Backtesting ESG Ratings

Christophe Boucher*, Wassim Le Lann†, Stéphane Matton‡, Sessi Tokpavi§

May, 2021

Abstract

Sustainable investing is growing fast and investors are increasingly integrating environmental, social and governance (ESG) criteria. However, ESG ratings are derived using heterogeneous data and non-standardised methodologies and so are quite divergent across providers, and this suggests that a formal statistical procedure is needed to evaluate the accuracy of any given ESG rating system. This paper develops a backtesting procedure that evaluates how well these extra-financial metrics help in predicting whether ESG risks will materialise, as given by increased idiosyncratic volatility. Technically, the inference is based on extending the conditional predictive ability test of Giacomini and White (2006) to a panel data setting. We apply our methodology to two ESG rating systems from Sustainalytics and Asset4 for Europe, North America and the Asia-Pacific region. The results show that the null hypothesis of lacking informational content of ESG ratings is rejected in Europe and North America, where there is less disagreement between the two ratings, while the results are mixed for the Asia-Pacific region, where there is more disagreement. Furthermore, applying the test only to firms with convergent ESG ratings leads to the null hypothesis being rejected for all three regions. Beyond providing insights into the accuracy of each of the ESG rating systems, these results suggest that information gathered from several ESG rating providers should be cross-checked before ESG is integrated into investment processes.

JEL Codes: G10, G17, C12, C33.

Keywords: Backtesting, ESG ratings, ESG risks, ESG-related events, Idiosyncratic realised volatility, Test of equal predictive power, Panel data, Consensus ESG ratings.

*christophe.boucher@parisnanterre.fr, EconomiX-UPL, CNRS, University Paris Nanterre

†Corresponding author. w.lelann@outlook.fr, University of Orléans, CNRS, LEO, FRE 2014, Orléans, France.

‡stephane.matton@parisnanterre.fr, EconomiX-UPL, CNRS, University Paris Nanterre

§sessi.tokpavi@univ-orleans.fr, University of Orléans, CNRS, LEO, FRE 2014, Orléans, France.

1 Introduction

Is there actually any informational content in the various existing ESG rating systems? Is this informational content related to what it is supposed to measure, which is the varying degrees of exposure to ESG risks of the companies being rated? This article aims to provide a statistical methodology that will allow these questions to be answered. To do this we develop a backtesting procedure to check for the informational content of ESG ratings about ESG risks. Our test proceeds by evaluating how well these extra-financial metrics help in predicting whether ESG risks will materialise, as given by increased idiosyncratic realised volatility, and this beyond the information conveyed by traditional financial variables.

There is obviously great interest in this issue, and ESG is among the best known acronyms by far today in the financial world and beyond it. This can be linked to the growing interest in the three facets of sustainable investing that emerged following the 2008 financial crisis as companies started developing ethical standards and best practices summarised as ESG. Today ESG ratings increasingly shape the investment decisions of investors. According to Bloomberg, ESG assets are on track to exceed $53 trillion by 2025, representing more than a third of the $140.5 trillion in projected total assets under management.[1]

The global craze for responsible investment has by now led to an abundant and rich literature that has tried, with mixed results, to evaluate how sustainable investment impacts market variables, and asset prices in particular. Some studies have found that ESG has a positive impact on asset prices (Mozaffar et al., 2016; Amel-Zadeh and Serafeim, 2018; Dyck et al., 2019; Hartzmark and Sussman, 2019), and Mozaffar et al. (2016) for instance present evidence that firms doing well on ESG issues outperform firms doing poorly on these issues. Amel-Zadeh and Serafeim (2018) reaffirm that ESG ratings have a material impact on asset prices and more specifically on the cost of capital, as investors expect higher return on equity for companies with strong ESG performance. Dyck et al. (2019) also demonstrate that engagement by investors has a positive impact on ESG performance and ultimately on financial returns, especially in countries where ESG issues are important. A study of US mutual funds flows confirms that investors find value in sustainability as a positive predictor of future returns (Hartzmark and Sussman, 2019). Arguing the other side though are some works (Riedl and Smeets, 2017; Pastor et al., 2020; Pedersen et al., 2020) based on the impact of investor preferences on the dynamics of asset prices (Fama and French, 2007), which report that ESG practices have either a negative or a positive impact on asset prices. Considering investor preferences for ESG, Riedl and Smeets (2017) notice that investors are

[1] https://www.bloomberg.com/professional/blog/esg-assets-may-hit-53-trillion-by-2025-a-third-of-global-aum/

willing to accept lower expected returns and higher management fees for holding companies with strong ESG performance. Pastor et al. (2020) model investor preferences for ESG in a mean-variance framework and show that in equilibrium, assets considered green generally have lower expected returns but provide greater utility and offer the ability to hedge against climate risk. They also introduce an ESG-factor that reacts to unexpected change in ESG, then conclude that green assets outperform when a positive shock hits this factor. Pedersen et al. (2020) extend the mean-variance-ESG framework by adding a third type of investor who is unaware of the ESG performance of firms. How the ESG ratings affect expected returns then depends on the wealth of this third investor.

Although this literature is interesting and provides useful information on the link between sustainable investment and asset price dynamics, it does not allow us to assess whether the available rating systems are effective in the sense that they are a thermometer of ESG risks. This gap in the literature is all the more worrying as the correlations between the ratings of the various available providers are weak. Indeed, the divergence of ESG ratings has been widely documented (Chatterji et al., 2009; Semenova and Hassel, 2015; Chatterji et al., 2016; Berg et al., 2020), and Berg et al. (2020) find for instance that correlations between the ESG ratings of providers are on average 61% for a set of five different ESG providers, whereas the credit ratings from the main agencies exhibit on average a correlation of 99%. They further explore the source of this divergence by splitting it into three components and looking at scope, or the selection of ESG categories to be measured; measurement, or how the ESG categories are assessed; and weight, or the importance given to each category. They observed that measurement explains more than 50% of the total divergence.[2] Billio et al. (2019) even notice that heterogeneity in ESG ratings can lead to completely opposite opinions on one and the same company and disperses the ESG preferences of investors.

Against this background, our paper introduces a statistical inferential procedure that allows the backtesting of ESG ratings. The test is articulated around the idea that being involved in ESG-related events, as evidenced by increased idiosyncratic volatility, is a sign of being exposed to ESG risks or to the failure of a firm to manage those risks. As a consequence, ESG ratings should have significant power in predicting ESG events. This idea can be found in Champagne et al. (2019) and Serafeim and Yoon (2021), who examine the link between extra-financial performance or ESG ratings, and the likelihood of adverse ESG events. Their analysis is based on the hypothesis that firms with strong extra-financial per-

[2]Unlike credit ratings, ESG ratings are most often created mainly from non-standardised information and are not regulated. Methodologies can be opaque and proprietary, leading to substantial rating divergence. International organisations like the WWF have warned against viewing ESG ratings as a meaningful indicator of how an investment strategy contributes to achieving sustainability goals. See e.g., WWF response to the Call for Feedback on the TEG interim report on EU Climate Benchmarks and ESG disclosures of benchmarks (Rendlen and Weber, 2019).

formance in the form of good environmental externalities, and good employee relationships and governance are less likely to face ESG events like environmental problems, violations of regulatory standards, employee claims, social conflicts, or boycotts and negative media campaigns than firms with poor extra-financial performance are.

On the empirical side, Champagne et al. (2019) use a logistic regression to check whether a firm's extra-financial performance during a given year helps significantly in anticipating ESG events the following year. Using a unique database of adverse corporate ESG events between 2001 and 2013, they indeed observe that an increase of one unit in a firm's rating reduces its probability of facing adverse events during the following year by 8%, and this result holds even after controlling for the impact of financial performance variables such as the market-to-book-value ratio, the firm's size, or the volatility of its shares. A similar approach can be found in Serafeim and Yoon (2021), who investigate whether ESG ratings predict future ESG news and the associated market reactions. Using a firm-day panel dataset they find that the latest outstanding consensus ESG rating is associated with future ESG news, but the link diminishes for firms over which there is large disagreement among raters.

Our contribution is related to these two works, but differs in many points. First, these papers used proprietary tools to identify and measure ESG risks from ESG incidents, but we use here an objective measure that is not subject to data revisions and the kind of backfill bias that comes from ex-post adjustments by providers of databases of ESG incidents. The measure is an approximation of the materialisation of ESG risks found from an increase in the idiosyncratic volatility of a firm's stock. So an ESG rating correctly measures ESG risks in our approach if its high values that are associated with good environmental, social and governance practices correspond to low levels of idiosyncratic volatility in the shares of firms, and its low, or bad, values correspond to high volatility, once the effects of traditional financial variables are taken into account. Further reinforcement for our choice is that ESG incidents themselves also seem divergent across providers. The rank correlations between ESG incidents from Sustainalytics and Asset4 for instance are weak at 43% for Europe, 43% for North-America and 34% for the Asia-Pacific region.[3] Second, we test for the informational content of the ESG ratings in an out-of-sample environment, taking a dynamic forward looking approach that fits with the practice of institutions, which are supposed to revise their ratings over time to take account of the information available on environmental, social and governance practices. Third, as we will discuss later, our approach allows for the possible misspecification of the model used to measure the link between ESG ratings and

[3]These figures are computed over the period from January 2010 to October 2018.

ESG risks, while in the two earlier papers, the existence of this link depends deeply on the correct specification of their econometric models.

It is worth stressing that our framework is based on the assumption that increased idiosyncratic volatility is expressed as the materialisation of ESG events. This hypothesis is empirically verified by the market stress that generally accompanies actual ESG events when they happen, with a sharp increase in idiosyncratic risk. An example occurred in September 2015 when the German car manufacturer Volkswagen came under fire after admitting that defective devices were installed on 11 million vehicles to cheat on emissions tests. This failure in good governance practices led to market stress and high levels of variability in prices as the share price fell by 40% in just a few days. Moreover, some other works have also reported a link between ESG risks and idiosyncratic volatility (Jo and Na, 2012; Bouslah et al., 2013; Sodjahnin et al., 2017; Hoepner et al., 2018; Albuquerque et al., 2019; Ilhan et al., 2019). Jo and Na (2012) for instance note that companies with lower leverage and high ESG ratings are best at capturing the benefits of ESG performance in order to reduce idiosyncratic risk.[4]

Technically, our inferential procedure for checking for the informational content about ESG risks in the ESG ratings is based on extending the conditional predictive ability test of Giacomini and White (2006) to a panel setting. Under weak assumptions including cross-sectional dependencies between the idiosyncratic volatilities of firms' shares, we derive the Gaussian asymptotic distribution of the test statistic. Monte Carlo simulations conducted under different types of model misspecification show that our test has good small sample properties. Most notably, the test has good sizes and increasing powers with n as the number of firms and T as the sample length.

We conduct empirical applications to illustrate our methodology, using two leading ESG rating systems, Sustainalytics and Asset4, for Europe, North America and the Asia-Pacific region. Our results show that the null hypothesis of ESG ratings lacking informational content about ESG risks is rejected for Europe and North America, where there is less disagreement between the two ratings, while the results are mixed for the Asia-Pacific region, where there is more disagreement. Furthermore, applying the test only to firms with convergent ESG ratings leads to the null hypothesis being rejected for all three regions. These results can be linked to that highlighted by Serafeim and Yoon (2021). They use a different approach to analyse the link between ESG ratings and ESG risks as measured by ESG events, and show that the consensus rating predicts future news, but its predictive ability diminishes for firms over which there is large disagreement among raters. From a

[4]Note that these works do not provide a formal test to check for the informational content of ESG ratings about ESG risks, as it is the purpose in our article.

practical point of view, our results provide crucial information for portfolio managers who apply sustainable management for their funds, as we show that it is necessary to cross-check the information gathered from several providers of ESG ratings before integrating ESG into the management process.

The rest of the article is organised as follows. Section 2 describes our backtesting procedure for ESG ratings, focusing on the formulation of the null hypothesis, the construction of the test statistic and the analysis of its asymptotic distribution. Section 3 simulates the small sample properties of the test statistic under various settings, and empirical applications are considered in Section 4. The last section concludes the paper.

2 The backtesting procedure

This section gives a description of the backtesting procedure for evaluating statistically the informational content about ESG risks in the ESG ratings. In the first part, we fix the notations and clearly define the null hypothesis of interest, while in the second part we provide the test statistic and its asymptotic distribution for inference.

2.1 Notations and the null hypothesis

To formulate the null hypothesis of our test, we take an investment universe with n traded firms, and let $r_{i,s}$ be the daily returns on stock i at time s. The daily returns on J factors like size, value or momentum are also available and are reported by the literature to drive the cross-sectional variations of the returns on the stocks. Let $r_{j,s}$ denote the return on factor j at time s, with $j = 1, ..., J$. As the philosophy of our test is to check for the informational content that ESG ratings give on ESG risks as given by increased idiosyncratic volatility, we use a factor model to compute the values of this volatility, yielding

$$r_{i,s} = \alpha_i + \sum_{j=1}^{J} \beta_{i,j} r_{j,s} + \epsilon_{i,s}, \tag{1}$$

with $i = 1, ..., n$, α_i as the alpha of stock i, $\beta_{i,j}$ as its exposure to factor j, and $\epsilon_{i,s}$ as the residual returns. Estimating model (1) with the ordinary least squares (OLS) method provides estimates $\widehat{\alpha}_i$ and $\widehat{\beta}_{i,j}$ of the parameters α_i and $\beta_{i,j}$, and hence the daily time series of the estimated residual returns $\widehat{\epsilon}_{i,s}$, $s = 1, ...S$, with S the number of days in the sample. From the daily residual returns, we may compute the monthly idiosyncratic realised volatility as follows

$$IRV_{i,t} = \sum_{s_k=1}^{v_t} \widehat{\epsilon}^2_{i,s_k}, \tag{2}$$

with t as the index of the month, v_t as the number of daily observations in month t, and $\widehat{\epsilon}_{i,s_k}$ as the s_k^{th} daily residual returns within month t. For each firm i, we thus obtain a time series of monthly idiosyncratic realised volatility of length T.

Let $x_{i,t}$ be a vector of length p in which the elements are innovations on p balance sheet variables that measure the financial strength of firm i for the month t. Examples of such variables are dividend yield, sales over assets, debt over assets, or the quick ratio. They measure different facets of a firm's solvency including its size, returns, risk, liquidity, debt and leverage. Innovations can be obtained through autoregressive filtering on raw balance-sheet variables, or simply as deviations from the long-term average. Finally, the value of an ESG rating is available for each firm i at month t and we denote it as $\omega_{i,t} \in \mathbb{R}$. This can be a global ESG rating measuring environmental, social and governance issues, or only one of these three components.

Now let $m_{i,t+\tau}^{(0)} = \mathbb{E}(IRV_{i,t+\tau} \,|x_{i,t})$ be the unknown expected value of idiosyncratic realised volatility of firm i at time $t+\tau$, conditional on its financial strength as measured by innovations $x_{i,t}$ in balance-sheet variables, with τ as a given forecast horizon. We can use a given predictive model, whether parametric, semi-parametric or non-parametric, to forecast $m_{i,t+\tau}^{(0)}$. The forecast we denote $\widehat{m}_{i,t+\tau}^{(0)}(\widehat{\beta}_{t,b_t}^{(0)})$ is based on the information set available at time t for all firms, so $\mathcal{F}_t^{(0)} = \{IRV_{i,s}, x_{i,s}, s = t - b_t + 1, ..., t, i = 1, ..., n\}$, where b_t refers to the size of the estimation sample and $\widehat{\beta}_{t,b_t}^{(0)}$ collects all the estimated parameters. In a parametric model like a linear regression, $\widehat{\beta}_{t,b_t}^{(0)}$ is the vector of the estimates of the unknown parameters. Otherwise, it corresponds to whatever semi-parametric or non-parametric estimators are used to forecast $m_{i,t+\tau}^{(0)}$.

Let $m_{i,t+\tau}^{(1)} = \mathbb{E}(IRV_{i,t+\tau} \,|x_{i,t}, \omega_{i,t})$ be defined as $m_{i,t+\tau}^{(0)}$, but with the conditional set extended to $\omega_{i,t}$, so $\mathcal{F}_t^{(1)} = \{IRV_{i,s}, x_{i,s}, \omega_{i,s}, s = t - b_t + 1, ..., t, i = 1, ..., n\}$. In other words, $m_{i,t+\tau}^{(1)}$ is the expected value of idiosyncratic realised volatility of firm i at time $t+\tau$, conditional on its financial states as given by $x_{i,t}$ and also on its ESG rating as given by $\omega_{i,t}$. We denote $\widehat{m}_{i,t+\tau}^{(1)}(\widehat{\beta}_{t,b_t}^{(1)})$ as the forecast value at time $t+\tau$.

Suppose that we produce T_0 out-of-sample forecasts of both the expected values $m_{i,t+\tau}^{(0)}$ and $m_{i,t+\tau}^{(1)}$ for each firm, so $\widehat{m}_{i,t+\tau}^{(0)}(\widehat{\beta}_{t,b_t}^{(0)})$ and $\widehat{m}_{i,t+\tau}^{(1)}(\widehat{\beta}_{t,b_t}^{(1)})$, $i = 1, ..., n$, $t + \tau = 1, ..., T_0$. With a loss function at hand that we denote $\mathcal{L}(.)$, we can evaluate the predictive performance of each model, generating two panels of losses as $\mathcal{L}_{i,t+\tau}^{(0)} \equiv \mathcal{L}_{i,t+\tau}^{(0)}(IRV_{i,t+\tau}, \widehat{m}_{i,t+\tau}^{(0)}(\widehat{\beta}_{t,b_t}^{(0)}))$ and $\mathcal{L}_{i,t+\tau}^{(1)} \equiv \mathcal{L}_{i,t+\tau}^{(1)}(IRV_{i,t+\tau}, \widehat{m}_{i,t+\tau}^{(1)}(\widehat{\beta}_{t,b_t}^{(1)}))$, where again $IRV_{i,t+\tau}$ is the idiosyncratic realised volatility of firm i at time $t+\tau$. From these panels, let $\Delta\mathcal{L}_{i,t+\tau} = \mathcal{L}_{i,t+\tau}^{(1)} - \mathcal{L}_{i,t+\tau}^{(0)}$ be the panel of loss differentials. For each firm, the loss differentials can be averaged across time,

yielding

$$\mu_i(\widehat{\beta}^{(0)}_{t,b_t}, \widehat{\beta}^{(1)}_{t,b_t}) = T_0^{-1} \sum_{t+\tau=1}^{T_0} \Delta\mathcal{L}_{i,t+\tau}. \tag{3}$$

Hence the null hypothesis of equal predictive ability of the two forecasting models can be stated as

$$\mathbb{H}_0 : \mathbb{E}(\mu_i(\widehat{\beta}^{(0)}_{t,b_t}, \widehat{\beta}^{(1)}_{t,b_t})) = 0, i = 1, 2, ..., \tag{4}$$

with the alternative hypothesis defined as

$$\mathbb{H}_1 : \mathbb{E}(\mu_i(\widehat{\beta}^{(0)}_{t,b_t}, \widehat{\beta}^{(1)}_{t,b_t})) < 0, i = 1, 2, ... \tag{5}$$

This null hypothesis calls for several remarks. First, when it holds, it means that including the ESG rating in the information set does not help for forecasting idiosyncratic realised volatility. In consequence, we should conclude that the ESG rating system investigated is void of information about idiosyncratic realised volatility taken as the materialisation of ESG risks. Under the alternative hypothesis, the expectation of $\mu_i(\widehat{\beta}^{(0)}_{t,b_t}, \widehat{\beta}^{(1)}_{t,b_t})$ across firms is negative, which means that considering the ESG rating in forecasting idiosyncratic realised volatility overall gives real benefit across all firms and times.

Second, in contrast to the traditional framework for comparing predictive ability in Diebold and Mariano (1995) and West (1996), we can observe that the null hypothesis involves $\mu_i(\widehat{\beta}^{(0)}_{t,b_t}, \widehat{\beta}^{(1)}_{t,b_t})$, which depends on $\widehat{\beta}^{(0)}_{t,b_t}$ and $\widehat{\beta}^{(1)}_{t,b_t}$, which are the estimated values of the parameters instead of their population values. As discussed by Giacomini and White (2006), this helps preserve the finite sample behaviour of the estimators in the evaluation procedure, hence reflecting the effect of estimation uncertainty on the relative performance of the forecasts, and allowing nested forecasting models to be compared like in our framework. However, they underline that adopting such a framework means remembering that the null hypothesis does not check the equal predictive ability of the competing models, but rather of the forecasting methods, where these methods include the models as well as the estimation procedures and the possible choices of estimation window.

This last remark means that some care is required in applying our test procedure to check for the validity of the null hypothesis in (4). First, the size of the estimation window should be kept fixed in the rolling window procedure ($b_t = b$) to ensure that parameter uncertainty does not vanish asymptotically. This naturally rules out an expanding window forecasting scheme, but allows for iterated or fixed schemes. Second, we should retain the same forecasting model and scheme and the same estimation window length to generate the forecasts $\widehat{m}^{(0)}_{i,t+\tau}(\widehat{\beta}^{(0)}_{t,b})$ and $\widehat{m}^{(1)}_{i,t+\tau}(\widehat{\beta}^{(1)}_{t,b})$. This is an important requirement, as it guarantees that the two forecasts diverge only by the set of information used, $\mathcal{F}^{(0)}_t$ or $\mathcal{F}^{(1)}_t$, the first of which excludes data on the ESG ratings for all firms.

Lastly, the choice of the loss function $\mathcal{L}(.)$ is straightforward as the forecast is for expectations of idiosyncratic realised volatility. Examples of this function are the squared error loss and the absolute error loss.

2.2 Test statistic and asymptotic distribution

In this section, we provide the test statistic for checking for the null hypothesis of a lack of informational content in an ESG rating system as expressed in (4). To do this we use the literature on comparing predictive ability in panel data settings (Davies and Lahiri, 1995; Timmermann and Zhu, 2019; Akgun et al., 2019). This literature considers extending the traditional predictive accuracy test for time series (Diebold and Mariano, 1995; West, 1996; Giacomini and White, 2006, etc) to a panel framework and it provides a test for overall equal predictive ability, meaning for all cross-sectional and time units as specified in (4), and also tests for joint equal predictive ability across cross-sectional units or time clusters.

The test statistic for our null hypothesis of overall equal predictive ability is based on the sample mean of loss differentials over time and units, so

$$\bar{\mu}_{n,T_0} = (nT_0)^{-1} \sum_{i=1}^{n} \sum_{t+\tau=1}^{T_0} \Delta\mathcal{L}_{i,t+\tau} = n^{-1} \sum_{i=1}^{n} \mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)}), \tag{6}$$

and is given by

$$\mathcal{T}_{n,T_0} = \frac{\bar{\mu}_{n,T_0}}{\bar{\sigma}_{n,T_0}/\sqrt{nT_0}}, \tag{7}$$

where

$$\bar{\sigma}_{n,T_0} = n^{-1} \sum_{i=1}^{n} \sigma_{i,T_0}^2, \tag{8}$$

and $\sigma_{i,T_0}^2 = var(\sqrt{T_0}\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)}))$ is the long run variance of the ith time series of loss differentials. For the asymptotic distribution of the test statistic in (7), we need the following assumptions

Assumption 1 *For a given forecast horizon $\tau \geq 1$ and estimation window size $b < \infty$, suppose that (i) $\{(IRV_{i,t}, x'_{i,t}, \omega_{i,t})', t = 1, ..., T_0\}$ for a given i is mixing with ϕ of size $-r/(2r-2)$, $r \geq 2$, or α of size $-r/(r-2)$, $r > 2$; (ii) $\mathbb{E}\left|\Delta\mathcal{L}_{i,t+\tau}\right|^{2r} < \infty$ for all t and a given i; (iii) $\sigma_{i,T_0}^2 = var(\sqrt{T_0}\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)})) > 0$ for all T_0 sufficiently large and a given i.*

Assumption 2 *$\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)})$, $i = 1, ..., N$ are independent, and $\mathbb{E}(\left|\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)})\right|)^{2+\delta} < C < \infty$ for some $\delta > 0$ for all i. $\bar{\sigma}_{n,T_0}^2 = n^{-1}\sum_{i=1}^{n} \sigma_{i,T_0}^2 > \delta' > 0$ for all n sufficiently large.*

Assumption 1 includes regularity conditions for the validity of Theorem 4 in Giacomini and White (2006). These conditions ensure that the test statistic for the unconditional

predictive ability applied to a fixed cross-sectional unit converges to a standard Gaussian distribution under the null hypothesis (4), so

$$\mathcal{T}_i = \frac{\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)})}{\sigma_{i,T_0}/\sqrt{T_0}} \xrightarrow[T_0\to\infty]{\mathcal{D}} N(0,1). \tag{9}$$

Assumption 2 assumes the independence between the n random variables $\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)})$, $i, ..., n$, meaning the average values over time of the loss differentials for each firm. This assumption allows the Central Limit Theory (CLT) applied to independent and heterogeneous random variables (White, 2001, Theorem 5.10) to hold, and eases the derivation of the asymptotic distribution of our test statistic in (7). Note however that this is not a strong assumption in our framework. Indeed, as discussed by Akgun et al. (2019), cross-sectional dependencies between forecast loss differentials averaged over time should arise when the forecast errors of different cross-section units, in this case firms, are affected by common global shocks. If such a statement can be true for forecasting macro-economic variables, where forecast errors have the same direction and magnitude across countries when the world is in some states like economic crisis, it is very unlikely that the errors in forecasting idiosyncratic realised volatility correlate across firms, as they are by nature a specific measure for each firm. The following proposition provides the asymptotic distribution of the test statistic in (7).

Proposition 1 *Under the null hypothesis of a lack of informational content about ESG risks in ESG ratings as stated in (4), and if Assumptions 1-2 hold, we have that*

$$\mathcal{T}_{n,T_0} = \frac{\bar{\mu}_{n,T_0}}{\bar{\sigma}_{n,T_0}/\sqrt{nT_0}} \xrightarrow[T_0,n\to\infty]{\mathcal{D}} N(0,1). \tag{10}$$

Thus we reject the null hypothesis when $\mathcal{T}_{n,T_0} < z_\eta$ with z_η the quantile of order η of the standard Gaussian distribution, and η the nominal significance level. The proof of Proposition 1 is straightforward, as we may note that under $\mathbb{H}_0$,

$$\sqrt{nT_0}\bar{\mu}_{n,T_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sqrt{T_0}\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)}), \tag{11}$$

with $\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)})$ as defined in (3). For a fixed i, if Assumption 1 holds, $\sqrt{T_0}\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)}) \xrightarrow[T_0\to\infty]{\mathcal{D}}$ ψ_i, with $\psi_i \sim N(0, \sigma_{i,T_0}^2)$, and $\sigma_{i,T_0}^2 = var(\sqrt{T_0}\mu_i(\widehat{\beta}_{t,b}^{(0)}, \widehat{\beta}_{t,b}^{(1)}))$. See Theorem 4 in Giacomini and White (2006). Hence the rest of the proof proceeds by noting that under Assumption 2, the CLT for heterogeneous but independent variables (White, 2001, Theorem 5.10) holds and $(1/\sqrt{n})\sum_{i=1}^{n}\psi_i \xrightarrow[T_0,n\to\infty]{\mathcal{D}} N(0, \bar{\sigma}_{n,T_0}^2)$, where again $\bar{\sigma}_{n,T_0}^2 = n^{-1}\sum_{i=1}^{n}\sigma_{i,T_0}^2$.

Note that to compute our test statistic $\mathcal{T}_{n,T_0}$ in (7), we need a consistent estimate $\widehat{\bar{\sigma}}_{n,T_0}^2$ of $\bar{\sigma}_{n,T_0}^2$. Under the assumption of cross-sectional independence of loss differentials, it follows

that $\widehat{\bar{\sigma}}^2_{n,T_0} = n^{-1}\sum_{i=1}^{n}\widehat{\sigma}^2_{i,T_0}$, where $\widehat{\sigma}^2_{i,T_0}$ is a suitable HAC estimator of the long-run variance σ^2_{i,T_0} of the ith time series of loss differentials, with

$$\widehat{\sigma}^2_{i,T_0} = T_0^{-1}\sum_{t+\tau=1}^{T_0}\Delta\mathcal{L}^2_{i,t+\tau} + 2[T_0^{-1}\sum_{j=1}^{p_{T_0}} w_{T_0,j} \times \sum_{t+\tau=1+j}^{T_0}\Delta\mathcal{L}_{i,t+\tau}\Delta\mathcal{L}_{i,t+\tau-j}], \tag{12}$$

and $\{p_{T_0}\}$ is a sequence of integers such that $p_{T_0} \to \infty$ as $T_0 \to \infty$, $p_{T_0} = o(T_0)$, and $\{w_{T_0,j} : T_0 = 1, 2, ...; j = 1, ..., p_{T_0}\}$ is a triangular array such that $|w_{T_0,j}| < \infty$, $T_0 = 1, 2, ...$; $j = 1, ..., p_{T_0}$, $w_{T_0,j} \to 1$ as $T_0 \to \infty$ for each $j = 1, ..., p_{T_0}$ (Andrews, 1991).

3 Small sample properties

In this section we use a realistic simulation framework to analyse the small sample properties of the test. We begin by describing the simulation setup and then provide results for the sizes and the powers of the test under different forms of misspecification for the forecasting method retained.

3.1 Simulation setup

Our simulation setup proceeds by first simulating a vector $x_{i,t}$ of length $p = 10$ of innovations in balance-sheet variables that measure the financial strength of firm i at time t, with $t = 1, ..., T$ and $T \in \{120, 180, 240\}$ as the sample size corresponding to 12, 15 and 20 years of monthly data. To have a realistic setup, these p variables are generated from a multivariate Gaussian distribution with mean vector $\bar{x}$ and covariance matrix Ω calibrated using real data (see Appendix A for details about the calibration). With the vector $x_{i,t}$ of length p ready at hand, we generate the logarithm of idiosyncratic realised volatility for firm i given by

$$\log(IRV_{i,t+1}) = c_i^\star + x'_{i,t}\beta_i^\star + \gamma\omega_{i,t} + u_{i,t+1}, \tag{13}$$

with $u_{i,t+1}$ following a standard Gaussian distribution, $c_i^\star$ as the constant term and $\beta_i^\star$ as a vector of parameters of length p. Note that we allow for heterogeneity across firms with specific values for the parameters for each firm. The values of $c_i^\star$ are generated as follows

$$c_i^\star = c^\star + U(-|\frac{c^\star}{10}|; |\frac{c^\star}{10}|), \tag{14}$$

with $U(a; b)$ as a uniform random variable over the set $[a, b]$. The same perturbation principle is used to generate each component of the vector $\beta_i^\star$, with

$$\beta^\star_{i,j} = \beta_j^\star + U(-|\frac{\beta_j^\star}{10}|; |\frac{\beta_j^\star}{10}|), \tag{15}$$

$j = 1, ..., p = 10$. The reference values $c^\star$ and $\beta^\star$ of the parameters are calibrated using real data (see Appendix A for details).

In equation (13), $\omega_{i,t}$ is the ESG rating, which for firm i and at each date t is generated from a uniform distribution over the set $[0,1]$, and $\gamma \in \mathbb{R}_-$ is a parameter. Note that our null hypothesis holds for $\gamma = 0$, since the ESG rating does not have any predictive content for idiosyncratic realised volatility considered as the materialisation of ESG risks. With γ diverging from zero, the null hypothesis does not hold, because high lagged values of the ESG rating decrease idiosyncratic realised volatility.

Based on our design and for each Monte Carlo replication, with n and T fixed, the above simulation design is run for the n firms, with $n \in \{100, 250, 500\}$. This leads to a pure heterogeneous panel for idiosyncratic realised volatility $IRV_{i,t}$, the $p = 10$ innovations in balance-sheet variables $x_{i,t}$, and the ESG score $\omega_{i,t}$, with $i = 1, ..., n$ and $t = 1, ..., T$.

3.2 Sizes and powers under a medium level of misspecification

For each Monte Carlo replication, we use the generated variables $IRV_{i,t}$, $x_{i,t}$ and $\omega_{i,t}$, $i = 1, ..., n$, $t = 1, ..., T$ and a fixed forecasting method to generate the forecast of $m^{(0)}_{i,t+1} = \mathbb{E}(IRV_{i,t+1}\,|x_{i,t})$ and $m^{(1)}_{i,t+1} = \mathbb{E}(IRV_{i,t+1}\,|x_{i,t}, \omega_{i,t})$, so $\widehat{m}^{(0)}_{i,t+1}(\widehat{\beta}^{(0)}_{t,b})$ and $\widehat{m}^{(1)}_{i,t+1}(\widehat{\beta}^{(1)}_{t,b})$ with b the estimation sample that we set to $b = [0.75T]$, and $[a]$ the integer part of a. This means that we use the first 75% of the T observations for each firm as the estimation sample, and generate T_0 forecasts corresponding to the last 25% of the observations, meaning $T_0 = [0.25T]$ and $T = T_0 + b$.

The forecasts for both models are obtained using pooled OLS regression models. This means that both forecasting models are misspecified, because the true panel structure of the data is heterogeneous across units. Besides, there is another form of misspecification that arises because the true data generating process uses a linear form for the *logarithm* of idiosyncratic realised volatility (see Eq. 13), while the pooled OLS regression models are fitted for the *raw* values of the same variable. Our goal is to evaluate how robust our inferential procedure is to these two levels of misspecification, which we call medium in comparison to another more severe form of misspecification that we will consider next. It may be recalled that the asymptotic behaviour of our test statistic under the null hypothesis suggests that with $\gamma \in \mathbb{R}_-$ in (13) diverging from zero, the null hypothesis is more likely to be rejected for $T_0, n \to \infty$, or equivalently, $T, n \to \infty$.

Figure 1 displays the rejection frequencies of the null hypothesis with respect to the parameter γ for a given couple (n, T), with the nominal significance level set to 5%. The rejection frequencies are computed over $1,000$ simulations. Overall the test exhibits very good small sample properties, and we observe that the rejection frequencies for all couples (n, T) are close to 5% for $\gamma = 0$ and increase monotonically as γ diverges from 0.

We also observe that for a fixed n and $\gamma < 0$ the powers increase with T. Indeed, for $n = 100$ and $\gamma = -0.25$, the rejection frequencies for T of 120, 180 and 240 are 39.10%, 53.30% and 61.00% respectively. The same behaviour is observed for a fixed T and $\gamma < 0$ with the powers increasing with n. For instance with $T = 120$ and $\gamma = -0.25$, the rejection frequencies for $n = 100$, 250 and 500 are respectively 39.10%, 71.50% and 91.30%. Hence our inferential procedure exhibits very good small sample properties. Figure B.1 in Appendix B displays the rejection frequencies for the same simulation setup using the absolute error loss function. We can observe similar small sample properties, offering proof that our test is robust to the loss function.

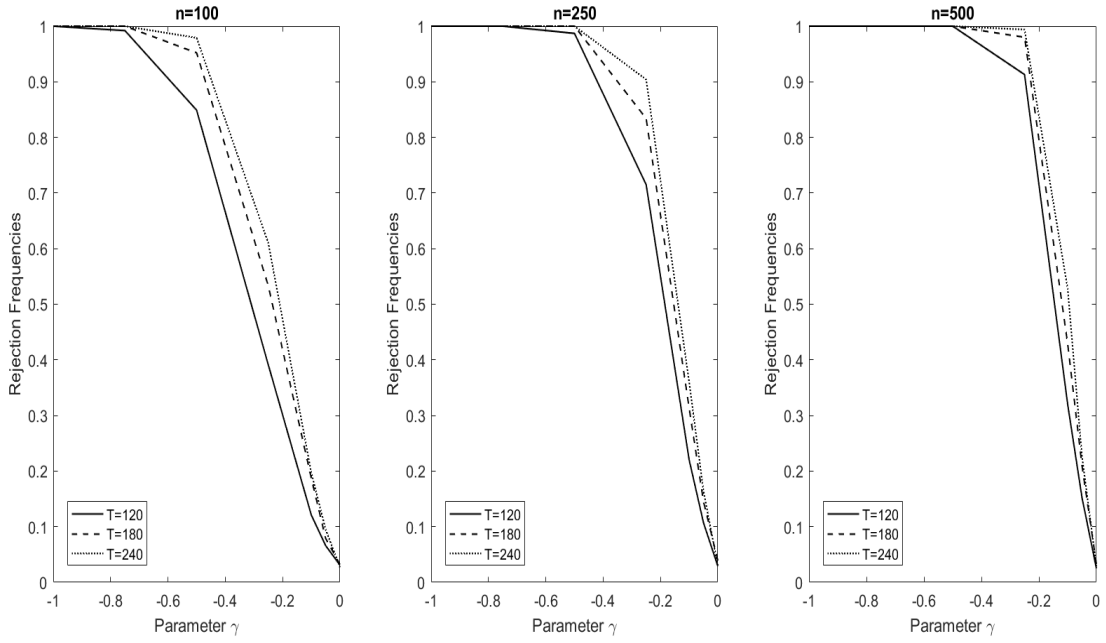


Figure 1: Rejection frequencies under a medium level of misspecification with the squared error loss function

3.3 Sizes and powers under a high level of misspecification

We now consider a configuration that will help us evaluate the properties of the test with respect to the choice of balance-sheet variables. In the last subsection we assumed that the user of the test includes in the forecast models all the $p = 10$ innovations in the balance-sheet variables that enter the specification of the true model, but we make here the assumption that only some of these variables are retained. In each Monte Carlo replication, the following two pooled OLS models are estimated to compute out-of-sample forecasts $\widehat{m}_{i,t+1}^{(0)}(\widehat{\beta}_{t,b}^{(0)})$ and $\widehat{m}_{i,t+1}^{(1)}(\widehat{\beta}_{t,b}^{(1)})$ of $m_{i,t+1}^{(0)} = \mathbb{E}(IRV_{i,t+1}\,|x_{i,t})$ and $m_{i,t+1}^{(1)} = \mathbb{E}(IRV_{i,t+1}\,|x_{i,t}, \omega_{i,t})$, so

$$IRV_{i,t+1} = c + \widetilde{x}_{i,t}'\beta + v_{i,t+1}^{(0)}, \tag{16}$$

$$IRV_{i,t+1} = c + \widetilde{x}'_{i,t}\beta + \omega_{i,t}\gamma + v^{(1)}_{i,t+1}, \tag{17}$$

with $v^{(0)}_{i,t+1}$ and $v^{(1)}_{i,t+1}$ as the error terms and $\widetilde{x}_{i,t}$ as a vector with $p/2$ randomly chosen financial variables from the $p = 10$ relevant ones as its elements, and $\widehat{\beta}^{(0)}_{t,b} = (\widehat{c}, \widehat{\beta}')'$, $\widehat{\beta}^{(1)}_{t,b} = (\widehat{c}, \widehat{\beta}', \widehat{\gamma})'$. Assessing the small sample properties of the test with this additional form of misspecification is of great interest because such misspecification could probably arise in empirical applications where users are very likely to be wrong in their choice of the financial variables that matter.

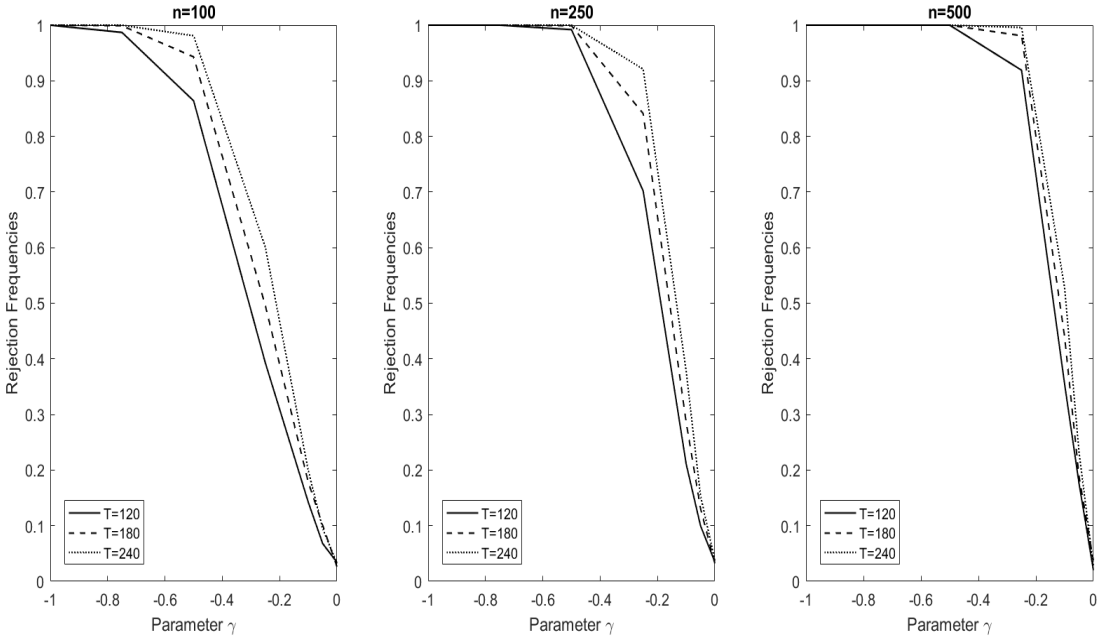


Figure 2: Rejection Frequencies under a high level of misspecification with the squared error loss function

Figure 2 displays the rejection frequencies over $1,000$ simulations. We observe that the proposed test is robust to this form of misspecification. Indeed, the rejection frequencies are similar to those displayed in Figure 1, suggesting that making a mistake in the choice of financial variables is not harmful. Results available from the authors upon request show that the robustness holds even when the misspecification is more pronounced as only a quarter of the financial variables of interest are chosen. The robustness to the choice of the loss function can be seen in Figure B.2 in Appendix B.

4 Empirical applications

This section illustrates our backtesting procedure using real datasets. We apply our methodology to two popular providers of ESG ratings, Sustainalytics and Asset4, over three universes from North America, Europe and the Asia-Pacific region. We first describe our

datasets and the related variables, and then conduct inferences to evaluate the informational content of each of the rating systems.

4.1 Description of the datasets and variables

The dataset for each of the three universes contains information for n firms at a monthly frequency over a period ranging from January 2010 to October 2018, giving a total of $T = 106$ months. The North America, Europe and Asia-Pacific datasets gather information on respectively $n = 326$, $n = 238$ and $n = 217$ firms. This deep panel structure ensures a lot of power for our backtesting methodology (see Monte Carlo simulations), with a total of $34,556$, $25,228$ and $23,002$ pooled observations for the North America, Europe and Asia-Pacific universes.

4.1.1 Information on ESG data

Table 1 displays pooled descriptive statistics of the ESG ratings for the two providers over the three universes. We may note that for both providers, higher values of the ESG ratings indicate lower ESG risks.

Table 1: Pooled descriptive statistics of the ESG ratings

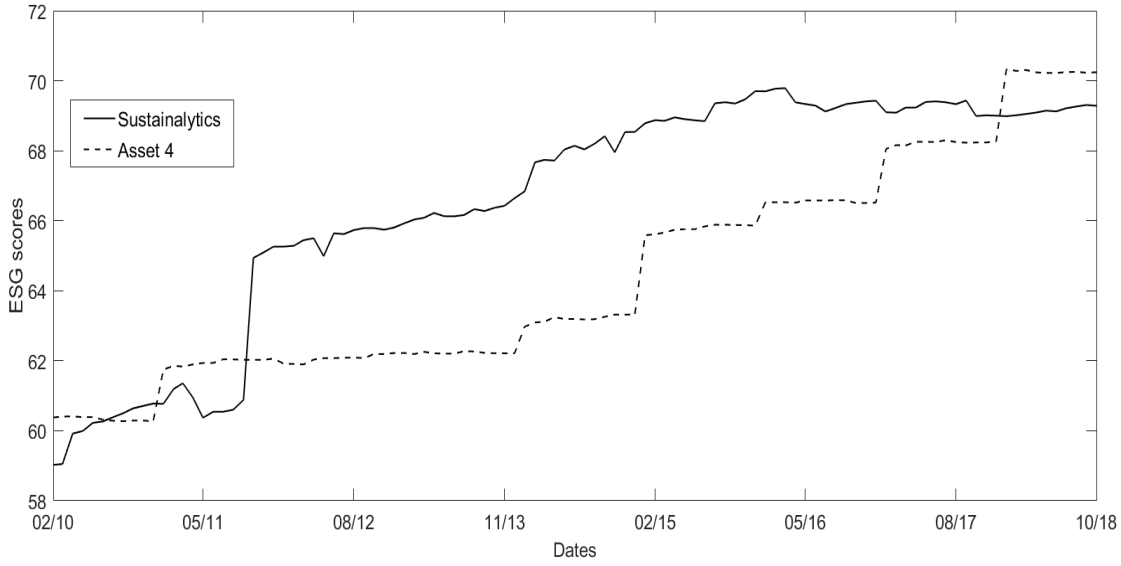
	Min.	Max.	Mean	Median	Std.
			Europe		
Sustainalytics	36.0000	89.6900	66.5310	67.3000	9.6449
Asset4	5.4700	94.1500	64.4389	66.1300	15.7645
			North America		
Sustainalytics	33.0000	88.0000	59.0831	59.0000	8.6864
Asset4	2.4700	94.7700	54.4304	56.5200	18.8691
			Asia-Pacific		
Sustainalytics	32.0000	90.0900	58.5848	59.0000	8.3848
Asset4	2.3500	90.2700	53.3590	56.1900	18.2707

Notes: The table displays pooled descriptive statistics of the ESG ratings for the two providers (Sustainalytics and Asset4) over the three universes. The datasets contain monthly observations over the period from January 2010 to October 2018, giving a total of 106 months. The North America, Europe and Asia-Pacific datasets contain information on respectively $n = 326$, $n = 238$ and $n = 217$ firms. Min. refers to minimum, Max. to maximum, and std. to standard deviations.

The average values of the ESG ratings for the Europe universe are 66.53 for Sustainalytics and 64.43 for Asset4. This means the central statistics are similar for both providers, as is confirmed by the values of the median of 67.30 for Sustainalytics and 66.13 for Asset4 for the Europe universe. This stylised fact holds for the other two universes. However, the

Asset4 ESG ratings have more variability across time and firms as given by the values of the standard deviations and ranges. The standard deviations of the Asset4 ESG ratings for instance are approximately twice as high as those for Sustainalytics.

Figure 3: Dynamics of the cross-sectional means of the ESG ratings: Europe



Source: The figure displays the evolution over time of the cross-sectional means of the ESG ratings for the two providers considered (Sustainalytics and Asset4). The dataset contains monthly observations for $n = 238$ firms from January 2010 to October 2018, giving a total of 106 months.

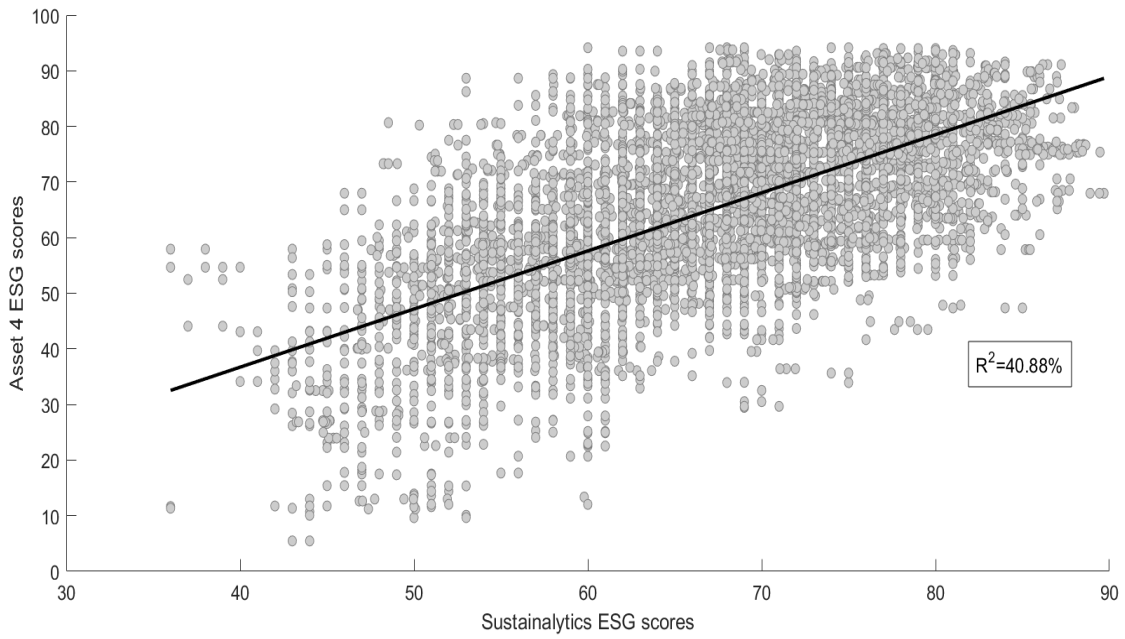
Figure 3 displays the evolution over time of the cross-sectional averages of the ESG ratings for the two providers in the Europe universe. We observe growth over time in the cross-sectional averages, which suggests a tendency towards upward revisions of the ESG ratings for firms. If we retain the hypothesis that the ESG ratings are well measured by the two providers,[5] this is proof of an overall trend of improvement over time in the corporate behaviour of the firms in Europe for best practices for environmental, social and governance issues. Figures B.3 and B.4 in Appendix B show the same dynamics in the other two universes of North America and Asia-Pacific.

To evaluate the link between the two rating systems, Figure 4 displays the scatter plot of the pooled ESG ratings from the two providers for the Europe universe. The figure also displays the fitted least square regression line, along with the adjusted R-squared, which is equal to 40.88%. Hence the link across firms and time between the two ESG ratings is weak, though it is positive. As already underlined, this has been highlighted many times in the literature and constitutes the main motivation of our paper, which proposes, in a context of

[5]Remember that this is the hypothesis that our inferential procedure is designed to check.

limited convergence, a formal testing procedure for evaluating the informational content of ESG rating systems. The phenomenon is not only European and is also highlighted for the other two universes as shown by Figures B.5 and B.6 in Appendix B. The trend is of the same order for the North America universe with an R-squared of 46.46%, but we observe a more pronounced divergence in the Asia-Pacific universe with an R-squared of only 32.65%.

Figure 4: Relation between the Sustainalytics and Asset4 ESG ratings: Europe



Source: The figure displays the scatter plot that shows the graphical relation between the ESG ratings for the two providers considered (Sustainalytics and Asset4). The dataset contains monthly observations for $n = 238$ firms from January 2010 to October 2018, giving a total of 106 months.

4.1.2 Information on idiosyncratic volatility

In this sub-section, we provide information on the target idiosyncratic volatility of asset returns. To build this variable for each firm i, we collect daily stock returns $r_{i,s}$ over our period of investigation from January 2010 to October 2018, with a total of $2,304$ observations. For each universe, we also collect the daily returns $r_{m,s}$ of the MSCI stock index over the same period, using MSCI Europe, MSCI USA and MSCI Pacific for the Europe, North America and Asia-Pacific universes. Residual returns are thus extracted assuming that the Capital Asset Pricing Model (CAPM) holds, with

$$r_{i,s} = \alpha_i + \beta_i r_{m,s} + \epsilon_{i,s}, \tag{18}$$

where α_i is the alpha of the stock, β_i is the beta or exposure of the stock to the market, and $\epsilon_{i,s}$ is the innovation or residual return for stock i at day s. With the daily residual returns,

we compute monthly idiosyncratic realised volatility as follows

$$IRV_{i,t} = \sum_{s_k=1}^{v_t} \widehat{\epsilon}^2_{i,s_k}, \tag{19}$$

with t the index of the month, v_t the number of daily observations in month t, and $\widehat{\epsilon}_{i,s_k}$ the s_k^{th} fitted residual returns within month t. For each firm i in a given universe, we obtain a time series of monthly idiosyncratic realised volatility of length 106, which thus matches the monthly frequency and the length of the ESG data analysed in the previous sub-section.

Table 2: Pooled descriptive statistics of idiosyncratic realised volatility

	Min (%)	Max (%)	Mean (%)	Median (%)	Std (%)
Europe	0.0133	24.6450	0.4634	0.2927	0.7064
North America	0.0099	50.6339	0.4970	0.2632	0.9217
Asia-Pacific	0.0236	33.3244	0.5416	0.3697	0.7303

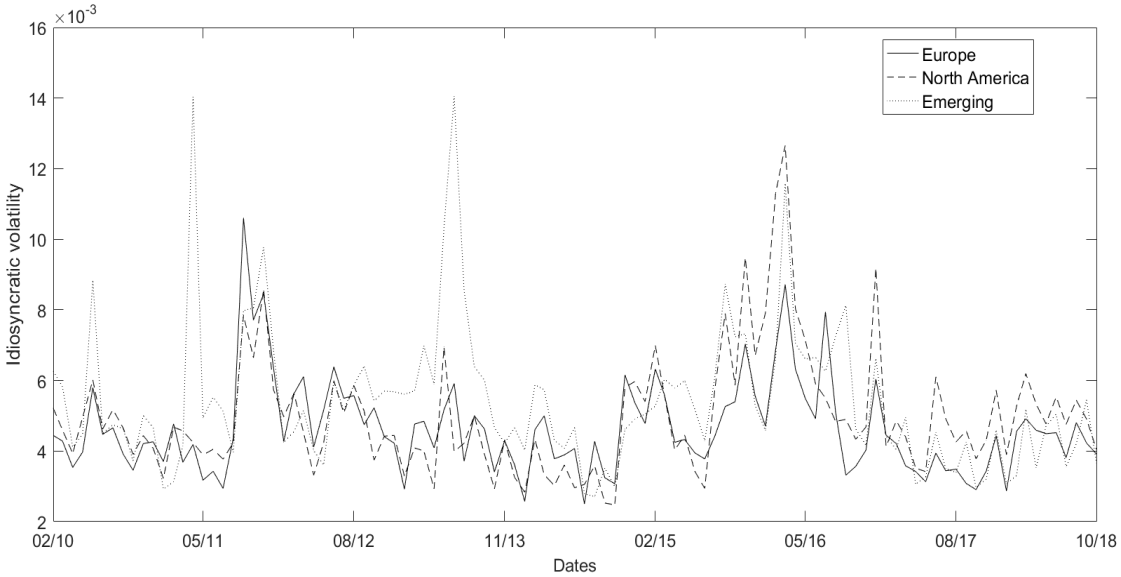
Notes: The table displays pooled descriptive statistics of monthly idiosyncratic realised volatilities for the three universes. Idiosyncratic realised volatilities are computed from residual asset returns from the CAPM. The datasets contain monthly observations from January 2010 to October 2018, giving a total of 106 months. The North America, Europe and Asia-Pacific datasets contain information on respectively $n = 326$, $n = 238$ and $n = 217$ firms. Min. refers to minimum, Max. to maximum, and std. to standard deviations.

Table 2 displays the pooled descriptive statistics of monthly idiosyncratic volatilities for the three universes. The Asia-Pacific universe appears as the one where firms have on average the highest levels of idiosyncratic volatility. In terms of dispersion, the North America universe has more variability in the measure of the volatility of residual returns, as given by the values for the standard deviation and the range.

To get an overhead view of the monthly series of idiosyncratic realised volatilities, Figure 5 displays the evolution over time of the cross-sectional means of monthly idiosyncratic realised volatilities. We observe the typical dynamics, with volatility clusters that nevertheless seem less pronounced because we are dealing with idiosyncratic volatility, and not total volatility which includes the systematic part.

It may be recalled that our testing procedure is designed to test the informational content of the ESG ratings by checking whether they have predictive power for the materialisation of ESG risks as measured by increased idiosyncratic realised volatility. Hence the relation that the test aims to validate is in the direction of high ESG ratings leading to low idiosyncratic volatilities and low ratings to high volatilities. So before we apply the testing procedure formally, Figures 6 and 7 try to illustrate whether there is such a relationship in the Europe universe. These figures report the distribution of the lagged values of the ESG ratings

Figure 5: Dynamics of the cross-sectional means of idiosyncratic realised volatility



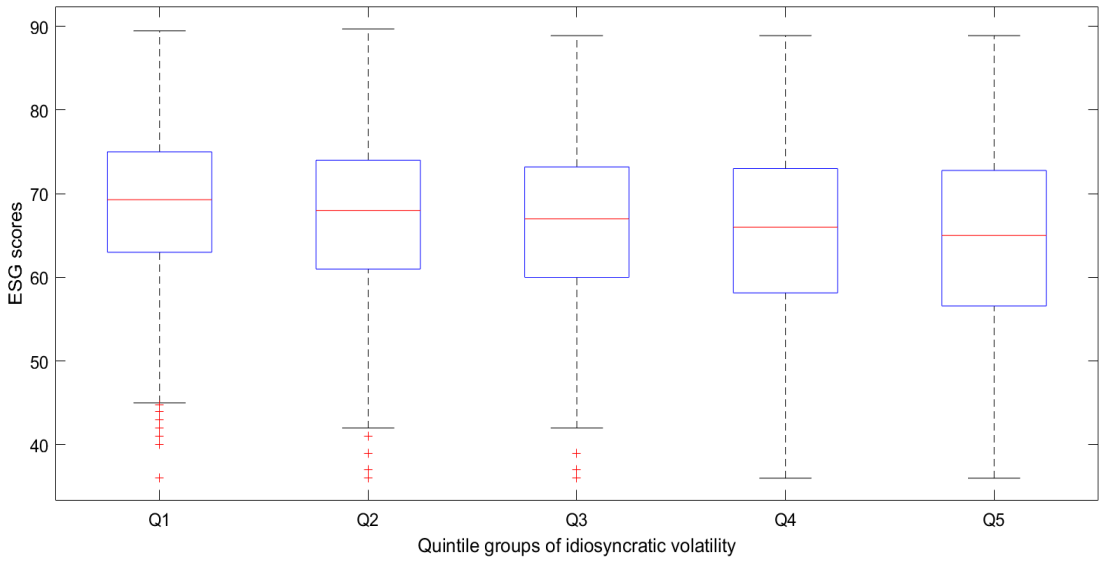
Source: The figure displays the evolution over time of the cross-sectional means of monthly idiosyncratic realised volatilities. Idiosyncratic realised volatilities are computed from residual stock returns from the CAPM. The datasets contain monthly observations from January 2010 to October 2018, giving a total of 106 months. The North America, Europe and Asia-Pacific datasets contain information on respectively $n = 326$, $n = 238$ and $n = 217$ firms.

(Figure 6 for Sustainalytics and Figure 7 for Asset4) by idiosyncratic volatility quintiles. Overall we observe that a negative relation arises, with high values of lagged ESG ratings associated with low idiosyncratic volatilities, while the median values of the lagged ESG ratings decrease with the order of the quintiles. Robustness across the universes is confirmed in Appendix B, with Figures B.7 and B.8 for the North America universe, and B.9 and B.10 for the Asia-Pacific universe.

However, and as already underlined, idiosyncratic volatility measuring the variability of residual returns can be driven by many factors, including innovations in financial balance-sheet variables. The effects of these financial factors can dampen the link between the lagged values of the ESG ratings and the idiosyncratic volatilities as reported in Figures 6 and 7 above, and Figures B.7-B.10 in Appendix B. What our inferential procedure tries to check is the existence of the predictive power of ESG ratings cleaning out those confounding effects.

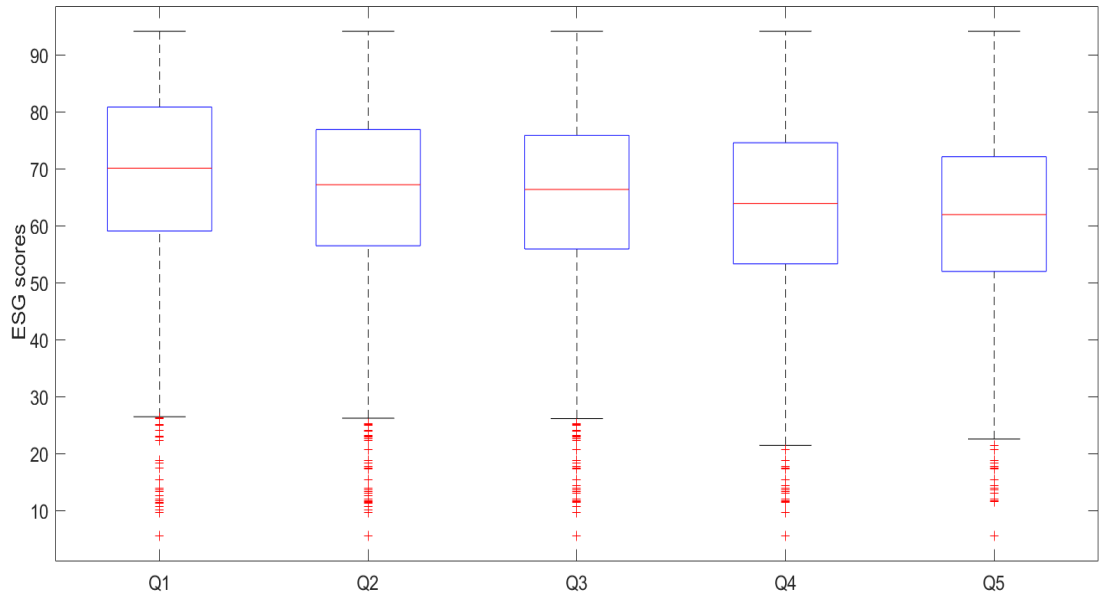
To this end, we retain the $p = 10$ balance sheet variables for which the monthly observations are available for all firms over the three universes and the timespan considered. These variables are tax burden, interest burden, operating margin, asset turnover, leverage, current ratio, net debt to earnings before interest, taxes, depreciation, and amortisation (EBITDA), capital expenditure (Capex) to depreciation, current assets, and current liabil-

Figure 6: ESG ratings by idiosyncratic volatility quintiles: Sustainalytics (Europe)



Source: For the Europe universe, the figure displays the means of the Sustainalytics ESG ratings within the five groups defined by the quintiles of idiosyncratic volatility computed with residual asset returns from the CAPM. The dataset contains monthly observations for $n = 238$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure 7: ESG ratings by idiosyncratic volatility quintiles: Asset4 (Europe)



Source: For the Europe universe, the figure displays the means of Asset4 ESG ratings within the five groups defined by the quintiles of idiosyncratic volatility computed with residual asset returns from the CAPM. The dataset contains monthly observations for $n = 238$ firms from January 2010 to October 2018, giving a total of 106 months.

ities (see Table 3 for a complete description of these variables). Innovations are extracted for each of these financial variables and for each firm by centering the raw values on the time average.

Table 3: Description of balance-sheet variables

Variables	Ratios	Description
Tax Burden	Net Income/Pretax Income	Profits retained after taxes
Interest Burden	Pretax Income/EBIT	Profits retained after interest paid
Operating Margin	EBIT/Revenue	Return on sales
Asset Turnover	Revenue/Total Assets	Revenue generated by own resources
Leverage	Total Assets/Total Equity	Measure of financial leverage
Current Ratio	Current Assets/Current Liab.	Measure of short-term resources
Net Debt to EBITDA	Net Debt/EBITDA	Capacity to finance debt
Capex to Dep.	Capex/Depreciation	Rate at which assets are renewed
Current Assets	Current Assets/Total Assets	Measure of short-term resources
Current Liab.	Current Liab./Total Liab.	Measure of short-term liabilities

Notes: The table gives the description of the balance-sheet variables retained. Innovations in these variables are used to control for the impact of financial factors when assessing the predictive contents of ESG ratings on the idiosyncratic volatility of a firm's assets.

4.2 Backtest results

Using the three categories of variables defined above as ESG ratings, idiosyncratic volatility and innovations in balance-sheet variables, we compute our test statistics in (6-8) and make inference for the predictive content of the two ESG rating systems considered. To predict the target idiosyncratic volatility variable, we consider a pooled OLS regression for the two models needed to run our testing procedure, which are the model that contains only innovations in the $p = 10$ financial variables, and the model that extends this set to include the lagged values of the ESG ratings.

In line with our out-of-sample testing environment, we consider two different forecasting schemes: (i) a fixed forecasting scheme where the first 75% of the total $T = 106$ months for each firm are used to estimate both models, and the forecasts are computed over the last 25% of observations, which are considered as the test sample; (ii) a rolling-window forecasting scheme with the forecasts computed by moving the estimation sample forward by including one more month and excluding the first, giving different estimation samples with the same fixed size of $b = [0.75T]$.

Table 4 displays the outcomes of the test for each provider of the ESG ratings over the three universes. The test statistics are computed using the squared error loss. For each forecasting scheme, the column "Model 0" displays the Mean Squared Error (MSE),

averaged across time and firms, for the model with only the innovations in the balance-sheet variables. The column "Sustainalytics" reports the same statistic when the information set is extended to include the Sustainalytics ESG ratings and "Asset4" includes instead the Asset4 ratings, followed by the values of the test statistics.

Table 4: Backtest of ESG ratings: results for squared error loss and idiosyncratic returns from CAPM

	Fixed window			Rolling window		
Europe						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MSE ($\times 10^{-5}$)	4.2572	4.1824	4.1591	4.2328	4.1668	4.1454
Test statistic		-5.8860^{***}	-5.2311^{***}		-7.6024^{***}	-8.7052^{***}
North America						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MSE ($\times 10^{-5}$)	8.4727	8.4685	8.4266	8.1224	8.1172	8.1108
Test statistic		-0.5990	-3.0246^{***}		-1.7375^{**}	-1.8911^{**}
Asia-Pacific						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MSE ($\times 10^{-5}$)	2.6769	2.6762	2.6729	2.6857	2.6881	2.6917
Test statistic		-0.0839	-1.6804^{**}		0.7032	3.3857

Notes: The table displays the predictive accuracy as measured by the mean squared error (MSE) and the test statistic associated with the null hypothesis of a lack of informational content in the ESG ratings for predicting the idiosyncratic volatility of asset returns. Idiosyncratic volatilities are computed from the residual asset returns from the CAPM. "Model 0" indicates that only innovations in the balance-sheet variables are used for the prediction. The columns "Sustainalytics" and "Asset4" are where the information set includes innovations in the balance-sheet variables together with the ESG ratings of Sustainalytics or Asset4. The datasets contain monthly observations from January 2010 to October 2018, giving a total of 106 months. The North America, Europe and Asia-Pacific datasets include information on respectively $n = 326$, $n = 238$ and $n = 217$ firms. *, ** and *** indicate statistical significance at the 10%, 5% and 1% nominal risk levels respectively.

Introducing the ESG ratings to the information set for the Europe universe increases the predictive power as measured by the MSE. Indeed the MSEs decrease when moving from "Model 0" to the other two models, and the decreases are statistically significant at the 1% significance level. This result holds for both forecasting schemes. The same conclusion is reached for the North America universe, except for the Sustainalytics ESG ratings in the fixed forecasting scheme. In this case, the test statistic is equal to -0.5990 and higher than the critical values at the usual nominal significance levels. Finally, for the Asia-Pacific universe, only the Asset4 ESG ratings in the fixed forecasting scheme lead to the rejection of the null hypothesis of no informational content at the 5% nominal risk level.

To evaluate the sensitivity of the test to the choice of loss function, Table 5 displays

the results of the test using the absolute error loss function. The presentation is similar to that in Table 4. Remember that this loss function is more robust to outliers than the squared error loss function. The results obtained go more in the direction of rejecting the null hypothesis. Indeed apart from the result for Asset4 in the Asia-Pacific universe using a rolling window forecasting scheme, all the other results conclude with the rejection of the null hypothesis at the 1% significance level.

Table 5: Backtest of ESG ratings: results for absolute error loss and idiosyncratic returns from CAPM

	Fixed window			Rolling window		
Europe						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MAE ($\times 10^{-3}$)	3.3834	3.2674	3.2277	3.2562	3.1683	3.1232
Test statistic		-9.7559^{***}	-10.3447^{***}		-10.7613^{***}	-14.7787^{***}
North America						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MAE ($\times 10^{-3}$)	4.2856	4.2696	4.1762	3.8955	3.8909	3.8560
Test statistic		-3.4380^{***}	-9.4377^{***}		-2.6272^{***}	-10.8805^{***}
Asia-Pacific						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MAE ($\times 10^{-3}$)	3.4339	3.4006	3.4201	3.3736	3.3631	3.3879
Test statistic		-3.6987^{***}	-5.0650^{***}		-4.2258^{***}	8.8552

Notes: The table displays the predictive accuracy as measured by the mean absolute error (MAE) and the test statistic associated with the null hypothesis of a lack of informational content in the ESG ratings for predicting the idiosyncratic volatility of asset returns. Idiosyncratic volatilities are computed from the residual asset returns from the CAPM. "Model 0" indicates that only innovations in the balance-sheet variables are used for the prediction. The columns "Sustainalytics" and "Asset4" are where the information set includes innovations in the balance-sheet variables together with the ESG ratings of Sustainalytics or Asset4. The datasets contain monthly observations from January 2010 to October 2018, giving a total of 106 months. The North America, Europe and Asia-Pacific datasets include information on respectively $n = 326$, $n = 238$ and $n = 217$ firms. *, ** and *** indicate statistical significance at the 10%, 5% and 1% nominal risk levels respectively.

Taken together, the results in Tables 4 and 5 are strong evidence that both rating systems are informative about the ESG risks associated with increased idiosyncratic volatility. This statement seems very robust for the Europe and the North America universes, and to some extent less robust for the Asia-Pacific universe. In the next sub-section we conduct additional empirical investigations to check for the robustness of our results.

4.3 Robustness to factor models

Here we evaluate the sensitivity of our results to the choice of factor model used to compute the target idiosyncratic realised volatility variable. We thus extend the CAPM model and consider a multifactorial model. This extension is anchored to the findings of academic research into the existence of common risk factors beyond the market index. This strand of the literature, which can be dated back to the seminal work of Fama and French (1992), has discovered many market variables or factors that may be able to explain the cross-sectional variations of stock returns. These include the size and value factors in Fama and French (1992) and the momentum factor in Jegadeesh and Titman (1993).

Table 6: Backtest of ESG ratings: results for absolute error loss and idiosyncratic returns from a multifactorial model

	Fixed window			Rolling window		
Europe						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MAE ($\times 10^{-3}$)	3.2231	3.1064	3.0699	3.0980	3.0087	2.9705
Test statistic		-9.9708^{***}	-10.3733^{***}		-11.1452^{***}	-14.7828^{***}
North America						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MAE ($\times 10^{-3}$)	4.0757	4.0576	3.9696	3.7148	3.7083	3.6727
Test statistic		-3.4373^{***}	-9.0788^{***}		-3.6558^{***}	-11.1500^{***}
Asia-Pacific						
	Model 0	Sustainalytics	Asset4	Model 0	Sustainalytics	Asset4
MAE ($\times 10^{-3}$)	3.2745	3.2419	3.2694	3.2221	3.2136	3.2421
Test statistic		-3.8828^{***}	-5.0681^{***}		-3.8904^{***}	9.2743

Notes: The table displays the predictive accuracy as measured by the mean absolute error (MAE) and the test statistic associated with the null hypothesis of a lack of informational content in the ESG ratings for predicting the idiosyncratic volatility of asset returns. Idiosyncratic volatilities are computed from the residual asset returns from a multifactorial model. "Model 0" indicates that only innovations in the balance-sheet variables are used for prediction. The columns "Sustainalytics" and "Asset4" are where the information set includes innovations in the balance-sheet variables together with the ESG ratings of Sustainalytics or Asset4. The datasets contain monthly observations from January 2010 to October 2018, giving a total of 106 months. The North America, Europe and Asia-Pacific datasets include information on respectively $n = 326$, $n = 238$ and $n = 217$ firms. *, ** and *** indicate statistical significance at the 10%, 5% and 1% nominal risk levels respectively.

To consider the multifactorial model, we extend the CAPM model in (18) by adding investable factors identified in the literature to drive the cross-sectional variations of the stock's returns. For the Europe and the North America universes these are the MSCI Small/Large Capitalisation factor, which approximates the size anomaly, the MSCI Value/Growth factor

associated with the value premium, the MSCI Momentum factor, the MSCI quality factor, and the MSCI Minimum Volatility factor. The lack of data for the Asia-Pacific universe means we consider three factors beyond the market, these being the MSCI Small/Large Capitalisation factor, the MSCI Value/Growth factor, and the MSCI Minimum Volatility factor.

Table 6, which displays the backtest results using the absolute loss error, shows similar results. Both rating systems are informative about the occurrence of ESG risks associated with increased idiosyncratic realised volatility, with this trend holding for the Europe and the North America universes in all configurations, and in some configurations for the Asia-Pacific universe.

4.4 Disagreement between raters and the informational content of the ESG ratings

What our results suggest is that both rating systems are robustly informative about ESG risks in the Europe and the North America universes where there is the highest levels of consensus between the ESG ratings. Indeed it may be recalled that the R-squared for the linear regression between the two ratings is equal to 40.88% for the Europe universe, and 46.46% for the North America universe (see Figures 4 and B.5). By contrast, this statistic is 32.65%, meaning the consensus is lower, for the Asia-Pacific universe (see Figure B.6), where the existence of informational content in the ESG ratings seems less robust in all the testing configurations considered above, particularly when the squared error loss function is used (see Table 4).

This result can be linked to that highlighted by Serafeim and Yoon (2021). They use a different approach to analysing the link between ESG ratings and ESG risks as measured by ESG-related events and show that the consensus in the rating predicts future news, but its predictive ability diminishes for firms about which there is large disagreement between the raters. Moreover, the relation between news and market reaction is moderated by the consensus in the ratings. When there is a high level of disagreement between the raters, the relation between the news and the market reactions weakens, while the rating with the most predictive power predicts future stock returns.

To check for this stylised fact, we want to replicate the backtesting results in Table 4 by restricting each universe to only firms with a high level of consensus between the two raters. The level of consensus for each firm is simply approximated by the rank correlation between the time series of the ESG ratings for the two raters. We retain only firms with positive and statistically significant correlations. To deal with the issue of multiple testing, the nominal risk level to gauge the statistical significance of each correlation is set to $0.01/n$ (Bonferroni

correction), with n the number of firms. This filtering process leads to a total of 94, 122 and 82 firms for the Europe, North America and Asia-Pacific universes, respectively.

Table 7: Backtest of convergent ESG ratings: results for squared error loss and idiosyncratic returns from CAPM

	Fixed window			Rolling window		
Europe universe						
	Model 0	Sustainalytics	Asset 4	Model 0	Sustainalytics	Asset 4
MSE (10^{-5})	5.5023	5.3689	5.3949	5.4886	5.3817	5.3701
Test statistic		-5.1546^{***}	-5.3136^{***}		-4.5809^{***}	-7.2548^{***}
North America						
	Model 0	Sustainalytics	Asset 4	Model 0	Sustainalytics	Asset 4
MSE (10^{-5})	5.0211	4.9813	4.9278	4.4019	4.3709	4.3869
Test statistic		-2.3400^{***}	-4.0960^{***}		-2.8232^{***}	-1.9315^{**}
Asia-Pacific						
	Model 0	Sustainalytics	Asset 4	Model 0	Sustainalytics	Asset 4
MSE (10^{-5})	3.4825	3.4502	3.4655	3.3936	3.3550	3.3826
Test statistic		-1.8987^{**}	-1.9499^{**}		-3.0666^{***}	-4.0811^{***}

Notes: The table displays the predictive accuracy as measured by the mean squared error (MSE) and the test statistic associated with the null hypothesis of a lack of informational content in the ESG ratings for predicting the idiosyncratic volatility of asset returns. Idiosyncratic volatilities are computed from the residual asset returns from the CAPM. "Model 0" indicates that only innovations in the balance-sheet variables are used for prediction. The columns "Sustainalytics" and "Asset4" are where the information set includes innovations in the balance-sheet variables together with the ESG ratings of Sustainalytics or Asset4. The datasets contain monthly observations from January 2010 to October 2018, giving a total of 106 months, and are restricted to only firms for which there is consensus (convergence) about the ESG ratings. The North America, Europe and Asia-Pacific datasets include information on respectively $n = 94$, $n = 122$ and $n = 82$ of those firms. *, ** and *** indicate statistical significance at the 10%, 5% and 1% nominal risk levels respectively.

Table 7 displays the same information as in Table 4, restricting the sample of firms to only those with consensus as to their ESG ratings. Unlike the results in Table 4, we note that the null hypothesis of a lack of informational content in the ESG ratings is rejected in all configurations. Backtest results for the opposite samples, that is, the samples including firms with divergent ESG ratings are displayed in Table 8. These samples are about firms with negative and statistically significant correlations between ESG ratings for the two providers, and include 23, 50 and 26 firms for Europe, North America and Asia-Pacific universes, respectively. Contrary to the results of Table 7, we observe here that the null hypothesis of lack of informational content of ESG ratings is not rejected in many configurations, in particular for the Asia-Pacific universe.

One can argue that the lack of rejection of the null hypothesis in Table 8 can result

Table 8: Backtest of divergent ESG ratings: results for squared error loss and idiosyncratic returns from CAPM

	Fixed window			Rolling window		
Europe universe						
	Model 0	Sustainalytics	Asset 4	Model 0	Sustainalytics	Asset 4
MSE (10^{-5})	9.4666	9.0222	9.1050	7.5211	7.3669	7.2789
Test statistic		-1.5287^{*}	-1.2299		-2.1750^{**}	-3.3521^{***}
North America						
	Model 0	Sustainalytics	Asset 4	Model 0	Sustainalytics	Asset 4
MSE (10^{-5})	4.1935	3.9806	4.0355	3.7317	3.7230	3.6781
Test statistic		-1.3071^{*}	-1.1843		-0.1864	-1.7128^{**}
Asia-Pacific						
	Model 0	Sustainalytics	Asset 4	Model 0	Sustainalytics	Asset 4
MSE (10^{-5})	1.6024	1.5229	1.6215	1.5069	1.4231	1.5247
Test statistic		-1.0312	0.8812		-2.0205^{**}	2.1627

Notes: The table displays the predictive accuracy as measured by the mean squared error (MSE) and the test statistic associated with the null hypothesis of a lack of informational content in the ESG ratings for predicting the idiosyncratic volatility of asset returns. Idiosyncratic volatilities are computed from the residual asset returns from the CAPM. "Model 0" indicates that only innovations in the balance-sheet variables are used for prediction. The columns "Sustainalytics" and "Asset4" are where the information set includes innovations in the balance-sheet variables together with the ESG ratings of Sustainalytics or Asset4. The datasets contain monthly observations from January 2010 to October 2018, giving a total of 106 months, and are restricted to only firms for which there is no consensus (divergence) about the ESG ratings. The North America, Europe and Asia-Pacific datasets include information on respectively $n = 23$, $n = 50$ and $n = 26$ of those firms. *, ** and *** indicate statistical significance at the 10%, 5% and 1% nominal risk levels respectively.

from the small number of firms (n small) which induces a weak power of the test to detect informational content in ESG ratings over small samples. This argument is however not robust, because for the Asia-Pacific universe with $n = 26$ firms, the null hypothesis is rejected in only one out of four cases, whereas for the Europe universe with a similar size ($n = 23$), it is rejected in three out of four cases.

All these results are very interesting because they indeed confirm that of Serafeim and Yoon (2021), who found that consensus about the ESG rating predicts ESG risks, but the predictive ability diminishes for firms about which there is large disagreement among raters.

From a practical point of view, this result provides portfolio managers involved in the sustainable management of funds with the crucial information that it is necessary to cross-check information gathered from several ESG rating providers before integrating those ratings into the management process. The focal point from our results is that consensus about the ESG ratings is informative about ESG risks, while ESG ratings with disagreement are

less valuable from this viewpoint.

5 Conclusion

The contribution of this article is to propose a formal statistical procedure for checking for informational content about ESG risks in the ESG ratings. The test proceeds by evaluating how well these extra-financial metrics help in predicting whether ESG risks will materialise, as given by increased realised idiosyncratic volatility, and this beyond the information conveyed by innovations in traditional balance-sheet variables.

Technically, our inferential procedure for checking for informational content about ESG risks in the ESG ratings is based on extending the conditional predictive ability test of Giacomini and White (2006) to a panel setting. Under weak assumptions, including cross-sectional dependencies among the idiosyncratic volatilities of firm's stocks, we derive the Gaussian asymptotic distribution of the test statistic. Monte Carlo simulations conducted under different types of model misspecification show that the test has good small sample properties.

Empirical applications are conducted using two leading ESG rating systems, Sustainalytics and Asset4, for Europe, North America and the Asia-Pacific region. The results show that the null hypothesis of a lack of informational content about ESG risks in the ESG ratings is rejected for Europe and North America, where there is less disagreement between the two ratings, while the results are mixed for the Asia-Pacific region, where there is more disagreement. Furthermore, applying the test only to firms about which there is a high degree of consensus between the ESG ratings from the two providers leads to the null hypothesis being rejected for all three regions.

The results have important implications for investors and researchers. For investors, our backtest procedure provides a useful and practical framework for considering ESG rating providers before integrating the ratings into the investment process. Our results suggest prudence about the information content of ESG ratings when they diverge. For researchers, it is crucial when studying ESG risks and pricing to check properly the quality of ESG ratings before using them, especially when the ratings are divergent. Moreover, the link between ESG ratings and idiosyncratic volatility when the ratings are convergent suggests that ESG investing is not just an issue of the preferences of investors but that ESG ratings can also provide information about future fundamentals and risks.

A future application for investors could be to compare the ratings of competing ESG rating agencies, since our inferential procedure can be easily adapted to compare the informational content about ESG risks in the ESG ratings. This would help investors in

selecting one agency among several competing ones in non-nested comparisons, or in considering additional competing agencies to combine with their already existing ratings in nested comparisons.

References

O. Akgun, A. Pirotte, G. Urga, and Z. Yang. Equal predictive ability tests for panel data with an application to oecd and imf forecasts. Unpublished Manuscript, 2019.

R. Albuquerque, Y. Koskinen, and Z. Chendi. Corporate social responsibility and firm risk: Theory and empirical evidence. *Management Science*, 65:4451–4469, 2019.

A. Amel-Zadeh and G. Serafeim. Why and how investors use esg information: Evidence from a global survey. *Financial Analyst Journal*, 74(3):87–103, 2018.

D. W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858, 1991.

F. Berg, J. Koelbel, and R. Rigobon. Aggregate confusion: the divergence of esg ratings. *Massachusetts Institute of Technology*, 2020.

M. Billio, M. Costola, I. Hristova, C. Latino, and L. Pelizzon. Inside the esg ratings: (dis)agreement and performance. *Unpublished Manuscript*, 2019.

K. Bouslah, L. Kryzanowski, and B. M'Zali. The impact of the dimensions of social performance on firm risk. *Journal of Banking and Finance*, 37:1258–1273, 2013.

C. Champagne, F. Coggins, and A. Sodjahin. The performance of extra-financial ratings as measure of esg-risk. Unpublished Manuscript, 2019.

A. Chatterji, D. Levine, and M. Toffel. How well do social ratings actually measure corporate social responsibility? *Journal of Economics and Management Strategy*, 18(1):125–169, 2009.

A. Chatterji, K. Durand, D. Levine, and S. Touboul. Do ratings of firms converge? implications for managers, investors and strategy researchers. *Strategic Management Journal*, 37(8):1597–1614, 2016.

A. Davies and K. Lahiri. A new framework for analyzing survey forecasts using three-dimensional panel data. *Journal of Econometrics*, 68:205–228, 1995.

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

A. Dyck, V. Lins, L. Roth, and H. Wagner. Do institutional investors drive corporate social responsibility? *Journal of Financial Economics*, 131(3):693–714, 2019.

E. F. Fama and K. R. French. The cross-section of expected stock returns. *Journal of Finance*, 47(2):427–465, 1992.

E. F. Fama and K. R. French. Disagreement, tastes, and asset prices. *Journal of Financial Economics*, 83(3):667–689, 2007.

R. Giacomini and H. White. Tests of conditional predictive ability. *Econometrica*, 74(6): 1545–1578, 2006.

S. Hartzmark and A. Sussman. Do investors value sustainability? a natural experiment examining ranking and fund ows. *Journal of Finance*, 74(6):2789–2837, 2019.

A. Hoepner, I. Oikonomou, Z. Sautner, L. Starks, and X. Zhou. Esg shareholder engagement and downside risk. *Unpublished Manuscript*, 2018.

E. Ilhan, Z. Sautner, and G. Vilkov. Carbon tail risk. *The Review of Financial Studies*, 34 (3):1540–1571, 2019.

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

H. Jo and H. Na. Positive and negative corporate social responsibility, financial leverage, and idiosyncratic risk. *Journal of Business Ethics*, 1174(2):431–448, 2012.

K. Mozaffar, G. Serafeim, and A. Yoon. Corporate sustainability: First evidence on materiality. *The Accounting Review*, 91(6):1697–1724, 2016.

L. Pastor, R. Stambaugh, and L. Taylor. Sustainable investing in equilibrium. *Journal of Financial Economics*, 2020.

L. Pedersen, S. Fitzgibbons, and L. Pomorski. Responsible investing: The ESG-efficient frontier. *Journal of Financial Economics*, 2020.

B. Rendlen and B. Weber. Guidance note: Integrating esg factors into financial models for infrastructure investments. WWF Working Paper, 2019.

A. Riedl and P. Smeets. Why do investors hold socially responsible mutual funds. *Journal of Finance*, 72(6):2505–2550, 2017.

N. Semenova and L. Hassel. On the validity of environmental performance metrics. *Journal of Business Ethics*, 132(2):249–258, 2015.

G. Serafeim and A. Yoon. Stock price reactions to esg news: The role of esg ratings and disagreement. *Working paper 21-079, Harvard Business School*, 2021.

A. Sodjahnin, C. Champagne, F. Coggins, and R. Gillet. Leading or lagging indicators of risk? the informational content of extra-financial performance scores. *Journal of Asset Management*, 18(5):347–370, 2017.

A. Timmermann and Y. Zhu. Comparing forecasting performance with panel data. Unpublished Manuscript, 2019.

K. D. West. Asymptotic inference about predictive ability. *Econometrica*, 64:1067–1084, 1996.

H. White. *Asymptotic Theory for Econometricians*. Academic Press, 2001. Revised Edition.

A Appendix A: Details on the Monte Carlo simulations

In this Appendix we provide details about the simulations of innovations in the balance sheet variables for generating the small sample properties of the test (see Section 3). These variables are generated via a multivariate Gaussian distribution with mean vector $\bar{x}$ and covariance matrix Ω calibrated using real data. The dataset we use contains historical monthly values of $p = 10$ innovations in the balance-sheet variables for 238 European firms from January 2010 to October 2018, giving a total of 106 months.

Innovations are computed as deviations from the overall means. The balance-sheet variables are, in order: tax burden ratio, interest burden ratio, operating margin ratio, asset turnover ratio, leverage as measured by the ratio of total assets to total equity, current ratio as measured by the ratio of current assets to current liabilities, debt ratio, capex as measured by the ratio of capital expenditures to depreciation, current assets as measured by the ratio of current assets to total assets, current liabilities as measured by the ratio of current liabilities to total liabilities.

The mean vector is thus equal to

$$\bar{x} = [0.8137; 0.8333; 0.1391; 0.8265; 3.8713; 1.4031; 1.7466; 1.2779; 0.3634; 0.2880],$$

and the covariance matrix Ω equal to

$$\Omega = \begin{pmatrix} 4.098 & -0.061 & -0.007 & -0.003 & -0.003 & 0.008 & 0.209 & -0.023 & -0.001 & -0.001 \\ -0.061 & 17.732 & -0.008 & 0.037 & 5.704 & 0.057 & -0.136 & -0.021 & 0.017 & 0.007 \\ -0.007 & -0.008 & 0.012 & -0.025 & -0.369 & 0.010 & -0.018 & 0.025 & -0.005 & -0.006 \\ -0.003 & 0.037 & -0.025 & 0.284 & -0.559 & -0.025 & -0.358 & -0.067 & 0.042 & 0.037 \\ -0.003 & 5.704 & -0.369 & -0.559 & 5012.291 & -0.521 & 30.792 & 4.391 & -0.160 & -0.092 \\ 0.008 & 0.057 & 0.010 & -0.025 & -0.521 & 0.642 & -0.420 & 0.042 & 0.053 & -0.040 \\ 0.209 & -0.136 & -0.018 & -0.358 & 30.792 & -0.420 & 32.172 & 0.550 & -0.154 & -0.058 \\ -0.023 & -0.021 & 0.025 & -0.067 & 4.391 & 0.042 & 0.550 & 1.841 & -0.023 & -0.022 \\ -0.001 & 0.017 & -0.005 & 0.042 & -0.160 & 0.053 & -0.154 & -0.023 & 0.029 & 0.014 \\ -0.001 & 0.007 & -0.006 & 0.037 & -0.092 & -0.040 & -0.058 & -0.022 & 0.014 & 0.018 \end{pmatrix}.$$

For the simulation of the target variable of idiosyncratic volatility, we run a pooled OLS regression with the dependent variable being the logarithm of the monthly time series of idiosyncratic realised volatility over the same period (January 2010 to October 2018) for the 238 European firms. The explanatory variables are the innovations in the 10 balance-sheet variables as described above.

$c^\star$	$\beta_1^\star$	$\beta_2^\star$	$\beta_3^\star$	$\beta_4^\star$	$\beta_5^\star$	$\beta_6^\star$	$\beta_7^\star$	$\beta_8^\star$	$\beta_9^\star$	$\beta_{10}^\star$
-5.9165	0.0070	-0.0015	-0.8739	-0.0679	0.0048×10^{-2}	0.0941	0.0044	0.0605	0.1869	0.0824

For the $p = 10$ balance-sheet variables, the estimated coefficients are displayed above. These estimates are used to generate data for simulating the logarithm of idiosyncratic realised volatility, and applying the exponential function leads to the target variable.

B Appendix B: Additional Tables and Figures

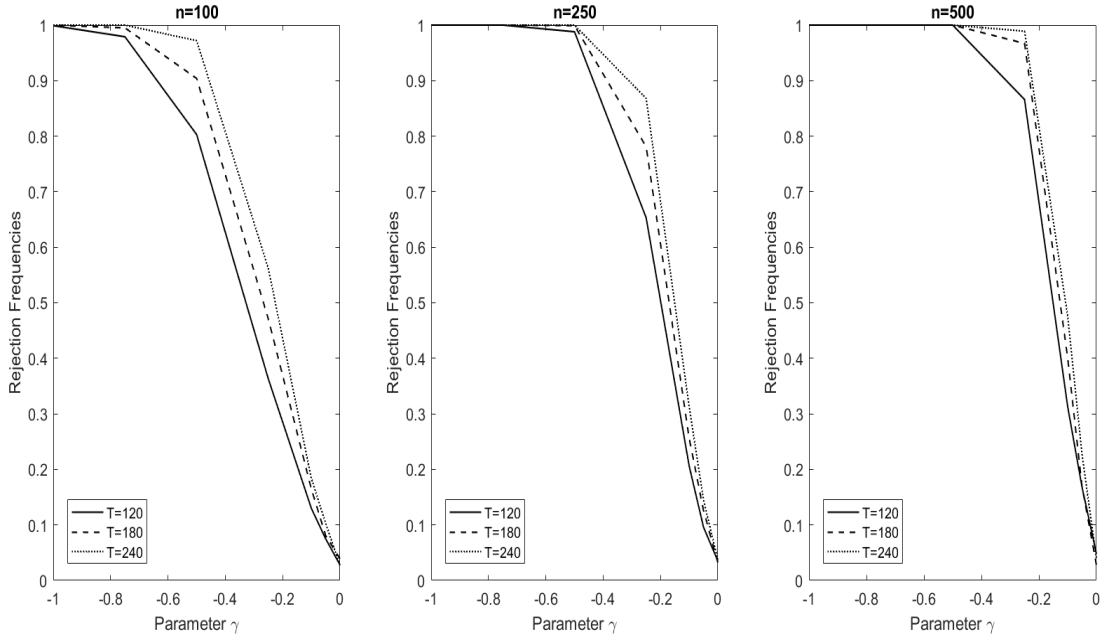


Figure B.1: Rejection Frequencies under a medium level of misspecification with the absolute error loss function

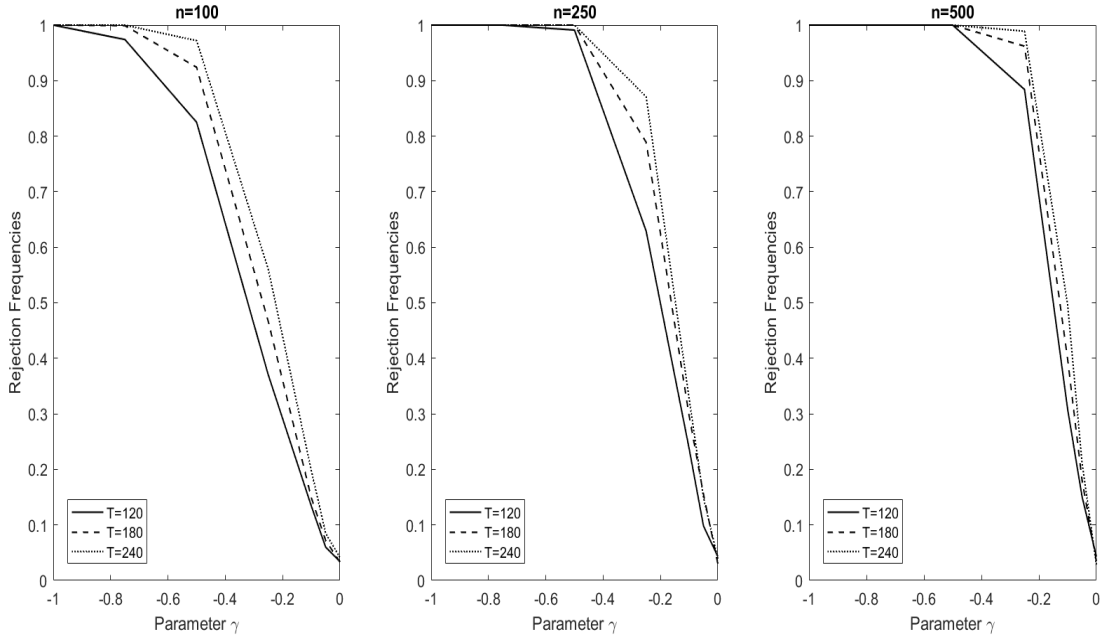
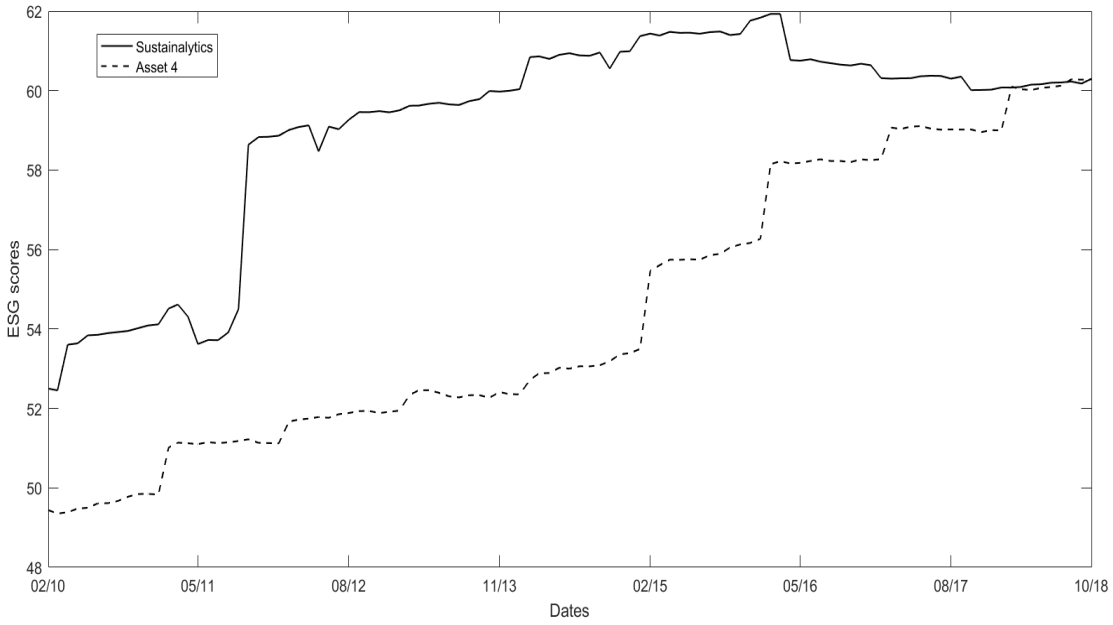


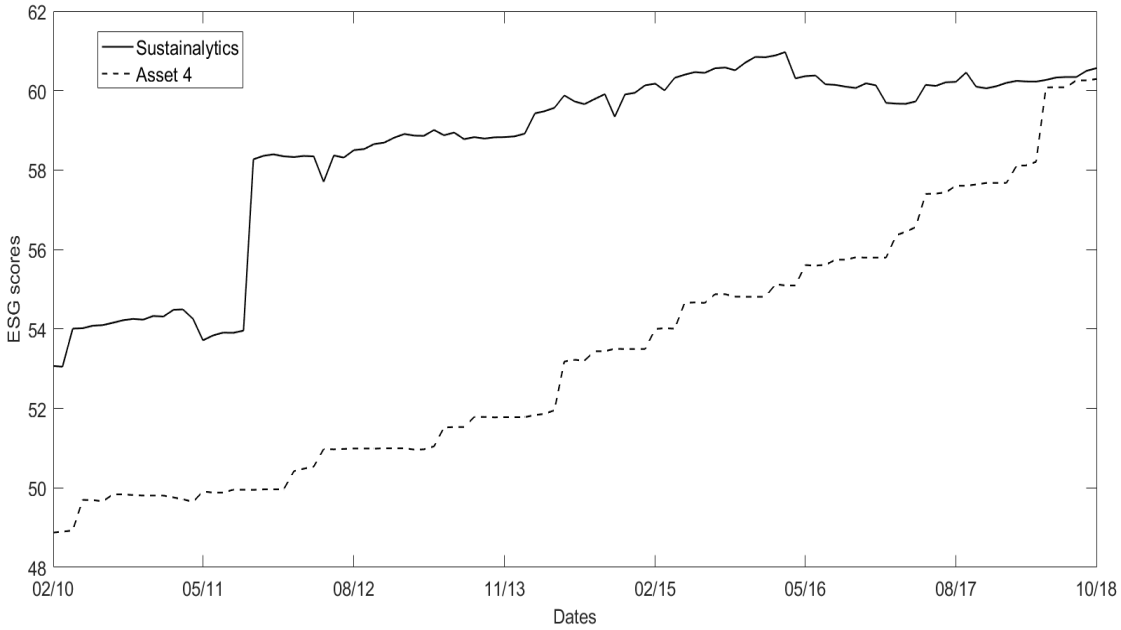
Figure B.2: Rejection Frequencies under a high level of misspecification with the absolute error loss function

Figure B.3: Dynamics of the cross-sectional means of the ESG ratings: North America



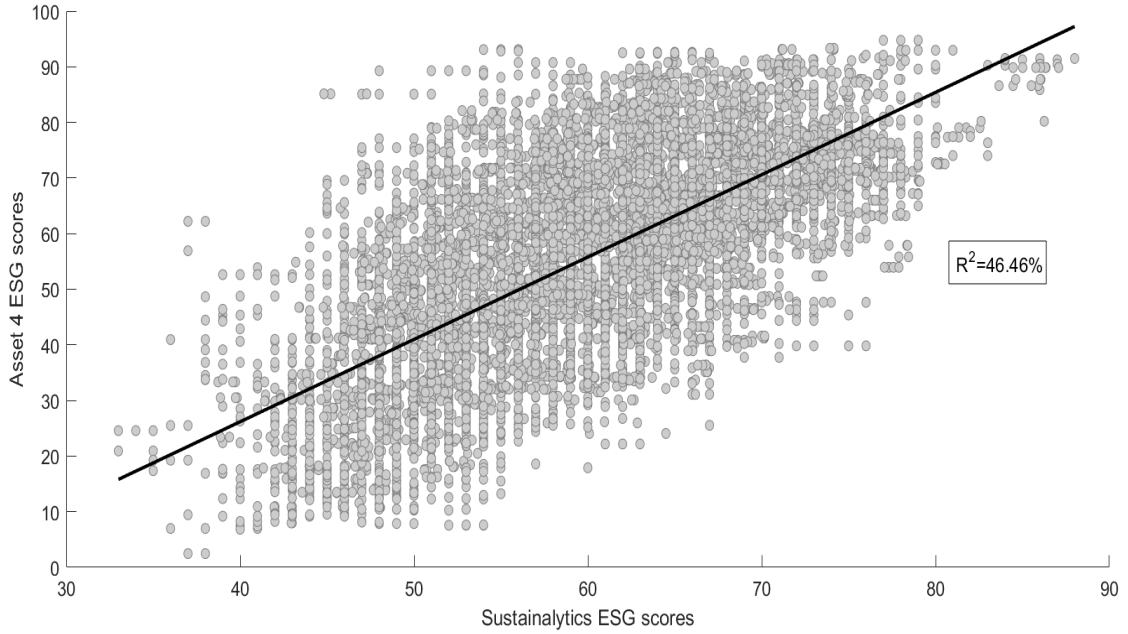
Source: The figure displays the evolution over time of the cross-sectional means of the ESG ratings for the two providers considered (Sustainalytics and Asset4). The datasets contain monthly observations for $n = 326$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure B.4: Dynamics of the cross-sectional means of the ESG ratings: Asia-Pacific



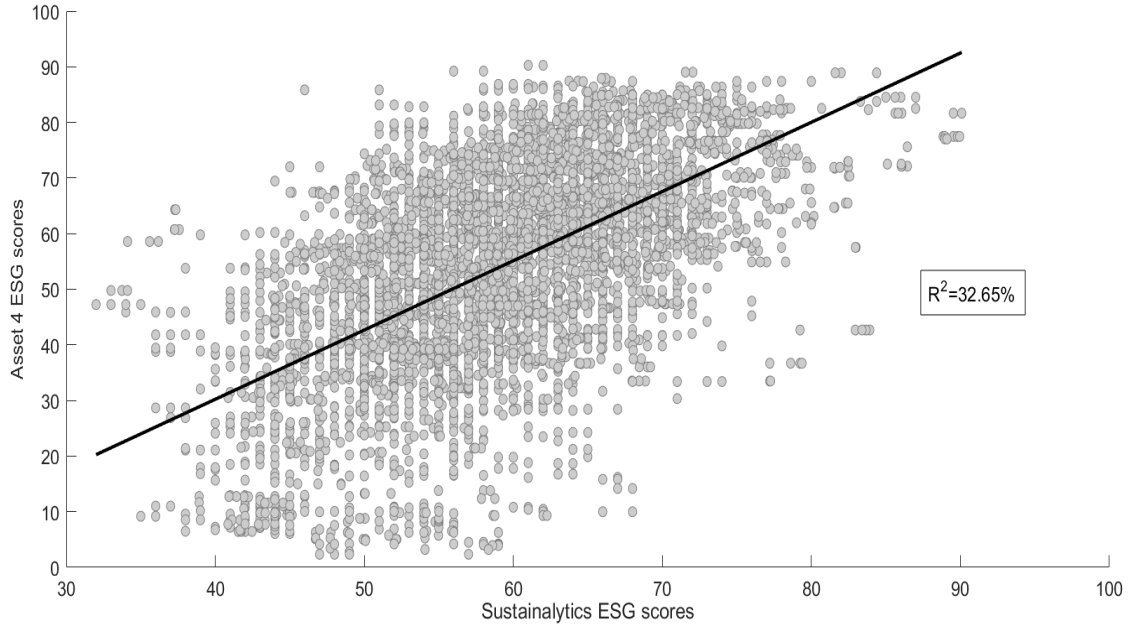
Source: The figure displays the evolution over time of the cross-sectional means of the ESG ratings for the two providers considered (Sustainalytics and Asset4). The datasets contain monthly observations for $n = 217$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure B.5: Relation between the Sustainalytics and Asset4 ESG ratings: North America



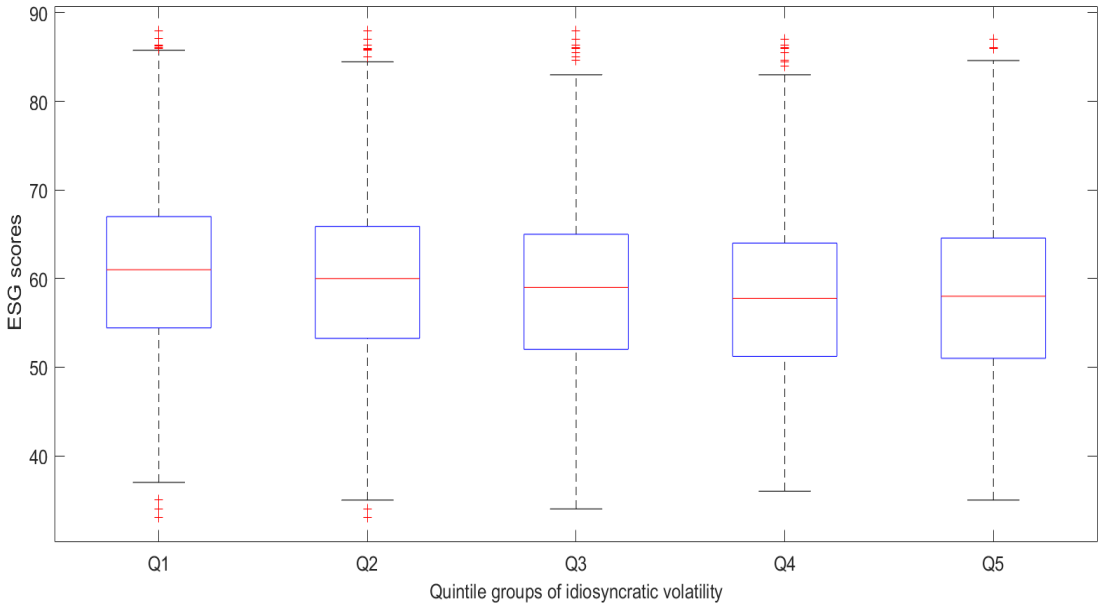
Source: The figure displays the scatter plot that shows the graphical relation between the ESG ratings for the two providers considered (Sustainalytics and Asset4). The datasets contain monthly observations for $n = 326$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure B.6: Relation between the Sustainalytics and Asset4 ESG ratings: Asia-Pacific



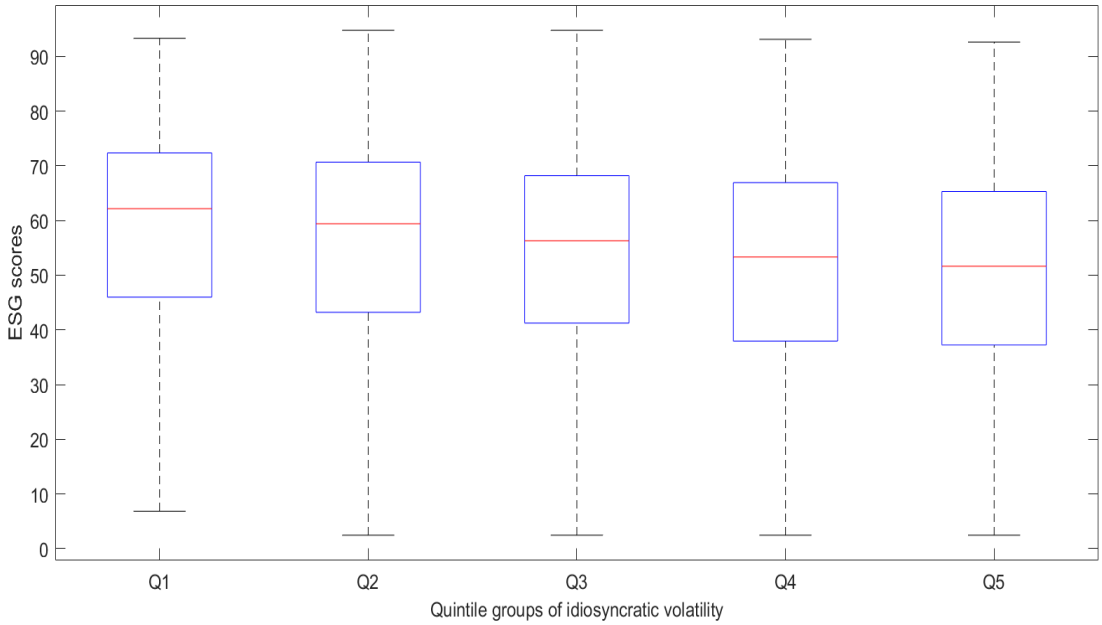
Source: The figure displays the scatter plot that shows the graphical relation between the ESG ratings for the two providers considered (Sustainalytics and Asset4). The datasets contain monthly observations for $n = 217$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure B.7: ESG ratings by idiosyncratic volatility quintiles: Sustainalytics (North America)



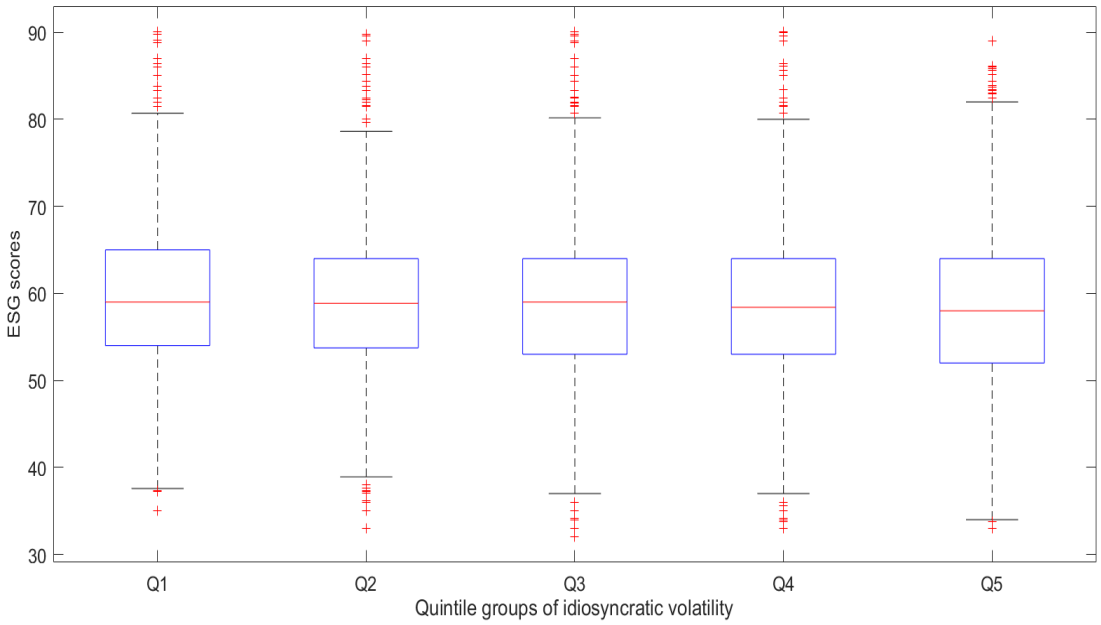
Source: For the North America universe, the figure displays the means of Sustainalytics ESG ratings within the five groups defined by the quintiles of idiosyncratic volatility computed with residual asset returns from the CAPM. The dataset contains monthly observations for $n = 326$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure B.8: ESG ratings by idiosyncratic volatility quintiles: Asset4 (North America)



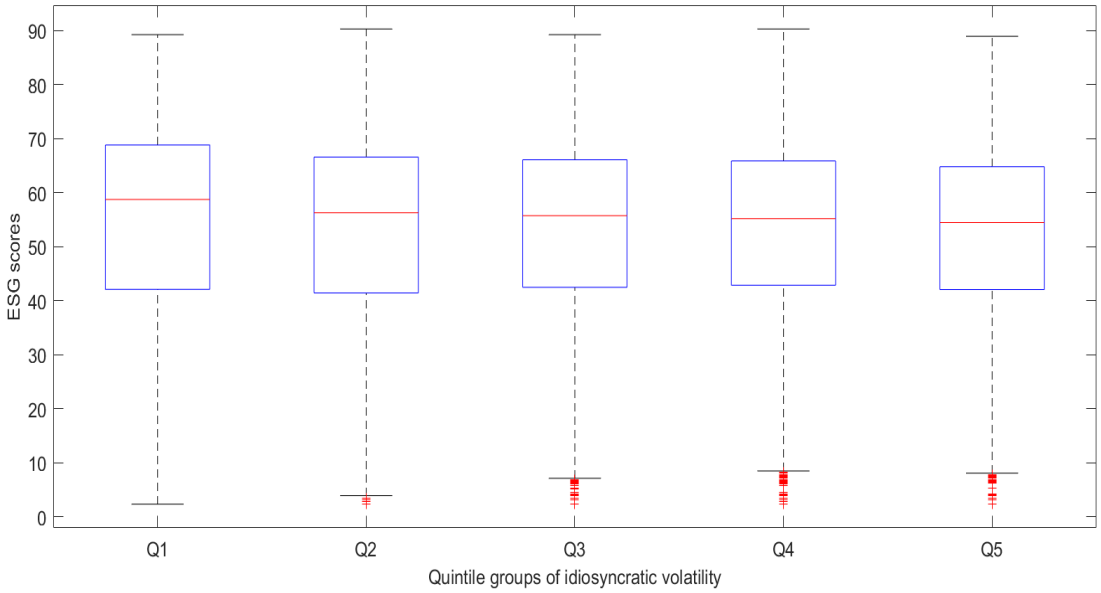
Source: For the North America universe, the figure displays the means of Asset4 ESG ratings within the five groups defined by the quintiles of idiosyncratic volatility computed with residual asset returns from the CAPM. The dataset contains monthly observations for $n = 326$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure B.9: ESG ratings by idiosyncratic volatility quintiles: Sustainalytics (Asia-Pacific)



Source: For the Asia-Pacific universe, the figure displays the means of Sustainalytics ESG ratings within the five groups defined by the quintiles of idiosyncratic volatility computed with residual asset returns from the CAPM. The dataset contains monthly observations for $n = 217$ firms from January 2010 to October 2018, giving a total of 106 months.

Figure B.10: ESG ratings by idiosyncratic volatility quintiles: Asset4 (Asia-Pacific)



Source: For the Asia-Pacific universe, the figure displays the means of Asset4 ESG ratings within the five groups defined by the quintiles of idiosyncratic volatility computed with residual asset returns from the CAPM. The dataset contains monthly observations for $n = 217$ firms from January 2010 to October 2018, giving a total of 106 months.