, in percentages), and annualized Sharpe ratio. Panel B reports abnormal returns (alpha, in percentages) from the factor models in Appendix D, with Newey and West (1987) t-statistics in parentheses (lag length of 3). Statistical significance at the 5% and 10% levels is denoted by $**$ and $*$, respectively.

	In-Sample (2002M8 – 2015M8)				Out-of-Sample (2015M9 – 2022M8)			
	T1–B1	T2–B2	T3–B3	T5–B5	T1–B1	T2–B2	T3–B3	T5–B5
Panel A: Summary Statistics								
Avg	0.234	0.064	0.039	0.031	0.265	0.091	0.062	0.000
Std	1.856	1.040	0.730	0.546	1.304	0.674	0.660	0.391
SR	0.437	0.213	0.185	0.198	0.705	0.470	0.326	0.000
Panel B: Abnormal Returns								
CAPMB	0.104	-0.008	0.033	0.086**	0.259	0.099	0.075	0.016
	(0.584)	(-0.104)	(0.420)	(2.035)	(1.443)	(1.152)	(0.907)	(0.665)
DEFTERM	0.154	-0.008	0.080	0.056*	0.194	0.085	0.057	0.033
	(0.961)	(-0.096)	(1.264)	(1.819)	(1.219)	(1.061)	(0.970)	(1.290)
CAPM	0.130	0.029	-0.018	0.044	0.242	0.117	0.075	0.050
	(0.842)	(0.378)	(-0.277)	(1.155)	(1.382)	(1.339)	(0.757)	(1.616)
HKM	0.135	0.027	0.001	0.050	0.264*	0.128*	0.094	0.054**
	(0.906)	(0.361)	(0.019)	(1.309)	(1.830)	(1.823)	(1.287)	(2.150)
HKMSF	0.190	0.050	0.009	0.034	0.229	0.088	0.046	0.015
	(1.419)	(0.756)	(0.154)	(1.009)	(1.358)	(1.020)	(0.543)	(0.392)
FF3	0.146	0.038	-0.013	0.047	0.256*	0.124*	0.092	0.052**
	(1.014)	(0.546)	(-0.219)	(1.263)	(1.723)	(1.765)	(1.387)	(1.976)
FF5	0.138	0.032	-0.004	0.041	0.279**	0.134**	0.104*	0.052**
	(1.015)	(0.438)	(-0.083)	(1.057)	(2.058)	(2.044)	(1.732)	(2.052)
FF5+MOMS	0.039	-0.218	-0.461**	-0.014	0.565	0.215	0.081	-0.020
+IVOL	(0.081)	(-0.724)	(-2.067)	(-0.096)	(0.921)	(0.750)	(0.334)	(-0.167)

exceeds that of the value-weighted stock market portfolio (0.671).

Moreover, in Panel B of Table 4, we compute the abnormal returns (i.e., alphas; in percentages) using popular factor models, with corresponding Newey and West (1987) t-statistics in parentheses. These models include: the Fama-French three-factor model (FF3; Fama & French, 1993) and five-factor model (FF5; Fama & French, 2015), the capital asset pricing model (CAPM; Sharpe, 1964) and the bond CAPM (CAPMB; Dickerson,

Mueller & Robotti, 2023), the Fama-French bond factor model (DEFTERM; Fama & French, 1993), the intermediary capital asset pricing model (HKM; He, Kelly & Manela, 2017) and its single-factor equivalent (HKMSF), and the FF5 model with the stock momentum factor (MOMS; Carhart, 1997) and the idiosyncratic volatility factor (IVOL; Campbell & Taksler, 2003). Appendix D provides more details on these factors, and its data collection procedure.

First, we observe that almost all abnormal returns are positive out-of-sample, and even larger than the corresponding average monthly excess returns in Panel A. In-sample, however, the results are a bit more mixed. On the other hand, we find that most abnormal returns—both in- and out-of-sample—are statistically insignificant at the 5% level. This finding is opposite to the strongly significant returns for US equities reported by Cong et al. (2024). That return-based anomalies for equities, such as momentum and short- and long-term reversals, do not always translate to the corporate bond market is not a surprise. While prior literature claimed that these return-based anomalies also earned large and significant abnormal returns in the corporate bond market,[26] a recent working paper by Dickerson, Robotti and Rossetti (2023) shows that these findings are almost completely attributable to the ex-post winsorization of returns and improper handling of MMN in TRACE-based transaction data. After addressing these issues, Dickerson, Robotti and Rossetti (2023) find that return-based anomalies generate negligible abnormal returns in the corporate bond market. Our findings contribute to this literature by showing that the return-predictability disagreement anomaly likewise fails to earn statistically significant abnormal returns.

5.4 Time-varying return predictability

In this section, we investigate the time-varying patterns in return predictability by considering macroeconomic indicators as potential splitting variables in *regime-switching* panel trees, in addition to the cross-sectional asset characteristics. These macroeconomic variables—defined in Table 1—partition the panel data along the time dimension. For this reason, and because our sample period is relatively short, we estimate a regime-switching panel tree on the full sample (2002M8–2022M8). Additionally, we increase the minimum number of monthly observations in each leaf node to 200 and allow the tree to split only once on a macroeconomic variable. Any further splits are restricted to the cross-sectional characteristics listed in Table 5 of Appendix C. The resulting panel tree is show in Figure 6.

[26] See, e.g., Jostova et al. (2013); Bali et al. (2017, 2021).

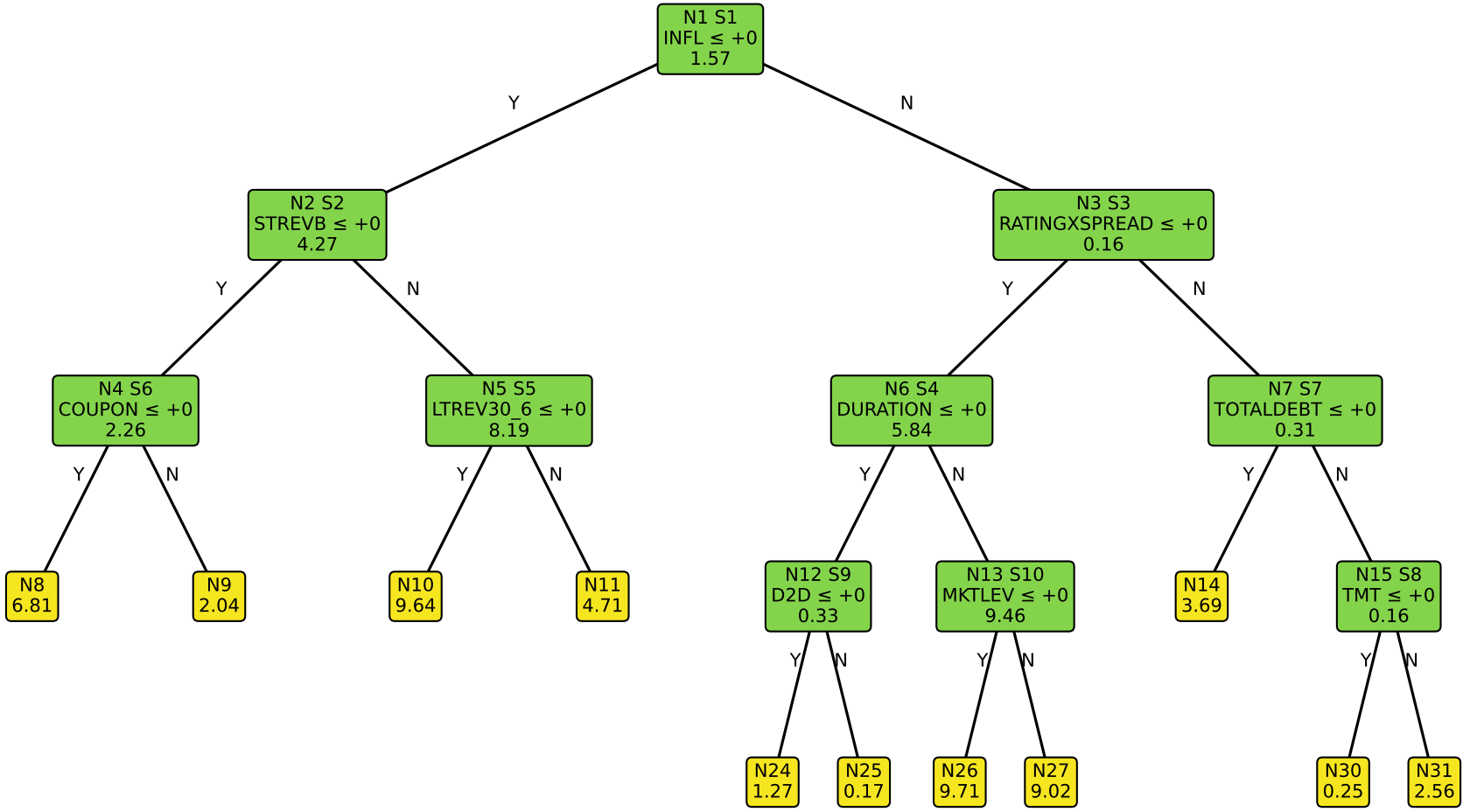


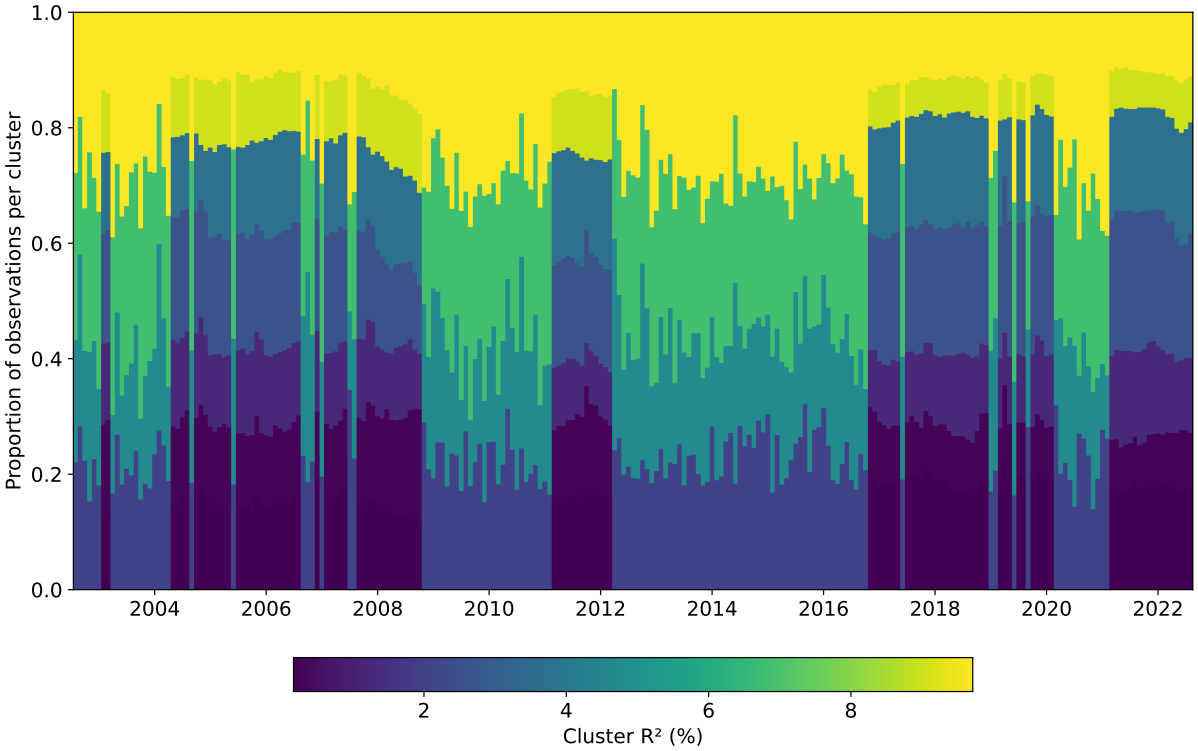
Figure 6: This figure illustrates the panel tree used to form clusters of bond-return observations with heterogeneous levels of predictability, measured by R^2. The tree is estimated using monthly data from 2002M8 to 2022M8, and splits on both cross-sectional asset characteristics (Table 5) and time-varying macroeconomic variables (Table 1). Each intermediate node (green) and leaf node (yellow) displays the node's ID, denoted by $N\#$, which is used to identify the node in this paper. Intermediate nodes additionally report the split rule (second line), which sends observations satisfying the rule to the left child node and the rest to the right child node, as well as the node's split order, denoted by $S\#$. Finally, all nodes display the in-sample R^2 statistic (in percentages; last line), estimated using the clusterwise Ridge regressions introduced in Section 3.

We observe that the first split separates the time dimension into a high- and low-inflation regime, using the rule: $\texttt{INFL} \leq 0$. Corporate bonds seem to be more predictable in the low-inflation regime. On the other hand, we find that the heterogeneity in return predictability is larger during periods of high inflation. More specifically, by considering the leaf nodes in Figure 6, we observe that the lowest (highest) R^2_C for the low-inflation regime is 2.04% (9.64%), but 0.17% (9.71%) for the high-inflation regime. Thus, return predictability varies not only across corporate bonds, but also across macroeconomic regimes.

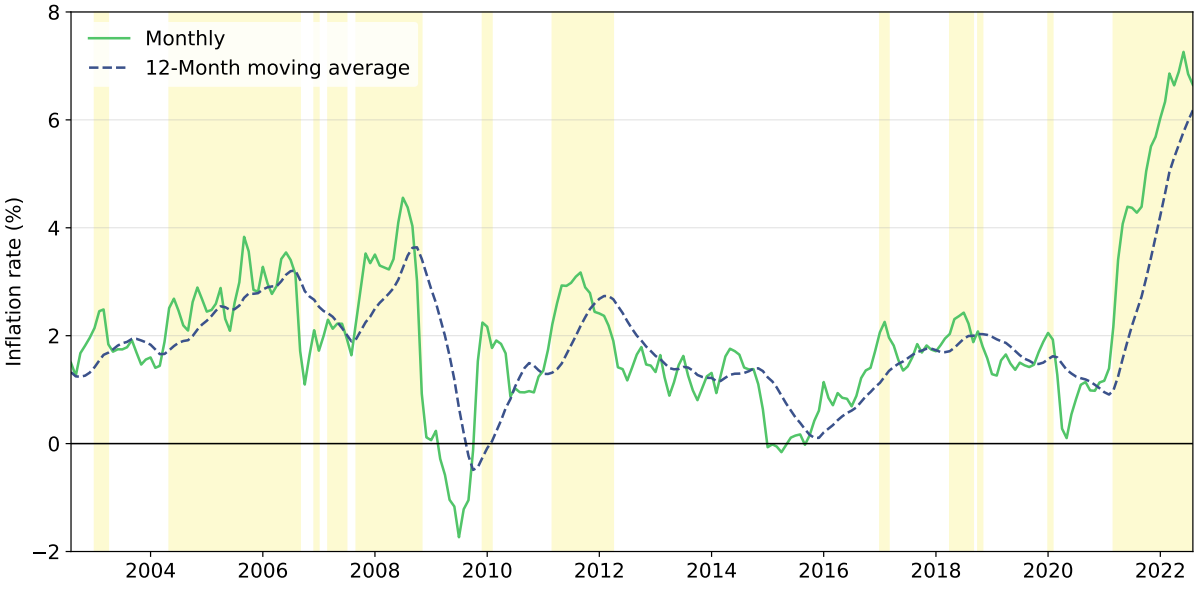
Moreover, the heatmap in Figure 7a illustrates the evolution of return predictability over time, estimated from the panel tree in Figure 6. The colors indicate the level of predictability, with lighter colors corresponding to higher R^2_C values. Furthermore, the height of each region reflects the proportion of observations with the corresponding level of return predictability. We observe that predictability declines during some major historical events in the corporate bond market. For instance, return predictability declined during the Global Financial Crisis (2007–2009), the European Sovereign Debt Crisis (2011–2012), the US-China trade war (2018), and the surge in inflation during 2021–2022 following pandemic-related supply chain disruptions and monetary stimulus (Reis, 2022).

As we mentioned above, the regimes are identified by the split rule: $\texttt{INFL} \leq 0$, which holds when the 12-month moving-average inflation is lower than its 10-year historical median value, as discussed in Section 4. Since it might be difficult to interpret when this regime is actually reached in practice, we plot the monthly inflation series, together with its 12-month moving average, in Figure 7b. By comparing these inflation series with the heatmap in Figure 7a, we find indeed that return predictability is lower when inflation is high. Interestingly, we observe that the regimes in the heatmap almost perfectly align with the shaded regions in Figure 7b—which highlights regions where the monthly inflation series exceeds its 2% target. In other words, we could alternatively say that corporate bonds are less predictable when inflation exceeds, or is close to exceeding, its 2% target.

That corporate bond returns are less predictable when inflation is high can potentially be explained by the increased uncertainty about future monetary policy, in particular interest rate rises. Clearly, when inflation rises—especially exceeding its long-term target of 2%—it becomes more likely that central banks need to intervene by increasing interest rates. For instance, in Figure 10 of Appendix F, we observe that periods with high interest rates are closely aligned with the low-predictability regimes in Figure 7a. Corporate bonds typically deliver fixed nominal coupons. Therefore, when inflation is high, the real value of these payments becomes highly uncertain (Kang & Pflueger, 2015). This implies that the link between monthly excess corporate bond returns and lagged asset characteristics weakens, and return predictability declines. For example, Campbell and Ammer (1993) decompose the excess returns of long-term bonds using a vector autoregressive model, and find that they are mainly driven by *news* about inflation.



(a) Time-varying return predictability for corporate bonds



(b) Inflation

Figure 7: Panel (a) of this figure illustrates the time variation in corporate bond return predictability estimated by the regime-switching panel tree in Figure 6. Colors indicate the level of predictability measured by R^2 (in percentages). The height of each region reflects the relative number of observations in each cluster over time. Panel (b) shows the monthly inflation series (solid line) together with its 12-month moving average (`INFL`; dashed). Shaded areas indicate periods when monthly inflation exceeds the 2% target.

Another explanation is again related to liquidity. We observe in Figure 7a that predictability declines during some stress periods, such as the Global Financial Crisis. These periods are associated with high illiquidity (Bao, O'Hara & Zhou, 2018; Friewald, Jankowitsch & Subrahmanyam, 2012). Following the same motivation as in Section 5.1, when illiquidity is high, corporate bonds returns can for a large part be explained by trading frictions in the over-the-counter corporate bond market rather than underlying fundamentals (e.g., Bao, Pan & Wang, 2011; Dick-Nielsen, Feldhütter & Lando, 2012).

While it is often argued that equity returns are more predictable during periods of high uncertainty and low economic activity (e.g., Dangl & Halling, 2012; Avramov, 2002; Cong et al., 2024), the evidence for (corporate) bonds seems to be more mixed. For example, Borup et al. (2024) show that US Treasury bonds are more predictable when economic activity is strong and uncertainty is low. Feng, He, Wang and Wu (2025) find that corporate bond returns are more predictable in bearish states of the economy—characterized by low liquidity, high stock market variance, and elevated risk aversion as measured by the risk-aversion index of Bekaert, Engstrom and Xu (2022). In contrast, we find that the predictability of corporate bond returns declines during periods of high inflation, which are typically associated with low liquidity and high stock market variance (see Figure 8 of Appendix D). However, when we take a closer look at the results of Feng, He, Wang and Wu (2025), we notice something interesting.[27] While Feng, He, Wang and Wu (2025) do not consider inflation as a regime variable, they do analyze the three-month US Treasury bill rate (`TBL`)—which is highly positively correlated with inflation—and show that the out-of-sample (OOS) R^2 is substantially larger when `TBL` is below its median (5.48%) than when it is above (3.81%). On the other hand, the spread in OOS R^2 between the low- and high-regime is considerably smaller for liquidity (4.61% vs. 4.67%) and the stock market variance (4.56% vs. 4.72%). This raises the question of whether the results of Feng, He, Wang and Wu (2025) are best interpreted as evidence that return predictability is higher in low-liquidity and high-volatility states, since their results indicate that the heterogeneity in return predictability is larger for different `TBL`-regimes. Instead, it would make more sense when Feng, He, Wang and Wu (2025) concluded that corporate bond returns are more predictable when `TBL` is low (i.e., below its median value). Since inflation and `TBL` are highly correlated, this would lead to the conclusion that corporate bonds are more predictable when inflation is low, which is consistent with our results.[28]

[27] We are particularly referring to Table 8 of Feng, He, Wang and Wu (2025, p.12).

[28] We also consider `TBL` as a potential macroeconomic splitting variable (Table 1). However, this variable is likely not selected by our regime-switching panel tree in Figure 6, as it does not capture the European Sovereign Debt Crisis (2011–2012). In contrast, this period is captured by an increase in inflation. This can be concluded after comparing the time-series plots of `TBL` and `INFL` in Figure 8 of Appendix D.

6 Conclusion

We study the cross-sectional and regime-switching patterns (i.e., *mosaics*) in the predictability of corporate bond returns. In contrast to prior literature, we assume that return predictability is a characteristic of corporate bonds—such as return volatility, which varies over time and cross-section. We cluster corporate bonds using the tree-based clustering procedure proposed by Cong et al. (2024), where 53 cross-sectional asset characteristics and eight time-varying macroeconomic predictors are considered as potential splitting variables.

Our findings are threefold. First, we show that corporate bonds with both low distance-to-default and low previous-month returns are substantially less predictable. Second, we find that the *predictability disagreement anomaly* of Cong et al. (2024) does not earn statistically significant abnormal returns in the corporate bond market, after accounting for popular bond and equity risk factors. Third, we show that corporate bonds exhibit weaker return predictability during periods of high inflation.

Finally, an interesting direction for future research is to explore different sources of forecast disagreement. Our analysis follows Cong et al. (2024) by focusing on disagreement arising from models trained on different datasets (i.e., cluster-specific versus global models), whereas Bali et al. (2023) emphasize an alternative dimension based on variation across machine learning model classes trained on the same data. Despite their opposing long-short portfolio strategy—the former longs high-disagreement stocks, while the latter longs low-disagreement stocks—both strategies deliver significant abnormal returns. An empirical comparison of these two forms of disagreement (i.e., dataset versus model class) would therefore be an interesting avenue for future research.

References

Altman, E. I., Hu, X. & Yu, J. (2022). Has the Evergrande debt crisis rattled Chinese capital markets? A series of event studies and their implications. *Finance Research Letters*, *50*, 103247.

Andreani, M., Palhares, D. & Richardson, S. (2024). Computing corporate bond returns: A word (or two) of caution. *Review of Accounting Studies*, *29*(4), 3887–3906.

Ang, A. & Bekaert, G. (2007). Stock return predictability: Is it there? *The Review of Financial Studies*, *20*(3), 651–707.

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

Asness, C. S., Moskowitz, T. J. & Pedersen, L. H. (2013). Value and momentum everywhere. *The Journal of Finance*, *68*(3), 929–985.

Avramov, D. (2002). Stock return predictability and model uncertainty. *Journal of Financial Economics*, *64*(3), 423–458.

Avramov, D., Cheng, S. & Metzker, L. (2023). Machine learning vs. economic restrictions: Evidence from stock return predictability. *Management Science*, *69*(5), 2587–2619.

Bai, J., Bali, T. G. & Wen, Q. (2019). Common risk factors in the cross-section of corporate bond returns. *Journal of Financial Economics*, *131*, 619–642.

Bali, T. G., Goyal, A., Huang, D., Jiang, F. & Wen, Q. (2020). Predicting corporate bond returns: Merton meets machine learning. *Georgetown McDonough School of Business Research Paper*, 20–110.

Bali, T. G., Kelly, B. T., Mörke, M. & Rahman, J. (2023). Machine forecast disagreement [Working Paper]. *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w31583 doi: 10.3386/w31583

Bali, T. G., Subrahmanyam, A. & Wen, Q. (2017). Return-based factors for corporate bonds. *Available at SSRN*.

Bali, T. G., Subrahmanyam, A. & Wen, Q. (2021). Long-term reversals in the corporate bond market. *Journal of Financial Economics*, *139*(2), 656–677.

Bao, J., Hou, K. & Zhang, S. (2023). Systematic default and return predictability in the stock and bond markets. *Journal of Financial Economics*, *149*(3), 349–377.

Bao, J., O'Hara, M. & Zhou, X. A. (2018). The Volcker Rule and corporate bond market making in times of stress. *Journal of Financial Economics*, *130*(1), 95–113.

Bao, J., Pan, J. & Wang, J. (2011). The illiquidity of corporate bonds. *The Journal of Finance*, *66*(3), 911–946.

Bartram, S. M., Grinblatt, M. & Nozawa, Y. (2025). Book-to-market, mispricing, and the cross section of corporate bond returns. *Journal of Financial and Quantitative Analysis*, *60*(3), 1185–1233.

Basu, S. (1977). Investment performance of common stocks in relation to their price-

earnings ratios: A test of the efficient market hypothesis. *The Journal of Finance*, *32*(3), 663–682.

Bekaert, G., Engstrom, E. C. & Xu, N. R. (2022). The time variation in risk appetite and uncertainty. *Management Science*, *68*(6), 3975–4004.

Bharath, S. T. & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *The Review of Financial Studies*, *21*(3), 1339–1369.

Bie, S., Diebold, F. X., He, J. & Li, J. (2024). Machine learning and the yield curve: Tree-based macroeconomic regime switching. *arXiv preprint arXiv:2408.12863*.

Borup, D., Eriksen, J. N., Kjær, M. M. & Thyrsgaard, M. (2024). Predicting bond return predictability. *Management Science*, *70*(2), 931–951.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. J. (2017). *Classification and regression trees*. Chapman and Hall/CRC.

Breiman, L. & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*(391), 580–598.

Campbell, J. Y. (1987). Stock returns and the term structure. *Journal of Financial Economics*, *18*(2), 373–399.

Campbell, J. Y. & Ammer, J. (1993). What moves the stock and bond markets? A variance decomposition for long-term asset returns. *The Journal of Finance*, *48*(1), 3–37.

Campbell, J. Y. & Taksler, G. B. (2003). Equity volatility and corporate bond yields. *The Journal of Finance*, *58*(6), 2321–2350.

Campbell, J. Y. & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies*, *21*(4), 1509–1531.

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

Chava, S. & Purnanandam, A. (2010). Is default risk negatively related to stock returns? *The Review of Financial Studies*, *23*(6), 2523–2559.

Chen, H., Cui, R., He, Z. & Milbradt, K. (2018). Quantifying liquidity and default risks of corporate bonds over the business cycle. *The Review of Financial Studies*, *31*(3), 852–897.

Choi, J. & Kim, Y. (2018). Anomalies and market (dis)integration. *Journal of Monetary Economics*, *100*, 16–34.

Chung, K. H., Wang, J. & Wu, C. (2019). Volatility and the cross-section of corporate bond returns. *Journal of Financial Economics*, *133*(2), 397–417.

Cong, L. W., Feng, G., He, J. & He, X. (2025). Growing the efficient frontier on panel trees. *Journal of Financial Economics*, *167*, 104024.

Cong, L. W., Feng, G., He, J. & Li, J. (2023). Sparse modeling under grouped heterogeneity with an application to asset pricing [Working Paper]. *National Bureau of Economic Research*. Retrieved from http://www.nber.org/papers/w31424 doi: 10.3386/w31424

Cong, L. W., Feng, G., He, J. & Wang, Y. (2024). Mosaics of predictability. *Available at SSRN*.

Correia, M., Richardson, S. & Tuna, İ. (2012). Value investing in credit markets. *Review of Accounting Studies*, *17*(3), 572–609.

Dangl, T. & Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, *106*(1), 157–181.

Dickerson, A., Julliard, C. & Mueller, P. (2024). The co-pricing factor zoo [Working Paper]. *UNSW Business School Research Paper*.

Dickerson, A., Mueller, P. & Robotti, C. (2023). Priced risk in corporate bonds. *Journal of Financial Economics*, *150*(2), 103707.

Dickerson, A., Robotti, C. & Nozawa, Y. (2024). Factor investing with delays [Working Paper]. *UNSW Business School Research Paper*.

Dickerson, A., Robotti, C. & Rossetti, G. (2023). Noisy prices and return-based anomalies in corporate bonds [Working Paper]. *UNSW Business School Research Paper*.

Dick-Nielsen, J. (2014). How to clean enhanced TRACE data [Working Paper]. *Copenhagen Business School*.

Dick-Nielsen, J., Feldhütter, P., Pedersen, L. H. & Stolborg, C. (2023). Corporate bond factors: Replication failures and a new framework. *Available at SSRN 4586652*.

Dick-Nielsen, J., Feldhütter, P. & Lando, D. (2012). Corporate bond liquidity before and after the onset of the subprime crisis. *Journal of Financial Economics*, *103*(3), 471-492.

Dor, A. B., Dynkin, L., Hyman, J., Houweling, P., van Leeuwen, E. & Penninga, O. (2007). Dtssm (duration times spread). *The Journal of Portfolio Management*, *33*(2), 77–100.

Duffie, D., Saita, L. & Wang, K. (2007). Multi-period corporate default prediction with stochastic covariates. *Journal of Financial Economics*, *83*(3), 635–665.

Elkamhi, R., Jo, C. & Nozawa, Y. (2024). A one-factor model of corporate bond premia. *Management Science*, *70*(3), 1875–1900.

Ellul, A., Jotikasthira, C. & Lundblad, C. T. (2011). Regulatory pressure and fire sales in the corporate bond market. *Journal of Financial Economics*, *101*(3), 596–620.

Elton, E. J., Gruber, M. J., Agrawal, D. & Mann, C. (2004). Factors affecting the valuation of corporate bonds. *Journal of Banking & Finance*, *28*(11), 2747–2767.

Evgeniou, T., Guecioueur, A. & Prieto, R. (2023). Uncovering sparsity and heterogeneity in firm-level return predictability using machine learning. *Journal of Financial and Quantitative Analysis*, *58*(8), 3384–3419.

Fama, E. F. & French, K. R. (1988). Dividend yields and expected stock returns. *Journal of Financial Economics*, *22*(1), 3–25.

Fama, E. F. & French, K. R. (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics*, *25*(1), 23–49.

Fama, E. F. & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, *33*(1), 3–56.

Fama, E. F. & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22.

Feldhütter, P. (2012). The same bond at different prices: Identifying search frictions and selling pressures. *The Review of Financial Studies*, *25*(4), 1155–1206.

Feng, G., Giglio, S. & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370.

Feng, G. & He, J. (2022). Factor investing: A bayesian hierarchical approach. *Journal of Econometrics*, *230*(1), 183–200.

Feng, G., He, J., Li, J., Sarno, L. & Zhang, Q. (2025). Currency return dynamics: What is the role of US macroeconomic regimes? *FEB-RN Research Paper*(72).

Feng, G., He, X., Wang, Y. & Wu, C. (2025). Predicting individual corporate bond returns. *Journal of Banking & Finance*, *171*, 107372.

Ferson, W. E., Sarkissian, S. & Simin, T. T. (2003). Spurious regressions in financial economics? *The Journal of Finance*, *58*(4), 1393–1413.

Foster, F. D., Smith, T. & Whaley, R. E. (1997). Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R2. *The Journal of Finance*, *52*(2), 591–607.

Friewald, N., Jankowitsch, R. & Subrahmanyam, M. G. (2012). Illiquidity or credit deterioration: A study of liquidity in the us corporate bond market during financial crises. *Journal of Financial Economics*, *105*(1), 18–36.

Gamba, A. & Saretto, A. (2013). Firm policies and the cross-section of cds spreads. *WBS Finance Group Research Paper*(191).

Gebhardt, W. R., Hvidkjaer, S. & Swaminathan, B. (2005). The cross-section of expected corporate bond returns: Betas or characteristics? *Journal of Financial Economics*, *75*(1), 85–114.

Giglio, S., Liao, Y. & Xiu, D. (2021). Thousands of alpha tests. *The Review of Financial Studies*, *34*(7), 3456–3496.

Golez, B. & Koudijs, P. (2018). Four centuries of return predictability. *Journal of Financial Economics*, *127*(2), 248–263.

Goulet Coulombe, P. (2024). To bag is to prune. *Studies in Nonlinear Dynamics & Econometrics*. (Forthcoming) doi: 10.1515/snde-2023-0030

Goulet Coulombe, P. & Göbel, M. (2023). Maximally machine-learnable portfolios. *Available at SSRN 4428178*.

Goyal, A. & Welch, I. (2003). Predicting the equity premium with dividend ratios. *Management Science*, *49*(5), 639–654.

Greenwood, R. & Hanson, S. G. (2013). Issuer quality and corporate bond returns. *The Review of Financial Studies*, *26*(6), 1483–1525.

Gu, S., Kelly, B. & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223–2273.

He, Z., Kelly, B. & Manela, A. (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics*, *126*(1), 1–35.

Hillegeist, S. A., Keating, E. K., Cram, D. P. & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, *9*(1), 5–34.

Hilscher, J. & Wilson, M. (2017). Credit ratings and credit risk: Is one measure enough? *Management Science*, *63*(10), 3414–3437.

Houweling, P. & Van Zundert, J. (2017). Factor investing in the corporate bond market. *Financial Analysts Journal*, *73*(2), 100–115.

Israel, R., Palhares, D. & Richardson, S. (2018). Common factors in corporate bond returns. *Journal of Investment Management*, *16*, 17–46.

Jegadeesh, N. & Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, *48*(1), 65–91.

Jessen, C. & Lando, D. (2015). Robustness of distance-to-default. *Journal of Banking & Finance*, *50*, 493–505.

Jostova, G., Nikolova, S., Philipov, A. & Stahel, C. W. (2013). Momentum in corporate bond returns. *The Review of Financial Studies*, *26*(7), 1649–1693.

Kang, J. & Pflueger, C. E. (2015). Inflation risk in corporate bonds. *The Journal of Finance*, *70*(1), 115–162.

Kelly, B., Palhares, D. & Pruitt, S. (2023). Modeling corporate bond returns. *The Journal of Finance*, *78*(4), 1967–2008.

Kelly, B. & Pruitt, S. (2022). Reconciling trace bond returns. *Available at SSRN 4069478*.

Kelly, B., Pruitt, S. & Su, Y. (2019). Characteristics are covariances: A unified model of risk and return. *Journal of Financial Economics*, *134*(3), 501–524.

Lin, H., Wang, J. & Wu, C. (2014). Predictions of corporate bond excess returns. *Journal of Financial Markets*, *21*, 123–152.

Lin, H., Wu, C. & Zhou, G. (2013). Predictability of corporate bond returns: A comprehensive study. *Available at SSRN*.

Lin, H., Wu, C. & Zhou, G. (2018). Forecasting corporate bond returns with a large set of predictors: An iterated combination approach. *Management Science*, *64*(9), 4218–4238.

Lintner, J. (1975). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. In *Stochastic optimization models in finance* (pp. 131–155). Elsevier.

Lo, A. W. & MacKinlay, A. C. (1997). Maximizing predictability in the stock and bond markets. *Macroeconomic Dynamics*, *1*(1), 102–134.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of Finance*, *29*(2), 449–470.

Nelson, C. R. & Kim, M. J. (1993). Predictable stock returns: The role of small sample bias. *The Journal of Finance*, *48*(2), 641–661.

Newey, W. K. & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, *55*(3), 703–708.

OECD. (2022). *Deteriorating conditions of global financial markets amid high debt* (OECD Business and Finance Policy Papers No. 15). Retrieved from https://doi.org/10.1787/89757fae-en doi: 10.1787/89757fae-en

Pástor, L. & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, *111*(3), 642–685.

Rapach, D. E., Strauss, J. K. & Zhou, G. (2013). International stock return predictability: What is the role of the united states? *The Journal of Finance*, *68*(4), 1633–1662.

Reis, R. (2022). *The burst of high inflation in 2021-22: How and why did we get here?* Bank for International Settlements, Monetary and Economic Department.

Rozeff, M. S. (1984). Dividend yields are equity risk premiums. *Journal of Portfolio Management*, 68–75.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, *19*(3), 425–442.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics*, *54*(3), 375–421.

Subrahmanyam, A. (2023). Keeping it simple: The disappearance of premia for standard non-market factors. *Available at SSRN 4584638*.

van Binsbergen, J. H., Nozawa, Y. & Schwert, M. (2025). Duration-based valuation of corporate bonds. *The Review of Financial Studies*, *38*(1), 158–191.

Vassalou, M. & Xing, Y. (2004). Default risk in equity returns. *The Journal of Finance*, *59*(2), 831–868.

Wang, Z. J., Zhang, H. & Zhang, X. (2020). Fire sales and impediments to liquidity provision in the corporate bond market. *Journal of Financial and Quantitative Analysis*, *55*(8), 2613–2640.

Welch, I. & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, *21*(4), 1455–1508.

A Programming code

All code used to obtain the results in this paper is developed by the authors, and is available on GitHub: https://github.com/trmulder/MosaicsOfPredictability. As an illustration, we present below the Python code for the `Tree` class (excluding the plotting method).

Tree.py

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

from predictability_tree.node import Node
from predictability_tree.model_maker import ModelMaker

class Tree:
    def __init__(
            self,
            R: np.ndarray,
            Z: np.ndarray,
            X: pd.DataFrame,
            model_maker: ModelMaker,
            split_values: list,
            max_depth: int,
            min_leaf_size: int
    ):
        self.R = R
        self.Z = Z
        self.X = X
        self.split_values = split_values
        self.max_depth = max_depth
        self.min_leaf_size = min_leaf_size

        self.depth = 0
        self.root = Node(R, Z, X, model_maker, depth=0, id=1)
        self.root.fit_model()
        self.nodes = [self.root]

    def fit(self, max_iter: int) -> None:
        for i in range(max_iter):
            opt_score = -np.inf
            opt_leaf = None
            opt_feature = None
            opt_value = None

            for leaf in self.get_leaf_nodes():
                if leaf.depth >= self.max_depth:
                    leaf.stop_splitting = True
                    continue

                if leaf.stop_splitting:
                    continue

                candidate_split = leaf.get_candidate_split(
                    self.split_values, self.min_leaf_size
                )

                if candidate_split is None:    # Leaf does not have a valid split
                    continue
```

```
                if candidate_split["score"] > opt_score:
                    opt_score = candidate_split["score"]
                    opt_leaf = leaf
                    opt_feature = candidate_split["split_feature"]
                    opt_value = candidate_split["split_value"]

            if opt_leaf is None:    # Tree does not have a valid split
                break
            self.grow(opt_leaf, opt_feature, opt_value, split_order=i+1)

    def grow(
            self,
            node: Node,
            split_feature: int,
            split_value: float,
            split_order: int
    ) -> None:
        # Check whether the node is a leaf:
        if not node.is_leaf:
            raise ValueError("Cannot split a non-leaf node.")

        # Make child nodes:
        child_left, child_right = node.split(split_feature, split_value)

        # Set ID of two child nodes
        child_left.set_id(2 * node.id)
        child_right.set_id(2 * node.id + 1)

        # Set split order of current node
        node.set_split_order(split_order)

        # Put child-nodes in the tree
        self.nodes.append(child_left)
        self.nodes.append(child_right)

        # Check whether we need to update the depth of the tree
        if self.depth < child_left.depth:
            self.depth = child_left.depth

    def predict(self, Z_new: np.ndarray) -> np.ndarray:
        R_pred = [self.root.predict(z) for z in Z_new]
        return np.array(R_pred).reshape(-1,)

    def get_cluster_labels(self, Z_new: np.ndarray) -> np.ndarray:
        labels = [self.root.label(z) for z in Z_new]
        return np.array(labels).reshape(-1,)

    def get_leaf_nodes(self) -> list[Node]:
        return [node for node in self.nodes if node.is_leaf]

    def get_nodes(self, id=None) -> list[Node]:
        if id is None:
            return self.nodes
        return [node for node in self.nodes if node.id == id]
```

B Tree-based clustering algorithm

This section presents the pseudocode to construct the panel trees that cluster bond-return observations into groups with heterogeneous levels of predictability. This algorithm can also be found in the work of Cong et al. (2024), and we present it here for completeness.

Algorithm 1 Tree-based clustering procedure

procedure PANEL TREES
Input: Stock returns $r_{i,t}$, firm characteristics $z_{i,t-1}$, market predictors x_{t-1}, and tree parameters.
Output: a tree architecture with a bunch of split rules that define clusters by panel characteristics.
    for i from 1 to num_iter **do** ▷ Loop over number of iterations
        if current depth $\geq d_{\max}$ **then**
            return
        else
            Search the tree, find all leaf nodes $\mathcal{N}$
            for each leaf node N in $\mathcal{N}$ **do** ▷ Loop over all current leaf nodes
                for each split candidate $\tilde{c}_{p,k,N}$ in $\mathcal{C}_N$ **do**
                    Partition data temporally in N according to $\tilde{c}_{p,k,N}$
                    if Left or right child node cannot satisfy minimal leaf size **then**
                        continue.
                    else
                        Obtain clusterwise return forecasts as in Eq. (2).
                        Calculate the cluster-based R_j^2 using Eq. (4).
                    end if
                end for
            end for
            Find the best leaf node and split rule that maximizes split criteria for this iteration

$$\tilde{c}_i = \underset{N \in \mathcal{N}, \tilde{c}_{p,k,N} \in \mathcal{C}_N}{\arg\max} |R^2_{\text{left}} - R^2_{\text{right}}|$$

            Compare globally for this iteration's split candidates among al leaf nodes.
            Split the node selected at the i-th split rule of the tree $\tilde{c}_i$.
        end if
    end for
    return
end procedure

Note: p, k, N in $\tilde{c}_{p,k,N}$ represent the p-th variable with the k-th value used for leaf node N.

C Corporate bond and equity characteristics

Table 5
This table defines the 53 bond- and equity-characteristics used in this paper. The variables are constructed by Dickerson, Robotti and Nozawa (2024), and the data is obtained from https://openbondassetpricing.com/machine-learning-data/. Panel A and Panel B report the corporate bond and equity characteristics, respectively.

Name	Description	Reference	Source
Panel A: Bond characteristics			
AGE	Bond age. The number of years the bond has been in issuance scaled by the tenor of the bond.	Israel, Palhares and Richardson (2018)	BAML/ICE
AVG12MSPREAD	12-month rolling moving average of bond option adjusted credit spreads skipping the prior month.	Elkamhi, Jo and Nozawa (2024)	BAML/ICE
BONDMOMIPR	6-month bond credit momentum skipping the prior month. Demeaned with duration-times-spread	Israel et al. (2018)	BAML/ICE
BONDPRICE	Price of corporate bond.	Bartram, Grinblatt and Nozawa (2025)	BAML/ICE
CONVEXITY	Bond convexity	-	BAML/ICE
COUPON	Annualized bond coupon payment in percentages	Chung, Wang and Wu (2019)	BAML/ICE
DSPREAD	First difference of bond option adjusted credit spread.	-	BAML/ICE
DTS	Duration-times-spread. Annualized bond duration multiplied by the bond option adjusted credit spread.	Dor et al. (2007)	BAML/ICE
DURATION	The derivative of the bond value to the credit spread divided by the bond value.	Israel et al. (2018)	BAML/ICE
EMPVALUE	Empirical bond value. The difference between the actual credit spread and the fair credit spread for each bond (in percentages). The fair spread is obtained from cross-sectional regressions of the log bond option adjusted credit spreads onto the log of bond ratings, bond credit return volatility, and duration. Finally, the variable is demeaned with duration-times-spread.	Israel et al. (2018)	BAML/ICE
FACEVAL	Face value of the bond. The amount outstanding in units.	Israel et al. (2018)	BAML/ICE
IMPLIEDSPREAD	Fitted spread used to estimate 33 value-characteristics.	Houweling and Van Zundert (2017)	BAML/ICE

Table 5
This table reports the 53 bond- and equity-characteristics... *(continued)*

Name	Description	Reference	Source
KURT	Kurtosis. Bond excess kurtosis computed using a rolling scheme of at least 12 monthly observations and at most 60.	-	BAML/ICE
LTREV30_6	Bond medium-term reversal. The rolling sum of the prior 30 months of corporate bond returns (minimum of 12 observations) minus the rolling sum of the most recent 6 monthly returns.	Subrahmanyam (2023)	BAML/ICE
LTREV48_12	Bond long-term reversal. The rolling sum of the prior 48 months of corporate bond returns (minimum of 12 observations) minus the rolling sum of the most recent 12 monthly returns (minimum of 6 observations).	Bali et al. (2021)	BAML/ICE
MOM3MSPREAD	Momentum 3-month log(SPREAD). The log of the spread 3 months ago minus the log of the current spread.	-	BAML/ICE
MOM6	Corporate bond momentum. The sum of the corporate bond returns over the last 6 months minus the current month's return.	Gebhardt, Hvidkjaer and Swaminathan (2005)	BAML/ICE
MOM6IND	Corporate bond portfolio industry momentum. The sum of the bond portfolio returns over the last 6 months minus the last month's return, where portfolios are formed using the 17 Fama-French industry classifications.	Kelly et al. (2023)	BAML/ICE
MOM6MSPREAD	Momentum 6-month log(SPREAD). The log of the spread 6 months ago minus the current log spread.	Israel et al. (2018)	BAML/ICE
MOM6XRTG	Momentum times bond rating. The sum of corporate bond returns over the last 6 months minus the prior month's return multiplied by the bond's numerical rating: $AAA = 1, \ldots, D = 22$.	Kelly et al. (2023)	BAML/ICE
RATING	Bond S&P rating. A numerical value, where $AAA = 1, \ldots, D = 22$.	Kelly et al. (2023)	BAML/ICE
RATINGXSPREAD	Rating times credit spread	-	BAML/ICE
SIZEB	Bond market capitalization. The number of corporate bonds outstanding (in units) multiplied by the price of the bond	Houweling and Van Zundert (2017)	BAML/ICE
SKEW	Skewness. The bond skewness computed using a 60-month rolling window scheme (minimum of 12 observations).	Kelly et al. (2023)	BAML/ICE

Table 5
This table reports the 53 bond- and equity-characteristics... *(continued)*

Name	Description	Reference	Source
SPRTOD2D	Spread-to-Distance-to-Default. Option-adjusted spread divided by one minus the CDF of the distance-to-default.	Kelly et al. (2023)	CRSP/COMP
SPREAD	Option-adjusted credit spread	Kelly et al. (2023)	BAML/ICE
SPREADVOL	Volatility of the first difference of the bond option adjusted credit spread. Computed using a rolling 24-month window, requiring a minimum of 12 observations.	-	BAML/ICE
STREVB	Bond short-term reversal. The previous month's corporate bond returns.	-	BAML/ICE
STRUCVALUE	Bond structural value. The difference between actual credit spread and fair credit spread (in percentages). The fair spread is derived from a cross-sectioanl regression of the log of bond option adjusted credit spreads onto the log of probability of default computed with firm-level distance-to-dfefauled. This variable is demeaned with duration-time-spread.	Israel et al. (2018)	BAML/ICE
TMT	Time to maturity	-	BAML/ICE
VALUE	Bond value. The percentage difference between the actual credit spread and the fair credit spread. The fair spread is derived from cross-sectional regressions of the log bond option-adjusted credit spread onto the 3-month change in spread, maturity, and credit rating.	Houweling and Van Zundert (2017)	BAML/ICE
SWAPSPREAD	Bond swap spread.	Kelly et al. (2023)	BAML/ICE
VaR	Historical 95% Value-at-Risk, computed using a 36-month rolling window (with a minimum of 12 observations).	Bai et al. (2019)	BAML/ICE
VIXBETA	VIX beta. The sum slope coefficients of lagged VIX and VIX in the time-series regression of corporate bond returns on the Fama-French 3-factors (*Mkt-RF*, *SMB* and *HML*), the default risk factor *DEF*, and the first difference in the CBOE VIX and lagged VIX. This model is estimated using a 60-month rolling-window scheme, where we require a minimum of 12 observations.	Chung et al. (2019)	BAML/ICE

Table 5
This table reports the 53 bond- and equity-characteristics... *(continued)*

Name	Description	Reference	Source
VOLATILITY	Bond return volatility.	Kelly et al. (2023)	BAML/ICE
YIELD	Bond yield-to-maturity	Gebhardt et al. (2005)	BAML/ICE
YLDTOWORST	Bond yield-to-worst.	-	BAML/ICE
Panel B: Equity characteristics			
BOOKLEV	Book leverage. Shareholder's equity plus long/short-term debt plus minority interest minus cash and inventories, divided by shareholder's equity minus preferred stock	Kelly et al. (2023)	COMP
BOOKPRC	Book-to-price. Shareholder's equity plus preferred stock divided by equity market capitalization.	Kelly et al. (2023)	CRSP/COMP
PROFCHANGE	Profitability change. The change in gross profitability over the last 5 years.	Asness, Frazzini and Pedersen (2019)	COMP
D2D	Distance-to-default.	Bharath and Shumway (2008)	CRSP/COMP
DEBTEBITDA	Debt-to-EBITDA. Total debt divided by EBITDA.	Kelly et al. (2023)	CRSP/COMP
EQTYVOL	Equity volatility. Computed using a 180-day rolling-window scheme.	Campbell and Taksler (2003)	CRSP
PROFITABILITY	Profitability. Sales minus cost-of-goods-sold, divided by assets.	Choi and Kim (2018)	COMP
GROSSPROFIPR	Duration-times-spread demeaned profitability.	Israel et al. (2018)	COMP/ICE
ME	Equity market capitalization	Choi and Kim (2018)	CRSP
MKTLEV	Market leverage. The sum of market capitalization, long/short-term debt, minority interest and preferred stock, minus cash and inventories, divided by market capitalization.	Kelly et al. (2023)	CRSP/COMP
MKTLEVIPR	Duration-times-spread demeaned market leverage.	Israel et al. (2018)	COMP/ICE
EP	Earnings-to-price. Net income divided by market equity	Correia, Richardson and Tuna (2012)	CRSP/COMP
OPERLVG	Operating leverage. Sales minus EBITDA, divided by EBITDA.	Gamba and Saretto (2013)	COMP
STOCKMOMIPR	Duration-times-spread demeaned momentum (sum of the past 6 months, excl. most recent month).	Israel et al. (2018)	COMP/ICE
TOTALDEBT	Total firm debt.	Kelly et al. (2023)	COMP

Table 5
This table reports the 53 bond- and equity-characteristics... *(continued)*

Name	Description	Reference	Source
TURNVOL	Turnover volatility. The quarterly standard deviation of sales divided by assets. The volatility is averaged over the preceding four quarters. Furthermore, the volatility for each quarter is calculated using the most recent 80 quarters (with a minimum of 10 observations).	Kelly et al. (2023)	CRSP/COMP

D Further information on data collection procedure

In this appendix, we supplement Section 4 with additional details on our data collection procedure. Appendix D.1 describes the construction of our panel dataset of monthly corporate bond returns. Appendix D.2 outlines the construction of our macroeconomic splitting variables (Table 1) and compares them with related literature. Finally, Appendix D.3 describes the equity and corporate bond risk factors used in this paper.

D.1 Corporate bond returns

In this appendix, we explain how the panel dataset of monthly corporate bond returns is constructed by Dickerson, Robotti and Rossetti (2023). The dataset can be downloaded from the *Open Source Bond Asset Pricing* (OSBAP) website,[12] or by using this link: `https://openbondassetpricing.com/wp-content/uploads/2024/07/WRDS_MMN_Corrected_Data_2024_July.csv`. While the corresponding `README` file can also be found on the OSBAP website, we provide all relevant details in this appendix for completeness.

As mentioned in Section 4, this dataset is constructed from the monthly bond data in the WRDS Bond Database. However, in practice we also make use of the Intercontinental Exchange (ICE) and the enhanced Trade Reporting and Compliance Engine (TRACE) databases under certain circumstances. More specifically, the monthly bond returns from the WRDS Bond Database are set to 100% for returns exceeding this threshold, and we use the returns from ICE or enhanced TRACE to correct these truncated returns (if available, and with databases ranked in this order). Therefore, we explain all these three databases in this appendix, and detail how each dataset is processed (e.g., which filters are used). Section D.1.1 discusses the WRDS Bond Database—the most import database in this paper. Section D.1.2 and Section D.1.3 explain the ICE and enhance TRACE databases, respectively.

D.1.1 Wharton Research Data Services (WRDS) Bond Database

The WRDS Bond Database is a pre-processed database of monthly corporate bond returns. It merges the monthly returns computed from intraday transaction data provided by the enhanced TRACE database with bond characteristics obtained from the Mergent Fixed Income Securities Database (FISD). In contrast to ICE, the WRDS Bond Database is publicly available, and therefore popular among researchers. Besides that, some empirical studies have shown that this database achieves similar results as the quote-based ICE database. For instance, Dickerson, Robotti and Rossetti (2023) show that their results are very similar under these two databases, after applying the filters of Andreani et al. (2024) to the WRDS Bond Database.

The WRDS Bond Database contains three distinct return series: `RET_LM5`, `RET_LDM`, and `RET_EOM`. Each series computes the bond return for month t using the most recent bond prices observed in month $t-1$ and month t. The key distinction among them lies in the criteria that determine when a return is considered valid. First, `RET_LM5` considers a return as valid if and only if the bond trades within the last 5 trading days of both month $t-1$ and month t. Second, `RET_LDM` requires the bond to be traded at the last trading day of both month $t-1$ and month t. Finally, `RET_EOM` only requires at least one trade to take place in both month $t-1$ and month t, such that these can be used to compute the return, but it does not matter when this trade takes place. In this paper, we use `RET_LM5` as the bond return variable.

As mentioned before, the returns in the WRDS Bond Database are truncated at 100%. In total, there are 94 truncated observations, which can be downloaded via https://openbondassetpricing.com/wp-content/uploads/2023/10/OSBAP_Return_Corrections.csv. For each of these truncated observations, we first investigate whether the corresponding return-observation is available in ICE (see Appendix D.1.2). If so, we use this as the monthly return observation. Otherwise we set it to the corresponding return obtained from enhanced TRACE (see Appendix D.1.3). In total, 91 of the 94 truncated bond returns are adjusted in this way and the remaining 3 observations are deleted.

Finally, we combine the WRDS Bond Database and Mergent FISD together, and apply the standard filters also used by other studies. These filters are presented in Table 6, together with the filters used for the enhanced TRACE database in Section D.1.3. Additionally, we use the filtering rules proposed by Andreani et al. (2024), and remove: (i) non-investment grade bonds with less than 100 million USD outstanding prior to the end of September 2016, and less than 250 million USD after this month; and (ii) investment-grade bonds with less than 150 million USD outstanding prior to the end of November 2004, and with less than 250 million USD outstanding after this month.

D.1.2 Intercontinental Exchange (ICE)

We use the Bank of America Merrill Lynch (BAML) database from ICE to calculate monthly corporate bond returns when the corresponding return observation from the WRDS Bond Database is truncated. This database contains the daily prices of constituent bonds from the Bank of America Investment Grade and High Yield Corporate Bond Indices. Since ICE is a quote-based database, monthly returns can be computed *exactly*, because a bond-price at the end of a month is always present. Furthermore, this database is also used to compute the equity and bond characteristics in Section 4.2.[14]

We also merge ICE with the Mergent FISD and apply the same filters as for the WRDS Bond Database, which are shown in Table 6. Besides that, we apply the following three filters: (i) keep bonds issued by firms located in the US (`Country == 'US'`); (ii) keep

only *corporate* bonds (`Ind_Lvl_1 == 'corporate'`); and (iii) keep bonds denominated in USD (`Currency == 'USD'`).

D.1.3 Enhanced Trade Reporting and Compliance Engine (TRACE)

Enhanced TRACE provides intraday transaction data on corporate bonds. In the situation that a return-observation from the WRDS Bond Database is truncated and the corresponding observation is not present in ICE, we use these intraday data to compute the monthly bond return. Given the transaction-based nature of the enhanced TRACE, exact end-of-day bond prices do not exists. For each day, we compute the bond price as the volume-weighted price of all transactions taking place that day. These daily prices are in turn used to compute monthly returns.

Besides that, we also apply a few filters to the enhance TRACE database. Specifically, we follow the cleaning procedure of van Binsbergen, Nozawa and Schwert (2025) by applying the TRACE filters presented in Table 6. In contrast to Dickerson, Mueller and Robotti (2023), we do not remove bonds that are close to default—identified by a bond price less than 5 USD or greater than $1,000$ USD. As mentioned in the `README` file for our dataset on the OSBAP website, this filter would lead to a look-ahead bias during out-of-sample portfolio performance analyses

Table 6
This table provides an overview of the data cleaning procedure for the enhanced TRACE and FISD corporate bond databases.

Rule	Description	TRACE/FISD filter
Remove bonds that do not trade on public exchanges in the US	Remove bonds issued under Rule 144A, issued through private placement, not traded in USD, and from which the issuers are not based in the US.	FISD: (1) `RULE_144A == 'N'` (2) `PRIVATE_PLACEMENT == 'N'` (3) `FOREIGN_CURRENCY == 'N'` (4) `COUNTRY_DOMICILE == 'USA'`
Remove special corporate bonds	Remove mortgage backed or asset backed bonds, equity linked or convertible bonds, structured notes, and agency backed bonds.	FISD: (1) `ASSET_BACKED == 'N'` (2) `CONVERTIBLE == 'N'` (3) `BOND_TYPE !%in% 'X'` where `'X' ==` (i) Agency, muni, or government bonds: {`TXMU, CCOV, CPAS, MBS, FGOV, USTC, USBD, USNT, USSP, USSI, FGS, USBL, ABS, O30Y, O10Y, O3Y, O5Y, O4W, CCUR, O13W, O52W, O26W`} (ii) Agency-backed bonds: {`ADEB, AMTN, ASPZ, EMTN, ADNT, ARNT`
Remove bonds with a variable coupon rate		FISD: `COUPON_TYPE != 'V'`
Remove bonds with a maturity of less than one year		
Remove labelled bonds	Remove all transactions that are classified: locked-in, when-issued, or have a special sales condition.	TRACE: (1) `lckd_in_ind != 'Y'` (2) `wis_fl != 'Y'` (3) `(sale_cndtn_cd == 'None') | (sale_cndtn_cd == '@')`
Remove all trades that have a two-day settlement period or longer		TRACE: (1) `days_to_sttl_ct == '002'` (2) `days_to_sttl_ct == '001'` (3) `days_to_sttl_ct == '000'` (4) `days_to_sttl_ct == 'None'`
Remove small transactions	Remove intraday transactions with daily par volume less than 10,000 USD.	TRACE: `entrd_vol_qt >= 10000`
Remove cancelled transactions	Also adjust for transactions that are corrected or reversed.	TRACE: see Dick-Nielsen (2014).
Remove bonds with a special interest payment structure	Bonds for which the payment structure is 'N/A', 'undocumented by FISD', 'bi-monthly' and 'Variable Coupon'	FISD: `INTEREST_FREQUENCY !%in% c(-1, 13, 14, 15, 16)`
Remove bonds for accrued interest cannot be computed	Remove bonds with missing `DAY_COUNT_BASIS, COUPON, COUPON_TYPE, OFFERING_DATE, DATED_DATE` and `INTEREST_FREQUENCY`.	

D.2 Macroeconomic splitting variables

In Section 4.3, we introduce nine macroeconomic variables that serve as time-varying splitting variables in the regime-switching panel trees. This appendix provides a more detailed discussion of their construction (Appendix D.2.1) and compares our set of variables with those used in related literature (Appendix D.2.2). For example, we adopt the same macroeconomic variables as Cong et al. (2025). While the authors have made their datasets publicly available, their coverage extends only through December 2021. Since our sample period runs from August 2002 to August 2022, we reconstruct all macroeconomic variables for consistency with our extended horizon. The reconstructed series exhibit high correlations with those of Cong et al. (2025), exceeding 0.80 for every variable except `NI`. Time series plots for our macroeconomic variables can be found in Figure 8.

D.2.1 Dataset construction

As described in Section 4.3, the macroeconomic predictors—except the liquidity measure, `LIQ`—are constructed with data from Amit Goyal's website.[29] We use this dataset to compute the variables defined in Table 1 as follows. Note that all predictors below are subsequently standardized as detailed in Section 4.3.

Default yield spread (`DFY`). Taken directly from Goyal's dataset (`dfy`), this variable measures the yield difference between BAA- and AAA-rated corporate bonds.

Term spread (`TMS`). Available in the dataset as `tms`, this variable is defined as the yield difference between long-term government bonds and the three-month US Treasury bill.

Stock variance on daily S&P 500 returns (`SVAR`). In the dataset of Amit Goyal, this variable (`svar`) represents the monthly stock market variance and is computed as the sum of squared daily S&P 500 returns. To better capture persistent economic regimes, we follow Cong et al. (2025) and define `SVAR` as the 12-month moving average of `svar`.

Inflation (`INFL`). This variable is defined as the 12-month moving average of the lagged *Consumer Price Index (All Urban Consumers)* (CPI). Following Welch and Goyal (2008), we use the one-month lagged CPI series because inflation numbers are usually published with a one-month lag by the Bureau of Labor Statistics. Thus, `INFL` is constructed as the 12-month moving average of `infl` from Goyal's dataset.

Dividend yield of S&P 500 (`DY`). This variable is not directly available in Goyal's dataset, but needs to be constructed ourselves using other variables in his dataset. Following Welch and Goyal (2008), we define the dividend yield as the difference between the logarithm of dividends and the logarithm of lagged prices, $\texttt{DY}_t = log(\texttt{D12}_t) - log(\texttt{index}_{t-1})$. `D12` is the 12-month rolling sum of dividends paid on the S&P 500 index, and `index` is

[29] https://docs.google.com/spreadsheets/d/1OIZg6htTK6OwtnCVXvxAujvG1aKEOVYv/edit?usp=sharing&ouid=113571510202500088860&rtpof=true&sd=true

the S&P 500 price index, both available in Goyal's dataset.

Earnings-to-price ratio of S&P 500 (EP). Also this variable needs to be computed ourselves. We use the earnings variable (E12) and the S&P 500 price index (index) provided in Goyal's dataset to construct EP. Specifically it is defined by $\texttt{EP}_t = log(\texttt{E12}_t) - log(\texttt{index}_t)$, where E12 is the 12-month rolling sum of earnings for the S&P 500.

Net equity expansion of NYSE listed stocks (NI). Provided in Goyal's dataset as nits, this variable is defined as the ratio of the 12-month rolling sum of net equity issues by NYSE-listed firms to the total market capitalization of NYSE-listed stocks.

Three-month US Treasury bill rate (TBL). We use the corresponding 3-month Treasury bill rate from Goyal's dataset (tbl) for TBL.

Pástor-Stambaugh aggregate liquidity (LIQ). Unlike the other variables, LIQ is not available in Goyal's dataset. As noted in Section 4.3, we obtain it from Robert Stambaugh's website.[30]

D.2.2 Comparison with Cong et al. (2025)

In this appendix, we compare our macroeconomic variables to those used by Cong et al. (2025). Our aim is to replicate their variables and extend the sample period through August 2022. Ideally, our variables should closely resemble those in Cong et al. (2025). However, Cong et al. (2025) is not fully transparent about the precise data sources and exact variable constructions. They only state that data from Amit Goyal's website and the Federal Reserve Economic Data (FRED) are used, but for several variables, either source could be appropriate. The same ambiguity applies to definitions. For example, there are multiple standard measures of inflation, but Cong et al. (2025) do not specify which one they adopt. Consequently, we combine the available information with conventions from related literature to construct our macroeconomic variables, as detailed in Section D.2.1. Our objective is not necessarily to reproduce the exact same variables, but rather to construct comparable measures that we expect to be highly correlated.

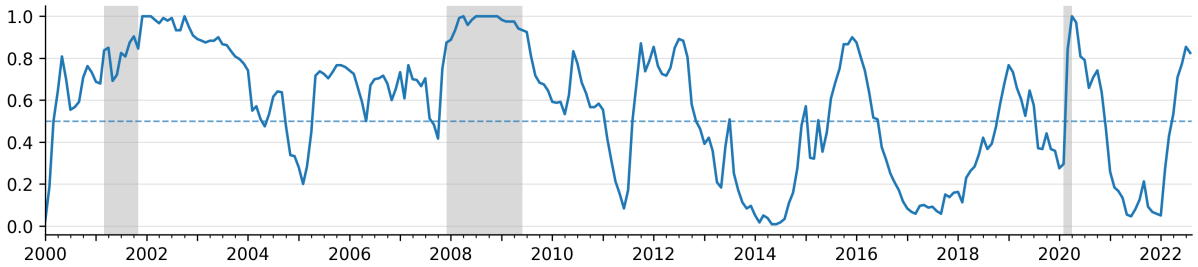
Table 7 reports summary statistics for the two datasets over the overlapping period from December 1982 to December 2021. We find that most of our macroeconomic variables are highly correlated with those provided by Cong et al. (2024). Specifically, pairwise correlations exceed 0.8 for all variables except NI. Sample means and standard deviations are also very similar across the two datasets. The divergence for NI—the net equity expansion variable—likely stems from differences in definition. As explained in Section D.2.1, we use the ntis variable from Amit Goyal's website, defined as the net equity expansion of NYSE-listed stocks. In contrast, Cong et al. (2025) define NI as the net equity issuance of the S&P 500, a series we could not locate in either Amit Goyal's dataset or FRED. For this reason, we follow related literature (e.g., Welch and Goyal (2008)) and use ntis

[30] https://finance.wharton.upenn.edu/~stambaug/liq_data_1962_2024.txt

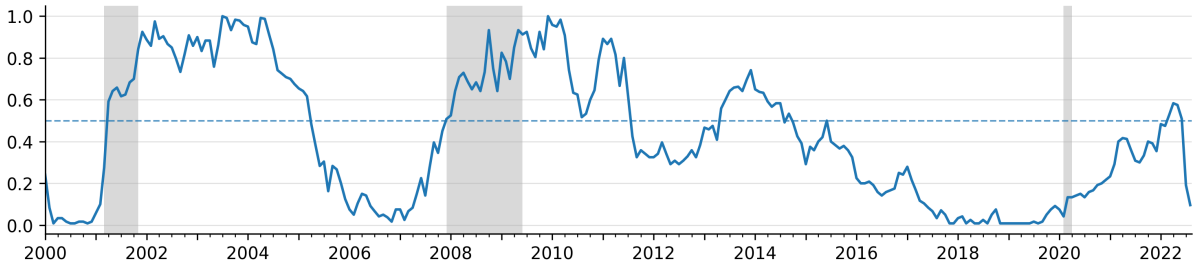
from Goyal's dataset to define NI.

Table 7
This table presents the pairwise correlations between the macroeconomic variables from this paper (Mulder) and Cong et al. (2025) (Cong). For each dataset, this table also reports sample means and standard deviations. All statistics are estimated over a period from December 1982 to December 2021. Rows are sorted by correlation in descending order.

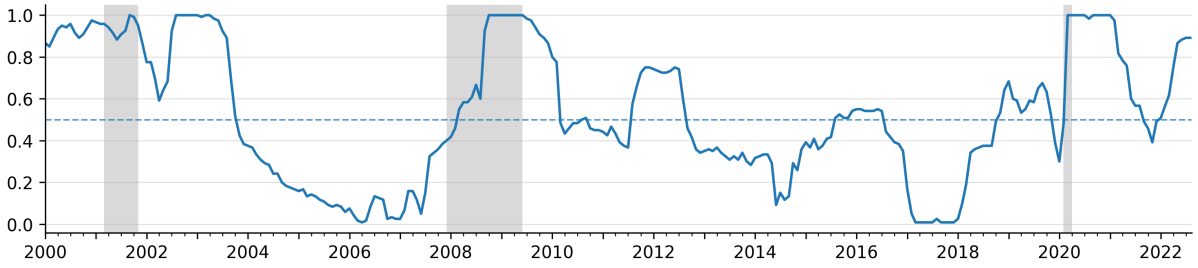
Variable	Obs.	Correlation	Mean (Mulder)	Mean (Cong)	SD (Mulder)	SD (Cong)
TBL	481	0.96	0.36	0.36	0.29	0.30
SVAR	481	0.95	0.52	0.52	0.32	0.32
TMS	481	0.93	0.49	0.49	0.31	0.31
DY	481	0.93	0.37	0.36	0.33	0.32
DFY	481	0.91	0.45	0.46	0.32	0.32
EP	481	0.87	0.35	0.35	0.30	0.32
ILL	481	0.87	0.58	0.58	0.33	0.33
INFL	481	0.85	0.39	0.39	0.30	0.30
NI	481	0.40	0.42	0.49	0.30	0.36



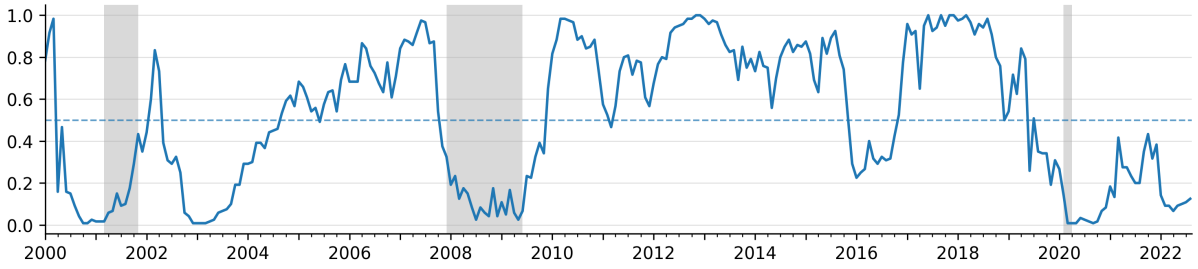
(a) `DFY`



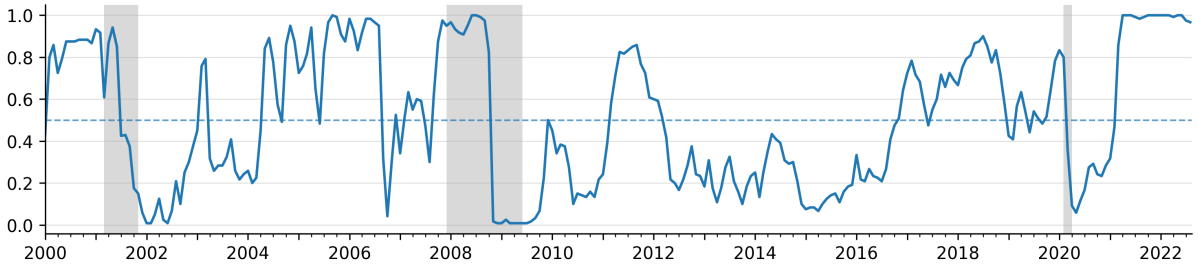
(b) `TMS`



(c) `SVAR`



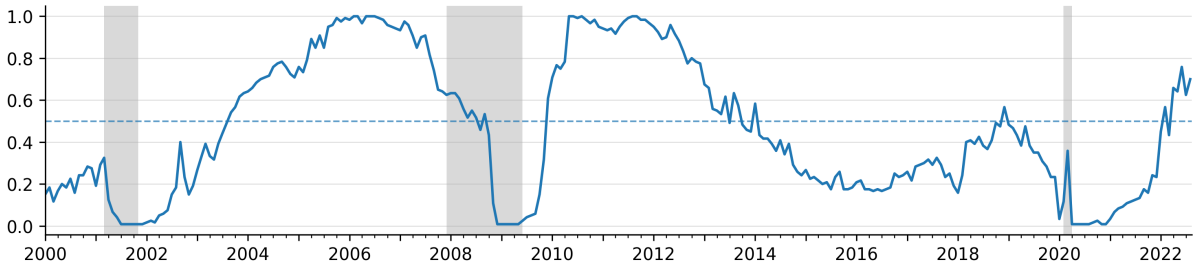
(d) `LIQ`



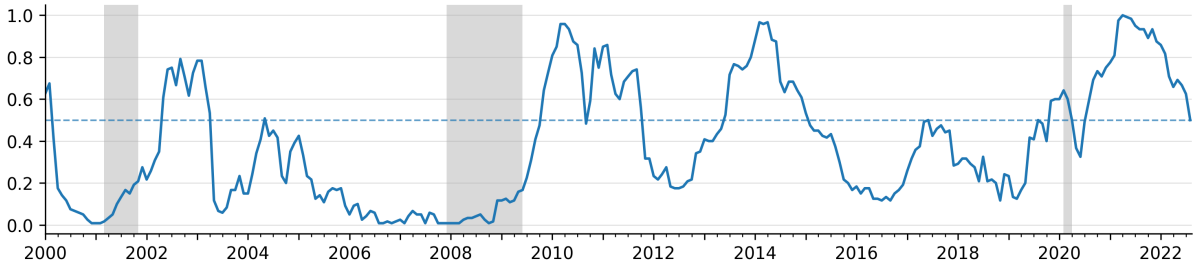
(e) `INFL`

Figure 8: Time series plots for the macroeconomic splitting variables defined in Table 1 from January 2000 to August 2022. Grey shading indicates NBER recessions. A dashed horizontal line at 0.5 indicates the splitting threshold used in the regime-switching panel trees.

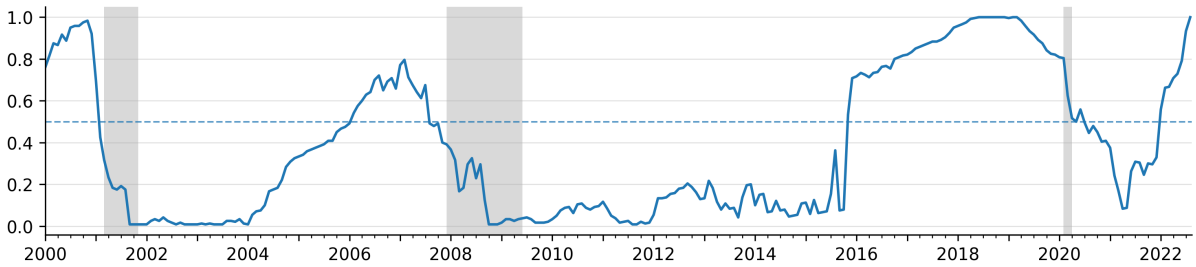
(f) DY



(g) EP



(h) NI



(i) TBL

Figure 8: Time series plots for the macroeconomic splitting variables... *(continued)*

D.3 Factor models

In this appendix, we describe the factor models used to construct Table 4 in Section 5. The data are downloaded from the *Open Source Bond Asset Pricing* website,[12] specifically from this link: https://openbondassetpricing.com/wp-content/uploads/2024/11/excess_factor_zoo.csv. This dataset corresponds to the working paper of Dickerson, Julliard and Mueller (2024). In Panel B of Table 4, we compute the abnormal returns (α; in percentages) using a time-series regression that is defined as follows

$$r_t^{(i)} = \alpha^{(i)} + \boldsymbol{\beta}^{(i)\intercal}\mathbf{f}_t + \varepsilon_t^{(i)},$$

where $r_t^{(i)}$ is the return on the long-short portfolio Ti–Bi at month t, $\mathbf{f}_t$ is a column-vector of risk factors at month t, $\alpha^{(i)}$ is the abnormal return of the corresponding long-short portfolio, and $\varepsilon_t^{(i)}$ is a residual term. The factor models used in Table 4 are defined below.

Capital asset pricing model (CAPM). This single-factor model is introduced by Sharpe (1964) and Lintner (1975), and considers the value-weighted excess return on the stock market (MKTS) as a risk factor. The stock market is defined as all common stocks (i.e., with share codes of 10 and 11) listed on the NYSE, AMEX, and NASDAQ. The risk-free rate, used to compute the excess returns, is the one-month US Treasury bill rate. The CAPM is typically used to price stocks. However, van Binsbergen et al. (2025) show that the CAPM can also price the duration-adjusted returns of corporate bonds, which motivates the inclusion of this model in this paper.

Bond capital asset pricing model (CAPMB). This single-factor model uses the value-weighted excess returns on the corporate bond market portfolio (MKTB) as a risk factor. Dickerson, Mueller and Robotti (2023) show that the CAPMB provides superior cross-sectional pricing performance compared to other traded- and nontraded-factor models. Again, the excess returns are computed in excess of the risk free rate, which is defined as the one-month US Treasury bill rate.

Default and term-structure model (DEFTERM). This two-factor model is introduced by Fama and French (1993). The first factor, DEF, captures risk associated with shifts in economic regimes that raise the likelihood of corporate defaults. It is measured as the difference between the return on a market portfolio of long-term corporate bonds and the return on long-term US government bonds. The second factor, TERM, reflects risk related to unexpected changes in interest rates. It is defined as the difference between the return on long-term government bonds and the one-month US Treasury bill rate.

Intermediary capital models (HKM/HKMSF). This two-factor model is introduced by He et al. (2017). Besides the stock market factor, MKTS, the model also includes a factor related to intermediary capital risk (CPTLT). Finacial intermediation has an important role in the—over-the-counter—corporate bond market, which motivates the use of HKM in this paper. Following Dickerson, Mueller and Robotti (2023), we also consider the

single-factor version of this model (HKMSF), which only uses CPTLT as a risk factor.

Fama-French three-factor model (FF3). This is the three-factor model introduced by Fama and French (1993). It extends the CAPM by considering two additional equity risk factors: small-minus-big (SMB) and high-minus-low (HML). Both factors are constructed from the value-weighted returns on the 2×3 size (i.e., market equity) and book-to-market sorted portfolios, which can be downloaded from Kenneth French's website.[31] More specifically, the HML factor is defined as the equal-weighted average return on the two high book-to-market portfolios minus the equal-weighted average return on the two low book-to-market portfolios. The SMB factor is defined as the equal-weighted average return on the three low market-cap portfolios minus the equal-weighted average on the three high market-cap portfolios.

Fama-French five-factor model (FF5). This is the five-factor model of Fama and French (2015), which extends the three-factor model with two factors: conservative-minus-aggressive (CMA) and robust-minus-weak (RMW). Both factors are constructed as long-short portfolios, in particular from the value-weighted returns of 2×3 characteristic-sorted portfolios, in the same spirit as SMB and HML. The CMA factor is defined by the equal-weighted return on the two aggressive investment portfolios minus the equal-weighted return on the two conservative investment portfolios, from the 2×3 Fama-French portfolios sorted on size and investment. Similarly, the RMW factor is defined as the equal-weighted return on the two robust operating profitability portfolios minus the equal-weighted return on the two weak operating profitability portfolios, from the 2×3 Fama-French portfolios sorted on size and operating profitability.

FF5 with stock momentum and idiosyncratic volatility (FF5+MOMS+IVOL). We also consider the FF5 model augmented with the stock momentum and idiosyncratic volatility factors. The stock momentum risk factor (MOMS) is introduced by Carhart (1997). It is defined as the return spread between the top and bottom decile portfolios, sorted on the cumulative returns over the 12 past months while excluding the most recent month. Furthermore, the idiosyncratic volatility factor (IVOL) is introduced by Campbell and Taksler (2003), and is computed by aggregating a firm-level measure of idiosyncratic equity volatility each month. For each firm, we estimate a time-series regression of daily stock returns on the FF3 factors, compute the residuals, and take their standard deviation within each month. The IVOL factor is then defined as the monthly average of these firm-level standard deviations.

[31] https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

E Additional results

In this appendix, we supplement Section 5 with additional results. Specifically, we present four *mosaics of predictability*, analogous to those in Figure 5 of Section 5, for the variables `SKEW`, `SPREAD`, `SPRTOD2D`, and `VaR`. These variables are considered because Feng, He, Wang and Wu (2025) identify them as important predictors of individual corporate bond returns. The purpose of this appendix is to demonstrate that being an important predictor of returns does not necessarily imply a strong relation to the heterogeneity in return predictability. The results are reported in Figure 9. We observe that the predictors are indeed weakly related to the heterogeneity in return predictability.

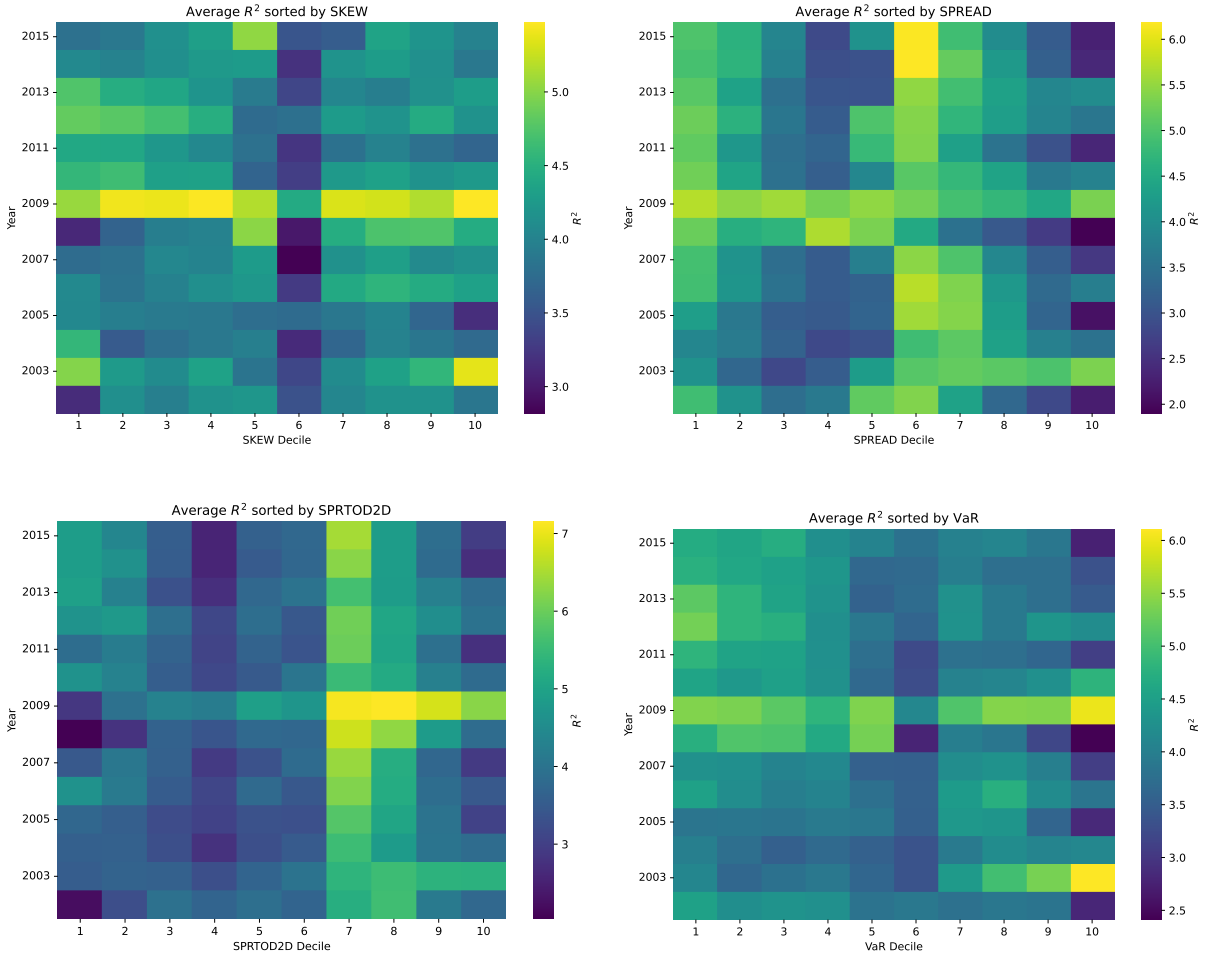


Figure 9: This figure presents four heatmaps of average in-sample R^2 under the panel tree in Figure 4 and when sorting observations into deciles of bond characteristics. The top panels show the average R^2 values for asset-return observations sorted annually into deciles of return skewness (`SKEW`, left) and option-adjusted credit spread (`SPREAD`, right). The bottom panels show similar heatmaps for spread-to-distance-to-default (`SPRTOD2D`, left) and Value-at-Risk (`VaR`, right). Lighter colors correspond to higher R^2 values, as shown by the colorbar. The R^2 values are estimated with Ridge regressions as clusterwise predictive models.

F Time-series plot of Federal Funds Effective Rate

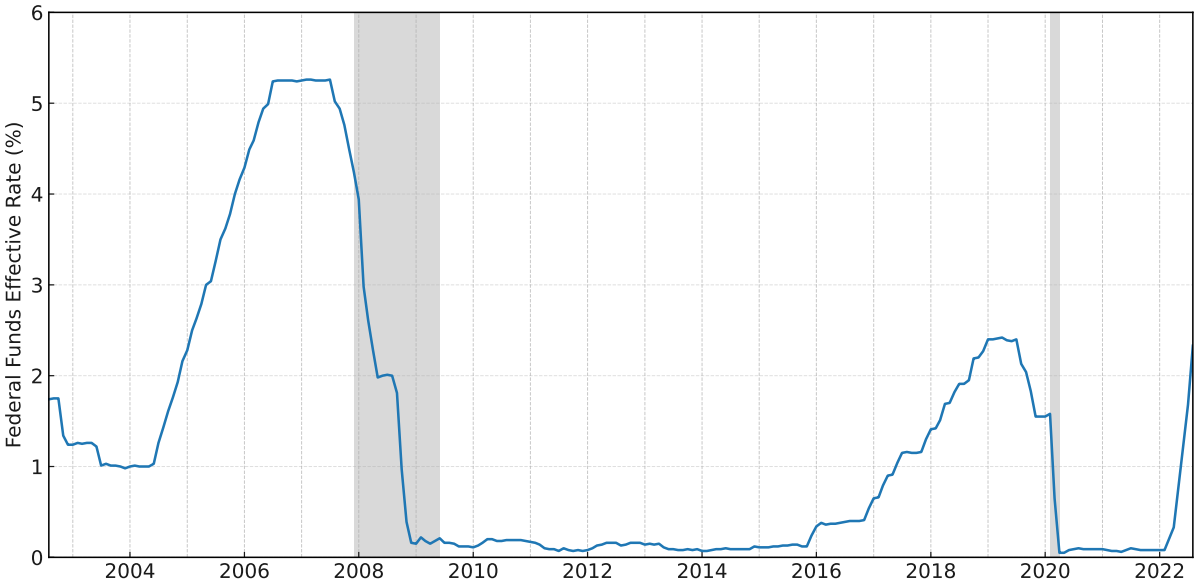


Figure 10: This figure plots the Federal Funds Effective Rate from 2002M8 to 2022M8. The grey shading indicates NBER recessions.

G Random forest results

In this appendix, we report the results of our tree-based clustering algorithm when random forests (RFs) are used as clusterwise predictive models. As noted in Section 5, estimating panel trees with RFs is computationally demanding. For instance, the results presented in this appendix require approximately 12 hours of computation, whereas the corresponding panel tree with Ridge regressions in Section 5 can be estimated in about 90 minutes. Due to this computational burden, we do not conduct an extensive hyperparameter search. Instead, RFs are estimated using fixed hyperparameters: (i) the minimum leaf size is set to 5% of the total sample size; (ii) at each split, $\sqrt{\texttt{num\_features}}$ variables are considered as candidate splits; and (iii) the forest consists of 500 trees. Small deviations from these settings—such as increasing the number of trees or reducing the minimum leaf size—substantially increase computation time and may render the problem infeasible on a standard computer.

The remainder of this appendix is structured similar as Section 5. First, we present the panel tree in Appendix G.1 that considers only the cross-sectional asset characteristics in Table 5 of Appendix C as potential splitting variables. In Section G.2, we compare the clusterwise models with global predictive models. Finally, in Section G.3, we present the RF results regarding the *predictability disagreement anomaly* of Cong et al. (2024).

G.1 The mosaic of return predictability

Figure 11 presents the panel tree fitted to the training data from 2002M8 to 2015M8. When we compare this to the panel tree with Ridge regressions as clusterwise predictive models presented in Figure 4 of Section 4, we observe the following. First, the global RF model (i.e., fitted to all observations grouped together) has a higher in-sample R^2 (1.89%) than the global Ridge regression (1.64%). This means that the RF fits the in-sample data better than the Ridge regression, which can potentially be explained by the non-linear character of the former model. Second, while the panel tree in Figure 4 considers 11 unique predictors from Table 5 of Appendix C as splitting variables, the panel tree in Figure 11 considers nine unique splitting variables, six of which are also considered in Figure 4. Thus, both panel trees use the same small set of predictors as splitting variables. Third, the panel tree in Figure 11 uses "$\texttt{STREVB} \leq 0.2$" as the first split rule, while this is considered as the optimal second split when Ridge regressions are used in Figure 4.

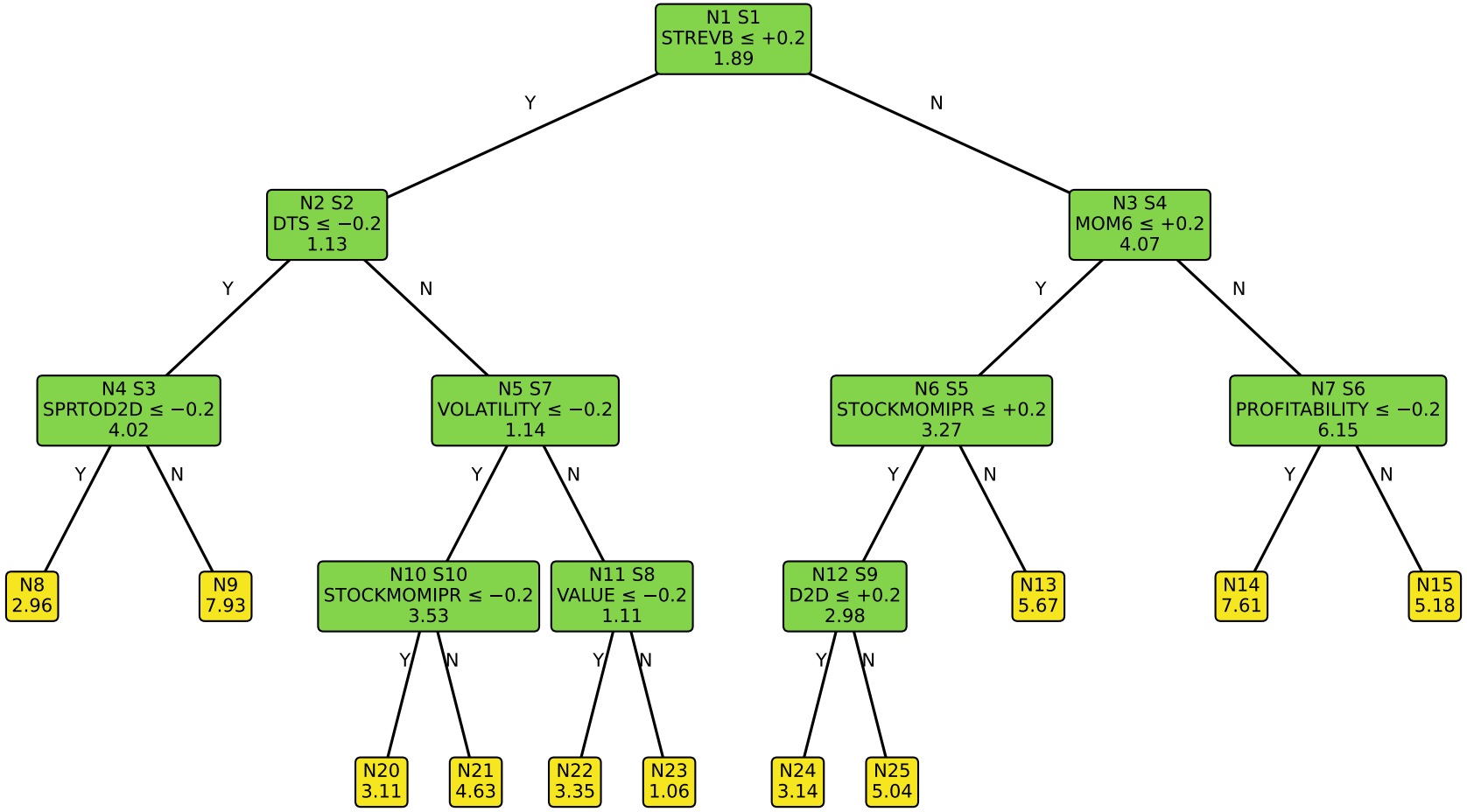


Figure 11: This figure illustrates the panel tree used to form clusters of bond-return observations with heterogeneous levels of predictability, measured by R^2. The tree is estimated using monthly data from 2002M8 to 2015M8, and splits on cross-sectional asset characteristics that are rank-standardized to the interval $[-1, 1]$ (see Appendix C). Each intermediate node (green) and leaf node (yellow) displays the node's ID, denoted by $N\#$, which is used to identify the node in this paper. Intermediate nodes additionally report the split rule (second line), which sends observations satisfying the rule to the left child node and the rest to the right child node, as well as the node's split order, denoted by $S\#$. Finally, all nodes display the in-sample R^2 statistic (in percentages; last line), estimated using the clusterwise random forests introduced in Section 3.

Furthermore, since the panel tree fitted with RFs is different from the panel tree in Figure 4 of Section 5, the resulting *mosaics of return predictability* also differ, as is shown in Figure 12. In the top-left panel, we observe that corporate bonds with high previous-month returns (i.e., `STREVB` > 0.2) are still better predictable than those with low previous-month returns. We further observe that the level of return predictability is fairly homogeneous over `RATING`-sorted deciles of corporate bonds in the bottom-left panel. However, the positive relationship between distance-to-default (i.e., `D2D`) and return predictability is almost completely gone. Also in the bottom-right panel, we observe that not the bonds with both low distance-to-default and low previous-month returns are substantially less predictable anymore, but just the bonds with low previous-month returns in general.

Furthermore, in Figure 13, we report similar heatmaps that illustrate how return predictability varies over characteristics-sorted groups of corporate bonds. Specifically, the heatmaps consider the characteristics: `SKEW`, `SPREAD`, `SPRTOD2D` and `VaR`. We observe that these variables are not particularly related to the degree of return predictability, which we also concluded from the mosaics estimated with Ridge regressions in Figure 9 of Appendix E. Interestingly, we do observe that the mosaics estimated by RFs (Figure 13) and Ridge regressions (Figure 9) are very similar.

All in all, we conclude that the panel tree, and resulting mosaics of predictability, are very similar when estimated with either RFs or Ridge regressions. The main difference with our analysis in Section 5 is that the relationship between return predictability and distance-to-default has vanished in this appendix. However, given the results presented in the sections below, we do believe that the results in Section 5 are more robust. Not only align these results more with our expectations, but they are also obtained after applying careful hyperparameter tuning. In contrast, in this appendix, we estimate the clusterwise RFs with fixed hyperparameters. Future research should investigate whether the results in this appendix become more similar to the Ridge regressions results in Section 5 after careful hyperparameter tuning is applied to the RFs. However, as mentioned before, these cannot be done on a standard computer.

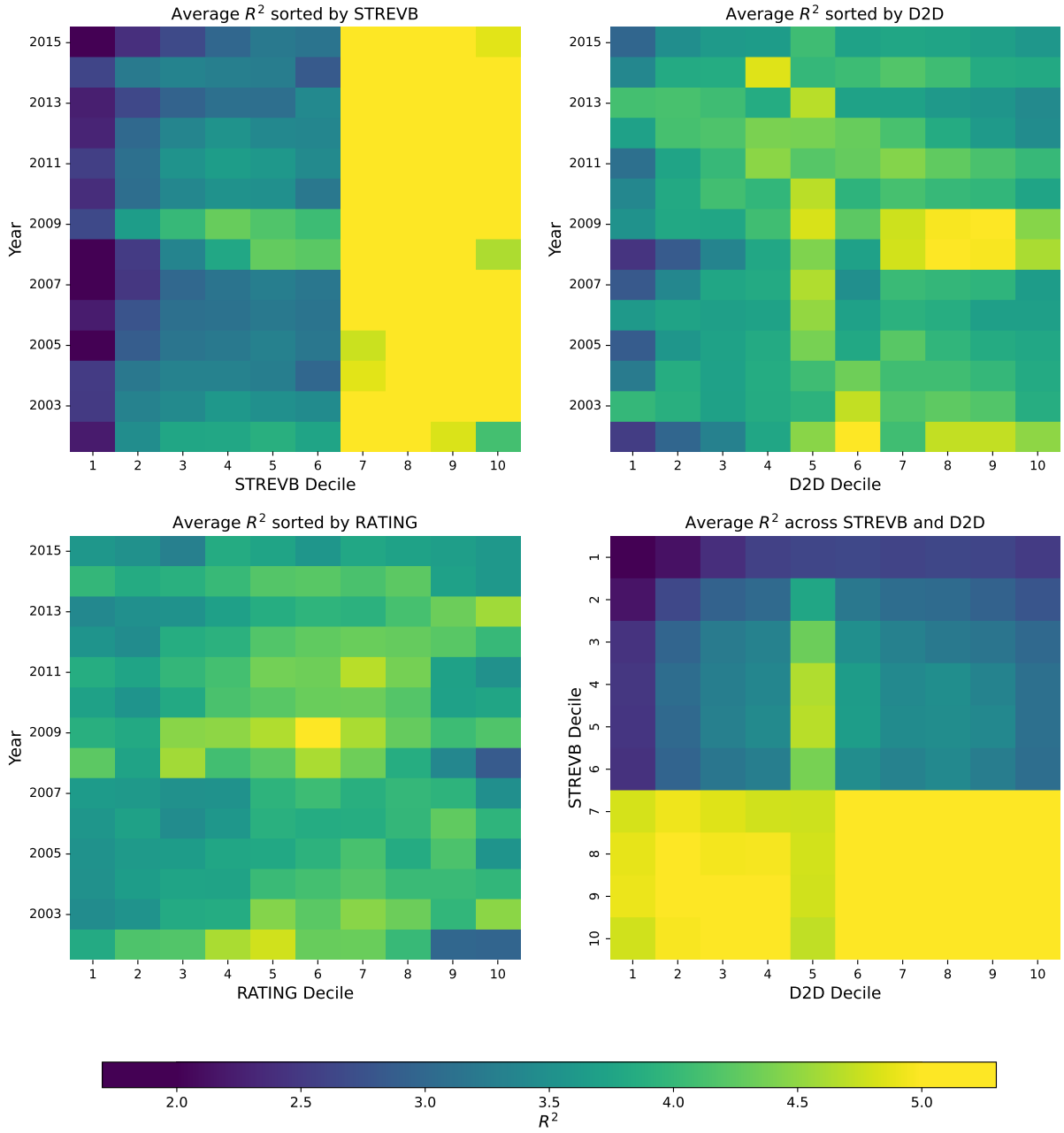


Figure 12: This figure presents four heatmaps of average in-sample R^2 under the panel tree in Figure 11 and when sorting observations into deciles of bond characteristics. The top panels show average R^2 for asset-returns observations sorted annually into deciles of short-term bond reversal (`STREVB`, left) and distance-to-default (`D2D`, right). The bottom left panel shows results for annual rating deciles, while the bottom right panel shows R^2 for 10×10 bivariate-sorted groups based on the first two splitting variables in Figure 4: `D2D` and `STREVB`. Lighter colors correspond to higher average R^2 values, as shown by the colorbar. Note that higher rating deciles are associated to lower-rated bonds, as can be seen from the definition of `RATING` in Table 5 of Appendix C.

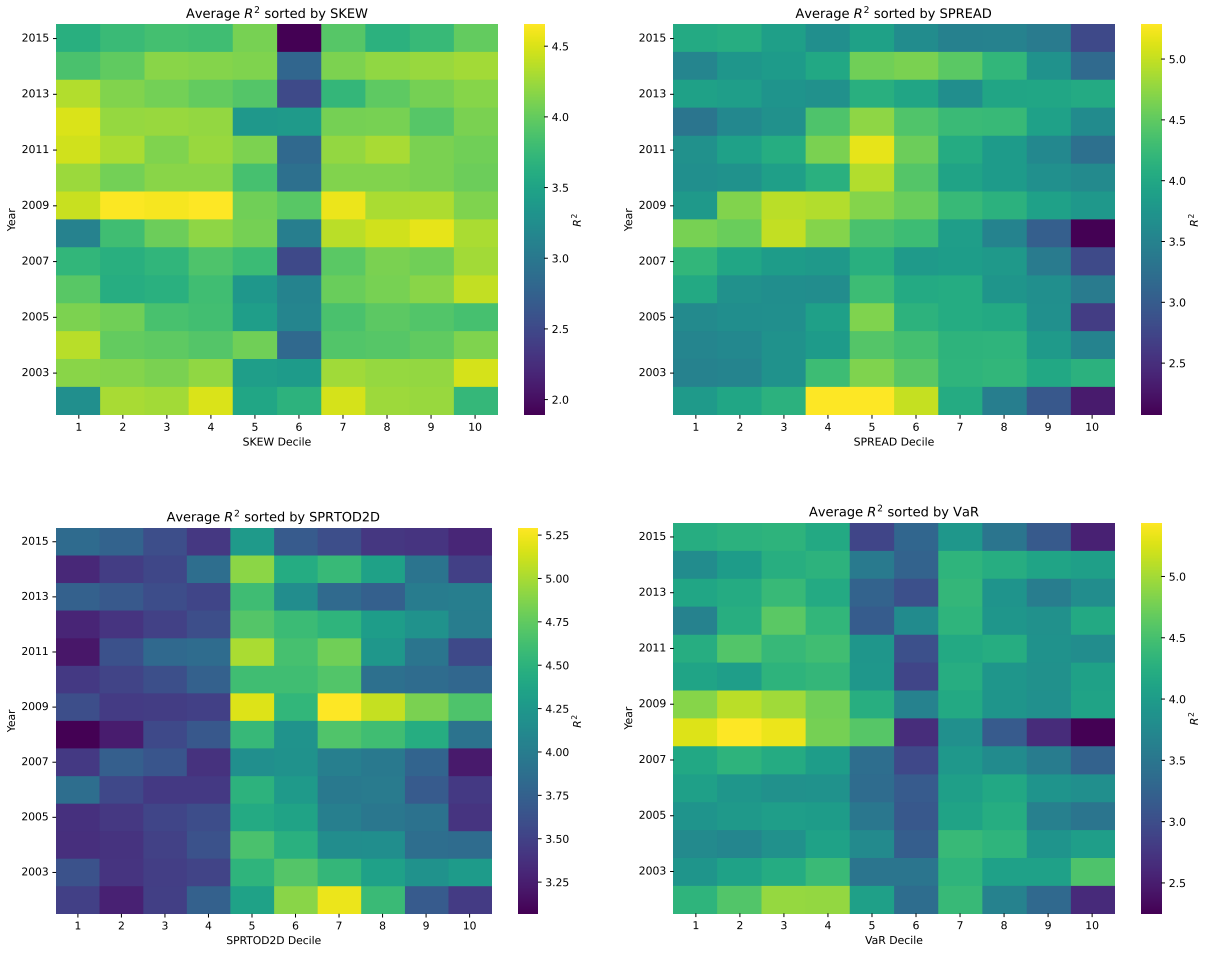


Figure 13: This figure presents four heatmaps of average in-sample R^2 under the panel tree in Figure 11 and when sorting observations into deciles of bond characteristics. The top panels show the average R^2 values for asset-return observations sorted annually into deciles of return skewness (`SKEW`, left) and option-adjusted credit spread (`SPREAD`, right). The bottom panels show similar heatmaps for spread-to-distance-to-default (`SPRTOD2D`, left) and Value-at-Risk (`VaR`, right). Lighter colors correspond to higher R^2 values, as shown by the colorbar. The R^2 values are estimated with random forests as clusterwise predictive models.

G.2 Predictive performance of clusterwise versus global models

In this appendix, we present some summary statistics for the clusters identified by the panel tree in Figure 11, and investigate whether there is a relationship between return predictability and investment outcomes when RFs are used as clusterwise predictive models. The results are reported in Figure 8. Panel A provides the same summary statistics as Table2 in Section 5, which include: (i) the number of observations (Obs.); (ii) the level of return predictability estimated by the corresponding clusterwise RF (R^2_C; in percentages); the level of return predictability estimated by the corresponding global RF (R^2_G; in percentages); and (iv) the difference between these two measures ($R^2_{CMG} = R^2_C - R^2_G$). The rows in Table 8 are ordered in descending order of R^2_C.

First, we observe that the bottom three clusters (N23, N24, and N8) contain substantially more observations than the others. Since these clusters also have the lowest R^2_C, this is already a first sign that the RFs are overfitting the training data. In contrast, in Table 2 of Section 5, we observe that large clusters are more evenly distributed when Ridge regressions are used as clusterwise predictive models. Second, we find that the clusterwise levels of return predictability (R^2_C) closely align with the global measures (R^2_G) in Table 8. More specifically, the differences between the two measures (i.e., R^2_{CMG}) always lies between 0 and 2, compared with a range of 0 to 7 for the Ridge results in Table 2. This narrow gap indicates that the clusterwise RFs have little specificity compared to the

Table 8
This table report summary statistics for each cluster identified by the panel tree in Figure 11. The sample period is from 2002M8 to 2015M8, and random forests are used as predictive models in the leaves of the tree. Panel A shows the number of observations in each leaf, the return predictability (R^2, in percentages) from cluster-specific (R^2_C) and global (R^2_G) predictive models, and their difference ($R^2_{CMG} = R^2_C - R^2_G$). Panel B reports the average monthly excess return (in percentages) and the annualized Sharpe ratio for the equal-weighted (EW) and value-weighted (VW) portfolios of clustered asset-return observations. Rows are sorted in descending order of R^2_C.

	Panel A				Panel B			
Leaf	Obs.	R^2_C	R^2_G	R^2_{CMG}	Avg_{EW}	SR_{EW}	Avg_{VW}	SR_{VW}
N9	34,798	7.98	6.17	1.80	0.34	1.25	0.33	1.19
N14	35,798	7.59	5.67	1.92	0.67	1.21	0.54	0.98
N13	31,407	5.68	4.59	1.08	0.68	1.21	0.62	1.06
N15	39,362	5.19	4.02	1.17	0.65	1.30	0.53	1.01
N21	14,654	4.62	3.54	1.08	0.50	0.93	0.50	0.88
N25	23,050	4.60	2.84	1.76	0.78	1.09	0.59	0.93
N22	30,851	3.33	2.13	1.20	0.61	0.91	0.54	0.79
N20	12,391	3.15	2.08	1.07	0.38	0.67	0.37	0.57
N8	95,954	2.97	2.15	0.82	0.23	0.75	0.23	0.72
N24	48,819	2.69	2.32	0.37	0.64	1.03	0.52	0.90
N23	86,131	1.06	0.33	0.72	0.56	0.66	0.54	0.68

global RF models.

Moreover, Panel B of Table 8 report the investment outcomes for the clusterwise portfolios. More specifically, equal-weighted (EW) and value-weighted (VW) clusters are constructed, and their average excess returns (Avg; in percentages) and annualized Sharpe ratios (SR) are presented in Panel B. As in Section 5, we conclude that there is not a strong relationship between return predictability and portfolio performance. In fact, the weakly positive relationship observed for Ridge models largely disappears—or even reverses—for RFs in Table 8. For example, the cluster with the highest R^2_C (N9) has an EW average monthly excess return of 0.34%, whereas the cluster with the lowest R^2_C (N23) has an average return of 0.56%. Yet the corresponding SRs are 1.25 and 0.66, respectively, indicating that the high SR of the most predictable cluster stems from lower volatility (i.e., denominator) rather than higher returns—again a potential sign of overfitting.

Finally, Table 9 presents the in- and out-of-sample return predictability estimates of the clusterwise RF models. As in Table 3 of Section 5, Panel B shows that the level of return predictability estimated over the full out-of-sample period (2015M9–2022M8) do not align with the in-sample estimates. More specifically, the most (least) predictable cluster in-sample are not the most (least) predictable out-of-sample. Again, we observe that this divergence is largely driven by the final year of the evaluation window (2021M9–2022M8).

Table 9
This table presents R^2 values (in percentages; Eq. (4)) estimated from the global predictive model and clusterwise models from the cross-sectional panel tree in Figure 11, both in- and out-of-sample. More specifically, we define four different samples of observations: (i) *Aggregate*, all forecasts combined; (ii) *High*, the forecasts for observations ending up in clusters N9, N14 or N13 in Figure 11 (i.e., with highest in-sample R^2_C); (iii) *Low*, the forecasts for observations ending up in the clusters N23, N24 or N8 (i.e., with lowest in-sample R^2_C); and (iv) *Medium*, the forecasts for all observations not ending up in either sample *High* or *Low*. Panel A presents the R^2 values estimated by a random forest fitted to all in-sample observations grouped together (i.e., a global model), but evaluated on all samples separately to produce the forecasts. Panel B reports similar results, but now the clusterwise models—random forests fitted to all observations in the same cluster—are used to produce the forecasts.

	In-sample	Out-of-sample		
	2002M8–2015M8	2015M9–2022M8	2015M9–2021M8	2021M9–2022M8
Panel A: Global forecasts				
Aggregate	1.92	-0.50	2.05	-14.08
High	5.34	0.08	5.04	-18.28
Medium	2.76	-0.01	3.68	-14.70
Low	0.50	-0.90	0.70	-12.24
Panel B: Clusterwise forecasts				
Aggregate	3.04	-0.32	2.29	-14.28
High	7.23	-0.76	5.82	-25.11
Medium	3.98	0.19	4.46	-16.77
Low	1.37	-0.56	0.58	-8.66

G.3 An anomaly related to predictability disagreement

Table 10 reports the summary statistics (Panel A) an abnormal returns (Panel B) for the *predictability disagreement anomaly* when RFs are used as clusterwise predicitive models. Similar as for the Ridge regression results in Table 3 of Section 5, the anomaly does not yield statistically significant abnormal returns in the corporate bond market. However, in contrast to the Ridge regression case, the results in Table 9 also lack economic significance.

Table 10
This table reports summary statistics (Panel A) and abnormal returns (Panel B) for long–short portfolios based on the return-predictability disagreement measure R^2_{CMG} from Table 8. Clusters are first sorted in descending order of R^2_{CMG}, and value-weighted portfolios are constructed for each cluster. For each $i \in \{1, 2, 3, 5\}$, we form an equal-weighted long position in the top i clusters (highest predictability disagreement) and an equal-weighted short position in the bottom i clusters (lowest predictability disagreement), yielding the Ti–Bi long–short portfolio. Clusters are estimated using in-sample data from 2002M8 to 2015M8 and correspond to the panel tree in Figure 11. Results are shown separately for the in-sample and out-of-sample periods. Panel A reports the average monthly excess return (Avg, in percentages), standard deviation (Std, in percentages), and annualized Sharpe ratio. Panel B reports abnormal returns (alpha, in percentages) from the factor models in Appendix D, with Newey and West (1987) t-statistics in parentheses (lag length of 3). Statistical significance at the 5% and 10% levels is denoted by $**$ and $*$, respectively.

	In-sample				Out-of-sample			
	T1–B1	T2–B2	T3–B3	T5–B5	T1–B1	T2–B2	T3–B3	T5–B5
Panel A: Summary statistics								
Avg	0.004	-0.099	0.014	0.050	0.070	-0.012	0.006	0.068
Std	2.178	1.624	1.088	0.894	1.755	1.335	0.708	0.623
SR	0.006	-0.210	0.046	0.194	0.138	-0.030	0.027	0.377
Panel B: Abnormal returns								
CAPMB	0.214*	0.143*	0.125*	0.127**	0.132	0.041	0.025	0.074
	(1.823)	(1.858)	(1.927)	(2.094)	(1.010)	(0.463)	(0.408)	(1.136)
DEFTERM	-0.024	-0.066	-0.004	0.031	0.177	0.089	0.044	0.078
	(-0.167)	(-0.724)	(-0.071)	(0.642)	(1.076)	(0.731)	(0.577)	(1.109)
CAPM	0.121	0.023	0.093	0.106*	0.306*	0.169	0.096	0.112*
	(0.745)	(0.192)	(1.343)	(1.824)	(1.872)	(1.217)	(1.512)	(1.773)
HKM	0.130	0.017	0.087	0.100	0.327**	0.180	0.102*	0.110*
	(0.839)	(0.137)	(1.184)	(1.593)	(2.232)	(1.378)	(1.712)	(1.758)
HKMSF	0.049	-0.047	0.049	0.075	0.144	0.051	0.037	0.090
	(0.312)	(-0.407)	(0.745)	(1.380)	(0.713)	(0.301)	(0.450)	(1.385)
FF3	0.138	0.032	0.098	0.110**	0.322**	0.179	0.100*	0.107*
	(1.025)	(0.310)	(1.573)	(2.057)	(2.173)	(1.380)	(1.648)	(1.772)
FF5	0.095	0.012	0.086	0.099*	0.327**	0.170	0.104*	0.100*
	(0.625)	(0.108)	(1.257)	(1.701)	(2.239)	(1.354)	(1.710)	(1.743)
FF5+MOMS	0.864*	0.843**	0.447*	0.148	0.393	0.428	0.269	0.162
+IVOL	(1.664)	(2.237)	(1.689)	(0.624)	(0.471)	(0.768)	(0.734)	(0.423)

H Full sample analysis

In this appendix, we replicate the analysis in Section 5 for the full sample (2002M8–2022M8), instead of splitting the dataset into a training and test set. The resulting panel tree is presented in Figure 14, which uses Ridge regressions as clusterwise predictive models. We observe that the first splitting variable is `STREVB` (i.e., short-term bond reversal), and the resulting left child node ($N2$) considers `D2D` (i.e., distance-to-default) to partition the observations further. Interestingly, the panel tree in Figure 4 of Section 5—which is estimated using only the training data (2002M8–2015M8)—considers `D2D` as the first splitting variable, and uses `STREVB` to partition the observations in the resulting left child node. Thus, nodes N1 and N2 of the panel trees in Figure 4 and Figure 14 are the same, only is the order of the split rules reversed. Furthermore, in contrast to the panel tree in Figure 4, we observe that node N2 of the full-sample P-Tree is not chosen as the second split (i.e., S2), but instead as the fifth split in our clustering algorithm. The full-sample panel tree in Figure 14 considers instead `YLDTOWORST` (i.e., bond yield-to-worst) as the optimal second split.

However, besides these differences in the tree-structure, we do observe that the full-sample clustering results are very similar to those presented in Section 5. Figure 9 replicates Figure 5 of Section 5, but now under the full-sample panel tree in Figure 14. We can make the same four conclusions from the results in these heatmaps. First, the returns of corporate bonds with high previous-month returns (i.e., `STREVB`) are more predictable. Second, corporate bonds with high distance-to-default (i.e., `D2D`) exhibit stronger return predictability. Third, return predictability is fairly homogeneous across rating-sorted deciles of corporate bonds. Fourth, in particular the returns on corporate bonds with low `D2D` *and* low `STREVB` are substantially less predictable.

Analogous to Table 2 in Section 5, Table 11 report the summary statistics (Panel A) and portfolio performance (Panel B) for the clusters identified by the full-sample panel tree in Figure 14. As in Section 5, we observe a relatively weak relationship between clusterwise predictability and portfolio performance. More specifically, while the relationship between return predictability and the annualized Sharpe ratio (SR) seems to be positive, the relationship the with average excess returns on the clusterwise portfolios (Avg, in percentages) is less clear.

All in all, we can conclude that the full-sample results reported in this appendix align with the results in Section 5. While the full-sample panel tree is (slightly) different from the panel tree in Figure 4 of Section 5, the estimated *mosaics of predictability* in Figure 5 and Figure 9 are very similar.

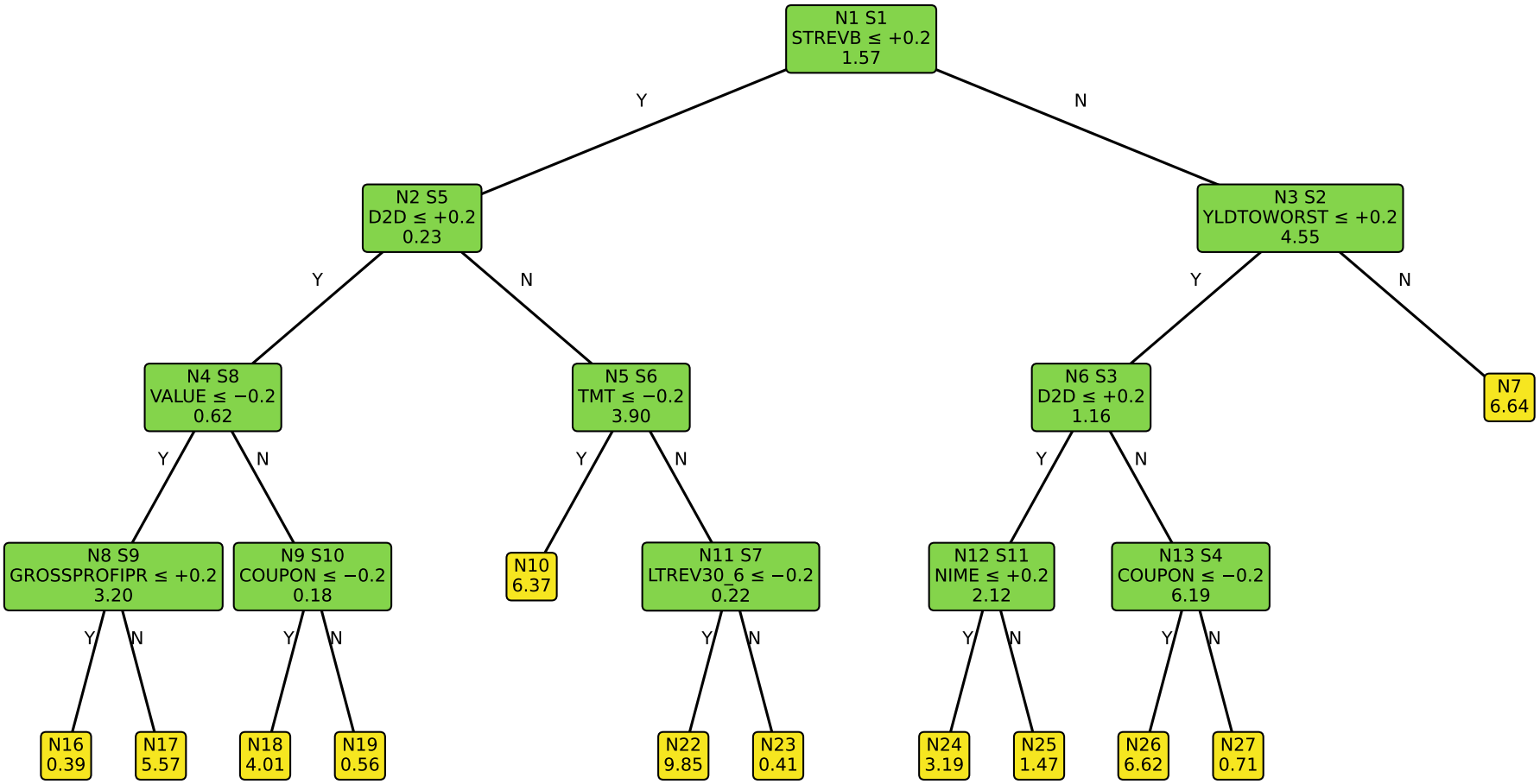


Figure 14: This figure illustrates the panel tree used to form clusters of bond-return observations with heterogeneous levels of predictability, measured by R^2. The tree is estimated using monthly data from 2002M8 to 2022M8, and splits on cross-sectional asset characteristics that are rank-standardized to the interval $[-1, 1]$ (see Appendix C). Each intermediate node (green) and leaf node (yellow) displays the node's ID, denoted by $N\#$, which is used to identify the node in this paper. Intermediate nodes additionally report the split rule (second line), which sends observations satisfying the rule to the left child node and the rest to the right child node, as well as the node's split order, denoted by $S\#$. Finally, all nodes display the in-sample R^2 statistic (in percentages; last line), estimated using the clusterwise Ridge regressions introduced in Section 3.

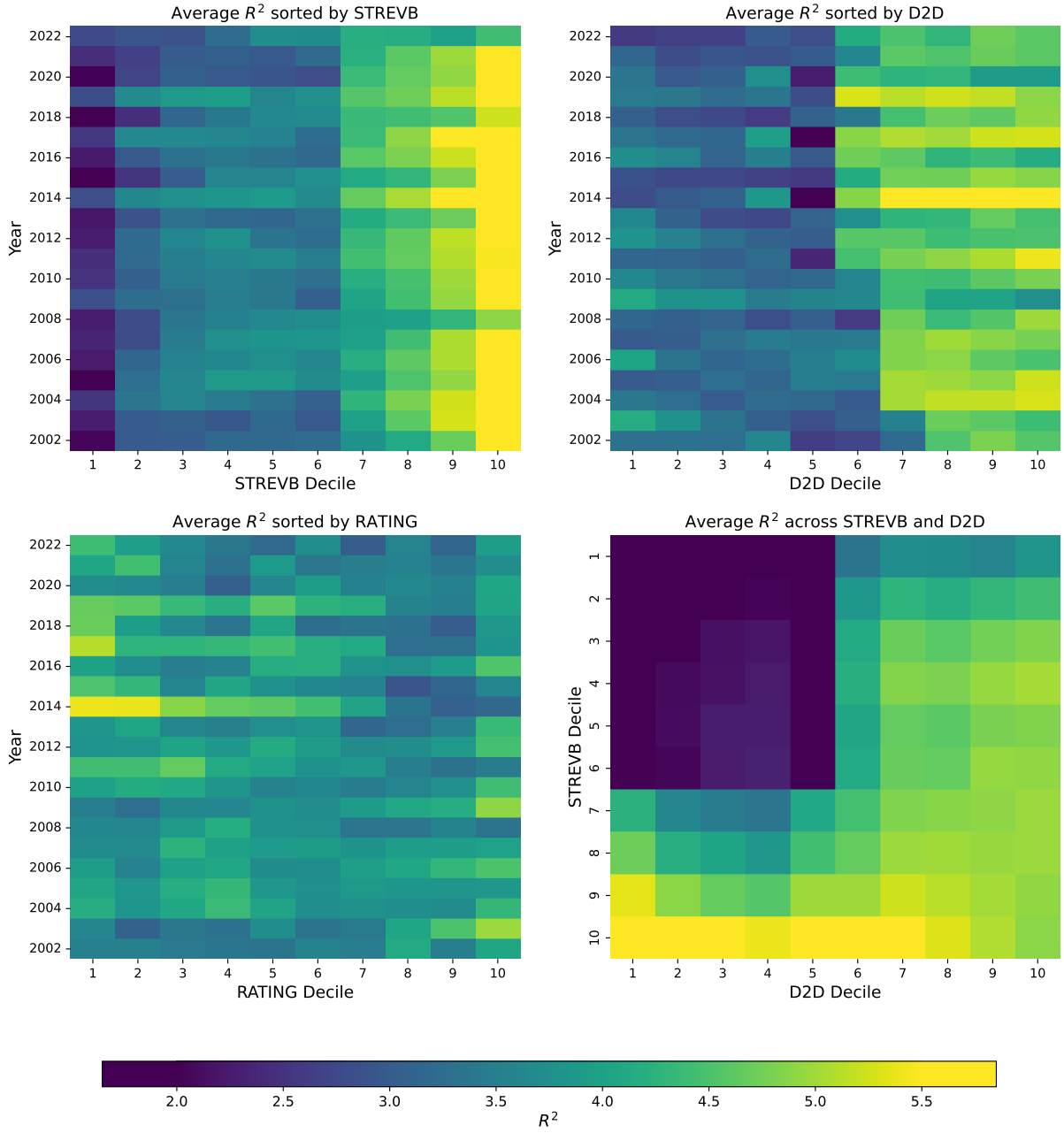


Figure 15: This figure presents four heatmaps of average in-sample R^2 under the panel tree in Figure 14 and when sorting observations into deciles of bond characteristics. The top panels show average R^2 for asset-returns observations sorted annually into deciles of short-term bond reversal (`STREVB`, left) and distance-to-default (`D2D`, right). The bottom left panel shows results for annual rating deciles, while the bottom right panel shows R^2 for 10×10 bivariate-sorted groups based on the first two splitting variables in Figure 4: `D2D` and `STREVB`. Lighter colors correspond to higher average R^2 values, as shown by the colorbar. Note that higher rating deciles are associated to lower-rated bonds, as can be seen from the definition of `RATING` in Table 5 of Appendix C.

Table 11

This table report summary statistics for each cluster identified by the panel tree in Figure 14. The sample period is from 2002M8 to 2022M8, and Ridge regressions are used as predictive models in the leaves of the tree. Panel A shows the number of observations in each leaf, the return predictability (R^2, in percentages) from cluster-specific (R^2_C) and global (R^2_G) predictive models, and their difference ($R^2_{CMG} = R^2_C - R^2_G$). Panel B reports the average monthly excess return (in percentages) and the annualized Sharpe ratio for the equal-weighted (EW) and value-weighted (VW) portfolios of clustered asset-return observations. Rows are sorted in descending order of R^2_C.

	Panel A				Panel B			
Leaf	Obs.	R^2_C	R^2_G	R^2_{CMG}	Avg_{EW}	SR_{EW}	Avg_{VW}	SR_{VW}
N22	43,311	9.85	1.25	8.60	0.39	0.61	0.40	0.61
N7	157,692	6.64	4.63	2.01	0.72	0.92	0.59	0.77
N26	58,690	6.62	0.15	6.47	0.29	0.65	0.24	0.48
N10	106,410	6.37	-3.99	10.35	0.18	0.67	0.18	0.67
N17	26,220	5.57	2.23	3.34	0.30	0.52	0.27	0.49
N18	64,789	4.01	1.65	2.36	0.27	0.47	0.23	0.45
N24	42,475	3.19	-0.36	3.55	0.29	0.65	0.27	0.57
N25	40,043	1.47	-1.21	2.68	0.22	0.52	0.21	0.46
N27	39,882	0.71	0.64	0.06	0.31	0.61	0.30	0.55
N19	116,954	0.56	0.07	0.49	0.30	0.35	0.30	0.40
N23	101,348	0.41	0.48	-0.08	0.38	0.59	0.39	0.59
N16	57,915	0.39	0.90	-0.51	0.19	0.38	0.15	0.30